Language Learnability Analysis of Hindi: A Comparison with Ideal and Constrained Learning Approaches

Sandeep Saini & Vineet Sahula

Journal of Psycholinguistic Research

ISSN 0090-6905

J Psycholinguist Res
DOI 10.1007/s10936-019-09641-2

Springer

Journal of Psycholinguistic Research
https://doi.org/10.1007/s10936-019-09641-2

Language Learnability Analysis of Hindi: A Comparison with Ideal and Constrained Learning Approaches

Sandeep Saini[1] · **Vineet Sahula**[2]

Abstract
Native language acquisition is one of the initial processes undertaken by the human brain in the infant stage of life. The linguist community has always been interested in finding the method, which is adopted by the human brain to acquire the native language. Word segmentation in one of the most important tasks in acquiring the language. Statistical learning has been employed to be one of the earliest strategies that mimic the way an infant can adapt to segment a lot of different words. It is desired that the language learnability theories be universal in nature and work on most, if not all the languages. In the present work, we have analyzed the learnability of Hindi, the most popular Indian language, using ideal (universal) and constrained Bayesian learner models. We have analyzed the learnability of the language using unigram and bigram approaches by considering word, syllables, and phonemes as the smallest unit of the language. We demonstrate that Bayesian inference is indeed a viable cross-linguistic strategy and works well for Hindi also.

Keywords Language acquisition · Language learnability · Bayesian learners · Hindi Language

Introduction

The human brain has an innate power to hear sound signals in the environment and get meaningful words from these speech signals. These words, when spoken or written with predefined rules (grammar), make a language. Infants, in their early stage of life, start grabbing the different sounds being produced in the environment around them (Floccia et al. 2016). These sounds are human-generated language words as well as different kinds of noises present in the environment. Infant's brain has to differentiate between useful and

✉ Sandeep Saini
sandeep.saini@lnmiit.ac.in

Vineet Sahula
sahula@ieee.org

[1] Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jaipur 302031, India

[2] Department of Electronics and Communication Engineering, Malaviya National Institute of Technology, Jaipur 302017, India

Published online: 01 April 2019

non-useful sounds of verbal communication to build its vocabulary in native language (Gural et al. 2015) and (Halpern 2016). Osokina Gural et al. (2015) has shown in his work that how a child starts acquiring the most spoken words by her caregivers and respond to the most asked questions only in the initial stages of her childhood. The brain is acquiring the most frequently spoken words faster and remembers those for a long time. Thus a lot of sounds which are new to the child, might not be useful in the first instance. When we discuss specifically the native language learning, then parents or caregivers are the principal trainers for most of the infants. They speak words to get the attention of an infant or direct him or her to do some actions. Infant's brain has to divide these speech signals into meaningful words to trigger the actions from different parts of the body, to respond properly to these speech signals. This creates several questions for linguists and psychologists. How does the speech segmentation take place inside an infant's brain? How does an infant's brain decode these sounds and the brain process these signals to trigger actions in different parts of the body? Which words or phrases are learned faster or better by the infants? Can they learn all the words with equal easiness? Can any child learn any language? All these questions have been the topics of research for linguists and recently computer scientists as well. There has been a lot of research on each of the topics in the computational world as well. Speech segmentation, language acquisition models, probabilistic learners, and universal learners are the computational models to support the theories proposed by the linguists.

In this work, we would focus on evaluating language learning strategies in Indian languages, specifically in Hindi. Hindi is having more than 300 million native speakers and more than 800 million people have one or more form of knowledge about the language (Graddol 2004). Hindi is mostly spoken in the northern part of India and mostly understood in the majority of the Indian territory. Hindi finds its presence in countries having Indian immigrants as well. We have evaluated the learnability of Hindi based on the learner's recall and precision skills. Pearl (2014) have explored and evaluated the language learning strategies in 7 languages, namely English, German, Spanish, Italian, Farsi, Hungarian and Japanese. Learnability of any language is measured in terms of the F-Score. Precision and Recall capability of the learner model is used to calculate the value of the F-score. In this analysis, the F score shows how accurately the words are comprehended and learned by the learner.

The article is organized as follows. Part 2 provides an introduction to various language acquisition theories proposed by linguists. Part 3 explores the computational models to evaluate these acquisition strategies. In part 4, we provide the details of data and pre-processing techniques used to filter the Hindi data for learnability analysis. We discuss the results in part 5 and conclude the paper in part 6.

Language Acquisition Theories

In this part, we have discuss some of the major language acquisition theories available in the literature. These theories are mainly based on the observations made on a group of infants under supervised environment. Hyams (2012) states that "The supreme issue in linguistic theory is to explain how a child can acquire any human language." An infant, coming from any social or economic background of the world, shows the same skills to sit, grip, walk and talk. There are no serious differences in acquiring these skills in the early stages of human life. It is observed the style of sitting, walking, grabbing, and displaying emotions are almost the same in any infant across the globe, irrespective of social or economic conditions. But, the only differentiating feature in the initial stages of life is the acquisition of a native language. The

process of language learning is a continuous process in which a person goes through his entire life. We focus on the early stages of life in the process of language acquisition. At this stage, we can divide the language acquisition process into two phases, namely, the perception of the language and production of the language. Taha (2017) and Felser and Drummer (2017) have worked on the language learnability for native and non-native languages.

The process of language comprehension in the infant's brain can be divided into three stages.

- *Segments* In this process, the infant brain can grab some of the vowel and consonant sounds in the age of 5–7 months. During the age of 9–11 months the infant develops phonetic skills and the ability to distinguish between different sounds. The brain develops to filter out non-native language sounds from 12 months onwards.
- *Super segments* At this stage, infants can learn the syllables which are stressed more by the source, parents in most of the cases.
- *Lexicons* In this process, infants identify proper names in the age group of 4–6 months. They develop the capability to segment words from the sentences in the age group of 9–12 months. A typical infant is having a memory of around 100 words by the age of 15 months, and it grows to around 200 words by 18 months.

The second important aspect of language learning is the production of the language. Children try to reproduce the sounds which they hear. In the initial stages from 3 to 5 months, there is an onset of babbling. In this stage, infants show the first manifestation of phonology. They can recognize and respond to particular sounds. In the initial stage only vowels related sounds are generated and then slowly some consonants are produced by the age of 8 months. From 10th months onwards the child starts producing native language words. Most of these words are comprised of one or two syllables. These may not be proper words belonging to the dictionary of the language, but somewhat close to the actual words. This stage is also critical to examine language generation process. After regular rectification from the parents, the child produces meaningful words from 12 months onwards. Later infants build a vocabulary of the native language and are able to interact with short answers. A summary of many of the stages in perception about the language and generation process is shown in Fig. 1.

A large number of the latest natural language processing applications focus on the human brain-inspired architectures and related theories for better results. These architectures provide models to replicate the language comprehension and processing in the human brain with the help of computational and cognitive models. Other research groups are working towards Natural language programming as well. To achieve this, language translation from natural language to machine language is necessary. Therefore, a lot of work is underway to map the human language learning strategies with machine learning strategies to develop the systems which would process the instructions in a way our brain processes. In the next part, we provide an overview of some computational models and methods to evaluate language learning principles adopted by the infant's brain.

Evaluation Models

Language learning strategies mentioned in the previous part have been proposed and developed, with the assumption that they are universal in nature and would work in case of multiple languages. These strategies are mainly implemented for human beings in some language

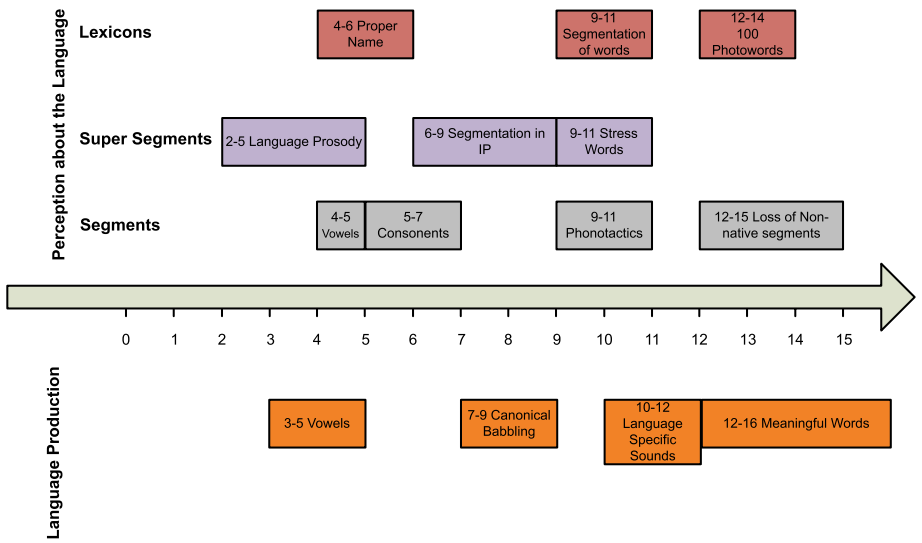

Fig. 1 Various stages in the infant life to understand and generate the native language

learning centers. The results for language learnability are generated from the data obtained from these evaluations. These evaluation results depend on human data. The whole process takes a lot of time and efforts in setting up the experiment and evaluation of the results. In this part, we explore some of the computational models and methods which can be used to evaluate language acquisition capabilities of a person. Most of these strategies have been tested on one language and very rarely cross-language analysis is performed. Demonstration of the cross-linguistic success of the strategies is very important in the infant stages of human life. Most of the theories presumed that the human being has innate properties to acquire any language as a newly born infant.

Many of the computational models for evaluation of language learning strategies are based on the Bayesian theory. The Bayesian theory provides a very robust probability based learning approach relying on the existing belief of the learner and the incoming inputs. The complete process involves generating data from the Child-directed speech and get the inference from this data about the learnability of a particular language. In most cases, the input signal is a speech signal. The speech signal is a continuous signal with a lot of overlapping frequencies in a tight bandwidth. This signal has to be synthesized and segmented into words before we start the processing. Speech segmentation can be performed by Bayesian segmentation approaches explained below.

Bayesian Segmentation

We have investigated a version of the Bayesian segmentation strategy that is, universal in nature, thus has language-independent properties. The sound generated in any language is comprised of a morpheme. A morpheme is the smallest possible grammatical unit in a language. Bayesian segmentation is able to adapt to an incremental gain in the belief and self-updating of new values of existing belief. This model would be best suited for our target

audience as well, which are infants in the age group of 6–12 months. At this age, the child is not having any prior knowledge of the language, and he learns from the sounds being produced around him. One benefit of Bayesian learning strategies is that they explicitly distinguish between the learner's pre-existing beliefs, the prior: P (h) and how the learner evaluates incoming data, the likelihood: $P(d|h)$. This information is combined using Bayesian theorem (1) to generate the updated beliefs of the learner the posterior: $P(h|d)$.

$$P(h|d) \propto P(d|h) \cdot P(h) \tag{1}$$

The Bayesian segmentation strategy was originally described by Goldwater et al. (2009) (GGJ). Considering the constraint that infants possess limited knowledge of language structure at the relevant age, authors have described two simple generative models.

1. A Unigram assumption
2. A Bigram assumption

Unigram model assumes independence between different words and thus effectively believes that word tokens are randomly selected. To encode this assumption in the model, GGJ assume that the observed sequence of words $w_1 \ldots w_n$ is generated sequentially using a probabilistic generative process. To choose the identity of *i*th word in the Unigram case, Goldwater et al. use Eqs. (2) and (3), where the probability of the current word is a function of how often it has occurred earlier.

$$P(w_i|w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w_i) + \alpha \cdot P_0 \cdot (w_i)}{i - 1 + \alpha} \tag{2}$$

$$P_0 = P(w = x_i \ldots x_m) = \prod_j P(x_j) \tag{3}$$

Here, P_0 is a base distribution specifying the probability that a novel word will consist of particular units $x_1 \ldots x_m$. $n_{i-1}(w_i)$ is the number of times word w_i appears in the previous $i-1$ words, α is a free parameter of the model which encodes how likely the novel word is to be generated. The mode suggests that a word with smaller length will be the better learned. Since children are more familiar and comfortable with smaller words, hence learnability is better.

Bigram model does not assume independence of consecutive words, but it rather assumes that consecutive words are related to each other, and if they appear once, then there is a higher probability that they will appear again in the corpus. Thus a word is generated based on the identity of the word that immediately precedes it.

$$P(w_i|w_1 = w', w_1 \ldots w_{i-2}) = \frac{n_{i-1}(w', w_i) + \beta \cdot P_1 \cdot (w_i)}{n_{i-2}(w') + \beta} \tag{4}$$

$$P_i(w_i) = \frac{b_{i-1}(w_i) + \gamma \cdot P_0 \cdot (w_i)}{b - 1 + \gamma} \tag{5}$$

Here $n_{i-1}(w', w_i)$ defines the number of times a particular Bigram (w_0, w_i) has occurred in the first i−1 words, $n_{i-2}(w_0)$ provides the information on the number of times the word w_0 occurs in the first $i-2$ words, $b_{i-1}(w_i)$ is the number of Bigram types which contain w_i as the second word, b is the total number of Bigram types which have existed before. P_0 was already defined in Eq. (3), and β and γ are free model parameters.

The role of β and γ is very similar to the role of α in Unigram model. They control the introduced bias towards fewer Biagrams (β) towards fewer unique lexical items as the second word in a Bigram (γ).

Bayesian Inference

To infer the learnability of a language based on the data generated from the child-oriented speech, different types of learners are designed. We have employed Bayesian learners in our experiment which has two categories.

Ideal learner processes data in a batch, and thus it assumes that there is a perfect memory available for the learner. This learner has enough processing resources to search for potential segmentations exhaustively. After an exhaustive search, it selects optimal segmentation. This learner heavily relies on the availability of a huge corpus and works well if the learner has prior knowledge of the language. *BatchOpt* is the notation used for this ideal learner.

Constraint learner In the practical world, infinite memory (corpus) is not available, and infants are also not having prior knowledge of the language. Thus a lot of constraints need be put on the learner as well. Three major types of constraint learners implemented in this work are as follows.

1. Online Optimal: *OnlineOpt*
   This learner processes data incrementally. The learner has enough processing resources to search for potential segmentations exhaustively, and it selects optimal segmentation.
2. Online Sub-optimal: *OnlineSubOpt*
   This constraint learner also processes data incrementally and have enough processing resources to search for potential segmentations exhaustively. It differs from the OnlineOpt learner in its segmentation selection mechanism, and it selects segmentation probabilistically.
3. Online Limited Working Memory: *OnlineMem*
   This learner also processes data incrementally like other two constraint learners. However, it has limited working memory, hence it cant do an exhaustive search. This learner focuses more on recent data (recency bias). It also selects optimal segmentation.

A summary of the properties of all ideal and constraint learners is compiled in Table 1.

Data Preparation for Learnability Analysis

We have analyzed language learning models for "Hindi" language. We have considered two types of dataset for this analysis. Initially, we have tested the evaluation models on speech dataset as analyzed by Pearl (2014). Later we have also considered text corpus as well. Speech and text corpus for Hindi have been obtained from the following datasets.

1. Speech dataset from Electro Medical and Speech Technology Laboratory (EMST), Department of Electronics and Electrical Engineering at Indian Institute of Technology Guwahati (Haris et al. 2012). This data is availabe in four phases namely, IITG MV Phase-I, Phase-II, Phase-III and PhaseIV. Phase-I dataset is reported to have been

Table 1 Summary of Bayesian learners

Learner	Parameters	Learning assumptions		
		Online processing	Sub-optimal decisions	Recency effects
BatchOpt	Iterations = 20,000	No	No	No
OnlineOpt	N/A	Yes	No	No
OnlineSubOpt	N/A	Yes	Yes	No
OnlineMem	Sample uttrance = 20,000	Yes	No	Yes

prepared from 100 subjects over two sessions in an office environment involving multiple sensors, multiple languages, and different speaking styles. We have used this phase only for our experimentation.

2. Hindi Speech Corpus from TDIL: Technology Development for Indian Languages Programme, India Hindi Speech Corpus (2010).
3. English–Hindi parallel corpus from Institute for Language, Cognition, and Computation, University of Edinburgh (ILCC) (2011).
4. Institute of Formal and Applied Linguistics (UFAL) at the Computer Science School, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic (Bojar et al. 2014).
5. Center for Indian Language Technology (CFILT), IIT Bombay (2010).

All these datasets are exhaustive with an abundant variety of words. Table 2 provides the information regarding the number of words and sentences in each of those dataset.

We have pre-processed and tested the evaluation model on two categories of datasets in different ways. The experimental setup and process are explained in Fig. 2. Speech data is initially processed through Hidden Markov Model Toolkit (HTK) (Young et al. 2002). The toolkit provides satisfactory results for Hindi as well (Kuamr et al. 2014). The resultant words are stored in text files. Once we have the text corpus, then we have considered the following three different approaches to evaluate the learnability of the language under consideration.

1. Learnability analysis is performed on the text corpus directly by considering unigram and bigram models for text corpus.
2. For this experiment, we have converted words into syllables and learnability analysis is performed on the text corpus directly by considering unigram and bigram models for the syllabified corpus.

Table 2 Details of EMST, TDIL, ILCC, UFAL and CFILT Hindi datasets

	EMST	TDIL	ILCC	UFAL	CFILT
No of sentences	945	7885	41,396	237,885	1,492,827
No of words	3675	48,297	245,675	1,048,297	20,601,012
No of unique words	1342	1673	8342	21,673	250,619

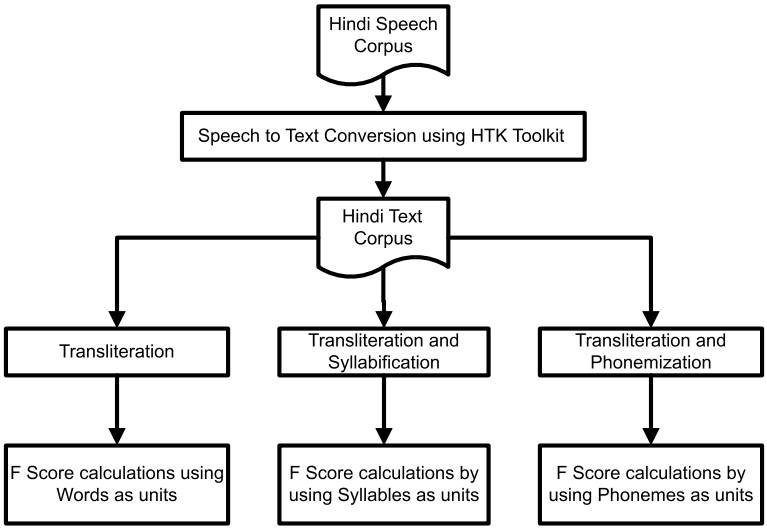

Fig. 2 Preprocessing and evaluation of different kinds of dataset

3. Phonemes are considered as the unit of language acquisition. for this experiment, we have converted words into phonemes and by considering unigram and bigram models for the phonemic corpus.

This data for the Hindi language is generally written in Devanagari script. We faced difficulties in applying Bayesian learner to the Devanagari script. Thus, the first step in the pre-processing of the data is to transliterate the files into Roman script (Gupta et al. 2012). The algorithm has two major tasks to perform:

1. Provide single correct transliterated word in English script (Roman) if there is only a single word in Hindi script which could match with the original word.
2. Provide two-three most probable options of the transliterated text, in the decreasing order of probability, if there are more than one words in Hindi script which could match with the original word.

Transliteration from Devanagari to Roman using this algorithm is found to be 97.24% accurate and quite reliable for our study. We have transliterated the dataset for further processing and computation.

Initial learning results by considering Unigram and Bigram learner models on words didn't show the expected results as discussed in the results part. It is suggested (Phillips and Pearl 2015) that syllables are basic units of learning in infant's language acquisition process. We have syllabified after transliterated the dataset by using a rule-based approach (Eddington et al. 2013) and (Weerasinghe et al. 2005).

For the third assumption, we have phonemicized our data using The Festival Speech Synthesis System (Clark et al. 2007). As a whole it offers full text to speech through a number APIs: from shell level, though a Scheme command interpreter. We have converted the words to Phonemes using the Festival TTS system (Black and Taylor 1997) and the results obtained after all pre-processing techniques are shown in Fig. 3.

Results and Discussions

Language acquisition evaluation depends on the different metrics proposed by different authors. In this process, we focus on how precisely the learner is able to recall the learned words/syllables. The precision of the learner is defined in (6).

$$Precision = \frac{\#\text{ of identified true word tokens}}{\#\text{ of all identified word tokens}} \tag{6}$$

Recall is the capability of the learner to recollect the learnt words, which is defined in (7).

$$Recall = \frac{\#\text{ of identified true word tokens}}{\#\text{ of all true word tokens}} \tag{7}$$

These two scores, which would always range between 0 and 1, are typically combined into a single score by using the harmonic mean, referred to as the F-score.

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

The F score would also range between 0 and 1, and when multiplied by 100 to results in percentage. Higher values of precision and recall imply that the model is able to provide better results. And, as we can see the relation of precision and recall for generation of the F score, the higher precision and recall values will result in higher F score as well. A higher value of the F score implies the better learnability of the learner model for a particular language. This metric is simple to compute and so is a convenient way to compare the results of various modeled segmentation strategies across studies.

As an example, if the correct sequence of words is "Rahul has entered the classroom," and instead, we find "Rahul has enter ed the class room" We have five words in true segmentation and seven in the found segmentation. Out of these three words match exactly in both the sequences. Therefor Precision would be 24.85% (3/7), the recall would be 50% (3/6), and F score would be 46.2%.

We have tested the learnability of Hindi using both the speech datasets, i.e., from EMST and TDIL. Both these datasets are tested for Unigram and Bigram models. For each of the model, all four Learner techniques (BatchOpt, OnlineOpt, OnlineSubOpt, and OnlineMem) have been employed. The F score values obtained after this experiment for English and Hindi languages have been compiled in Table 3. These scores illustrate that the Bigram

Devanagari Input	Roman Transliteration	Syllabified Output	Phoneme Representation
एक	ek	ek	eh-k
लड़का	ladaka	la-da-ka	l-ax-d|aa-k|ax
सुबह	subaha	su-ba-ha	s-ah-b|ax-hh|ax
सुबह	subaha	su-ba-ha	s-ah-b|ax-hh|ax
दौड़ने	daudane	dau-da-ne	d-ao-d|ey-n
को	ko	ko	k-ow
जाया	jaaya	jaa-ya	jh-ay|ax
करता	karta	kar-ta	k-aa-r-t|ax
था	tha	tha	th-ax

Fig. 3 Devanagari input, transliterated output, syllabified representation and phoneme representation of sample words

model used by the constraint learner leads to better learnability of English as well as Hindi. The results are the same for both the datasets for learnability of Hindi.

The results described in Table 3 have been obtained by considering words as a unit of language. Since words are not considered as the units of speech, and hence we have examined the learner mechanisms on syllable level as well as phonemes level datasets. Many authors (Werker and Tees 1984; Jusczyk and Derrah 1987; Eimas 1999) have considered phonemes as the basic units of language. While a few others Swingley (2005), Lignos and Yang (2010) and Gambell and Yang (2006) have considered syllables as the units of the language. We have considered both the approaches and demonstrated that both the approaches could be considered for language learnability. The dataset under consideration is converted to the syllabified format and phonemicized formats respectively for analysis. Tables 4 and 5 provide the F score computation after following these two modifications in the provided datasets.

We have also analyzed text corpus with the same evaluation model for which the results are presented in Table 6.

In all these results, i.e. Tables 3, 4, 5 and 6, the best F scores are highlighted with the bold numbers. In Table 3, we compare unigram and bigram modeling using four different learners for English as well as Hindi assuming word as a unit of the language. The best score for English is obtained using a constraint learner of the type *Online-Mem*. This result implies that the English acquisition is better if the learner possesses recent knowledge and successive words have strong dependence on the previous words. We could expect this as the language have such strong relation between consecutive words. In the same table, we have obtained better learnability using EMST dataset as it is comprised of stories and the words are closer to infant vocabulary. Hindi on the other hand, provides the best F score for Bigram model combined with the ideal (BatchOpt) learner for both datasets. This implies that prior knowledge of the language and relation between consecutive words increases the learnability of the languages. A better was expected for both the languages using Bi-gram model. In both the languages, there are a lot of words which exist in pairs. Also, a lot of verbs and helping verbs are made of two or more words. Thus, we could expect that the results of the bigram model would be better than the unigram model. This is also validated in both languages. In Hindi, the number of such pairs is large, and most of these case studies for the verbs are having two or more words. Therefore, the prior knowledge of the language and relation

Table 3 F score calculated for English and Hindi languages based on Unigram and Bigram based ideal and constraint learners

Model	Learner	F score		
		English CHILDES (MacWhinney 2000)	Hindi (EMST) (Haris et al. 2012)	Hindi (TDIL) India Hindi Speech Corpus (2010)
Unigram	BatchOpt	0.531	0.459	0.429
	OnlineOpt	0.588	0.586	0.559
	OnlineSubOpt	0.637	0.589	0.534
	OnlineMem	0.551	0.605	0.591
Bigram	BatchOpt	0.771	**0.814**	**0.801**
	OnlineOpt	0.751	0.759	0.734
	OnlineSubOpt	0.778	0.785	0.642
	OnlineMem	**0.879**	0.741	0.714

Table 4 F scores calculated for English and Hindi languages using syllables as a unit of language

Model	Learner	F score		
		English CHILDES dataset (MacWhinney 2000)	Hindi (EMST) (Haris et al. 2012)	Hindi (TDIL) India Hindi Speech Corpus (2010)
Unigram	BatchOpt	0.535	0.665	0.695
	OnlineOpt	0.573	0.625	0.715
	OnlineSubOpt	0.598	0.689	0.723
	OnlineMem	0.563	0.617	0.671
Bigram	BatchOpt	0.782	**0.854**	**0.845**
	OnlineOpt	0.734	0.728	0.759
	OnlineSubOpt	0.728	0.756	0.742
	OnlineMem	**0.888**	0.750	0.759

Table 5 F scores calculated for English and Hindi languages using phoneme as a unit of language

Model	Learner	F score		
		English CHILDES dataset (MacWhinney 2000)	Hindi (EMST) (Haris et al. 2012)	Hindi (TDIL) India Hindi Speech Corpus (2010)
Unigram	BatchOpt	0.571	0.659	0.672
	OnlineOpt	0.580	0.662	0.687
	OnlineSubOpt	0.584	0.679	0.689
	OnlineMem	0.588	0.686	0.685
Bigram	BatchOpt	0.731	**0.834**	**0.830**
	OnlineOpt	0.745	0.767	0.784
	OnlineSubOpt	0.748	0.787	0.757
	OnlineMem	**0.894**	0.783	0.754

Table 6 F scores calculated for different text corpus of Hindi languages using word as a unit of language

Model	Learner	F score		
		ILCC dataset [12]	UFAL (Bojar et al. 2014)	CFILT [14]
Unigram	BatchOpt	0.453	0.543	0.543
	OnlineOpt	0.487	0.554	0.554
	OnlineSubOpt	0.472	0.565	0.556
	OnlineMem	0.498	0.537	0.565
Bigram	BatchOpt	**0.831**	**0.839**	**0.840**
	OnlineOpt	0.675	0.697	0.678
	OnlineSubOpt	0.756	0.737	0.772
	OnlineMem	0.794	0.753	0.723

between consecutive words increases the learnability of the Hindi and this is the reason for Ideal learner providing better results for the language.

Tables 4 and 5 provide results with some improvements in the values of F score. In both these tables, the bigram model outperforms unigram, and the same learner (as in Table 3) provides the best scores for the respective language. The improvement in scores was also expected as syllables and phonemes are actual units of a language in comparison to just words.

Table 6 is a compilation of our experiment by taking the text data directly and applying the above-mentioned models and learners for Hindi. In this experiment, the results are uniform for the Hindi language. Depending on the size and the content of the dataset, we have obtained different values of the F scores. Larger datasets generally result in better training and testing of these learners.

Conclusion and Future Work

In this work, we have analyzed the learnability of Hindi using language learning models based on Bayesian inference. Bayesian learners along with different n-gram models are employed to calculate the learnability metrics for the languages. Hindi learnability is analyzed with two standard speech datasets along with three text datasets. It has been observed that unlike English, Hindi is the best learnable with the ideal learner approach. The first set of results were obtained by considering word as a unit of spoken language. Further investigations were performed by syllabification and phonemization of the dataset. These results establish that syllables and phonemes are better units to learn the language especially at the early stage of human life.

Acknowledgements This research is partially supported by the project under SMDP-C2SD-ERP-1000110086, Department of Electronics and Information Technology, Ministry of Communication & IT, Government of India at Malaviya National Institute of Technology (MNIT), Jaipur. We thank MNIT's computer labs for setting up the experiment and also the LNMIIT's GPU services in simulations to obtain the results.

Compliance with Ethical Standards

Funding The research is not funded by any external project/agency other than the LNMIIT Jaipur and MNIT Jaipur, India. This research is partially supported by the project under SMDP-C2SD-ERP-1000110086, Department of Electronics and Information Technology, Ministry of Communication & IT, Government of India at Malaviya National Institute of Technology (MNIT), Jaipur.

Conflict of Interest The work is supported by the LNM Institute of Information Technology, Jaipur and Malaviya National Institute of Technology (MNIT), Jaipur, India only. The details are mentioned in funding section. We have no conflict of interest to disclose.

References

Black, A. W., & Taylor, P. A. (1997). The festival speech synthesis system: System documentation. Technical Report HCRC/TR-83. Scotland: Human Communciation Research Centre, University of Edinburgh. Avaliable at http://www.cstr.ed.ac.uk/projects/festival.html.

Bojar, O., Diatka, V., Rychlỳ, P., Stranák, P., Suchomel, V., Tamchyna, A., & Zeman, D. (2014). Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC* (pp. 3550–3555).

Clark, R. A. J., Richmond, K., & King, S. (2007). Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, *49*(4), 317–330.

Cognition Institute for Language and Indic Multi-parallel Corpus Computation, University of Edinburgh (2011). http://homepages.inf.ed.ac.uk/miles/babel.html.

Eddington, D., Treiman, R., & Elzinga, D. (2013). Syllabification of american english: Evidence from a large-scale experiment. Part ii. *Journal of Quantitative Linguistics*, *20*(2), 75–93.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *The Journal of the Acoustical Society of America*, *105*(3), 1901–1911.

Felser, C., & Drummer, J.-D. (2017). Sensitivity to crossover constraints during native and non-native pronoun resolution. *Journal of Psycholinguistic Research*, *46*(3), 771–789.

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Luche, C. D., Durrant, S., et al. (2016). British english infants segment words only with exaggerated infant-directed speech stimuli. *Cognition*, *148*, 1–9.

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty*. Unpublished manuscript.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Graddol, D. (2004). *The future of language* (Vol. 303). Washington, DC.: American Association for the Advancement of Science.

Gupta, K., Choudhury, M., & Bali, K. (2012). Mining hindi-english transliteration pairs from online hindi lyrics. In *LREC* (pp. 2459–2465).

Gural, S. K., Kecskes, I., Gillespie, D., Rijlaarsdam, G. C. W., Ter-Minasova, S. G., Karasik, V. I., et al. (2015). Word collocations as language knowledge patterns: A study of infant speech. *Procedia–Social and Behavioral Sciences*, *200*, 353–358.

Halpern, M. (2016). How children learn their mother tongue: They dont. *Journal of Psycholinguistic Research*, *45*(5), 1173–1181.

Haris, B. C., Gayadhar Pradhan, A., Misra, S. R. M. P., Das, R. K., & Sinha, R. (2012). Multivariability speaker recognition database in indian scenario. *International Journal of Speech Technology*, *15*(4), 441–453.

Hyams, N. (2012). *Language acquisition and the theory of parameters* (Vol. 3). Berlin: Springer Science & Business Media.

IIT-Bombay Hindi Corpus (2010). http://www.cfilt.iitb.ac.in/downloads.html.

India Hindi Speech Corpus. TDIL: Technology Development for Indian Languages Programme (2010). http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=268&lang=en

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*(5), 648.

Kuamr, A., Dua, M., & Choudhary, T. (2014). Continuous hindi speech recognition using gaussian mixture hmm. In *2014 IEEE students' conference on electrical, electronics and computer science (SCEECS)* (pp. 1–5).

Lignos, C., & Yang, C. (2010). Recession segmentation: Simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 88–97). Vancouver: Association for Computational Linguistics.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). London: Psychology Press.

Pearl, L. (2014). Evaluating learning-strategy components: Being fair (commentary on ambridge, pine, and lieven). *Language*, *90*(3), e107–e114.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132.

Taha, H. (2017). How does the linguistic distance between spoken and standard language in arabic affect recall and recognition performances during verbal memory examination. *Journal of Psycholinguistic Research*, *46*(3), 551–566.

Weerasinghe, R., Wasala, A., & Gamage, K. (2005). A rule based syllabification algorithm for sinhala. In *Natural language processing–IJCNLP 2005* (pp. 438–449). Berlin: Springer.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., et al. (2002). The htk book. *Cambridge University Engineering Department*, *3*, 175.