# Acquisition of Language Symbols

Pankaj Prateek
Advisor: Dr. Amitabha Mukerjee
{pratikkr, amit} @cse.iitk.ac.in
Department of Computer Science and Engineering,
IIT Kanpur, India

April 18, 2013

---

**Abstract**

*Using the "baby designer enterprise", we work with the objective of learning grounded Hindi symbols based on experience. The computational categories of tight-fit and loose-fit emerge from the functional constraints of the problem as abstractions. Eventually when the agent interacts with language systems, the labels for these abstractions are learnt. In the experiments, the functional distinctions of tight and loose fit are learnt in terms of the radii of the peg and the hole. Different participants were asked to describe the interaction between the peg and the hole in unconstrained Hindi, and the frequencies of words related to the concept were determined. The results show that "tight" and "loose" emerge as labels for the tight and loose-fit concept. The native words for Hindi like "तंग" and "कसा" are not so frequently used showing the influence of English on Hindi.*

## 1 Previous Work

The baby designer model learns patterns in an apprenticeship situation. When presented with a set of functional constraints and variable set governing them, it explores the design space, using domain-general learning algorithms to discover patterns in the better performing designs. These patterns get transformed to chunks in case they

occur frequently. In this process, knowledge of language and labels for such concepts and patterns is not required. On exposure to language, these implicit associations get transformed to rules in the symbolic space, thus incorporating labels.

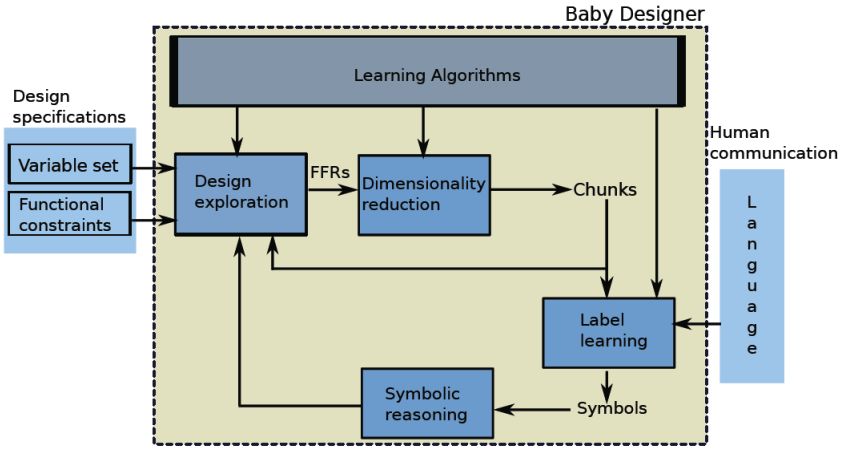


Figure 1: Baby Designer Model (Image taken from [2])

This was shown by Mukherjee and Dabbereu in the paper "Using Symbol Emergence to Discover Multi-Lingual Translations in Design" [4]. They focused on the situation where the image schema is ontologically prior to the label, i.e., the schema is available before its label is known.

Using English and Telugu, they showed that it is easy to learn linguistic labels for concepts, even when there is not much exposure to the language and the grammar.

## 2 Associating Labels

The association of a word $w_i$ with a concept $C_j$ can be measured using conditional Bayesian probability. But the direction of such an association is ambiguous. But


(a) peg and hole (b) peg-in-hole assembly

Figure 2: Peg-in-hole Assembly (Image taken from [2])

since only two concepts (tight and loose-fit) are involved,

$$\frac{p(\frac{C_T}{w})}{p(\frac{C_L}{w})} = \frac{p\frac{w}{C_T}.p(C_T)}{p\frac{w}{C_L}.p(C_L)}$$

The number of instances of $C_L$ and $C_T$ are almost the same in the set, so the direction of association had very less influence on the results.

For strongest association with $C_T$, compute $max_i \ \frac{p(\frac{w_i}{C_T}).p(C_T)}{p(\frac{w_i}{C_L}).p(C_L)}$. The inverse ratio is to be maximized to find the strongest association with $C_L$.

Let $n_t$ and $n_l$ be the total number of words in each narrative and the frequency with which the a word $w_i$ occurs be $k_t$ and $k_l$ respectively, then the conditional probability $p(w/C_t)$ is $k_t/n_t$. Similarly the conditional probability is calculated for $C_L$. The ratios are the computed, on which the results have been sorted.

# 3 Experiment I: STATE

This experiment focused at collecting spoken Hindi data in situations when the peg is already inserted in the hole. The pre-assembled setup was kept on a table and the participants were asked not to lift the assembly from the table.

### 3.1 Apparatus

A wooden block with 5 Holes (A (tapering), B, C, D, E) and 5 aluminium/steel pegs (1,2,3,4,5) of diameters (9mm, 12.25mm, 15.75mm, 19mm, 25mm).
Loose-fit situations A:1, C:5, D:3
Tight-fit situations A:2, B:4, E:3

### 3.2 Participants

12 IIT Kanpur students, all male, of ages 18-24, participated in the experiment. Apart from this, 8 IITK mess workers from Hall-3 also took part in generating narratives. Level of fluency and competance varied across the group, but not greatly. The sentence structures were retained, even if they were grammatically incorrect, as they were spoken.

Each participant was provided with the following instruction:
"यह एक छेद है और यह एक पैग है। यह पैग पहले से इस छेद में डाली हुई है। आपको इन दोनों के बीच हो रहे संपर्क का वर्णन करना है। यह ध्यान रखिये कि यह ब्लॉक इस टेबल से उठे नहीं और जितना हो सके हिन्दी के अलवा किसी और भाषा का प्रयोग ना करिये।"

There was no reference to tight or loose fit and the participants were asked to report various aspects of the interaction which they found important. The pegs were inserted in the holes and placed on the table and the participant were asked to not to lift the block from the table, so that they could experience the assembly rather than focussing on the insertion process. Loose and tight assemblies were provided to the participants alternatively.

## 4 Experiment II: ACTION

This experiment focused at collecting spoken Hindi data in situations where the participant is asked to actively insert the peg in the hole. There was no restriction as in the previous experiment.

State Profiled: [tight] corpus

| $Term$ | $f_T$ | $p(\frac{w}{C_T})$ | $f_L$ | $p(\frac{w}{C_L})$ | $f_{T,L}$ | $p(w)$ | $\frac{p(\frac{w}{C_L})}{p(\frac{w}{C_T})}$ |
|---|---|---|---|---|---|---|---|
| **Without Stemming** | | | | | | | |
| टाईट | 9 | 0.01711 | 1 | 0.00219 | 10 | 0.01018 | 7.80228 |
| रही | 7 | 0.01330 | 1 | 0.00219 | 8 | 0.00814 | 6.06844 |
| लग | 5 | 0.00950 | 1 | 0.00219 | 6 | 0.00610 | 4.33460 |
| महसुस | 4 | 0.00760 | 1 | 0.00219 | 5 | 0.00509 | 3.46768 |
| मूवमेंट | 3 | 0.00570 | 1 | 0.00219 | 4 | 0.00407 | 2.60076 |
| इसलिये | 3 | 0.00570 | 1 | 0.00219 | 4 | 0.00407 | 2.60076 |
| पर | 6 | 0.01140 | 2 | 0.00438 | 8 | 0.00814 | 2.60076 |
| जो | 5 | 0.00950 | 2 | 0.00438 | 7 | 0.00712 | 2.16730 |
| यहाँ | 5 | 0.00950 | 2 | 0.00438 | 7 | 0.00712 | 2.16730 |
| थोड़ा | 16 | 0.03041 | 7 | 0.01535 | 23 | 0.023421 | 1.98153 |
| **With Stemming** | | | | | | | |
| टाईट | 9 | 0.01711 | 1 | 0.00219 | 10 | 0.01018 | 7.80228 |
| महसुस | 4 | 0.00760 | 1 | 0.00219 | 5 | 0.00509 | 3.46768 |
| मूवमेंट | 3 | 0.00570 | 1 | 0.00219 | 4 | 0.00407 | 2.60076 |
| इसलिये | 3 | 0.00570 | 1 | 0.00219 | 4 | 0.00407 | 2.60076 |
| पर | 6 | 0.01140 | 2 | 0.00438 | 8 | 0.00814 | 2.60076 |
| जो | 5 | 0.00950 | 2 | 0.00438 | 7 | 0.00712 | 2.16730 |
| यहाँ | 5 | 0.00950 | 2 | 0.00438 | 7 | 0.00712 | 2.16730 |
| लग | 5 | 0.00950 | 2 | 0.00438 | 7 | 0.00712 | 2.16730 |
| थोड़ा | 18 | 0.03422 | 8 | 0.01754 | 26 | 0.02647 | 1.95057 |
| पोल | 6 | 0.01140 | 3 | 0.00657 | 9 | 0.00916 | 1.73384 |

Figure 3: State Profiled Tight Corpus

State Profiled: [loose] corpus

| $Term$ | $f_T$ | $p(\frac{w}{C_T})$ | $f_L$ | $p(\frac{w}{C_L})$ | $f_{T,L}$ | $p(w)$ | $\frac{p(\frac{w}{C_L})}{p(\frac{w}{C_T})}$ |
|---|---|---|---|---|---|---|---|
| **Without Stemming** | | | | | | | |
| लूस | 1 | 0.00190 | 4 | 0.00877 | 5 | 0.00509 | 4.61403 |
| घूमे | 1 | 0.00190 | 4 | 0.00877 | 5 | 0.00509 | 4.61403 |
| ताकत | 2 | 0.00380 | 7 | 0.015350 | 9 | 0.00916 | 4.03728 |
| कम्पेरीजन | 1 | 0.00190 | 3 | 0.00657 | 4 | 0.00407 | 3.46052 |
| पेग | 1 | 0.00190 | 3 | 0.006578 | 4 | 0.00407 | 3.46052 |
| साइज़ | 1 | 0.00190 | 3 | 0.006578 | 4 | 0.00407 | 3.46052 |
| इसमें | 2 | 0.00380 | 6 | 0.013157 | 8 | 0.00814 | 3.46052 |
| कम | 1 | 0.00190 | 3 | 0.006578 | 4 | 0.00407 | 3.46052 |
| बड़ा | 2 | 0.00380 | 6 | 0.013157 | 8 | 0.00814 | 3.46052 |
| ढीला | 3 | 0.00570 | 7 | 0.015350 | 10 | 0.01018 | 2.69152 |
| **With Stemming** | | | | | | | |
| लूस | 1 | 0.00190 | 4 | 0.00877 | 5 | 0.00509 | 4.61403 |
| कम्पेरीजन | 1 | 0.00190 | 3 | 0.00657 | 4 | 0.00407 | 3.46052 |
| पेग | 1 | 0.00190 | 3 | 0.00657 | 4 | 0.00407 | 3.46052 |
| कम | 1 | 0.00190 | 3 | 0.00657 | 4 | 0.00407 | 3.46052 |
| बड़ा | 3 | 0.00570 | 9 | 0.01973 | 12 | 0.01221 | 3.46052 |
| ढीला | 3 | 0.00570 | 7 | 0.01535 | 10 | 0.01018 | 2.69152 |
| पा | 3 | 0.00570 | 7 | 0.01535 | 10 | 0.01018 | 2.69152 |
| कारण | 1 | 0.00190 | 2 | 0.00438 | 3 | 0.00305 | 2.30701 |
| गैप | 1 | 0.00190 | 2 | 0.00438 | 3 | 0.00305 | 2.30701 |
| छेद | 2 | 0.00380 | 4 | 0.00877 | 6 | 0.00610 | 2.30701 |

Figure 4: State Profiled Loose Corpus

## 4.1 Apparatus

A wooden block with 5 Holes (A (tapering), B, C, D, E) and 5 aluminium/steel pegs (1,2,3,4,5) of diameters (9mm, 12.25mm, 15.75mm, 19mm, 25mm).

Loose-fit situations A:1, C:5, D:3
Tight-fit situations A:2, B:4, E:3

### 4.2 Participants

12 IIT Kanpur students, all male, of ages 18-24, participated in the experiment. Apart from this, 8 IITK mess workers from Hall-3 also took part in generating narratives. Level of fluency and competance varied across the group, but not greatly. The sentence structures were retained, even if they were grammatically incorrect, as they were spoken.

Each participant was provided with the following instruction:
"यह एक छेद है और यह एक पैग है। आपको यह पैग इस छेद में डालनी है और इन दोनों के बीच हो रहे संपर्क का वर्णन करना है। जितना हो सके हिन्दी के अलवा किसी और भाषा का प्रयोग ना करिये।"

As in the previous experiment, the tight and the loose fit situations were presented alternatively, and there was no reference to the tight and loose fit. There was no restriction as in the previous experiemnt and the particiants were asked to describe their experience of the interaction between the peg and the hole while the peg was actively inserted in the hole.

## 5 Results

The spoken English data which was collected was transcribed and categorized into tight-fit (E:3), (A:2), (B:4) and loose-fit situations (D:3), (A:1), (C:5). The frequency of each word in the corpus was then calculated. Many words from one text didn't appear in the second, so the words which have a high frequency in either corpus were focussed upon.

The results have been compiled in 3, 4, 5, 6.

**Action Profiled: [tight] corpus**

| *Term* | $f_T$ | $p(\frac{w}{C_T})$ | $f_L$ | $p(\frac{w}{C_L})$ | $f_{T,L}$ | $p(w)$ | $\frac{p(\frac{w}{C_T})}{p(\frac{w}{C_L})}$ |
|---|---|---|---|---|---|---|---|
| **Without Stemming** | | | | | | | |
| टाईट | 10 | 0.01336 | 1 | 0.00173 | 11 | 0.00829 | 7.72727 |
| इसको | 7 | 0.00935 | 1 | 0.00173 | 8 | 0.00603 | 5.40909 |
| को | 5 | 0.00668 | 1 | 0.00173 | 6 | 0.00452 | 3.86363 |
| रही | 9 | 0.01203 | 2 | 0.00346 | 11 | 0.00829 | 3.47727 |
| दोनों | 4 | 0.00534 | 1 | 0.00173 | 5 | 0.00377 | 3.09090 |
| सेम | 3 | 0.00401 | 1 | 0.00173 | 4 | 0.00301 | 2.31818 |
| गयी | 3 | 0.00401 | 1 | 0.00173 | 4 | 0.00301 | 2.31818 |
| थोड़ी | 3 | 0.00401 | 1 | 0.00173 | 4 | 0.00301 | 2.31818 |
| बिल्कुल | 3 | 0.00401 | 1 | 0.00173 | 4 | 0.00301 | 2.31818 |
| घर्षन | 5 | 0.00668 | 2 | 0.00346 | 7 | 0.00527 | 1.93181 |
| **With Stemming** | | | | | | | |
| टाईट | 10 | 0.01336 | 1 | 0.00173 | 11 | 0.00790 | 7.72727 |
| सेम | 3 | 0.00401 | 1 | 0.00173 | 4 | 0.00287 | 2.31818 |
| पे | 3 | 0.00401 | 1 | 0.00173 | 4 | 0.00287 | 2.31818 |
| साईस | 7 | 0.00935 | 3 | 0.00519 | 10 | 0.00718 | 1.80303 |
| बड़ा | 11 | 0.01470 | 5 | 0.00865 | 16 | 0.01149 | 1.70000 |
| ज्यादा | 6 | 0.00802 | 3 | 0.00519 | 9 | 0.00646 | 1.54545 |
| ईक्वल (equal) | 2 | 0.00267 | 1 | 0.00173 | 3 | 0.00215 | 1.54545 |
| फ्रिक्स्हन (friction) | 2 | 0.00267 | 1 | 0.00173 | 3 | 0.00215 | 1.54545 |
| पेग | 4 | 0.00534 | 2 | 0.00346 | 6 | 0.00431 | 1.54545 |
| इसलिये | 2 | 0.00267 | 1 | 0.00173 | 3 | 0.00215 | 1.54545 |

Figure 5: Action Profiled Tight Corpus

**Action Profiled: [loose] corpus**

| *Term* | $f_T$ | $p(\frac{w}{C_T})$ | $f_L$ | $p(\frac{w}{C_L})$ | $f_{T,L}$ | $p(w)$ | $\frac{p(\frac{w}{C_T})}{p(\frac{w}{C_T})}$ |
|---|---|---|---|---|---|---|---|
| **Without Stemming** | | | | | | | |
| लूस | 0 | 0 | 10 | 0.01730 | 10 | 0.00754 | Large Value |
| इसका | 1 | 0.00133 | 3 | 0.00519 | 4 | 0.00301 | 3.88235 |
| था | 1 | 0.00133 | 3 | 0.00519 | 4 | 0.00301 | 3.88235 |
| पे | 1 | 0.00133 | 3 | 0.00519 | 4 | 0.00301 | 3.88235 |
| बड़ा | 3 | 0.00401 | 7 | 0.01211 | 10 | 0.00754 | 3.01960 |
| जा | 4 | 0.00534 | 8 | 0.01384 | 12 | 0.00904 | 2.58823 |
| फिट | 1 | 0.00133 | 2 | 0.00346 | 3 | 0.00226 | 2.58823 |
| इसलिये | 1 | 0.00133 | 2 | 0.00346 | 3 | 0.00226 | 2.58823 |
| इसे | 1 | 0.00133 | 2 | 0.00346 | 3 | 0.00226 | 2.58823 |
| मतलब | 1 | 0.00133 | 2 | 0.00346 | 3 | 0.00226 | 2.58823 |
| **With Stemming** | | | | | | | |
| लूस | 0 | 0 | 10 | 0.01730 | 10 | 0.00718 | Very Large |
| पड़ा | 1 | 0.00133 | 6 | 0.01038 | 7 | 0.00502 | 7.76470 |
| को | 1 | 0.00133 | 5 | 0.00865 | 6 | 0.00431 | 6.47058 |
| दोनों | 1 | 0.00133 | 5 | 0.00865 | 6 | 0.00431 | 6.47058 |
| पूरा | 1 | 0.00133 | 4 | 0.00692 | 5 | 0.00359 | 5.17647 |
| उस | 2 | 0.00267 | 6 | 0.01038 | 8 | 0.00574 | 3.88235 |
| बिल्कुल | 1 | 0.00133 | 3 | 0.00519 | 4 | 0.00287 | 3.88235 |
| घर्षन | 2 | 0.00267 | 5 | 0.00865 | 7 | 0.00502 | 3.23529 |
| नहीं | 7 | 0.00935 | 17 | 0.02941 | 24 | 0.01724 | 3.14285 |
| फिट | 1 | 0.00133 | 2 | 0.00346 | 3 | 0.00215 | 2.58823 |

Figure 6: Action Profiled Loose Corpus

# 6 Discussion

It is demonstrated by the experiments that the linguistic labels for image schemas, like loose and tight-fit concepts in this case, can be learnt without much prior knowledge of either domain or language.
Words like "tight" and "loose" are readily associated with the [tight] and [loose] fits respectively, based on uninformed word associations from the Hindi language. Thus these words have top associations with the respective concepts.
Unlike shown by Mukerjee et.al.[] for Telugu language, Hindi is richly inflected, it has incorporated many foriegn words into the language with time. The can be seen from the fact that the Hindi native correspondences for "tight", viz., "तंग" and "कसा" were used only occasionally by the participants even after they were specifically told to adhere to Hindi in their narrations. The Hindi native for "loose", viz., "ढीला" had some comparable conditional probability as compared to the native for "tight". This was the same with the Hindi speaking population which had very less contact with English language.

# 7 Conclusion

The results show that linguistic labels can be learnt even after little exposure to the linguistic mapping. This is independent of the knowledge of the grammar or the language. The emergence of certain labels agree with Saussure's view [3] that symbols are a bipolar entity in which the label is coupled with an image schema.

**Acknowledgements**

I would like to show my sincere gratitude to my advisor, Professor Amitabha Mukerjee, for providing us the chance to work on this project, and for acting as a guiding light throughout the duration of the project. I would also like to thank Mr. Madan Dabbereu for his help throughout the project.

# References

[1] Madan Dabbeeru and Amitabha Mukerjee. Computational models of tacit knowledge. In *CIRP Design 2012*, pages 47–57. Springer, 2013.

[2] Madan Mohan Dabbeeru and Amitabha Mukerjee. Learning concepts and language for a baby designer. In *Design Computing and Cognition10*, pages 445–463. Springer, 2011.

[3] Ferdinand De Saussure. Nature of the linguistic sign. *Course In General Linguistics*, 1916.

[4] Amitabha Mukerjee and Madan Mohan Dabbeeru. Using symbol emergence to discover multi-lingual translations in design. ASME, 2010.

[5] SVP Gopi Srinath, Nikhil Joshi, Prabhat Mudgal, and Amitabha Mukerjee. Learning grounded semantics of hindi nouns from video surveillance and user commentary.