A semantics-first approach for word learning using visuo-linguistic corpus

by

Nikhil Sudhakar Joshi

A thesis submitted in partial fulfillment of the
requirements for the degree of
Masters of Technology

to the

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur
June 2011

CERTIFICATE

This is to certify that the work contained in the thesis entitled *"A semantics-first approach for word learning using visuo-linguistic corpus"* by *Nikhil Sudhakar Joshi* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Date: _____

Place:_____

Sign: _____

Dr. Amitabha Mukerjee

(Thesis Supervisor)

Professor

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

Kanpur, India

June, 2011

To My Parents & Grandparents

Acknowledgments

I would like to thank my supervisor, Prof. Dr. Amitabha Mukerjee, for his motivation, continuous encouragement, valuable guidance and support in my research efforts.

I am very much grateful to Prof. R M K Sinha for encouraging me to work on Indian Languages. I would like to thank Prof. Achla M Raina for her thought provoking lectures on Cognitive Linguistics.

I acknowledge the excellent work of P Guha in object tracking. I also acknowledge S V P Gopi Srinanth for his work on "unsupervised object discovery".

A special thanks to Niranjan Uppoor and Prabhat Mudgal who shared their views on my work from time to time.

I owe to my parents and my brother for their moral support and encouragement.

I am very grateful to My maternal Uncle (Mama) who always encouraged me to go for higher studies.

I am thankful to all my teachers for the knowledge they imparted to me.

I acknowledge all the subjects involved in data collection without which this work would not have been possible.

I acknowledge Research-I foundation, Dept. of Comp Sc. & Engg., IIT Kanpur for promoting this research.

Finally, I thank all my friends to make my stay at IIT, Kanpur a wonderful and memorable one.

Nikhil Sudhakar Joshi

Contents

List of Figures

List of Tables

Abstract

Acquiring words of a language consists of two aspects: a) having some conceptual categories, and b) associating these with linguistic units. We build on earlier work that demonstrates visual category learning from complex scenes to present a computational approach that attempts to learn words and phrases as labels for these visual categories. Given a multimodal corpus (complex 3D-scene with multiple narrative descriptions), we (a) first discover object categories with minimal supervision using foreground extraction, object tracking and object clustering (b) predict the visual saliency of the objects in the scene using a bottom-up attention model (c) discover motion concepts by clustering the trajectories of the tracked objects (d) Segment the utterances into smaller linguistic units (e) associate the linguistic units in the narrations with the salient objects in the scene to learn labels for object categories as well as motion concepts. We assume no prior domain knowledge either during visual or language analysis. Using a bi-modal (visuo-linguistic) corpus of a complex traffic video and widely varying narratives by different narrators, we show how linguistic units may be discovered for the object categories BICYCLE, TRUCK, and CAR and motion concepts LEFT-TO-RIGHT and TURN. We also show that the knowledge of word-boundaries is, though helpful, not a prerequisite for word-learning. We propose a mechanism to identify appropriate size of linguistic unit based on fragment analysis and unit-independence conjecture. By analyzing the word-concept associations over increasing narration exposure, we measure the confidence of the discovered labels in terms of consistent dominance. We find that the labels discovered for three object categories are consistently dominating. However, the labels discovered for the motion concepts do not show the consistent dominance. We argue that the consistent dominance of a label with respect to a conceptual category is necessary for granting it as a word for that category.

Chapter 1

Introduction

The problem of language acquisition has been of great interest to many disciplines including Linguistics, Psychology, Philosophy, Neurobiology, Cognitive science and Computer Science. From Panini [25] to Chomsky [7] to Tomasello, there have been many attempts to formalize the theory of language.

Language is one of the key-characteristics that distinguish humans from all other animals. There is a long-standing debate on whether the language is innate to humans or not.This debate has led to two different accounts of language. According to rationalists' account, language is considered to be prewired in humans and is distinguished from all other cognitive systems. For example, Chomsky [7] argues for the innateness of language based on the argument (known as "poverty of stimulus") that the child acquiring language has access to only positive examples (grammatical sentences), and very little corrective feedback. Thus, the Chomskyan framework focuses on the syntax of a language and is largely sceptical about semantics. So, learning a language in the view of rationalists' is learning a "generative syntax" for that language.

On the other hand empiricist views of acquisition such as the cognitive grammar proposed by Langacker [18] give a central role to semantics. Langacker considers grammar as conceptualization and formalizes it as a bipolar symbolic unit interconnecting the phonological pole (linguistic representation) and the semantic pole (conceptual representation). In the view of cognitive grammar, language is entrenched in the usage and linguistic representations get their meanings because of their usage with some conceptual entity. So, mapping linguistic representations to their conceptual referents is at the core of learning a language in the "cognitive grammar" view.

In most of the attempts to learn the mapping between linguistic representations and semantics, the semantics were often limited to logical forms such as predicate structures and λ-calculus [39, 16]. However, what the conceptual representation be is again a debatable issue. In [14], Harnad has posed the symbol grounding problem as how the meaning of meaningless symbols can be expressed in something but other than meaningless symbols.

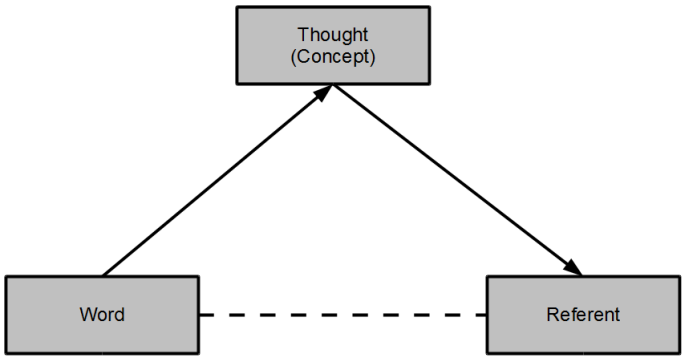

Figure 1.1: **Semiotic Triangle** based on the work of Odgen and Richards

Barsalou [2] proposes perception as one of the ways to ground the meaning of meaningless symbols and argues for the importance of perceptual schema as an abstract representation of concepts.

Mandler has argued, based on the work of a number of developmental psychologists (e.g. Quinn [26], Baillargeon [1], Spelke [34]), that some early notion of categories may be available to infants from age 2 months onwards, well before any phonetic understanding. This categorical discrimination, Mandler suggests, is based on the notion of image schema [20], which are largely perceptual in the young infant.

In this work, we attempt a computational simulation of a similar process, in which a preliminary perceptual concept is acquired first and is available when attempting to discover linguistic units that may be associated with it.

1.1 Grounded Semantic Models

In the classical work "Meaning of meaning" [24], Ogden and Richards opposed the tendency to confuse a "symbol" (or "word") with the thing or object that it refers to. They posited that this relationship works via an intermediary which they called "thought" (or "concept"), and they posited the process as operating in a triangle - the word symbolizes a concept, and the concept is an abstraction of an actual referent. Thus, the word "car" symbolizes the concept CAR, of which a specific observed car may be an image. Thus, in the triangle formed by these three entities, the link between the word and the referent is imputed, and not direct (Figure 1.1).

In the cognitive grammar view, all symbols are schematizations or structured abstractions of experiences. Even a word such as "truck" is an abstraction from the many ways the sound can be uttered (or the word written). All of these map to the same phonological pole, the word "truck", which is coupled to the semantic pole, the TRUCK. These two poles together constitute the symbolic unit [TRUCK], which we are trying to learn in this work.

In our descriptions of this model, we have used the term "word" (and sometimes

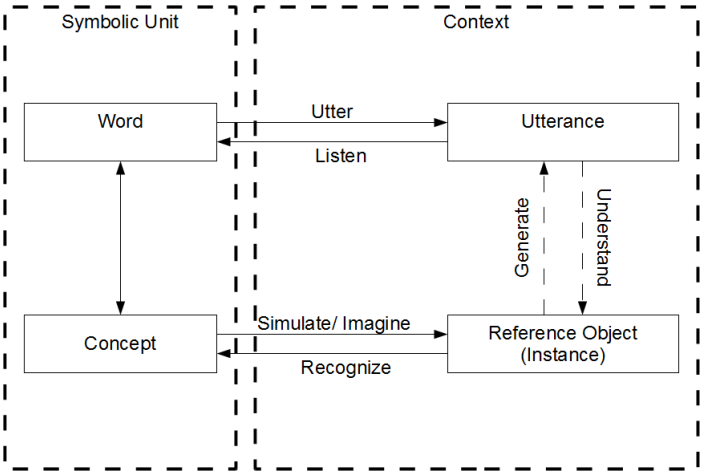

Figure 1.2: **Cognitive Grammar**:Processes

"label") to refer to what is called "linguistic unit", "lexeme", or "phonological pole". Similarly, we have used the term "concept" to refer to what has been called "image schema", "semantic pole", "meaning", etc.

The semiotic triangle of Ogden and Richards may now be elaborated in the cognitive grammar view. There are two processes related to the link between a specific utterance and a referent - language understanding and language generation (autoreffig:cognitive-process.

- **Understanding**: upon hearing the utterance, it is mapped to the appropriate word schema, which invokes to the concept. If the referent is actually present, it then is identified. In a purely linguistic context, the discourse situation is being simulated (imagined) so that the referent may be instantiated in the simulation, and its properties become computable for further statements.

- **Generation**: upon encountering an object, either in direct perception or in mental simulation, the object is being recognized and the word becomes available within the symbolic unit. Then, depending on the context and given the utterance schema of the individual, the prosodic motor functions are launched and the speech utterance is produced.

Both these processes take place within a certain sensorimotor and discourse context. Thus, the visual semantics we learn here, e.g. the appearance properties of *truck*, applies to the similar visual scenes and not to completely different contexts (e.g. toy trucks). However, it is not fixed to a very specific video scene . Once the initial symbol for [TRUCK] is formed, linguistic usage of "truck" will cue its properties and cause the semantic pole TRUCK to be suitably enriched with the new relations. But this again, is outside the scope of this work.

Here we restrict ourselves to the words (note that our use of this term includes larger linguistic units) associated with a specific conceptual entity. We are not considering how such symbols combine to form larger linguistic structures (a combination of a phonological string and a semantic structure); we are merely interested in the process by which the system may bootstrap an initial lexicon of grounded symbols. In the work of Harnad, it has been pointed out that the "concept" or semantics cannot be in terms of a set of symbols. Thus, the meaning cannot be defined in terms of a logical predicate which is defined using other predicates and so on, because this leads to an infinite regress. Thus, at least some of the concepts must be *grounded*, or defined in terms of structures outside the set of symbols, such as in the domain of sensorimotor experience. The objective of this work is to try to learn some such symbols, which may form part of the substrate on which the entire set of symbols is constructed.

We now note that many familiar *concepts* are extremely rich in associations. Thus, the concept for a TRUCK may include what it looks like (how this is represented is a key aspect of our work); but it would also have some notion of it's purpose (used to transport things), its internal structure (has wheels, has an engine in front and a lading area at back, etc.), and also many behaviours (pollutes, makes growling sounds, etc.). In fact, it is not easy to bound the list of such properties - it is an abstraction based on all our direct encounters with trucks, but also based on what we are told (linguistic information). For example, many of us may believe that "trucks bodies are made of steel," or that "drivers of trucks often drive for long stretches without sleep," though we are unlikely to have direct experiential evidence for such facts. These linguistic statements can be correlated to the concept TRUCK only via the link between the concept and the word "truck". Thus, language is a rich source for elaborating concepts - indeed, the vast majority of an adult's vocabulary today is learned via linguistic context [3].

Further, the content of a concept is dependent on language. The categories as experienced by any individual (what Wittgenstein has called "private language" [37]) are fluid – e.g. an infant familiar with dogs may perceive a calf as a large dog. The fact that she eventually learns otherwise is a result of the linguistic community using other terms for the calf. Thus an individual's concepts get anchored in terms of social norms for categories which are expressed in language.

Word learning involves segmenting words from a linguistic stream and associating these with concepts segmented from the "blooming, buzzing confusion" of the world. Acquiring the very first word-meaning associations is a serious challenge since the perceptual stream from which proto-concepts need to be learned may have many ambiguous referents, and the linguistic stream from which words are to be found may be long and discursive. Without the chunks of the world, it is hard to discover the linguistic units, but partitions in the world often reflect linguistic usage. At the same time, linguistic units themselves are

often hard to determine without conceptual structure, e.g. the definition of a phoneme, the basic phonological unit of language relies on the notion of a minimal pair - a single sound substitution that causes meaning to change.

Thus, we have a chicken and egg problem here. On the one hand, concepts must exist in order to associate a linguistic label with it, but on the other hand, the content of the concepts are crucially dependent on language. Learning the initial meaning-label map is often considered a bootstrapping problem [36]. Once a few initial associations are available, these can affect (and stabilize) the perceptual schema, and also help learn other associations. But how are the very first associations acquired? How then, does this process start? How can one bootstrap a system that wishes to build a semantic lexicon, where every unit has a map to some meaning? This is the problem we set out to tackle in this work.

There are three approaches to the initial word learning problem: a) Language-first or nativist: Since there may not be enough data for an infant to learn the mappings, both concepts, and their association with semantics is taken to be inborn [11]. b) In the Piagetian view, concepts are acquired, but these do not form until the end of the sensorimotor stage (about 18 months). Thus, semantics and language are learned around the same time [21]. c) Semantics-first: A preponderance of evidence indicates that babies are able to make many category distinctions before coming to language [21, 27]. These are then associated with words, which may be segmented based on prosodic cues.

Along with these associations, it is sometimes assumed that words can be segmented from the speech input based on prosody, pauses or other non-semantic cues alone. On the other hand, it is possible that knowledge of semantic classes can also help in this segmentation.

Most computational models of word learning follow Piagetian view, and one attempts to learn the label and the semantics at the same time [28, 29, 38]. Some models also consider the problem of partitioning words from speech [30, 38]. Others assume the existence of logical predicates for relations and attempt to instantiate a propositional view of language [32] based on manually constructed scene interpretations.

There is considerable evidence for some degree of perceptual distinction being available at the earliest stages of word learning, including object categories and spatial prepositions [27], event structures [1, 27], etc. It has been hinted that some of these categories may provide priors in learning language [21, 33]. One way of investigating such a possibility would be to try to construct computational simulations.

1.2 Acquiring symbols

A number of approaches have tried to construct such term-meaning associations from sensorimotor data [36, 12, 30, 23, 9]. However, the semantics in these approaches were

often limited to scenes with simple objects, and the learning was guided by considerable feedback. The linguistic input was in the form of "bag of words" or simpler sentences. Also, the semantics was often hand-coded or learnt in a supervised manner. In [32], Siskind tried to learn words in presence of referential uncertainty, however, the objects were simple and linguistic descriptions constrained.

In this work, we attempt to learn words of language as a coupling of a semantics, learned from a perceptual space, with a unit of language, discovered from a sequence of syllables or as phrases from a word-separated text. The emphasis on semantics is aligned to the cognitive view of grammar, but in this work, we make no attempts to discover any aspects of how words are combined to form larger strings, which would be the main goal of syntax. Instead, our aim is merely to discover an initial (relationally impoverished) semantics based on perception alone, which may serve as the "phonological pole" for the first cognitive symbols being acquired. Without this, it is clear that language acquisition, in the "cognitive grammar" view, cannot get off the ground.

In this work, we define the usage of language in terms of the perceptual experience. So, semantics considered in this work are purely visual. We consider learning objects and interactions from a complex 3D-scene and mapping them to words and phrases from free, minimally constrained language with full sentences describing the scene. Part of the mechanism for handling referential uncertainty is visual saliency, predicted using a bottom-up attention model. The salient objects are then associated with the co-occurring utterances in the narratives to learn the labels for the visual concepts.

For constructing visual models of objects and interactions, image sequences from a fixed camera, as typically used in surveillance scenarios, are considered. The stable patterns of background are first learned, and used to extract foreground blobs corresponding to the objects of interest. The object blobs are tracked across the frames and regions of occlusion are identified. Only unoccluded object appearances are considered for object learning. The foreground blobs are then projected to a feature space based on the "Pyramidal Histogram Of visual Words" (PHOW) approach [4]. The resulting PHOW descriptor for the blobs are then classified in an unsupervised manner, resulting in a number of object classes. For evaluation purposes, we label these clusters into seven known object categories based on user labels (the *ground-truth*): TEMPO, BICYCLE, MOTORCYCLE, TRUCK, HUMAN, CAR, and also a small category NOISE with object fragments and lighting effects etc. This is the only element which brings minimal supervision in our mechanism.

For every agent obtained by tracking object blobs across the frames, a trajectory is defined using the position and velocity of the blobs in the successive frames. These trajectories are clustered to obtain a number of motion classes. These are also labeled into five known categories based on user labels:LEFT-TO-RIGHT,RIGHT-TO-LEFT, TURN, CROSS and NOISE.

We note that the resulting models resemble what [36] have called the *conceptualizer*, which serves to recognize the input into one of several classes, but unlike in that work, the model here is learned and not programmed beforehand. Also, these models are similar to abstract perceptual schema proposed by Barsalou [2]. However, these are not as powerful as image-schema of Mandler [20] since they consider only visual appearance, and not the behaviour.

Our work is based on the availability of a bi-modal visuo-linguistic corpus. Such visuo-linguistic corpus consists of a visual scene and multiple narrations of the scene by number of subjects. We construct such a corpus by asking number of human subjects to narrate IITKGTV2 traffic video [10] in Hindi with minimal restrictions on their speech. The narrations collected are then manually transcribed and time-stamped at sentence boundaries as well as long pauses. These transcribed and time-stamped narratives along with the visual categories discovered earlier form a visuo-linguistic corpus which is the basis of all our experimentation.

For the word association task, we use the visuo-linguistic corpus described above. The objects and trajectories in visual focus, as identified by the bottom-up (task independent) attention model, are aligned with poly-syllabic strings, words or phrases in the narrative. In case of poly-syllabic strings we merge the words across word boundaries to form a continuous stream of syllables and try to associate the poly-syllabic sequences within an utterance with the visual concepts. As we are dealing with the transcribed text, we approximate the notion of syllable as a vowel terminated string of characters. In case of phrases, we consider all possible k-grams of words or syllables as candidate labels without assuming any fixed length. Using fragment-analysis mechanism based on unit-independence conjecture, we discard lower length units which mostly occur as a part of larger length units. In the end, we assess the confidence of learnt associations in terms of consistent dominance by analyzing the nature of associations with increasing exposure to narrations.

1.3 Summary of Results

We are able to discover the names for three object classes with high visual purity viz. BICYCLE, TRUCK and CAR as the labels with strongest association both at poly-syllabic and word-level associations. Also, phrases like *bAe.N se dAe.N* (left to right), *geT kI taraf* (towards the gate) are also discovered for the motion classes LEFT-TO-RIGHT and TURN respectively.

In order to estimate the confidence of an association, we evaluate the stability of the concept-word association as new narrations are considered. We analyze the association strengths of these discovered labels with respective visual categories incrementally by providing increasing number of narrations and measure the confidence of these acquired labels

in terms of consistent dominance. We find that the labels *sAikal* (bicycle), *Trak* (truck) and *kAr* (car) are dominant labels for the respective categories BICYCLE, TRUCK and CAR consistently over the set of narrations. However, the label *bAe.N se dAe.N* (left to right) and *geT kI taraf* (towards the gate) fail to dominate for the category LEFT-TO-RIGHT and TURN consistently over period of time. We argue that the consistent dominance of a label with respect to the particular category is necessary for the label to be granted as a term for that category. Based on the notion of consistent dominance, we claim that labels *sAikal*, *Trak* and *kAr* have established themselves as the labels for the respective categories whereas *bAe.N se dAe.N* (left to right) and *geT ki taraf* (towards the gate) have not. Similar results obtained both at poly-syllabic and word level associations, show that the knowledge of word-boundaries may not be a prerequisite for the early word learning.

During association, we remove units that are very frequent in general discourse, which are assumed to be non-relevant to this context. However, no part of speech, phrasal structure or other syntactic knowledge is used at any step. Also, no morphological knowledge is assumed. Thus we use no stemming, though the language tested, *Hindi*, is highly inflected.

Our unsupervised approach to both vision and language implies two important scalability advantages. Since we use no knowledge of the camera placement or the types of objects in the scene, the visual analysis is potentially applicable to a wide range of scenes. Also, since we use no knowledge of the syntax of the target language, it is possible to use the approach to other languages as well. Since the terms learned are grounded in the visual domain, it can be flexibly related to new input situations. The discovered objects and their linguistic labels also address an important practical problem in the context of multimedia retrieval where content pertaining to user's linguistic query are to be retrieved in multimedia documents. Learning visuo-linguistic mappings is fundamental to building systems that can respond to user queries via linguistic means, reporting the ongoing activity in the scene. In a context such as India, it is important to be able to respond in local languages such as Hindi. The approach outlined here constitutes the first step in this direction.

Chapter 2

Symbol Learning Framework

In this chapter, we describe the framework for learning symbols as word-concept pairs using bimodal visuo-linguistic corpus. The framework uses a complex 3D-video and multiple narrations of the same video by number of speakers. Based on the semantics-first approach, we learn from 3D-video various object categories first. Then we align the objects in the video with the utterances from the narratives based on the time-stamps. Using a bottom-up attention model, we find the most salient objects in the scene. Finally, we associate linguistic units in the narrative with co-occurrent salient objects in the video. The unit having maximal association with a given object category according to an association measure consistently over period of time is taken as a word for that category. This consistent dominance with increasing narration exposure gives us high confidence in the associations learned. Next, we describe this framework in detail.

2.1 Overall Framework

The Symbol learning framework (Figure 2.1) consists of four major modules (a) Discovering visual concepts from a complex 3D-scene (b) Attention model (c) Linguistic segmentation (d) Label Association task. The framework assumes the availability of a bimodal visuo-linguistic corpus consisting of visual scene with multiple narrations describing the scene. For discovering visual concepts, image sequences from a fixed camera, as typically used in surveillance scenarios, are considered. The stable patterns of background are first learned, and used to extract foreground blobs corresponding to the objects of interest. The object blobs are tracked across the frames and regions of occlusion are identified. Only unoccluded object appearances are considered for object learning. The foreground blobs are then projected to a feature space based on the "Pyramidal Histogram Of visual Words" (PHOW) approach [4]. The resulting PHOW descriptor for the blobs are then classified in an unsupervised manner, resulting in a number of visual classes. section 2.3 describes the process of object discovery in more detail. The trajectories of the tracked objects

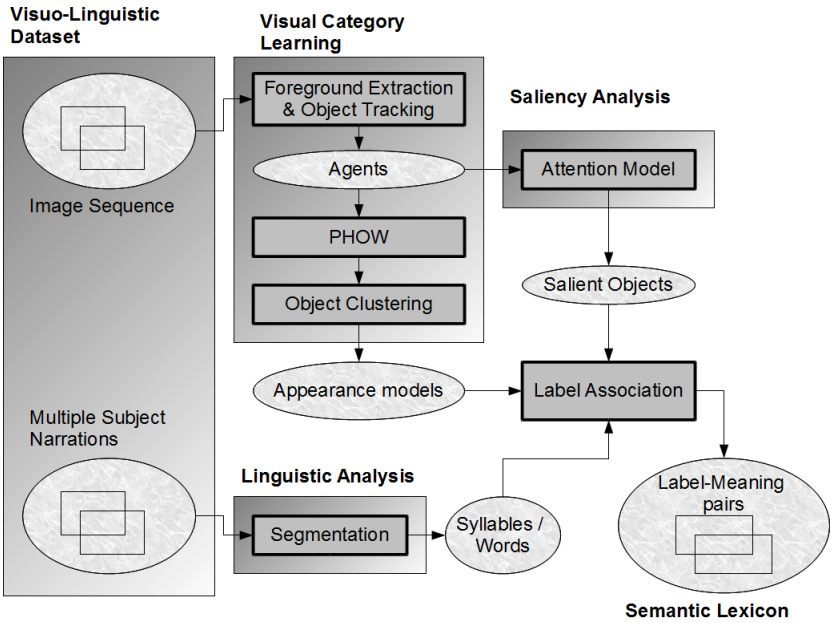

Figure 2.1: **Symbol Learning Framework**: Major components of the framework

are then clustered based on the position and velocity vectors to learn motion concepts in an unsupervised manner. The details of how motion concepts are learnt are covered in section 2.4. A task-independent bottom-up attention model is used to predict the visual saliency of the objects in the scene. As there is no way to determine the attentive focus of the speakers (no gaze or gesture information being available), this visual saliency is used to decide the attentive focus. The attention model used is described in section 2.5. For the label association task, we first compile a set of narratives by asking number of adults to describe the object and activities in free and minimally constrained language. The narratives are first manually transcribed and then time-stamped at sentence boundaries and long pauses. The transcribed time-stamped narratives (text) are input to the system. The process of compiling and time-stamping these narrations used to construct the visuo-linguistic bimodal data-set is described in detail in chapter 3. Finally, we segment the sentences in the narratives into sequence of linguistic units. Assuming that the linguistic focus follows the visual focus and both the speaker and listener follow the visual focus, we try to associate the linguistic units segmented from transcribed speech with the objects in the attentive focus. The process of association is described in section 2.7. The end result is word-concept pair or a grounded symbol.

2.2 Visuo-linguistic data-set

As described above, our symbol learning framework is based on the availability of a bimodal visuo-linguistic corpus. Such a data-set consists of a visual scene and multiple

narrations of the scene by number of subjects. For the purpose of our experimentation, we use IITKGTV2 traffic video ([10]). We show the video to number of adults and ask them to narrate the objects and activities in free and minimally constrained language. We choose local language "Hindi" as the language of our experimentation. As voices of speakers are not so clear and a good speech recognizer for Hindi is still an issue, we prefer to deal with transcribed text. The recorded narrations are manually transcribed and then time-stamped at sentence boundaries as well as long pauses. Using the time-stamps, we align the utterances with the visual context. Finally, we form a corpus with subject ID identifying the speaker, starting frame, ending frame and the utterance uttered by the speaker during interval specified by starting and ending frames.

From visual analysis of the video, we are able to discover objects and motion categories whereas from narratives we are able to get linguistic information. Such a data is the similar to the data available to a child in the early period of language learning.

2.3 Unsupervised Object Discovery

In recent years, supervised learning for visual object categories has been able to distinguish hundreds of classes of objects with high accuracies [4, 22]. The critical step in these approaches is to project the images onto a set of patterns, called "words", so that each image is characterized as a distribution on the words. This class of approaches, known as "bag of words" after similar approaches in document analysis, classify novel images based on their similarity to the trained models. In [35], these ideas are extended to unsupervised object classification. Here the object images (foreground blobs from surveillance video) have the advantage that these are relatively tightly cropped around the region of interest. (Figure 2.2). Foreground blobs are then tracked to identify the same agent across contiguous frames - sample views of some agents are shown in Figure 2.3. As can be seen, the results are very noisy owing to occlusions, shadows, tracking errors, agent appearance changes etc. The tracking step considers substantially overlapping sequences of blobs. Only where an agent is isolated is the blob considered for modeling its appearance. The pyramidal histogram of words (PHOW) approach [4] is used, based on computing the SIFT operator [19] on a very large number of points (100K) on these blobs. These are clustered to obtain a code-book of 300 "words". Next, each foreground blob in a tracked agent is projected onto these words, and the agent is modeled as a probability distribution on the space of words (estimated by the histogram).

Using a Bhattacharya distance metric, the histograms are clustered using k-means (results reported for $k = 30$). Figure 2.4 shows blobs of agents from some of the clusters formed for k=30. This results in an over-segmentation of the category space, and to evaluate the effectiveness of the clusters, the agents are manually categorized into seven *ground-truth*

Figure 2.2: **Segmentation is free, but noisy**. A frame from a traffic video, and the extracted foreground blobs. Blobs like the tempo-car occlusion are identified as occlusions during tracking. Only isolated tracks are considered for object discovery.

Figure 2.3: **Agents as sequences of isolated foreground blobs.** Bottom row (agent 130): the sequence is initially tracking a car - but after it exits, it is erroneously mapped to a motorcycle.

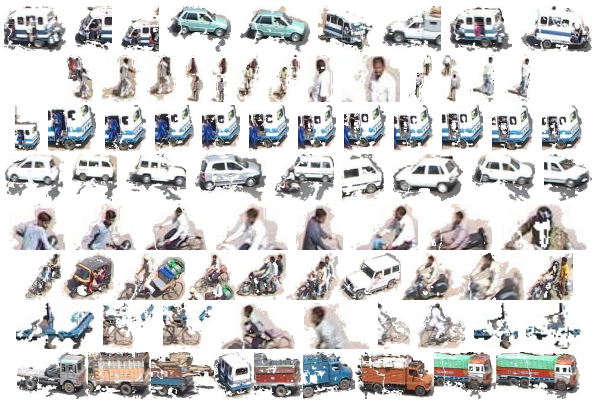

Figure 2.4: **k-means ($k = 30$) clusters** Clusters C0, C10, C16, C19, C21, C26, C27, C29. Representative views from all agents in each class are shown. The membership of these clusters can be seen in Table 2.1. Whereas C10 and C19 are relatively clean classes, C27 has several noise agents, and C26 is a mixed class.

classes: TEMPO, BICYCLE, MOTORCYCLE, TRUCK, HUMAN, CAR, and also a small category NOISE with object fragments and lighting effects etc.

This brings some kind of supervision in our approach as we are making use of ground-truth.

The purity of each cluster is defined as the percentage of its dominant class. If N_k is the number of agents of ground-truth category k in a cluster C, then purity of the cluster C is given by

$$m = argmax_k(N_k(C)) \ , \ P(C) = \frac{N_m(C)}{\|C\|}$$

Ground-truth of the dominant object class in a cluster is assigned as the ground-truth of that cluster. Overall purity of the classification is given by

$$Purity = \frac{\sum_{C_i} N_m(C_i)}{\sum_{C_i} \|C_i\|}$$

The average purity of the clusters obtained by this process is 76.5%. By training the model with a $N - m$ of agents and testing with the remaining M, a cross-validation accuracy of 70.8% (for $M = 5$) is obtained.

Classes with many agents (e.g. human, bicycle and motorcycle), have a number of clusters. Some of the clusters appear to have fine-graded semantic significance - e.g. the class C16 of Tempo as seen in Figure 2.4 may correspond to "passengers getting off from tempo". While such classes were not marked in the ground-truth, this type of discrimination may actually be important in detecting activity such as humans getting on or off a tempo. Another unusual cluster is C26 (third row from bottom of Figure 2.4), which has a very poor correlation with our ground-truth classes (purity of 25%), but one may interpret the semantics of this class as "a vehicle going at an angle to the bottom-left". Such a category may be arising here because SIFT is sensitive to gradient histograms, and appears to have discovered a coherent class of agents that have high gradients corresponding to this type of orientation, rather than a particular class of vehicles. Similarly, the cluster C21 is a group of agents ground- labeled HUMAN who are either on a motorbike or a bicycle, but the vehicle is not visible in the most of the frames (4_{th} row from bottom in Figure 2.4). Some other clusters are less meaningful; e.g. cluster C27 (second row from bottom), is mostly noise.

In Table 2.1 we tabulate the purities of the different clusters formed (black = only a single category, white = completely heterogeneous). Table 2.1 also shows one more category T of transition agents. A top-down cluster refinement procedure is used on the object clusters discovered to identify the transition agents such as agent 130 (Figure 2.3) in order to improve object models [35]. These agents are denoted by T. As can be seen, the purity of object categories HUMAN, BICYCLE, TRUCK/LORRY and CAR is quite high (more than 80%), whereas the purity of MOTORCYCLE is moderate and that of TEMPO is very poor. The seven object categories thus discovered with the help of ground-truth are then

Class:# agents	Cluster	Purity	Distribution
H:52 (81%)	C1	12/13	12H,1X
	C2	7/8	7H,1N
	C4	8/9	8H,1C
	C10	7/9	7H,2N
	C11	4/6	4H,1X,1N
	C13	5/8	5H,1X,1B,1N
	C14	2/4	2H,1T,1M
	C21	6/6	6H
M:36 (73%)	C3	3/3	3M
	C8	8/9	8M,1X
	C9	11/15	11M,2T,1X,1B
	C22	6/6	6M
	C23	3/5	3M,2B
	C24	2/2	2M
	C26	2/8	2M,2B,1X,1T,1R,1C
B:32 (88%)	C5	5/5	5B
	C6	1/2	1B,1X
	C7	2/3	2B,1X
	C15	2/2	2B
	C20	5/5	5B
	C28	7/8	7B,1T
T:21 (56%)	C0	8/16	8T,4X,2C,1L,1R
	C16	1/1	1T
	C17	1/2	1T,1B
	C18	1/1	1T
	C25	4/7	4T,2C,1N
L:12 (83%)	C12	4/5	4L,1C
	C29	7/8	7L,1T
C:16 (90%)	C19	9/10	9C,1X
N:8 (50%)	C27	2/4	2N,1B,1H

Table 2.1: **Purity and ground-truth distribution object clusters.** Purity of a cluster = degree to which it is dominated by a single ground-truth class. Clusters C0 to C29. (H: HUMAN, M: MOTORCYCLE, B: BICYCLE, T: TEMPO, L: TRUCK/LORRY, C: CAR, N: NOISE, T: Transition agents)

Figure 2.5: **Sample Trajectories** traced by agent 120 and agent 156. Both agents are moving from left to right and then crossing the road upwards.

used to learn the language labels.

2.4 Learning trajectories

For every agent tracked across the frames, the path followed by the agent during its appearance in the scene defines its trajectory. Formally, a trajectory T_a of agent a is defined as an ordered set of (t_i, f_i) for $i = 1, 2, ..., n$. Mathematically,

$$T_a = \{(t_i, f_i) | i = 1, 2, ..., n\}$$

where f_i is the set of features describing the agent at time t_i. The features f_i may be position, velocity, acceleration, orientation etc. of agent a at time t_i. In this work, we consider only position and velocity of the agent as the set of features. So, for us, a trajectory of an agent a is given as

$$T_a = \{(t_i, x_i, y_i, vx_i, vy_i) | i = 1, 2, ..., n\}$$

where x_i, y_i are the co-ordinates of agent a and vx_i, vy_i are its velocity components in x and y directions at time t_i.

Figure 2.5 shows the sample trajectory traced by agent 120 and 156 during their appearance in the scene. Each of the points on the path (marked with red line) defines the trajectory of the agent.

To reduce the dimension of the trajectory, we choose 10 distinct frames at regular intervals (10 distinct points on the path traced). So, four components i.e. x, y co-ordinates and vx, vy in each of these frames define a 40-dimensional vector of trajectory for each of 192 agents obtained during object discovery. All positions of an agent are taken relative to its position in the starting frame. So, each agent is assumed to start its trajectory at the same point, the origin O(0,0). This avoids the misclassification of trajectories due to locational bias. We cluster all these 192 trajectories into seven clusters using simple *k-means* algorithm with Euclidean distance as the distance measure.

Figure 2.6 and Figure 2.7 show the various agent blob sequences for some of the agents in trajectory clusters C1 and C2 respectively. The blobs are taken from the ten frames selected for the purpose of modeling the trajectories. As can be seen from Figure 2.6,

Figure 2.6: **Trajectory Cluster C1**: Blob sequences of agents 9, 20 and 45 in the selected 10 frames. All the three agents are going from right to left.

Figure 2.7: **Trajectory Cluster C2**: Blob sequences of agents 120, 130 and 147 in the selected 10 frames. Agent 130 is noisy due to tracking errors.

the agents 9 (bicyclist), 20 (car) and 45 (truck) are going from right to left in the scene. Figure 2.7 shows the blob sequence of agent 120 (first row) who initially comes from left towards right and then at some point turns upwards to cross the road. Figure 2.5 shows the traced trajectory of agent 120. The last row of Figure 2.7 shows a white van (agent 147), which initially turns towards right coming from the bottom and then moves from left to right.

For evaluation purpose, we marked the ground-truth of these trajectories as one of the five categories: LEFT-TO-RIGHT (LR), RIGHT-TO-LEFT (RL), TURN (T), CROSS (C) and NOISE (N). A category NOISE is used to mark the trajectories which can not fit into any of the other four categories. It also contains wrongly tracked agents. e.g. many times the two vehicles crossing each other are tracked as same agents due to high overlap during the transition period (agent 130 in the middle row of Figure 2.7). However, the two agents belong to two different kinds of trajectories but are considered to be a part of single trajectory, as we consider only a single trajectory per agent. Figure 2.9 shows the representative trajectories for each of ground-truth categories. The frames shown are the final frames of the trajectory. The agent of trajectory is present at the end of red-line tracing the trajectory.

Each cluster is a representative of the ground-truth category to which plurality of trajectories in that cluster belong. The purity of each cluster is calculated in the same way as it is calculated for object clusters. Table 2.2 shows the distribution of ground-truth

Figure 2.8: **Misclassified trajectories**: agent 143 turning belongs to C4 (RL), a noisy trajectory of agent 58 in C6.

Figure 2.9: **Representative Trajectories**:LEFT-TO-RIGHT (agent 125),RIGHT-TO-LEFT (agent 9),TURN (agent 54),CROSS (agent 109), NOISE (agent 130)

Ground-Truth	**LR**	**RL**	**T**	**C**	**N**	**Total**	**% Purity**
Cluster							
C1 (RL)	0	20	0	0	1	21	95
C2 (LR)	15	0	1	0	1	17	88
C3 (LR)	20	0	2	0	1	23	87
C4 (RL)	0	26	8	1	3	38	68
C5 (LR)	21	2	4	8	4	39	54
C6 (LR)	13	8	4	2	7	34	38
C7 (T)	0	3	14	3	0	20	70
Total	69	59	33	14	17	192	

Table 2.2: **Ground-Truth distribution of Trajectory clusters**: Distribution of ground-truth categories for each of seven trajectory clusters

categories for each of the seven trajectory clusters discovered. As shown there, out of seven clusters discovered, C1 and C4 are good representatives of ground-truth category RIGHT-TO-LEFT (RL), whereas C2, C3 and C5 are good representatives of LEFT-TO-RIGHT (LR). The clusters C6 though noisy overall represents LR. In addition to this, one cluster (C7) for the category TURN (T) is also discovered. The purity of clusters C1, C2, C3 is quite hight whereas C6 has very low purity. The purity of C4, C5 and C7 is moderate. The reason for lower purity of C4 is that many vehicles in the video come from right going towards left and then in between turn towards down in the video (e.g. agent 143 of C4 Figure 2.8). These agents have been classified under the ground-truth category of TURN (T), however during clustering they are classified in the cluster C4 containing mostly vehicles going from Right-To-Left. Similarly, many vehicles crossing the road come from left, move towards right and then cross the road (e.g. agent 156 in Figure 2.5). So, during clustering these are grouped together along with vehicles going from Left-To-Right reducing the purity of cluster C5. The very low purity of cluster C6 is mostly because of the Noisy trajectories. The noisy trajectories generally include small-sized human blobs which keep moving arbitrarily in the scene (e.g. agent 58 shown in Figure 2.8 is standing doing some hand movements).

2.5 Attention Model

We use an attention model to find the most salient part of the scene. Such a model tries to predict the part of the scene the human is most likely attend to. The words used in the description are more likely to refer to objects that are in perceptual focus, i.e. we assume that linguistic focus follows perceptual focus.

In general, attention combines bottom-up mechanisms (independent of task) with top-down mechanisms (task dependent). While a number of models are available for bottom-up attention, on both still [17] and dynamic [31] images, top-down attention is far more difficult to model owing to complexities in modeling the task. Also, in our context, linguistic commentaries were collected without providing any specific task, hence the role of top-down attention is limited, and we use a dynamic bottom-up model.

In our work, we have an advantage over traditional dynamic attention models since the objects of attention are already segmented and available as tracked sequences of segmented foreground blobs. These are the scene regions that are competing for attention. Unlike many computational models that consider saliency of pixels in the data, we are in a position to evaluate the saliency of the scene, objects i.e. segmented foreground region directly. Our attention model is based on the findings that a) Objects with higher speed are likely to be more salient, and b) Objects with a larger image size are more likely to be attended [17]. We ignore some other factors such as colour and texture, which are more relevant in still images; for image sequences, motion and size are more significant. In addition to the saliency map based on the above factors, we also need to construct a confidence map, based on how recently information was collected about the object. Objects which have not been attended for some time tend to decay in their confidence, and thus become more likely to be attended to. These aspects are combined in an overall saliency measure. For object blob j, this is given as

$$S_j = (1 - e^{-k\Delta t})(w_1 A_j + w_2 v_j)$$

where A_j is the image area (in pixels) and v_j is image speed (in pixels per frame) of the object j. Δt is the time elapsed since the object was last updated. Larger delays result in lower confidence and higher saliency, and the weights w_1 and w_2 reflect relative importance of object size and object image velocity. We set k,w_1 and w_2 all to 1.

2.6 Linguistic segmentation

Linguistic segmentation refers to breaking down the utterances into smaller linguistic units. However, what the smaller linguistic unit of break-up should be is a debatable issue. It is sometimes assumed that words can be segmented from the speech input based on prosody, pauses or other non-semantic cues alone. In such case, it is assumed that

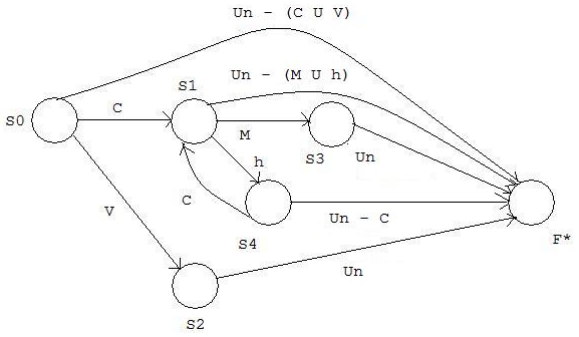

Figure 2.10: **FSM**: To identify syllabic units

word-boundaries are known before learning words from other non-semantic cues. So, an utterance can be segmented into words based on the knowledge of word-boundaries. On the other hand, it is possible that knowledge of semantic classes can also help in this segmentation. In this case, the knowledge of word boundaries is not assumed to be known before word-learning. Moreover, the linguistic labels learnt for semantic classes without assuming word boundaries themselves can lead to identification of word-boundaries.

To put some light on the two approaches, we allow segmentation of utterances in two types of linguistic units. In one case, we assume that the word-boundaries are known before word-learning based on prosodic and other non-semantic cues and make use of word-boundaries available in the transcribed speech to segment utterances into words. In the other case, we merge the words in the transcribed speech across word-boundaries. Then we find the syllables in the continuous utterance and break the utterance into a sequence of syllables. As we are dealing with transcribed speech, we approximate the notion of syllable to the vowel terminated string of characters. We use a simple FSM to identify a unit as a syllable. Figure 2.10 shows the FSM used. In Figure 2.10, U_n denotes the set of all Unicode characters, C denotes set of consonants, V denotes the set of vowels, M denotes the set of *mAtrAs* accompanying consonants in Hindi whereas h denotes *halant*. The state F^* is a failing state reaching which 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.

2.7 Associating language labels

Before the process of label association, what we have is the visual categories discovered earlier through visual analysis, the most salient agents per frame according to visual saliency predicted by attention model and the time-stamped sentences in the narration broken down into smaller linguistic units. From this input, we need to align the most salient objects in the video with co-occurrent linguistic units in the narratives. A mathematical model for label association task in described in chapter 4. It tries to associate the most

salient objects in the video with co-occurrent linguistic units uttered by number of speakers. Based on some association measure such as conditional probability, mutual information, we rank the labels in order of their relevance for each of the visual categories learnt.

Chapter 3

Bimodal Dataset for Vision and Language

In order to learn language with the help of visual context, it is important to have a linguistic description aligned with the visual context. Such a visuo-linguistic corpus is at the core of the language learning framework described in chapter 2. We construct a bimodal visuo-linguistic corpus, with minimal restrictions on both the visual or linguistic domains. We collected human narrations in Hindi from 44 adults on IITKGTV2 Traffic Video. The collected narrations are then aligned with the video frames. The aligned bimodal data-set is then used for the purpose of further experimentation. We have made this data-set publicly available for the future research on Vision and language [10]. The details of the methodology used to construct this data-set and the various properties of the data-set are explained in this chapter.

3.1 Visual scene and linguistic narrations

For the purpose of constructing a visuo-linguistic corpus, we use IITKGTV2 traffic [10] video shot from a static camera . The scene is a natural one as it is shot in the real traffic environment. Also, the scene consists of complex interactions of multiple agents present in the video. Typically, the scene consists of different kinds of vehicles, people, road and its premises. Typical interactions include the crossing and turning of vehicles, overtaking of a vehicle by the another, person riding or getting off a vehicle. The presence of multiple objects (sometimes as many as 20) in the scene simultaneously provides possibility of ambiguous reference. The scene is full of occlusions and no manual analysis or helpful camera angle is available to resolve these occlusions. The visual scene is recorded first as it forms the basis for linguistic narrative. The recorded video has a length of 4 minutes and 32 seconds (around 4.5 minutes). The video is then sampled at the rate of 24 frames per second to get an image sequence of 6538 frames. This image sequence is used for the visual processing

described in section 2.3. The sample frames of the video are shown Figure 3.1.

Figure 3.1: **Sample frames from the video** .Frame no.s 1200, 1255, 1300, 1350.

For the purpose of collecting narrations, we show the visual traffic scene to number of people and ask them to narrate the scene in Hindi. Hindi is morphologically rich and highly inflected language. It distinguishes from most of the western languages by using subject-object-verb (SOV) form. It is characterized by relatively free word order forming a loose and flexible syntax. Again, there are many dialects of Hindi that are spoken in various regions of India. Hindi-Urdu dialects vary in a continuum across most of North India and West Pakistan. These dialects differ in terms of the way Hindi is spoken, or the way certain words and constructions are used. Although the mother tongue of the respondents was in various dialects, the language spoken here was generally in the dominant lect, which is often known as *Standard Hindi*. Finally, in the country like India where a large number of people speak in Hindi, it is important to build systems that can understand and respond to users in Hindi. For all these reasons, we prefer Hindi as a language of our experimentation.

3.2 Collecting narrations

We collected human narrations in three phases. In the first phase, we collected narrations from 11 subjects without constraining the subjects about what they should talk. We call this phase as "free unconstrained narratives". In the second phase, we collected narrations from 20 subjects by asking them to focus on the objects and activities in the video with some initial feedback. We call this phase as "Feedback-based narratives". Finally, in the third phase, we collected narrations from 13 subjects by asking them to describe the scene as if they are describing it to a 2-years' baby. We call this phase as "child directed narratives". Each of these phases is described next in detail. In all three phases, the subjects were allowed to talk full sentences.

3.2.1 Free unconstrained narratives

In this phase, we showed the subject first 40 sec of the video in order to familiarize him with some context. Immediately after this, the subject was shown the entire video for around 4.5 minute and was asked to describe the scene in his own words in Hindi. The

specific instruction given was:

> " This is a traffic video. This video will be shown to you twice. First time, you will be shown this video for around 40 sec when you have to just watch it. Next time when this video will be shown to you for around 4.5 minutes, you have to describe the scene in your own words in Hindi. "

After giving this instruction, we showed the first 40 sec of the video in order to provide the subjects with some context. Then we showed the subject full video and recorded his speech.

As the language was not restricted in any way, this data-set is bit noisy in that the people often described peripheral things not very much related to scene. As some of people could figure out that this road is the one outside IITK, they often talked "yah IITK ke bAhar ke grAnT-Tra.nk roD kA dRushya hai" (This is the scene of Grant-trunk road outside IITK). One of the subject also said "yah grAnT Tra.nk roD sher shAh surI dvArA banAyI gayI thI" (This Grant-Trunk road was built by Sher-Shah-Suri).

In this phase, we collected narrations from 11 different subjects. Hereafter, the data collected in this phase is referred to as ADULT-1.

3.2.2 Feedback-based narratives

In this phase, we continued to show the subject first 40 sec of the video in order to familiarize the subject with some context. Immediately after this, the subject was shown the first 40 sec again and asked to comment on the people, vehicles and their activities in the video. Based on this sample narration, we gave the subject some suggestions. In the third step, we showed the entire video and asked subject to describe the people,vehicles and their activities in own words in Hindi. The specific instruction given was:

> " This is a traffic video. You will be shown this video first for 40 sec and you have to just watch it. Next time you will be shown the same 40 sec of the video. This time you have to describe the objects like people, vehicles and what they are doing in the video in Hindi. Then you may be given some feedback on your narration. Finally, you will be shown the full video of around 4.5 minutes. Considering the suggestions given if any, you have to describe the objects and what they are doing in Hindi."

- Now you are going to watch the video for next 40 sec.
  (Presented the first 40 sec of the video).
- Now you will be shown the video for next 40 sec and you have to describe the people, vehicles and what they are doing along with the video in your own words in Hindi.
  (Presented the first 40 sec of the video again and gave some suggestions after listening the narration).
- Now you are going to watch the full video for next 4.5 minutes. Considering the suggestions given, you have to describe people, vehicles and what they are doing along with the video in your own words in Hindi.
  (Presented the full video and recorded the narration).

Figure 3.2: **Picture of Baby**: Used during Child directed narratives

After giving first instruction, we showed first 40 sec of the video. In the second step, we showed first 40 sec of the video again and asked the subject to comment. Based on this sample commentary, sometimes we provided some feedback in order to make the narrator to focus on events in the video rather than the broader context. Suggestions given may be "Talk about what is being shown in the scene and not about the notion of traffic in your mind.", "Don't talk about the dog, the bench, tea-shop etc. Instead focus on what is happening on the road". In the third step, we showed the full video, and recorded the narration.

In this phase, we collected narrations from 20 different subjects. Hereafter, the data collected in this phase is referred to as ADULT-2. These 20 narrations were collected in two different sub-phases. We refer to the two sub-datasets as ADULT-2-1 and ADULT-2-2. ADULT-2-1 consisted of 9 narrations whereas remaining 11 narrations were part of ADULT-2-2. However, instructions given in both these sub-phases were the same.

3.2.3 Child directed narratives

In this phase, we continued with the method of Feedback-based narratives. However, we asked the subjects to speak about the scene as if they were speaking to a 2-years' baby. To make experiment a little realistic, we kept the picture shown in Figure 3.2 alongside the screen so that narrator would feel as if a baby is watching the video. The specific instruction given was:

> " This is a traffic video. You will be shown this video first for 40 sec and you have to just watch it. Next time you will be shown the same 40 sec of the video. This time you have to describe the objects like people, vehicles and what they are doing in the video in Hindi as if you are describing it to a 2-years' baby who is watching this video. Then you may be given some feedback on your narration. Finally, you will be shown the full video of around 4.5 minutes. Considering the suggestions given if any, you have to describe the objects and what they are doing in Hindi as if you are describing it to a 2-years' baby who is watching this video."

- Now you are going to watch the video for next 40 sec.
  (Presented the first 40 sec of the video).

- Now you will be shown the video for next 40 sec and you have to describe the people, vehicles and what they are doing along with the video in your own words in Hindi as if you are describing it to a 2-years' baby who is watching this video.
  (Presented the first 40 sec of the video again and gave some suggestions after listening the narration).
- Now you are going to watch the full video for next 4.5 minutes. Considering the suggestions given, you have to describe people, vehicles and what they are doing along with the video in your own words in Hindi as if you are describing it to a 2-years' baby who is watching this video.
  (Presented the full video and recorded the narration).

After giving the first instruction, we showed first 40 sec of the video. In the second step, we showed first 40 sec of the video again and asked the subject to comment. After listening to his commentary, we gave some suggestions in order to reduce the noise. In addition to the suggestions given in Feedback-based narratives, the suggestions given in this phase prominently included a suggestion like "Describe the scene as if you are describing it to a baby". In the third step, we showed the full video, and recorded the narration.

In this phase, we collected narrations from 13 different subjects. Hereafter, the data collected in this phase is referred to as CDS.

3.3 Post-processing of narrations

After collecting these narrations, we transcribed them into text using "Kamraj" Unicode Hindi converter [15]. While transcribing the utterances, every two consecutive words were separated by space to maintain word boundary. The post-positions were generally treated as separate words and hence were separated from the content words they were attached to. However, the morphological variations were preserved and transcribed as it is without separating them from their roots. The compound words like *sAikalwAlA* and *moTarsAikal* were written together. The narrations were transcribed as it is without correcting them for any grammatical errors. The care was taken to follow uniform writing style to avoid the transliteration variations. The transcribed narrations were time-stamped along with the video at sentence boundaries. We also broke the narration at pauses longer than 1.5 sec. For every segment of the narration, we note down the starting and ending frames of the video.The reason for choosing sentence boundaries for segmentation is that the sentence can be regarded as the unit of describing an event in the scene. The transcription and time-stamping of narratives were done manually. For recording and time-stamping the narration, we make use of "Microsoft Movie Maker". After, this post-processing of the narrations, finally we construct a data-set which consists of subject-ID identifying the subject, start-frame, end-frame and the narration of the specified subject during the period specified by start-frame and end-frame. Table 3.1 shows the snapshot of the data-set. The

Speaker ID	Start Frame	End Frame	Utterance
ADULT1-7	1216	1375	aur ek dIlaks bas gayI abhI bAe.N se dAe.N or
ADULT1-3	1201	1392	is saD.ak pe Traifik pulis kI vyavasthA bhI dikhAi nahi paDatI
ADULT1-10	1208	1228	bAikwAle log hai.n
ADULT2-2	1254	1308	kuch moTarsAikal aur jAte aA rahI hai.n
ADULT2-14	1225	1288	aur usake pIche pIche moTarsAikalwAle aAe hai.n
ADULT2-15	1217	1275	moTarsAikal pe sawAr ek yAtrI
CDS-3	1201	1277	yah ek moTarsAikal
CDS-6	1210	1288	aur yah Tempo kI taraha hakate hai.n
CDS-11	1220	1274	moTarsAikal sTArT huI calanA

Table 3.1: **Snapshot of visuo-linguistic corpus**: Transcribed descriptions with time-stamps

descriptions shown belong to the portion of the video shown in Figure 3.1.

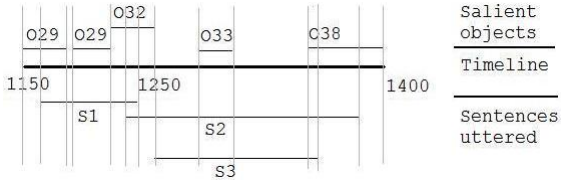

Figure 3.3: **C**o-occurring sentences and salient objects in a given time-interval

Based on the time-stamps, sentences from the narrations are then aligned with the most salient objects in the scene as predicted by attention model. Figure 3.3 shows the alignment of salient objects and some of the sentences in the narration. Figure 3.1 shows the sample frames of the visual scene during the timeline shown in Figure 3.3. Table 3.2 shows the sentences S1, S2 and S3 in Figure 3.3 with their gloss. Sentences indicated with * are grammatically incorrect, but are given as uttered.

3.4 Subject Information and Dataset properties

We collected the information from subjects regarding their age, sex, first language, second language, place where they spent first few years of their lives, first language of their parents etc. This information is summarized in Table 3.3 where S represents the number of subjects, N_{h1} the number of subjects with Hindi as their first language, N_{h2} the number of subjects with Hindi as their second language, $N_{h'}$ the number of subjects with Hindi as

	Sentence	Interval
S1	ek bAik gayI abhI	1158 -1224
	One bike go+past now.	
	A bike went now.	
S2	sAiD me.n sAikal rikshA pe ek ADamI caDhA	1216-1382
	Side [on] one cycle rickshaw [on] one man climb+past	
	A man climbed on a cycle rickshaw on the side (of the scene).	
S3	* sAikal bAik Aye jA rahe hai.N.	1239 -1354
	Bicycles bikes come+pp go+pp are.	
	Bicycles, bikes are coming and going.	

Table 3.2: Sentences uttered by different speakers and their time-lines

Dataset	S	N_{h1}	N_{h2}	$N_{h'}$	N_m	N_f
ADULT-1	11	8	3	0	8	3
ADULT-2	20	17	1	2	19	1
CDS	13	12	1	0	12	1
All	44	37	5	2	39	5

Table 3.3: **Subject Information**: S: Total # of speakers, N_{h1}: # of speakers with Hindi as first language, N_{h2}: # of speakers with Hindi as second language, $N_{h'}$: # of speakers with first and second languages other than Hindi, N_m: # of male speakers, N_{h1}: # of female speakers

neither the first language nor second language, N_m the number of male subjects whereas N_f represent the number of female subjects. All 44 subjects were college students (39 male; 5 females) with their ages between 18-27 years. As can be seen from Table 3.3, 37 out of 44 subjects mentioned their first language as Hindi. The subjects were from various areas of the country like Uttar Pradesh, Delhi, Rajsthan, Punjab, West Bengal etc. and hence were speaking various dialects of Hindi such as Avadhi, Bhojapuri, Bundelkhandi etc. We also analyzed the datasets collected for different statistical properties which are summarized in Table 3.4. In Table 3.4, S represents the number of subjects in the data-set, A_s the average number of sentences spoken, A_p the average number of pauses taken, A_{st} the average sentence length in time (s), A_{st} the average pause length in time (s), A_{sw} the average sentence length in words and W the total number of words in the data-set. As can be seen in the Table 3.4, ADULT-1 data-set consists of small number of longer sentences with small number of longer pauses. On the other hand, in CDS there are large number of sentences of smaller length. Also, the pause length is shorter. Analyzing the descriptions

Dataset	S	A_s	A_p	A_{st}	A_{pt}	A_{sw}	W
ADULT-1	11	51	11	4.05	6.02	8	4599
ADULT-2	20	77	16	2.9	3.66	7	10799
CDS	13	83	17	2.79	3.06	6	6991
All	44	72	15	3.15	4.09	7	22389

Table 3.4: **Dataset Statistics**: S: Total # of speakers, A_s: Average # of sentences, A_p: Average # of pauses, A_{st}: Average sentence length in seconds, A_{st}: Average pause length in seconds, A_{sw}: Average sentence length in words, W: Total # of words

mentioned in the narrations, we find that the lexical choice and linguistic constructions varied widely across the subjects. Thus the same event may be described as "gADI dAe.N se bAe.N or gayI" (car went from right to left), "blaik kalar kI gADI gayI" (black car went) "ek sa.NTro gayI"(one Santro [car-make] went) etc. More importantly, perspectives varied tremendously; thus, for the same time interval in the video, subjects said: "ek kAr aAyI" (One car came), "vah saD.ak krOs kar rahA hai" (He is crossing the road) etc. Also, the commentaries include considerable peripheral descriptions: "yaha jI TI roD sher shAh surI dvArA banayI gayI thI (this GT raod was built by Sher Shah Suri), "usane dekhA bhI nahI be.nc kI taraf" (He didn't even look at the bench) etc. Even when we asked the narrators to focus on the people, vehicles and their activities during instructions, some of narrators described considerable peripheral descriptions: "bIc me.n koI DivAiDar nahI.n hai" (There is no divider in the middle), "pArki.ng ke liE yahA.N par kuch hai nahI.n" (There is nothing for parking here) etc.

We also analyze the frequency of different words relevant to the video in different datasets. Table 3.5 lists some of the important words and their frequencies in each of the data-set. As can be seen, the word *Trak* is used heavily as compared to its synonymous word *lauri*. In fact, the word *lauri* is used only in data-set ADULT-2-1. Similarly, *bAik* is more used as compared to *moTarsAikal* and *skUTar* in general. ADULT-1 uses *bAik* more often than *moTarsAikal* whereas in data-set ADULT-2-2 both are used equally. The terms *kAr* and *gADI* are also present in good proportion. In ADULT-2-2, the use of *kAr* is much higher than the use of *gADI* whereas CDS contains them in almost equal proportion. Most of the occurrences of *bAe.N se dAe.N* are present in ADULT-2-1 data-set only. *dAe.N kI taraf* is present only in ADULT-2-1. So, ADULT-2-1 data-set is rich in directional descriptions.

	adult-1	Adult-2-1	ADULT-2-2	Adult-2	CDS	ALL
OTo	13	43	29	72	89	174
Tempo	53	48	107	155	60	268
sAikal	65	91	181	272	155	492
bAik	33	53	65	118	65	216
moTarsAikal	5	23	65	88	42	135
skUTar	11	18	41	59	22	92
Trak	44	55	107	162	117	323
lauri	0	19	0	19	0	19
bas	24	26	45	71	52	147
aAdamI	22	50	76	126	66	214
aurat	2	8	6	14	7	23
sAikalwAlA	10	9	18	27	15	52
rikshAwAlA	0	5	10	15	8	23
bAikwAlA	0	2	3	5	11	16
moTarsAikalwAlA	1	2	8	10	1	12
kAr	18	24	76	100	57	175
gADI	18	26	17	43	42	103
vain	6	17	14	31	13	50
jIp	5	13	23	36	15	56
bAe.N	19	30	35	65	17	101
dAe.N	19	48	29	77	10	106
bAe.N se dAe.N	3	15	9	24	0	27
dAe.N kI taraf	0	23	0	23	0	23
lefT	1	44	43	87	20	108
dAe.N se bAe.N	7	5	6	11	0	18
lefT kI taraf	0	8	0	8	0	8
mUD	2	7	14	21	17	40
geT kI taraf	8	6	4	10	0	18
krOs	21	47	45	92	57	170

Table 3.5: **Word frequencies**: Frequencies of some important words relevant to the video in different datasets

Chapter 4

Learning language labels

This chapter describes the label association algorithm and various experiments we performed with respect to label association for proto-concepts discovered in chapter 2. We experiment with various kinds of linguistic units, different association measures and different datasets. Typically, we assume the linguistic units to be contiguous (k-grams) at word and syllabic-level. We also experiment with units of different lengths combined to form phrases at word and syllabic level. We propose a mechanism to learn the appropriate units of correct size based on fragment analysis and unit-independence conjecture. We confirm the stability of learnt labels by analyzing the associations incrementally to assess the confidence in terms of consistent dominance. We analyze the behaviour of different association measures for label association task. We compare results on different datasets mentioned in chapter 3. Different configurations of label learning process are described in section 4.1. The subsequent sections describe the results of various experiments we performed based on the different configurations.

4.1 Label Association

Given the visual categories discovered earlier, the salient agents in the scene and the time-stamped narrations , we try to associate the language labels in the narratives with the co-occurrent salient objects in the scene. The label having maximum association with a given object category is taken as the label for the category.

Next, we define the mathematical model of label-association task:

Let, C be the set of concepts c_i for $i = 1, 2, ..., |C|$. Let, A be the set of agents a_j for $j = 1, 2, ..., |A|$. Let, S be the set of speakers s. With each agent $a_j \in A$, there is a concept associated to which it belongs. Let, $C(a_j)$ denote the concept associated with agent a_j. With each speaker $s \in S$, there is a set of utterances associated denoted by U_s. With each utterance $u \in U_s$, there is a time-interval associated. Let $t_s(u)$ and $t_e(u)$ denote the start and end time of utterance u. Also, with each utterance u, there is a list of linguistic

units associated. Let, $L(u) = \{l_i | i = 1, 2, ..., u\}$.

We say a concept c_i is attended at time t, if $\exists a \in A$ such that $C(a) = c_i$ and a is attended at time t. We say that a linguistic unit l is uttered by a speaker s at time t if $\exists u \in U_s$ such that $l \in L(u)$ and $t_s(u) < t < t_e(u)$.

We define following probabilities.

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

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

A continuum probability could have been used for attention, but cognitively attention is more of "winner-take-all" problem. Hence, we consider the attention probability to be binary.

Utterance probability of a linguistic unit l for the speaker s at time t

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

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

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

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 (number of occurrences) of label l in the narrative corpus.

Initially, we assume that all speakers attend to those agents which are visually the most salient. The visual saliency is predicted by the attention model. So, the attention probability does not depend on the speaker. Therefore, simplifying the notations,

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

With this assumption, we can rewrite the joint probability as

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

where

$$P(l|t) = \frac{1}{|S|} \sum_{s \in S} P(l|s,t)$$

is the fraction of number of speakers uttering label l at time t

Using joint probabilities $J(l, c)$, concept probabilities $P(c)$ and label probabilities $P(l)$, a suitable association measure $M(l, c)$ can be defined.

As can be seen, there are three main issues in the framework that need to be addressed : 1. What should be the linguistic unit of association? 2. What should be the association measure? 3. Which of the linguistic units should be associated? Besides these main issues there are some other issues from modeling point of view such as minimal supervision involved in merging the object clusters, usefulness and necessity of attention model used. Another issue is to decide when we can say that a label is learnt for the particular proto-concept. This section describes the various linguistic units, various association measures and various strategies we experimented with for associating labels with visual categories.

4.1.1 What should be the linguistic unit of association?

In most of the attempts to learn labels for the semantic categories, the basic unit of association is a "word". However, this assumes that the knowledge of word-segmentation is known before the word is learnt. The obvious question is how we can know the word-boundary when we don't know the word itself. Though there are evidences that the infants are sensitive to the acoustic properties and can identify the word-boundaries from non-semantic cues[8], it is not very clear whether the word-boundaries are known before learning the words or not. In fact, the knowledge of semantic categories may help in identifying word boundaries during the process of word-learning. To address these issues, we experiment mainly with two different kinds of linguistic units: Word-level unit and Poly-syllabic unit. In word-level association, we assume that the word-boundaries are somehow known and we treat word k-grams as the basic unit of association. In poly-syllabic unit, we assume no word-boundaries and merge all the words in an utterance across word-boundaries. We approximate the notion of a syllable to vowel-terminated string of characters as we are dealing with transcribed text and not with the speech directly. We find syllabic k-grams in the continuous utterance and associate these with the visual categories discovered.

Also, associating a single word with the semantic categories in some sense assumes that the semantic category can be described with the single word. In real world, semantic category may have a phrasal description. The size of the phrase is also likely to vary from one semantic category to the other. So, we need to allow a linguistic unit to vary in size. In both word and syllabic-level units, we further distinguish between two types of units: Fixed-length units and Variable length units. In fixed length units, we experiment with fixed value of k, whereas in variable length units, we allow linguistic units to vary in length and combine all fixed-length units for different values of length k. We call the

word-level units with variable length as phrases and syllabic-level units with variable length as syllabic-phrases.

The setting where we use word as basic unit of association is denoted by $W+$ whereas the setting where we use poly-syllabic unit as basic unit of association is denoted as $W-$. The corresponding settings at phrase level are denoted respectively by P_w and P_s.

4.1.2 Association measures

To find the maximally associated linguistic unit for a given visual category, we need an association measure which can rank the labels according to the degree of co-occurrence between the label and the visual category. A typical association measure should have following properties:

1. It should give high association values if the label and the visual category co-occur frequently.

2. It should penalize the labels which co-occur frequently with many categories whereas should prefer labels which co-occur frequently only with a particular category.

Various association measures we experimented with are described next.

Dominance Weighted Joint Probability

Dominance weighted joint probability is proposed by Guha and is described in [13]. To capture the dominance of a label for a particular concept over all other concepts, we look at the distribution of joint probability of a label over all concepts. The dominance weighted joint probability favours the peaky distributions as compared to the flat ones.

To calculate Dominance weighted joint probability, we multiply the joint probability with a term called as dominance weight.

First, we normalize the joint probability as follows:

$$NJ(l,c) = \frac{J(l,c)}{\sum_{c \in C} J(l,c)}$$

Then, we calculate, the dominance of label l for the concept c over all other concepts c as follows:

$$w(l,c) = \frac{1}{|C|-1} \sum_{x \neq c} (NJ(l,c) - NJ(l,x))$$

The dominance weights are then again normalized as

$$w_s(l,c) = \frac{w(l,c) - min_x\{w(l,x)\}}{max_x\{w(l,x)\} - min_x\{w(l,x)\}}$$

The normalization is required to avoid the negative values of weights. The weight $w_s(l,c)$ captures the peakiness of the distribution of a label over set of concepts. The weight

is high for the labels having peaky distribution, whereas it is low for the labels having flat distribution.

Finally, the dominance weighted joint probability is given as

$$DJ(l,c) = w_s(l,c) * J(l,c)$$

The dominance weighted joint probability favours the labels whose associations are concentrated around a particular concept and penalizes the labels having equal association with many concepts.

Conditional Probability

Conditional probability of a label l given a concept c is given as

$$P(l|c) = J(l,c)/P(c)$$

Conditional probability of a label given concept favours the concepts having rare occurrence but having sufficient co-occurrence with the label. However, it doesn't consider the distribution of the joint probability of the label over all concepts and hence fails to capture the second property of association measure.

Mutual Information

Mutual information of a label l and a concept c is given as

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

Mutual information favours the rare concepts and rare labels having sufficient degree of co-occurrence.

4.1.3 Which of the linguistic units should be associated?

Not every linguistic unit in the description may be good candidate to be appropriate labels of the visual categories. Some linguistic units can be very much specific to the context whereas some other may be general. The linguistic units which are specific to the visual context are more likely to be the labels for visual categories than the linguistic units which are used in general such as articles, common verbs, auxiliary verbs etc. To evaluate the effect of using context-specific knowledge, we make use of the word frequencies in Hindi Unicode Corpus provided by IIT, Bombay ([6]) (general corpus). We typically experiment with two types of settings:

1. We consider all linguistic units present in the narrative corpus for the association task without using any context specific knowledge. We denote this setting by $T-$.

2. We remove the most frequent k linguistic units of general corpus assuming these to be non-relevant to the current visual context. We denote this setting by $T+$.

L	**M**	**T**	**G**	**A**	**V**	**D**
W	DJ	$T+$	$G+$	$A+$	obj	ADULT-1
S	**CP**	$T-$	$G-$	$A-$	traj	ADULT-2
P_w	**MI**					CDS
P_s						ALL

Table 4.1: **Parameters of experimentation**: Different configurations tested for validating language model. L=Linguistic Unit, M=Association Measure, T=Top-K removed or retained, G=using ground-truth, A=Use of attention model, V=visual categories, D=Dataset

4.1.4 Using ground-truth and attention model

To assess if we can get rid of even the minimal supervision involved in the process due to use of ground-truth in merging the object clusters, we experiment with two different configurations. In one configuration, we use the seven ground-truth object categories obtained using ground-truth information. This setting is denoted as $G+$. In the other configuration, we use the thirty object clusters obtained directly through clustering of agents without using ground-truth. This setting is denoted as $G-$.

To assess if we really require the attention model to learn the labels, we experiment with again two different settings. In one setting denoted by $A+$, we use the attention model described in section 2.5 to predict the visual saliency. In the other setting denoted by $A-$, we don't use attention model for predicting the visual saliency, but assume that each agent present in the scene to be salient at that point of time.

Also, we try to learn labels for two kinds of visual categories viz. object categories (obj) and motion categories (traj).

Table 4.1 lists the different settings for each parameter used for the purpose of experimentation. Based on these parameters, we present the results of various experiments performed in the following sections. Unless mentioned otherwise, the results presented here mostly are according to conditional probability (CP), with top-k units removed (T+), using ground-truth (G+) and using attention model (A+). Each of the table presenting the results mentions the parametric configuration used. The parameter which is being varied is marked with an (*). The appropriate terms discovered as top-most labels are high-lightened as white text against dark black background. The relevant labels in top-3 (other than top-1) are high-lightened against gray background. Also, note that, the values of conditional probability (CP) mentioned are multiplied by 100, whereas the values of dominance weighted joint probability (DJ) and mutual information (MI) are multiplied by 1000. Also, while performing association, we have not considered first 1000 frames (around first 40 sec) and last 500 frames (around 20 sec) of the video.

(W, CP, T+, G+, A+, obj, ALL)						
	k = 1		**k = 2**		**k = 3**	
Concept (c)	*l*	**CP**	*l*	**CP**	*l*	**CP**
TEMPO	Tempo	4.46	ek Trak	2.52	dAe.N se bAe.N	1.08
	kAr	4.33	ek Tempo	2.16	Ai Ai TI	0.87
	pe	4.25	ek kAr	1.84	do OTo aur	0.79
BICYCLE	sAikal	1.95	ek sAikal	1.14	sAikal jA rahI	0.32
	moTarsAikal	0.79	aur sAikal	0.32	gais silinDar le	0.32
	pe	0.63	lefTsAiD	0.32	silinDar le ke	0.32
MOTORCYCLE	pe	8.60	ek Tempo	4.39	sAmAn le ke	1.45
	bAik	7.12	ek bAik	3.27	aur ek bAik	1.44
	Tempo	6.56	ek OTo	3.19	ek sAikal pe	1.03
TRUCK	Trak	17.29	ek Trak	10.67	se ek Trak	1.74
	pe	3.24	tIn sAikalwAle	2.01	ek Trak nikalA	1.47
	sAikal	2.84	Trak gayA	1.76	niilii ra.ng kI	1.25
HUMAN	saD.ak	7.50	krOs kar	3.93	krOs kar rahA	3.04
	krOs	6.68	ek Tempo	2.68	roD krOs kar	1.46
	roD	6.54	roD krOs	2.52	lAl sharT me.n	1.16
CAR	kAr	7.76	ek kAr	4.89	bAe.N se dAe.N	1.41
	gADI	3.99	ek gADI	2.31	kAr jA rahI	1.12
	nikalii	2.81	krOs kar	1.44	krOs kar rahA	1.11

Table 4.2: **Word-level Associations**: Top3 k-grams (k = 1 to 3) for six ground-truth categories according to Conditional Probability with top1000 removed using all-in-one dataset.Appropriate labels are discovered as top-1 unigrams for four of the categories and within top-3 unigrams for one more category.

4.2 Experimenting with Linguistic Unit (L)

4.2.1 Word-level Association (W)

In Word level association, we assume that the word boundaries are known to the system before learning the labels for object categories. So, we consider each k-gram of words to be separate linguistic unit for the association. Table 4.2 shows the top-3 labels for each of six ground-truth categories (G+) of objects (obj) for k = 1 to 3 separately according to Conditional probability of a label l given a concept c. As most frequent words in the general context are likely to be non-relevant to the current context, we remove 1000 most frequent words (T+) from consideration during association.

As can be seen from Table 4.2, the appropriate labels *Tempo* (tempo), *sAikal*

(bicycle), *Trak* (truck) and *kAr* (car) are discovered as top 1-grams for the categories TEMPO, BICYCLE, TRUCK and CAR respectively. A label like *bAik* also appears among top-3 1-grams for the category MOTORCYCLE. A label like *gADI* (car) which is synonymous to *kAr* also appears as the second most strongest label for the category CAR. Among 2-grams, phrases like *ek sAikal* (a bicycle), *ek Trak* (a truck), *ek kAr* (a car) appear to have the strongest association for the categories BICYCLE, TRUCK and CAR. The phrases like *Trak gayA* (truck went), *krOs kar* (crossing) describing the motions of the vehicles appear among top-3 2-grams . The phrases like *dAe.N se bAe.N* (right to left), *bAe.N se dAe.N* (left to right) appear among top-3 3-grams indicating the direction of motion.

4.2.2 Poly-syllabic Association (S)

In poly-syllabic association, we first merge the words across word-boundaries and then try to associate poly-syllabic sequences (referred to as *s-word* to mean "syllabic word) with the visual concepts. The notion of a syllable is explained in section 2.6. We find the association of all possible poly-syllabic words of length k in a continuous utterance without assuming the knowledge of word-boundaries. The poly-syllabic word having the strongest association with a category c is considered to be the label for that category.

Table 4.3 shows the top3- k-grams for $k = 2$ to 4 according to conditional probability. Here also, we remove 100 most frequent k-grams at syllabic level from consideration. The appropriate word *Trak* appears as the strongest label for the category TRUCK among 2-grams at syllabic-level. Also, the labels like *sAikal* (bicycle) , *ekTrak* (a truck) and *ekkAr* (a car) are discovered as the top 4-grams. As can be seen, the labels *sAi*, *ik*, *kal* at 2-gram level and *sAik*, *ikal* at 3-gram level are nothing but the parts of appropriate label *sAikal* for BICYCLE. The label *kAr* does not appear in top-3 labels at 2-gram level for CAR because it is among the 100 most frequent 2-grams in general corpus and hence is not considered for the association. The labels like *jArahAhai* (is going) appear among top-3 k-grams ($k = 4$) indicating the motion of the vehicle. This shows that the results of label association hold even in the absence of knowledge of word-segmentation. This in some sense is an indication of the fact that the knowledge of word-segmentation is not a prerequisite to the word learning.

4.2.3 Phrase-level Association

By looking at Table 4.2, we can note that, for the category of TRUCK, *Trak*, *ek Trak* and *se ek Trak* appear as the top-most 2-gram, 3-gram and 4-gram of words respectively. Also, from Table 4.3, we can note that *ik*, *sAik* and *sAikal* appear as the top-most 2-gram, 3-gram and 4-gram of syllables respectively for the category BICYCLE. Now the question is which among these labels is a true label for the particular category. From the nature of

(S, CP, T+, G+, A+, obj, ALL)						
	k = 2		**k = 3**		**k = 4**	
Concept (c)	l	**CP**	l	**CP**	l	**CP**
TEMPO	ik	12.23	taraf	6.81	sAikal	5.79
	jAr	9.23	rek	6.45	aurek	5.22
	kal	8.76	sAik	6.32	jArahAhai	4.52
BICYCLE	ik	3.4	sAik	3.06	sAikal	2.9
	sAi	3.06	ikal	2.9	eksAi	1.3
	kal	2.91	eksA	1.3	ksAik	1.3
MOTORCYCLE	ik	19.09	sAik	10.54	sAikal	9.24
	jAr	13.43	ikal	9.24	rsAik	7.21
	sAi	12.41	bAik	8.88	jArahAhai	6.02
TRUCK	Trak	19.23	ekTra	11.83	ekTrak	11.83
	kTra	11.83	kTrak	11.83	sAikal	6.41
	jAr	10.2	rahehai.n	8.61	jArahAhai	4.67
HUMAN	hIhai	14.37	saD.ak	7.66	sAikal	6.35
	jAr	10.86	aAdamI	7.62	jArahAhai	6.29
	kal	10.78	taraf	7.26	ekaAd	5.2
CAR	kkA	5.32	ekkA	5.32	ekkAr	5.15
	jAr	4.51	kkAr	5.15	ekaur	2.63
	hIhai	4.33	rahIhai	4.33	jArahIhai	2.46

Table 4.3: **Poly-syllabic Associations**: Top3 k-grams ($k = 2$ to 4) for six ground-truth categories according to Conditional Probability with top100 k-grams removed using all-in-one data-set. Appropriate labels appear to have strongest association for three object categories.

labels, it is clear that top-most labels at lower-level are generally substrings of the top-most labels at higher level. Analyzing association strengths of the above units w.r.t appropriate categories, one can say that most of the times *ik* and *sAik* co-occur with BICYCLE as a part of larger unit *sAikal* (because association strengths of the three are very close). On the other hand, as association strengths of *Trak* and *ek Trak* w.r.t TRUCK differ largely, there are many occasions where *Trak* co-occurs with TRUCK independently and not as a part of *ek Trak*.

Fragment analysis

We say that a k-gram l_k is a fragment with respect to an n-gram l_n $(n > k)$ for a category c if l_k is contained in l_n and $\frac{M(l_n,c)}{M(l_k,c)} > \tau$ where $\mathbf{M(l, c)}$ is the association-value between label l and category c and $0 < \tau < 1$ is some threshold.

We call l_k a fragment of l_n because most of the occurrences of l_k are also the occurrences of l_n where l_k is a part of l_n.

Unit Independence conjecture

A smaller k-gram l_k is independent of a higher n-gram l_n w.r.t. a concept c if l_k is not a *fragment* of l_n. Only those smaller k-grams l_k which are independent of all higher n-grams l_n w.r.t a concept c can be labels for c.

In other words, we can say that a smaller unit is independent of a larger unit containing lower unit as its substring w.r.t to some concept if the association of the two w.r.t that concept differ considerably. Otherwise, the lower unit is not independent one and hence can not be a label for that concept.

In subsection 4.2.1 and subsection 4.2.2 association, we assumed the size of linguistic unit to be fixed. However, the size of the label may vary from one category to the another. In fact, according to Cognitive grammar proposed by [18], any sequence (of any length) of phonemes can be a word provided that it is sufficiently usage-entrenched. To allow the candidate labels for an object category to be of any size and to make it more meaningful in the cognitive grammar view, we combine all k-grams for $k = 1$ to 4 together and associate each such k-gram or phrase with the object categories. The association is similar to the word-level and poly-syllabic association except for the fact that we identify fragments and remove them from consideration focusing only on non-fragments or independent units.

Associating word phrases

Table 4.4 shows the top3 phrases according to conditional probability and mutual information after removing 1000 most frequent words. Here we set $\tau = 0.9$. The appropriate labels for BICYCLE, TRUCK and CAR appear as the topmost labels whereas *bAik* appears

(P_w, **CP / MI , T+, G+, A+, obj, ALL**)				
	CP		**MI**	
Concept (c)	l	$M(l,c)$	l	$M(l,c)$
TEMPO	Tempo	4.46	kAr	7.41
	kAr	4.33	bAik	7.34
	pe	4.25	Tempo	6.54
BICYCLE	sAikal	1.95	sAikal	1.34
	ek sAikal	1.14	ek sAikal	0.96
	moTarsAikal	0.79	gais silinDar	0.53
MOTORCYCLE	pe	8.60	pe	12.88
	bAik	7.12	bAik	11.64
	Tempo	6.56	skUTar	8.99
TRUCK	Trak	17.29	Trak	15.01
	ek Trak	10.67	ek Trak	9.91
	pe	3.24	tIn sAikalwAle	2.37
HUMAN	saD.ak	7.50	saD.ak	27.90
	krOs	6.68	krOs	20.76
	roD	6.54	roD	18.19
CAR	kAr	7.76	kAr	9.30
	ek kAr	4.89	ek kAr	6.61
	gADI	3.99	gADI	4.38

Table 4.4: **Phrase-level Associations**: Top3 word k-grams ($k = 1$ to 4 combined) for all-in-one with top1000 removed. Four of the categories have appropriate terms as top-most label according to CP.

(P_s, **CP / MI, T+, G+, A+, obj, ALL**)				
	CP		**MI**	
Concept (c)	l	$M(l,c)$	l	$M(l,c)$
TEMPO	ik	12.23	ik	29.68
	jAr	9.23	jAr	21.35
	kal	8.76	kal	19.70
BICYCLE	sAikal	2.90	sAikal	2.81
	jAr	1.62	eksAi	1.60
	eksAi	1.30	ksAik	1.60
MOTORCYCLE	ik	19.09	ik	39.35
	D	15.08	D	28.61
	jAr	13.43	Tar	26.42
TRUCK	Trak	19.23	Trak	22.55
	ekTrak	11.83	ekTrak	14.70
	jAr	10.20	jAr	9.42
HUMAN	hAhai	14.37	hAhai	62.35
	D	13.85	D	53.54
	jAr	10.86	wAlA	46.14
CAR	ekkAr	5.15	ekkAr	9.38
	jAr	4.51	gADI	6.12
	rahIhai	4.33	rahIhai	5.05

Table 4.5: **Syllabic phrase level Association**: Top3 k-grams combined for $k = 1$ to $k = 4$ according to MI after removing top100 k-grams. Three of the object categories have strongest association with appropriate terms

as the second label for MOTORCYCLE. The label *Tempo* also appears as the topmost label for the category TEMPO according to conditional probability. This result is similar to the results of word-level association. However, using phrase-level association relieves the system from unnecessary assumption of fixed-length label allowing variable length phrases to be discovered as the labels for visual categories. The visual categories we are dealing with, however, do not have a phrasal label to show the strength of this approach.

Syllabic phrase association

Table 4.5 shows the top3 Poly-syllabic phrases for different object categories with top100 k-grams removed. Here, we set $\tau = 0.75$ considering the larger number of frequent combinations possible at syllabic-level. Appropriate labels *sAikal*, *Trak* and *ekkAr* appear as the strongest labels for the categories BICYCLE, TRUCK and CAR respectively. In addition

(W, M*, T-, G+, A+, obj, ALL)						
	DJ		**CP**		**MI**	
Concept (c)	l	$M(l,c)$	l	$M(l,c)$	l	$M(l,c)$
TEMPO	hai	1.11	hai	21.62	hai	21.50
	aur	0.90	aur	15.74	aur	17.04
	jA	0.52	se	10.35	se	10.32
BICYCLE	gais	0.03	ek	2.58	sAikal	0.86
	sAikal	0.02	hai	1.96	gais	0.45
	uspar	0.02	sAikal	1.95	silinDar	0.32
MOTORCYCLE	ek	1.40	ek	36.83	ek	31.58
	hai	1.22	hai	30.37	hai	28.32
	skUTar	0.76	aur	15.80	jA	14.90
TRUCK	Trak	0.41	ek	30.99	ek	12.21
	ek	0.25	hai	21.93	Trak	12.11
	Ta.Nkar	0.21	Trak	17.29	hai	8.52
HUMAN	hai	5.84	ek	34.11	hai	62.17
	ek	5.39	hai	31.32	ek	58.05
	rahA	3.61	aur	15.64	rahA	33.77
CAR	kAr	0.40	ek	18.20	kAr	7.43
	camcamAtI	0.23	hai	12.58	ek	6.29
	mahAshay	0.22	kAr	7.76	hai	4.12

Table 4.6: **Word-level Associations for different probability measures**: Top3 1-grams according to *Dominance weighted Joint Probability*, *Conditional probability* and *Mutual Information* measures without removing top1000 words. All three measures rank appropriate labels within top3 for three object categories. DJ and MI succeed in ranking appropriate labels as top-most labels for two of the categories whereas CP fails to do so.

to this, according to MI, *gADI* also appears as the second label for CAR. The label *kAr* is not appearing for CAR as top100 bigrams contain it as a common frequent bigram and hence gets removed from the consideration.

4.3 Comparing different association measures (M)

Table 4.6 and Table 4.7 show the top-3 labels for different association measures with and without removing top1000 words. As can be seen, dominance weighted joint probability (DJ) and mutual information (MI) have a tendency to favour rare co-occurrences of labels. There is one event in the scene where a person is carrying a gas cylinder over the bicycle. As these kinds of events which are surprising to the viewer are generally

attended to, the labels like *gais* (gas), *silinDar* (cylinder) appear for BICYCLE among top-3 labels according to DJ and MI. The relevant labels *sAikal* (bicycle), *Trak* (truck)and *kAr* (car)appear among top-3 labels according to all three probability measures respectively for the categories BICYCLE, TRUCK and CAR. In fact, MI ranks *sAikal* and *kAr* as the top-most label. Conditional probability (CP), however, ranks labels *ek* (one) and *hai* (is) higher than the relevant labels for these categories. Generally, the descriptions are often of the form *ek Trak jA rahA hai* (a truck is going), *ek kAr jA rahI hai* (a car is going) for the events where vehicles are going in the scene, these labels have high co-occurrence with many of the concepts.

However, after removing top1000 words from consideration, we get rid of the words like *ek* , *hai* as these are frequent in general corpus [6]. The most frequent words in the general corpus can be assumed not to be so much relevant to the current context and hence not considering them for the association is a valid assumption. Table 4.7 shows the top-3 1-gram for different probability measures. As can be seen, all of them rank the relevant labels at the top for the categories TRUCK and CAR. Though MI performs good even without removing top1000 words compared to conditional probability, after removing top1000 words, it fails to rank *sAikal* as the top-most label for BICYCLE. From the general observation we find that DJ and MI are likely to favour rare co-occurrences. The similar behaviour is observed in case of poly-syllabic associations though only word-level associations are shown here. This is the reason why we prefer conditional probability with top1000 words removed over mutual information.

4.4 Comparing results on different datasests (D)

Table 4.8 shows the top-3 1-grams for three different datasets according to Conditional probability respectively. The appropriate labels for BICYCLE, TRUCK and CAR are among top-3 even for these individual datasets. That means results hold even for the smaller datasets. This indicates that even a set of 10 narrations is sufficient to learn the labels for visual categories.

The reason why a relevant label doesn't appear for TEMPO in case of ADULT-1 data-set can be inferred from the analysis of attention model (section 4.7). As can be seen from Figure 4.6, the attention precision is very low for TEMPO in case of ADULT-1.

4.5 From Minimal supervision to totally unsupervised learning

So far we made use of ground-truth information to find coherent object categories. Using ground-truth (Human judgment), we grouped 30 object clusters obtained during un-

(W, M*, T+, G+, A+, obj, ALL)						
	DJ		**CP**		**MI**	
Concept (c)	l	$M(l,c)$	l	$M(l,c)$	l	$M(l,c)$
TEMPO	bAik	0.42	Tempo	4.46	mArutI	2.24
	kAr	0.40	kAr	4.33	bAik	1.92
	piilii	0.36	pe	4.25	kAr	1.72
BICYCLE	sAikal	0.02	sAikal	1.95	silinDar	0.16
	uspar	0.02	moTarsAikal	0.79	I.njin	0.14
	I.njin	0.02	pe	0.63	Dilaks	0.14
MOTORCYCLE	skUTar	0.76	pe	8.60	bAik	4.65
	bAik	0.70	bAik	7.12	pe	4.43
	a.ndar	0.54	Tempo	6.56	skUTar	4.30
TRUCK	Trak	0.41	Trak	17.29	Trak	6.22
	Ta.Nkar	0.21	pe	3.24	peTrol	1.47
	peTrol	0.16	sAikal	2.84	Ta.Nkar	1.13
HUMAN	saD.ak	2.74	saD.ak	7.50	saD.ak	12.51
	biThAke	2.12	krOs	6.68	krOs	7.06
	rikshAwAlA	1.84	roD	6.54	biThAke	5.81
CAR	kAr	0.40	kAr	7.76	kAr	3.61
	camcamAtI	0.23	gADI	3.99	gADI	1.46
	mahAshay	0.22	nikalii	2.81	nikalii	1.33

Table 4.7: **Word-level Associations for different probability measures with top1000 removed**

Word-level Associations for different probability measures with top1000 removed: Top3 1-grams according to *Dominance weighted Joint Probability*, *Conditional probability* and *Mutual Information* measures after removing top1000 words. DJ and MI are found to favour the rare co-occurrences whereas CP favours sufficiently co-occurring labels. All three measures ranked appropriate labels at the top for four of the object categories

(W, CP, T+, G+, A+, obj, D*)						
	ADULT-1		**ADULT-2**		**CDS**	
Concept (c)	l	**CP**	l	**CP**	l	**CP**
TEMPO	pe	4.75	Trak	6.13	kAr	5.23
	nikalA	4.07	kAr	5.77	moTarsAikal	4.45
	Trak	2.91	Tempo	5.71	OTo	4.37
BICYCLE	sAikal	1.3	sAikal	3.02	rikshA	0.95
	pe	1.3	moTarsAikal	1.38	sAikal	0.84
	rikshA	0.65	sAiD	0.71	krOs	0.55
MOTORCYCLE	pe	8.18	Tempo	10.95	pe	10.52
	roD	7.31	bAik	8.66	OTo	9.29
	bAik	5.84	pe	7.59	sAmAn	7.28
TRUCK	Trak	7.77	Trak	22.48	Trak	23.93
	Tempo	3.76	pe	5.07	roD	3.18
	sTrIT	3.45	sAiD	5.07	sAikal	3.13
HUMAN	saD.ak	8.07	Tempo	8.34	saD.ak	7.34
	krOs	8.03	pe	7.97	pe	6.46
	rikshewAlA	5.66	roD	7.79	roD	6.44
CAR	nikalii	4.35	kAr	10.74	kAr	6.07
	kAr	4.35	pe	3.68	gADI	5.37
	gADI	2.91	gADI	3.68	nikalii	3.05

Table 4.8: **Word level Associations for different datasets**: Top3 1-grams according to conditional probability for ADULT-1, ADULT-2 and CDS datasets (After removing top1000 words). Appropriate labels are discovered as top-most label for many of the object categories even for individual datasets despite smaller data size.

(W, CP / MI, T+, G-, A+, obj, ALL)				
Cluster (Concept)	l	**CP**	l	**MI**
C0 (T)	kAr	4.98	bAik	2.4
	bAik	4.96	moTarsAikal	2.13
	Tempo	4.5	kAr	2.01
C8 (M)	bAik	14.22	skUTar	4.57
	skUTar	12.66	bAik	3.7
	pe	12.53	pe	2.27
C15 (B)	sAikalwAle	8.82	sAikal	3.64
	sAikal	7.12	sAikalwAle	3.63
	dAe.N	6.85	dAe.N	1.67
C19 (C)	kAr	8.27	kAr	3.78
	gADI	4.05	gADI	1.4
	nikalii	2.85	nikalii	1.27
C22 (M)	roD	6.82	roD	0.41
	pe	2.68	khAlI	0.3
	skUTar	1.92	laDkI	0.29
C25 (T)	Tempo	18.33	Tempo	5.62
	pe	11.75	mUD	3.05
	sAikal	6.87	pe	2.75
C28 (B)	Tempo	12.36	Tempo	3.02
	sAikal	8.48	sAikalwAle	3.01
	sAikalwAle	6.27	mUD	1.83
C29 (L)	Trak	26.4	Trak	4.83
	pe	8.02	sAmAn	1.51
	sAmAn	5.95	Ore.nj	1.25

Table 4.9: **Word-level Association without ground-truth**: Top3 1-gram for some of 30 object clusters according to CP and MI (After removing top1000 words). Four of the clusters could get appropriate terms as top-most labels whereas four other clusters have appropriate terms within top-3 labels.

supervised object discovery into 7 different categories. This step brings some supervision into the model. To assess if we can get rid of this minimal supervision and make our approach totally unsupervised, we tried to associate labels with 30 clusters obtained directly from clustering of objects without using ground-truth categories. Considering 30 clusters separately reduces the coherency of the object categories as clusters that can be grouped together according to ground-truth are now considered as different object categories. Despite this fact, we could learn appropriate labels for some of the 30 clusters.

Table 4.9 show the top3 words for selected clusters after removing top1000 words. Only those clusters for which the appropriate labels could be found are shown in the results. We are able to learn appropriate labels as shown in Table 4.9 for four of the clusters: *bAik* for cluster C8 of BICYCLE, *kAr* for cluster C19 which is the only cluster of CAR, *Tempo* for cluster C25 of TEMPO and *Trak* for cluster C29 which is one of the two clusters of TRUCK. Appropriate labels *Tempo* for C0, *skUTar* for C22, *sAikal* and *sAikalwAle* for C15 and C28 appear among top3 labels. In fact, according to MI, *sAikal* is the top-most label for the cluster C15 whereas. Top3 labels for C8 also contain *skUTar* which is synonymous to top-most label *bAik*. Similarly, *gADI*, a synonym for *kAr* also appears in top3 labels for cluster C19. So, we are able to get relevant terms in top3 labels in case of 8 out of 30 clusters.

This shows that the use of ground-truth helps in learning appropriate labels by providing required coherence for the object categories. However, labels can be learnt even in the absence of such a ground-truth. If the object clusters of sufficiently high coherency can be obtained with improved object models and better clustering techniques, then we may be able to reproduce good results even without making use of ground-truth thus avoiding the minimal supervision involved in the process.

4.6 Incremental Analysis

In subsection 4.2.1, subsection 4.2.2 and subsection 4.2.3, we found that the appropriate labels for some of the object categories especially, BICYCLE, TRUCK and CAR appeared as the label with highest association. However, we can also note that these labels were also found to have highest association with appropriate categories even when smaller number of narrations were considered as shown in Table 4.8. So, the obvious question is whether these labels are really acquired and if acquired then at which point of time during the process. Word learning is not a one-stage-process but a continuous one. But can we quantify the notion of word-learning so that we can answer when a word can be said to be learnt? With this in mind, we tried to analyze the status of object-word associations after every k- narrations for $k = 1, 2, ..., n$ where n is the total number of narrations in the data-set. This process of incremental analysis is equivalent to the incremental word learning except that at every step k, the results are obtained considering all k narrations together

and not by updating the associations of step $k-1$ considering only k^{th} narration.

4.6.1 Effect of increasing usage on Label learning

We experiment with ADULT-2 data-set having 20 narrations and consider the association of various labels with various categories by considering k- narrations incrementally for $k = 1, 2, ..., 20$.

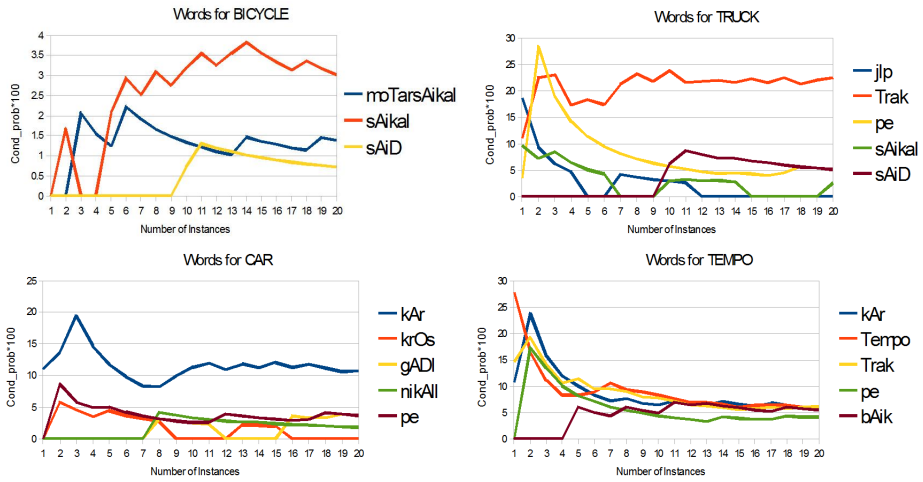

Figure 4.1: **Increasing usage**: Effect on word-level associations. The words *sAikal*, *Trak* and *kAr* appear to have stable associations with respect to categories BICYCLE, TRUCK and CAR whereas *Tempo* is not stable with respect to TEMPO

.

Figure 4.1 shows the effect of increasing usage on word-level association for some words with respect to various object categories. As can be seen, the association strengths (conditional probability on Y-axis) of appropriate labels *sAikal*, *Trak* and *kAr* respectively for categories BICYCLE, TRUCK and CAR are sufficiently high and consistent after first few narrations. There is some competition observed for the category BICYCLE for two labels *moTarsAikal* and *sAikal* for first few commentaries. In case of TEMPO, there is no single label which has dominating association strength. In fact, many labels are competing throughout the process and even after 20 narrations, there is no clear winner. So, we can say that the labels *sAikal*, *Trak* and *kAr* have established themselves as the labels for the categories BICYCLE, TRUCK and CAR respectively. However, the label *Trak*, despite being the topmost label, can not be said to have established itself as the label for TEMPO due to its instability.

Similar results are shown for poly-syllabic level association in Figure 4.2. The labels *sAikal* and *ekTrak* are dominating other labels consistently respectively for BICYCLE and TRUCK. In case of CAR, however, the label *ekkAr* doesn't seem to have a clear dominance over other labels. So, perhaps it is not yet established as a label for the category.

Note that only few important labels are shown for various categories in Figure 4.1 and Figure 4.2. Also, "zero" value on Y-axis indicates that either the association value is

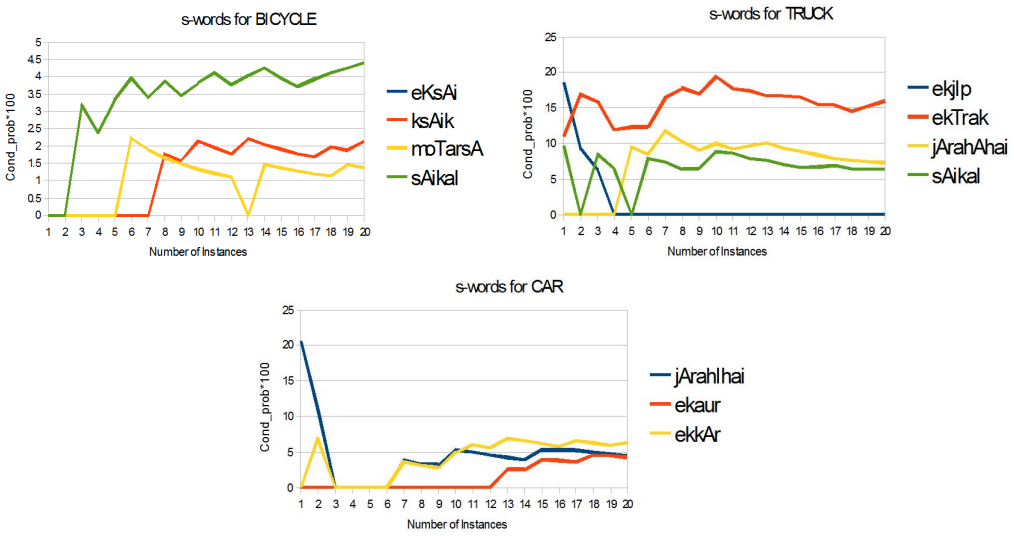

Figure 4.2: **Increasing usage**: Effect on poly-syllabic associations. The labels *sAikal*, *ekTrak* appear to have stable associations with respect to categories BICYCLE and TRUCK whereas *ekkAr* is not stable with respect to CAR

"zero" or the label is not among top 10 labels for that category at that point.

4.6.2 Random ordering and stability of label learning

In the above analysis, we used a fixed order of narrations. However, the stability of the labels may get affected if we consider the narrations in some other order. To confirm that stability of acquired labels is not incidental, we perform the same incremental analysis over random orders of randomly selected set of narrations. For this matter, we choose 9 narrations of ADULT-2-1 data-set and 6 narrations from CDS data-set. We experiment with three different random orderings of these 15 narrations. In random-order-1, we first randomly order 9 narrations from ADULT-2-1 followed by randomly order 6 narrations from CDS data-set. So, in this order, all narrations in ADULT-2-1 are preceding to the CDS narrations. In random-order-2, we randomly ordered all 15 narrations so that narrations from ADULT-2-1 and narrations from CDS may alternate. In random-order-3, we randomly order narrations but with a condition that all CDS narrations should precede the ADULT-2-1 narrations.

Figure 4.3 shows the effect of increasing usage for category TRUCK over the set of 15 narrations for three different random orders and their average effect. Figure 4.4 shows the effect of increasing usage averaged over the three random orders for the categories BICYCLE, TRUCK and CAR. As shown by these results, we can conclude that the labels *sAikal*, *Trak* and *kAr* are indeed stable enough for the categories BICYCLE, TRUCK and CAR respectively. This also confirms the robust acquisition of these labels for the corresponding concepts.

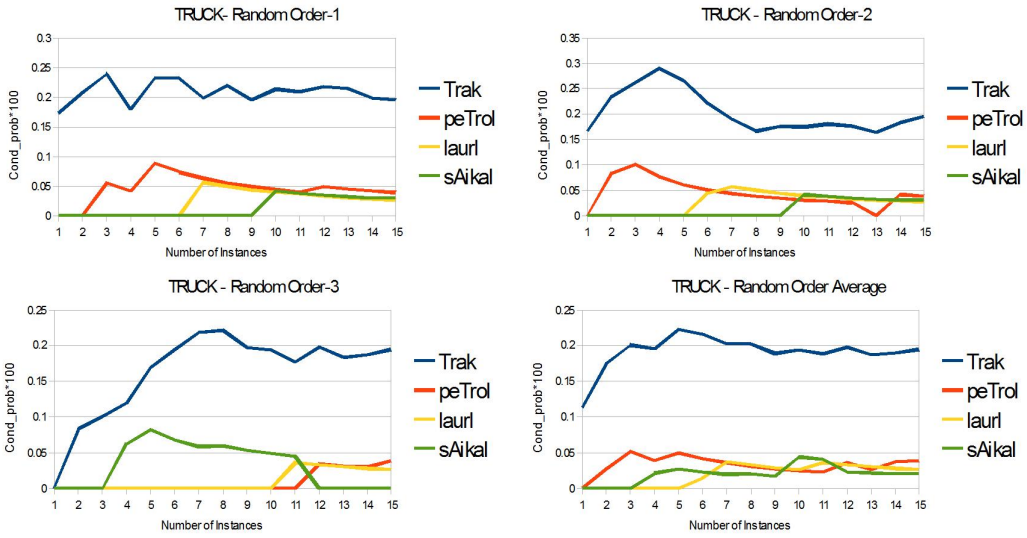

Figure 4.3: **Random usage**: Effect on word-level associations for TRUCK. *Trak* is consistently dominating other labels for TRUCK after few initial narrations

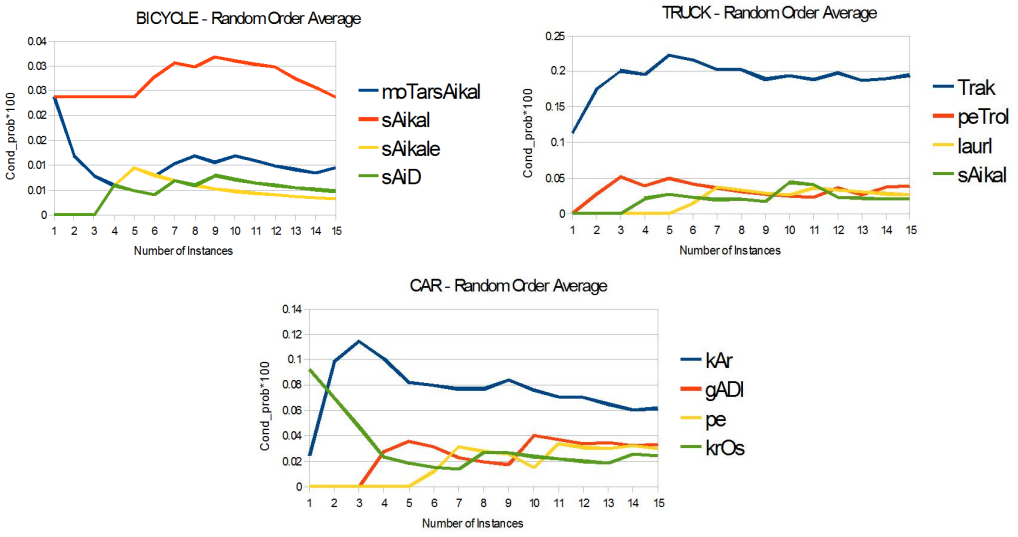

Figure 4.4: **Average Random usage**: Average Effect on word-level associations

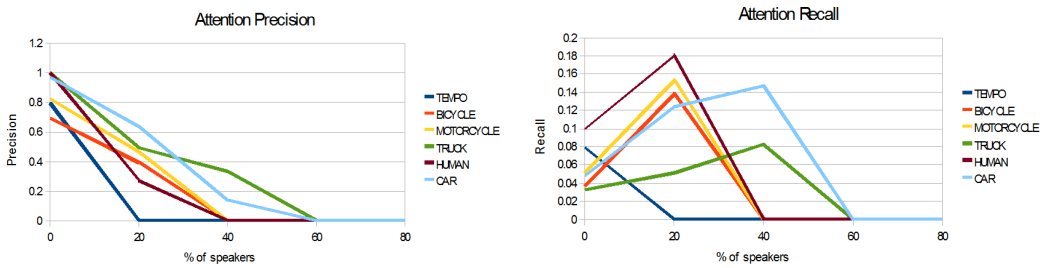

Figure 4.5: **Attention Precision and Recall**: Precision and Recall Vs % of speakers. Precision is good for TRUCK and CAR even for large fraction of subjects indicating that most of the subjects simultaneously attend to trucks and cars.

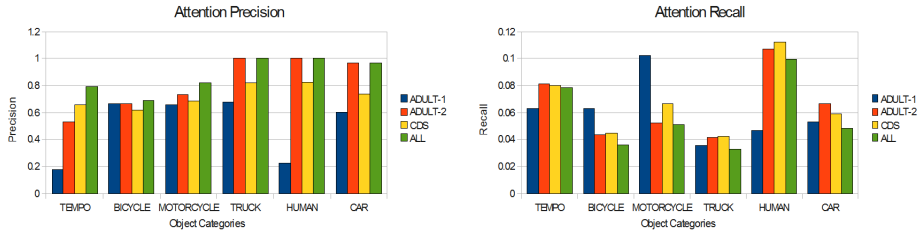

Figure 4.6: **Attention Precision and Recall**: Precision and Recall for various datasets

4.7 Evaluating the attention model (A)

We evaluate the attention model in two ways:

1. We try to evaluate how good the attention model conforms to human attention. To evaluate this performance of attention model, we analyze the Attention-Precision and Attention-Recall of various categories.

2. We try to evaluate the need of attention model in order to learn the language labels. For this, we compare the results of label association with and without using the attention model.

We call a concept c to be visually salient at time t if there exists an agent a belonging to concept c such that a is predicted to be the most salient at time t by the attention model. Also, we call a concept c to be linguistically salient at time t if there exist more than x% of speakers who utter a label relevant to concept c at time t.

Attention-Precision $P_A(c)$ for concept c is defined as the ratio of number of frames in which a concept c is both visually and linguistically salient to the number of frames in which a concept c is visually salient. Also, Attention-Recall $R_A(c)$ for concept c is defined as the ratio of number of frames in which a concept c is both visually and linguistically salient to the number of frames in which a concept c is linguistically salient.

We plot the Attention-Precision and Attention-Recall of various object concepts for various values of x, the percentage of speakers. Figure 4.5 shows the plots for Attention-Precision and Attention-Recall over entire data-set of 44 narrations. As can be seen, for

small values of x i.e. when a small % of speakers is considered for linguistic saliency, the precision of attention model is quite good. In fact, high precision values (near to 1) for TRUCK ,CAR and HUMAN when $x = 0$ indicate that there is at least one speaker in the set of speakers who talks about the object predicted to be visually most salient. However, as x increases, precision values decrease sharply as having attention of speakers on the same object becomes rare when number of speakers increase. So, attention model seems to be a good proposition when number of speakers is small say 8-10. However, considerable precision of TRUCK and CAR even for $x = 40$ indicate that there are large number of speakers who attend to the same object simultaneously and this object is visually most salient. So, the prediction of attention model for TRUCK and CAR seems to be good. The low values of recall suggest that the attention model fails to predict the speaker's attention many times. Also, talking about the objects even after the object moves out of the scene may have resulted into low recall by making denominator large. The reason why recall increases with the increase in x (% of speakers) is that as x increases number of frames in which a concept is linguistically salient decreases (many people talking about the same object simultaneously becomes rare as number of people increases).

Also, Figure 4.6 shows the Attention-Precision and Attention-Recall values of object categories for various datasets when $x = 0$.

To assess whether the use of attention model helps to learn the labels, we compare the word-level associations with and without using the attention model. When we use the attention model, the words in the narratives are associated only with the co-occurrent objects which are most salient according to attention model. When experimenting without attention model, we assume that every object in the visual scene to be equally salient and associate words in the narratives with every co-occurrent object in the video.

Table 4.10 shows the top3 1-gram of words for each of object categories with and without using the attention model. As can be seen, even without using attention model, appropriate words *sAikal*, *Trak* and *vain* are discovered for the categories BICYCLE, TRUCK and CAR respectively. Moreover, for the category CAR, all top3 labels are relevant. Even the results for the other categories are similar. Additionally, it can be noted that when attention model is not used, the label *Tempo* has much stronger association with TEMPO as compared to its competing labels *pe* and *OTo*. When the attention model was used the association strengths of all top3 labels for TEMPO were close showing that none of these labels was dominant. This shows that the use of attention model is not necessary for learning the labels for visual categories. In fact low attention-recall values suggest that, many times none of the objects in the scene are considered salient according to attention model. This makes much of the linguistic information unusable. With no attention model this information can be used to some extent.

(W, CP, T+, G+, A*, obj, ALL)				
	With Attention (A+)		**Without Attention (A-)**	
Concept (c)	l	**CP**	l	**CP**
TEMPO	Tempo	4.46	Tempo	9.71
	kAr	4.33	pe	5.79
	pe	4.25	OTo	5.51
BICYCLE	sAikal	1.95	sAikal	1.63
	moTarsAikal	0.79	moTarsAikal	0.69
	pe	0.63	pe	0.59
MOTORCYCLE	pe	8.60	pe	7.17
	bAik	7.12	bAik	6.25
	Tempo	6.56	roD	5.55
TRUCK	Trak	17.29	Trak	14.39
	pe	3.24	sAikal	4.40
	sAikal	2.84	pe	3.87
HUMAN	saD.ak	7.50	krOs	6.76
	krOs	6.68	roD	6.68
	roD	6.54	saD.ak	6.32
CAR	kAr	7.76	vain	7.89
	gADI	3.99	kAr	7.73
	nikalii	2.81	gADI	5.68

Table 4.10: **Word-level Association with and without attention model**: Top3 1-gram for some of object categories with and without using attention model. Appropriate labels are discovered for four of object categories irrespective of whether attention model is used or not. This suggests that the attention model is not required for learning labels

4.8 Learning labels for trajectories

We try to associate linguistic labels with seven trajectory clusters discovered. The process of label association is the same as it was in case of object-label association. Table 4.11 shows the top3 3-grams of words according to conditional probability for each of the seven clusters of trajectory with and without removing top1000 words. As can be seen, when we consider all 3-grams of words without removing top1000 words, most of the clusters got associated with units like *jA rahA hai*, *jA rahI hai*, *jA rahe hai.n* (is/are going) as all the trajectory clusters represent some kind of motion. A label like *bAe.N se dAe.N* (left to right) also appears at third position for the cluster C3.

After removing top1000 words, we get rid of these commonly occurring descriptions and *bAe.N se dAe.N* appears as the strongest label for cluster C3. Also, labels like *roD krOs kar*, *krOs kar rahA* appear for cluster C5 indicating the trajectory CROSS (C). We can note that cluster C5 contains many agents with ground-truth category CROSS (C) (Table 2.2). A label like *geT kI taraf* (towards the Gate) appears as the top-most label for C7. Many vehicles of the category TURN (T) are those which turn towards down from the middle of the video. As many narrators could predict the location where video was shot (though they were not explicitly told) and knew that there is gate of an educational institute towards the downside of the video scene, a description like *geT kI taraf* is relevant one. The descriptions like *aur ek bAik* (and a bike) and *ek aur Tempo* (one more tempo) are due to the presence of descriptions of objects which are inherently associated with trajectories.

Table 4.12 shows the top3 3-grams for seven trajectory clusters with top1000 words removed using ADULT-2-1 data-set. As can be seen, the label *bAe.N se dAe.N* (left to right) appears as the strongest label for clusters C2 and C3 whereas the label *dAe.N kI taraf* appears among top3 for clusters C5 and C6. The label *geT kI taraf* appears as the strongest label for cluster C7. The reason for better results with data-set ADULT-2-1 can be inferred from Table 3.5. One can note that the labels *bAe.N se dAe.N* and *dAe.N kI taraf* are profound in this data-set as compared to other datasets. Also, the reason for not learning the appropriate label like *dAe.N se bAe.N* (right to left) or *lefT kI taraf* (towards left) for trajectory categories C1 and C4 (representing RL) can be the low frequency of such labels in the corpus. In other words, the event of "vehicles going from right to left" is not commented as often as the event of "vehicle going from left to right".

Though the overall results are not very exciting, the success in getting phrases like *jA rahA hai*, *bAe.N se dAe-N* as top3 labels shows the ability of the system to learn verb phrases and motion directives. With improved object tracking and better techniques of trajectory and activity recognition, it should be possible to learn more verbs and motion directives with their associated action schema.

(W, CP, T*, G-, A+, traj, ALL)				
	Top 1000 retained (T-)		**Top 1000 removed (T+)**	
Concept (c)	$k = 3$	**CP**	$k = 3$	**CP**
C1	jA rahA hai	2.84	aur ek bAik	0.79
	jA rahe hai.n	1.52	krOs kar rahA	0.78
	jA rahI hai	1.39	geT kI taraf	0.61
C2	jA rahA hai	3.96	krOs kar rahA	2.24
	kar rahA hai	2.76	lAl sharT me.n	1.17
	jA rahI hai	2.31	roD krOs kar	0.92
C3	jA rahA hai	6.49	bAe.N se dAe.N	2.12
	jA rahI hai	2.39	pUch rahA hai	1.47
	bAe.N se dAe.N	2.12	pAr hotI hai	1.24
C4	jA rahA hai	3.35	roD krOs kar	2.18
	jA rahI hai	3.15	krOs kar rahA	2.16
	kar rahA hai	3.03	saD.ak krOs kar	0.84
C5	jA rahA hai	3.92	roD krOs kar	2.43
	jA rahI hai	3.62	krOs kar rahA	2.38
	kar rahA hai	2.95	geT kI taraf	1.54
C6	jA rahe hai.n	3.19	ek aur Tempo	0.75
	jA rahI hai	2.44	blaik kalar kI	0.74
	jA rahA hai	1.70	pe ek aAdamI	0.66
C7	jA rahI hai	5.28	geT kI taraf	1.36
	jA rahA hai	4.47	TI ke geT	1.20
	jA rahe hai.n	2.95	ke geT kI	1.12

Table 4.11: **Word-level Association for trajectory clusters**: Top3 3-gram for seven clusters of trajectories with and without removing top1000 words. Appropriate phrases are discovered for clusters C3 and C7 when top1000 words are removed.

(W, CP, T+, G-, A+, obj, ADULT-2-1)		
Trajectory	$k = 3$	**CP**
C1	purI khAlii hai	1.71
	saD.ak pUrI khAlii	1.71
	kuch log bAiks	1.43
C2	bAe.N se dAe.N	3.16
	lAl sharT me.n	2.73
	sharT me.n lAl	2.73
C3	bAe.N se dAe.N	4.44
	pUch rahA hai	3.96
	ek rikshAwAlA sIn	2.77
C4	roD krOs kar	4.62
	krOs kar rahA	4.47
	krOs kar rahI	2.19
C5	krOs kar rahA	4.67
	roD krOs kar	4.2
	dAe.N kI taraf	3.28
C6	kuch log roD	2.2
	dAe.N kI taraf	2.18
	ek kAlii gADI	1.82
C7	geT kI taraf	3.57
	Ai Ai TI	3.57
	Ai TI ke	3.06

Table 4.12: **Word-level Association for trajectories**: Top3 3-gram for 7 clusters of trajectories after removing top1000 words for ADULT-2-1 DATA-SET. Clusters C2, C3 and C7 have appropriate phrases as top-most labels.

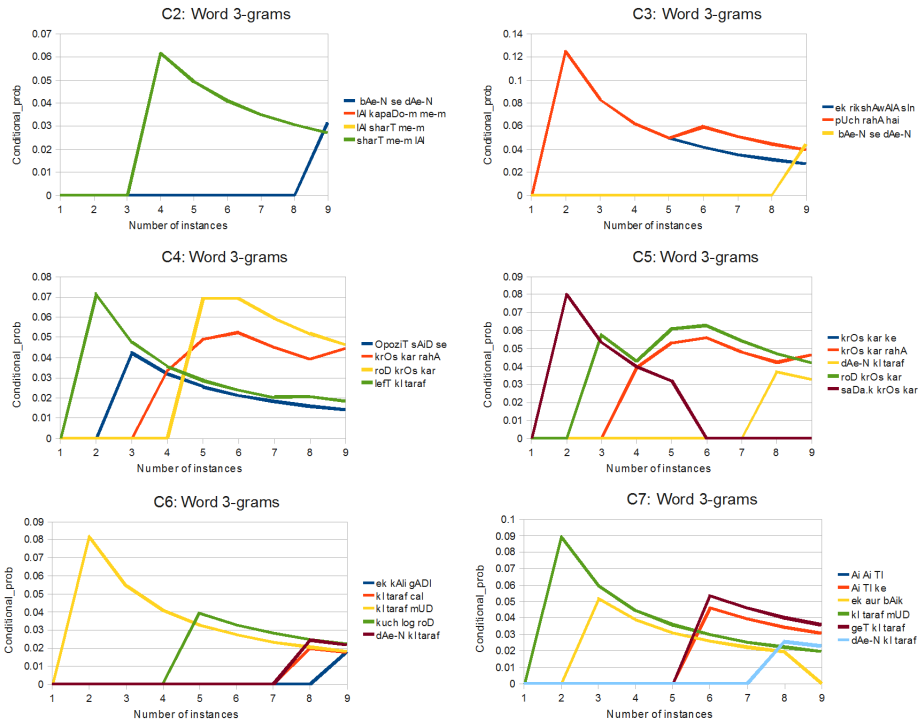

Figure 4.7: **Incremental Analysis of Trajectory labels**

4.8.1 Incremental analysis of trajectory labels

To confirm whether labels discovered as the strongest associations for the trajectory categories are really learnt or not, we performed incremental analysis of the associations of top few labels in similar way as we did it for confirming object label learning. Figure 4.7 shows the association strengths of top few labels for some trajectory categories. As can be seen, the top labels discovered for the clusters, C2, C3, C7 are not stable enough and do not show considerable difference as compared to the association strengths of competing labels. So, though, these labels are found to have maximal association, these can not be said to be acquired at this point as they do not show consistent dominance required.

However, given few more set of narrations which are rich in describing motion of the vehicles, the appropriate labels may show stable and sufficiently high association strength for the respective categories. One can ask the narrators to focus on directions of motion and use the narrations thus obtained to learn the relevant motion directives as discovered in the label association here.

4.9 Results Discussion

This chapter discussed the results of label association under varying assumptions and parameters. The results show that the labels *Tempo*, *sAikal*, *Trak* and *kAr* are discovered as the top-most label according to conditional probability measure for the object categories

TEMPO, BICYCLE, TRUCK and CAR respectively. Also, in case of poly-syllabic associations, we show that the labels *sAikal*, *ekTrak*, *ekkAr* are discovered for BICYCLE, TRUCK and CAR respectively even without the knowledge of word boundaries. The results hold even when we allow the variable length word-level and poly-syllabic phrases. In all these cases, we removed most frequent units in general corpus assuming them to be non-relevant and less likely to be labels of the visual categories. To assess the confidence of the discovered labels, we analyze their associations with increasing exposure to the narratives. We find that the labels *sAikal*, *Trak* and *kAr* at word-level are consistently dominating other labels showing more confidence in acquisition whereas the label *Tempo* fails to do so resulting in low confidence. Similarly, the incremental analysis of poly-syllabic associations shows that the labels *sAikal* and *ekTrak* have higher confidence in terms of consistent dominance whereas *ekkAr* has lower confidence. The reason for not learning the labels of other object categories is partly due to the poor visual categories. It can be noted from Table 2.1, that the purity of object category TEMPO is quite poor. The category MOTORCYCLE can be described by many labels such as *moTarsAikal*, *bAik*, *skUTar* etc. Same is the case with the category of HUMAN which is often described as *aAdamI* (man), *aurat* (woman) etc. Many times, the humans riding on the vehicles are referred to as *bAikwAlA* (motorcyclist), *sAikalwAlA* (bicyclist) etc. For these reasons the labels for TEMPO, MOTORCYCLE and HUMAN are not learnt.

We also show that the phrases like *bAe.N se dAe.N* and *geT kI taraf* are discoverable for motion categories LEFT-TO-RIGHT and TURN respectively. However, these labels have very low confidence in terms of consistent dominance. The reason why the label for the motion category RIGHT-TO-LEFT is not discovered during label association is that the events of vehicles going right to left are commented rarely as shown by the frequency of a relevant term *dAe.N se bAe.N* in the corpus of narrations (Table 3.5). Also, the concepts of LEFT-TO-RIGHT and RIGHT-TO-LEFT can be described in many ways in Hindi e.g. LEFT-TO-RIGHT can be described as *bAe.N se dAe.N*, *dAe.N kI taraf*, *dAe.N or*, *dAe.N taraf*. Also, *lefT* and *rAiT* are often used even by Hindi speakers for left and right instead of *bAe.N* and *dAe.N*. Due to the multiple possible ways of expressing the motion and not a single phrase being sufficiently used, the system could not gain enough confidence in discovered labels for motion categories.

The evaluation of attention model also shows that the prediction of visual saliency is good for the objects of TRUCK and CAR. However, owing to the similar results obtained for label associations with and without using attention model, we can say that such an attention model may not be necessary for label learning.

The study of different association measures show that conditional probability works reasonably well with most frequent linguistic units removed. Mutual information though works sufficiently good even without removing most frequent linguistic units, its

Variable	**Parameters** (L, M, T, G, A, V, D)	**Top-1**	**Top-3**
L	(W, CP, T+, G+, A+, obj, ALL)	4/6	2/6
	(S, CP, T+, G+, A+, obj, ALL)	3/6	0/6
	(P_w, CP, T+, G+, A+, obj, ALL)	4/6	3/6
	(P_s, CP, T+, G+, A+, obj, ALL)	3/6	1/6
M	(W, CP, T+, G+, A+, obj, ALL)	4/6	2/6
	(W, DJ, T+, G+, A+, obj, ALL)	4/6	3/6
	(W, MI, T+, G+, A+, obj, ALL)	3/6	3/6
T	(W, CP, T-, G+, A+, obj, ALL)	0/6	3/6
	(W, CP, T+, G+, A+, obj, ALL)	4/6	2/6
G	(W, CP, T+, G+, A+, obj, ALL)	4/6	2/6
	(W, CP, T+, G-, A+, obj, ALL)	4/30	5/30
A	(W, CP, T+, G+, A+, obj, ALL)	4/6	2/6
	(W, CP, T+, G+, A-, obj, ALL)	4/6	2/6
V	(W, CP, T+, G+, A+, obj, ALL)	4/6	2/6
	(W, CP, T+, G-, A+, traj, ADULT-2-1)	3/7	2/7

Table 4.13: **Summary of Results**: Number of categories for which the relevant terms appear in top1 and top3 labels in different parametric settings

tendency to prefer rare co-occurrences makes it unreliable especially with smaller data sets. Designing a good association measure which can take care of second property of association measures mentioned may relieve from the need of removing most frequent units.

Table 4.13 summarizes the results of label association. The number of categories for which relevant terms appear in top1 and in top3 are tabulated in Table 4.13 for each of parametric setting.

Chapter 5

Conclusion and Future Work

In this work, we propose a semantics-first approach for word-learning based on (a) Minimally supervised object discovery from a complex 3D-scene (b) Bottom-up attention model and (c) multiple human narrations describing the scene. Given the object categories discovered and visual saliency of these objects over the time, we demonstrate the ability of our system to learn nouns like *sAikal*, *Trak* and *kAr* for the object categories BICYCLE, TRUCK and CAR respectively. We confirm the success in learning words by analyzing the strength of associations with increasing number of narrations. We argue that the consistent dominance of association strength of label with a visual category over the other labels is desirable and can be taken as a confirmation of the word learning. The success in learning appropriate labels even without knowing word-boundaries shows that the knowledge of word boundaries may not be a prerequisite for early word-learning. Moreover, word boundaries can be automatically discovered with the labels learnt for semantic categories. Also, we propose a mechanism based on *fragment analysis* and *unit independence conjecture* to automatically learn labels of appropriate size without assuming a fixed length for the labels.

We also learnt the motion concepts like LEFT-TO-RIGHT, RIGHT-TO-LEFT, TURN and CROSS by clustering the trajectories of the objects discovered. We attempted to learn the labels for these motion concepts. Though labels like *bAe.N se dAe.N*, *geT kI taraf* appear to have highest association strengths with appropriate concepts LEFT-TO-RIGHT and TURN respectively, these labels failed to show the consistent dominance required.

The success of the semantics-first approach in learning words as a simple label association task shows the importance of preverbal conceptual development in simplifying the process of word learning. Also, as we assume no language or domain specific knowledge, the approach is likely to work independent of specific language and specific domain.

The word-meaning pairs thus obtained address many important questions in language understanding, language generation and content-based multimedia retrieval. Using word-meaning pairs, such as discovered in this work, would help to detect the various objects and activities in novel scenes and generate appropriate linguistic descriptions for them.

Also, it is possible to use such paired associations to answer linguistic queries with relevant multimedia documents containing objects and activities specified in the query.

We would like to extend the current work for different visual domains and different languages to prove the domain-independent and language-independent nature of the model claimed here. A larger goal is to integrate models of actions and motion-trajectories with the knowledge of nominals, and begin to attempt to build the kind of defeasible knowledge structures. Further, using the knowledge of nouns, we can replace the synthetic attention model with the linguistic attention model which is more realistic. Typically, if we know that *Trak* refers to the category TRUCK, we can always say that if *Trak* appears in the narration, it is the object of truck present in the scene which is being referred to. Such an attention model can be used to learn the verbs and motion directives from action models. Also, given the label *bAe.N se dAe.N* for LEFT-TO-RIGHT, *kAr* for CAR and a description like *gADI bAe.N se dAe.N jA rahI hai* when there is an object of CAR in the scene, we can infer that *gADI* is a synonym of *kAr* as both represent the same object concept CAR. Thus, using the knowledge learnt earlier we can dig out more and more knowledge of the system to build the knowledge structures rich in semantics.

The model presented here is not an incremental one as we are not updating the association values with each additional instance. However, in human learning, the process of word learning is incremental where each new usage revises the beliefs and the association between linguistic units and the concepts. We would like to extend our idea to model an incremental word learning process which is more realistic than the one presented here. The incremental analysis of label association is one step towards it. Also, we would like to quantify the necessary and sufficient conditions for word-learning. The notion of consistent dominance introduced in this thesis needs to be studied further and a precise mathematical quantification that can allow us to define the point where a label can be said to be learnt for a particular concept.

[5] make a distinction between two knowledge acquisition processes: In *robotic toil*, one learns the symbol grounding using direct sensorimotor experience as in this work, while in *symbolic theft*, knowledge structures are learned from language. For humans, the vast majority of our knowledge structures (and vocabulary) is learned using symbolic theft, though the resulting symbols remain grounded, because some initial symbols were directly grounded in sensorimotor terms. We believe that computational NLP systems must be able to evolve a process for symbolic theft, where a vast number of concepts are acquired from symbolic data, based on a small subset that has been acquired in a directly grounded manner. In the process, the system would build defeasible ontology, without any need for hand-coding. With some initial grounding of nouns and verbs, it should be possible for the system to form higher level concepts using the mechanism of *symbolic theft*. e.g. if we know the grounding of *kAr* as object category CAR, *bAe.N se dAe.N* as LEFT-TO-RIGHT

and *jA rahI hai* as IS-GOING, then it should be possible to infer a scenario of CAR GOING LEFT-TO-RIGHT with a description like "*kAr bAe.N se dAe.N jA rahI hai*".

Bibliography

[1] Renée Baillargeon and Su hua Wang. Event categorization in infancy. *TRENDS in Cognitive Sciences*, 6(2):85–93, February 2002.

[2] Lawrence W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609, 1999.

[3] Paul Bloom. *How Children Learn the Meanings of Words*. MIT Press,Cambridge, 2000.

[4] A. Bosch, A. Zisserman, and X. Munoz. Image Classification using Random Forests and Ferns. In *ICCV'07*, pages 1–8, 2007.

[5] A. Cangelosi, A. Greco, and S. Harnad. Symbol grounding and the symbolic theft hypothesis. In Cangelosi A and Parisi D, editors, *Simulating the evolution of language*, pages 191–210. Springer-Verlag NY, Inc., 2002.

[6] CFILT. Hindi unicode corpus, Center For Indian Language Technology. `http://www.cfilt.iitb.ac.in`, IITB.

[7] N. Chomsky. Syntactic Structures. Mouton, The Hague, Netherlands, 1985.

[8] Anne Christophe, Ariel Gout, Sharon Peperkamp, and James Morgan. Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics*, 31(3-4):585 – 598, 2003.

[9] Paul R. Cohen, Clayton T. Morrison, and Erin Cannon. Maps for verbs: the relation between interaction dynamics and verb use. In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1022–1027, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[10] CSE. Traffic Video Narratives. `http://www.cse.iitk.ac.in/users/vision/hindi/`, IITK.

[11] J. A. Fodor. Précis of the modularity of mind. *The Behavioral and Brain Sciences*, 8:1–42, 1985.

[12] P. Gorniak and D. Roy. Grounded semantic composition for visual scenes. *JAIR*, 21(1):429–470, 2004.

[13] P Guha. Unsupervised concept acquisition from surveillance video. In *PhD Thesis Report, Department of Computer Science and engineering, IIT, Kanpur*. 2010.

[14] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335 – 346, 1990.

[15] HiGopi. Kamraj Unicode Hindi converter. `http://www.higopi.com/ucedit/Hindi.html`.

[16] Michael Ringgaard Hiyan Alshawi, Pi-Chuan Chang. Deterministic Statistical Mapping of Sentences to Underspecified Semantics. In *Proceedings of the Ninth IWCS*, pages 15–24, 2011.

[17] Laurent Itti and Christof Koch. Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[18] Ronald Langacker. *Foundations of Cognitive Grammar*. 1987.

[19] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV, 1999*, volume 2, pages 1150 –1157, 1999.

[20] Jean M. Mandler. How to Build a Baby: II. Conceptual Primitives. *Psychological Review*, 99(4):587–604, 1992.

[21] Jean M. Mandler. Thought before language. *Trends in Cognitive Sciences*, 8(11):508–513, November 2004.

[22] J. Mutch and D.G. Lowe. Multiclass object recognition with sparse, localized features. In *IEEE CVPR*, volume 1, pages 11–18, 2006.

[23] T Oates, Z Eyler-walker, and P. Cohen. Toward Natural Language Interfaces for Robotic Agents: Grounding Linguistic Meaning in Sensors. In *In Proceedings of the 4th ICAA*, pages 227–228, 2000.

[24] Charles Kay Ogden and I.A. Richards. *The meaning of meaning*. Trubner & Co, London, 1923.

[25] Panini. *Ashtadhyayi*. 400 BC.

[26] P. C. Quinn. The categorization of above and below spatial relations by young infants. *Child Development*, 65:58–69.

[27] P.C. Quinn. Concepts are not just for objects: Categorization of spatial relation information by infants. In David H Rakison and Lisa M. Oakes, editors, *Early category and concept development: Making sense of the blooming, buzzing confusion*, pages 50–76. Oxford University Press, 2003.

[28] Terry Regier. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Bradford Books, September 1996.

[29] Deb K. Roy. Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3-4):353–385, July-October 2002.

[30] Deb K. Roy and Alex P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146, 2002.

[31] V K Singh, S Maji, and A Mukerjee. Confidence Based updation of Motion Conspicuity in Dynamic Scenes. In *Third Canadian CRV 2006*, page 13. IEEE, 2006.

[32] J M Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):1–38, 1996.

[33] L. Smith and C. Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

[34] E. S. Spelke and S. J. Hespos. Conceptual development in infancy: The case of containment.

[35] S V P Gopi Srinath. Segmentation for Free: Discovering Object Categories in Surveillance Videos. In *MTech Thesis Report, Department of Computer Science and engineering, IIT, Kanpur*. 2010.

[36] L. Steels and F. Kaplan. Bootstrapping grounded word semantics. In *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, chapter 3, pages 53–74. Cambridge University Press, 2002.

[37] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell Publishing, 1953.

[38] C. Yu and D.H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 1(1):57–80, 2004.

[39] Luke S. Zettlemoyer and Michael Collins. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *UAI*, pages 658–666, 2005.