Evolution and Overgeneralization Of Grammar

Term Project
CS784

Submitted by : Shruti Dube
Y3337

Acknowledgement

I would like to take this opportunity to thank the instructors for the course Dr. Harish Karnick and Dr. Achla M.Raina without whose guidance the this project could never be completed. I am very grateful to them for having spared their valuable time to clear my doubts. I would also like to thank them for giving me the opportunity to model a theory of language acquisition, to practically test the feasibility of a proposed theory of language acquisition and to employ my own ideas to the existing model to explain language acquisition more completely.

Contents

Introduction

Natural languages can be viewed broadly to consist of a combination of elements formed using rules of syntax and some highly language specific idiosyncratically missing elements which occur in the domain of the syntactically formed elements. This presence of ‘holes’ or exceptions, if they were, in the domain of allowable sentences prompts us to conclude that these rules of syntax are not universal, they have language specific violations and so they cannot be generalized. Learning of language cannot be done by simply looking at some sample sentences, generalizing rules on that basis and generating new sentences according to those rules. The approach to learn grammar proposed here employs the Simplicity Principle derived from the Kolmogorov Complexity

Theory using the Minimum Description Length (MDL) as a measure of complexity.

Idiosyncracies:

If we consider the English language, several of such exceptions are observable. Some examples are :

2) The most prominent is the one where the irregular form ‘went’ replaces the form ‘goed’ which is a natural deduction looking at some sample sentences of the language.

2) Similarly, the Dative Shift in which

John gave/donated a book to the library

is allowed and induces us to make the following error

John gave/*donated the library a book.

3) ‘***to be***’ deletion rule

The baby appears/seems to be happy
The baby appears/seems happy
The baby seems to be sleeping
*The baby appears/seems sleeping

4) Lexical constraints

strong/high/*stiff winds

strong/*high/*stiff currents
strong/*high/stiff breeze

5) Transitive/Intransitive: The oft proclaimed rule that transitive verbs involve both agents and patients while intransitives involve agents alone is not valid in the following

John broke the cup
The cup broke
John kissed Mary
*Mary kissed

6) We cannot interpret syntactic behaviour from semantics

John waved Mary goodbye
John waved goodbye to Mary
*John said Mary hello
John said hello to Mary

Simplicity Principle and the Kolmogorov Complexity:

Cognitive systems will always prefer simpler patterns over more complex ones. The length of the shortest program that regenerates the object is a natural measure for the complexity of the object. The length of the program is independent of the specific choice of the programming language. This length is known as the Kolmogorov complexity.

Simulation- Simplicity based Language Learning

Toy Language:

For the purpose of simulation, a toy language with the following structure was chosen:
It consisted of two syntactic categories, A and B each containing four words. The language also contained the exception element. Each sentence contained only two words, AB or BA

$S_1 = AB$
$S_2 = BA$
$A=\{ a_1,a_2,a_3,a_4\}$
$B=\{ b_1,b_2,b_3,b_4\}$
2) $= \{ (a_1),(a_2,b_2),(a_2,b_3),(a_2,b_4),(b_1,a_1),$
$(b_2,a_1),(b_3,a_1),(b_4,a_1) \}$

This language mimics the pattern of all the alternations cited above. Eg. Verbs in English can nominally occur either in a transitive or an intransitive context, but some are blocked from occurring in one or the other. This is emulated by the toy language where items in either categories can occur in principle in either the first or the second positions, but can be blocked from doing so by entries in the exceptions element.
All the grammars will be represented using codelengths

$$C= C(H)+C(D/H)$$

where C(H) represents the codelength necessary to specify the hypothesis, and C(D/H) is the codelength necessary to specify all the data that has been heard under that grammar.

Learning through Gambles:

The model is at the stage where the child has learnt the productive rules and makes overgeneralization errors. The task is to spot these ‘holes’. Learning proceeds by gambles. The learner listens to sample sentences from a grammar and then on their basis postulates an exception. In making this proposition, the learner must specify this as a part of the new hypothesis, coding which shall increase the complexity of the new hypothesis. But in doing so, it has reduced the number of sentences it can expect to encounter hence the codelength to code the data heard under the grammar decreases. Hence if the gamble is correct, the learner will eventually win back the number of bits it had taken to specify the exception, and if the sentence posited as an exception is encountered ever in the set of sample sentences, that gamble is abandoned.

Two Approaches:

2) There is a “super speaker” for all the listeners who knows the correct grammar entirely. All his utterances shall be completely correct and can have no exceptions

2)Transmission over generations: The listener listens to sentences spoken by a speaker, who himself was the listener some time ago.

The first approach, though fair enough for simulation purposes, is unrealistic.The second is closer to reality.It can be thought to be a model wherein a generation of parents instruct their successors about the exceptions in language. Children also posit exceptions on listening to their utterances.

Samples of a language were produced by a speaker and experienced by a learner agent. The speaker and the learner share the knowledge of word frequency. Learner agents initially have a completely regular hypothesis about the language. Whenever a new exceptionis posited, a new hypothesis is generated. This entails an increase in the codelength associated with each hypothesis, and a rescaling of the probabilities for the remaining sentences. The codelengths were calculated after exposure to 50,100,500,1000 sentences. This attempts to mimic the situation of the poverty of stimulus, where one never hears all sentences and receive no negative feedback

A variant of the toy language:

As a variant to the above rudimentary language, to increase the sample size and to verify the simplicity principle for a larger corpus of data, another language was considered in which there were 8 verbs and 8 nouns (instead of 4 and 4 as above). The exceptions element was common between the two languages. The two approaches cited above were used with this toy language as well.

Algorithmic Details:

All grammars are interpreted as matrices of probabilities. The words in the language were all arranged in a matrix according to the frequency of their occurrence. Using Zipf's law (the frequency of an object is equal to the inverse of it's rank when arranged in an order determined by their frequency), the probability of each word to occur as the first word and the second word was calculated as:

$$p = f/\sum f$$

where f is the frequency of that word and $\sum f$ is the sum of frequencies of n words in the distribution. To code a sentence number of bits needed

$$= \log_2(1/ (p(w_1)*p(w_2/w_1))$$

Initially there was a completely regular grammar. Then an exception was posited. The coding of an exception takes exactly the same number of bits, if it were to be encoded as a part of the data. Given that N sentences have been heard, a sentence x can be posited as an exception if the saving over those N sentences is greater than the cost needed to specify the exception,

$$\text{ie. } N\log_2(1/(1-p(x)))> \log_2(1/p(x))$$

which simplifies according to the Taylor's Expansion to

$$Np(x) > \log_2(1/p(x))$$

If the above MDL parameter is satisfied then the sentence is posited as an exception. A sentence which has been encountered in the data is never posited as an exception. After choosing an exception, a scaling up of the probabilities of the remaining sentences is done by a factor

$$= 1/(1-p(x))$$

The data codelengths for the new grammar are calculated using these new scaled up probabilities.

Results

1)First Approach :

The "super speaker" approach yields perfect results. With a sample size of 32 sentences, it was observed that the final grammar did have 11 exceptions and those 11 exceptions were the same as that present originally in the super speaker's language. ie. the learner successfully acquires the desired grammar. As was shown by Onnis Chater (2002) if we increase the number of sentences (varying it from 50 – 1000) each grammar was being exposed to, as more and more language is seen, the codelength associated with these grammars decreases and the grammar which is identical to the actual grammar possesses the smallest codelength. These findings are verified by the results.

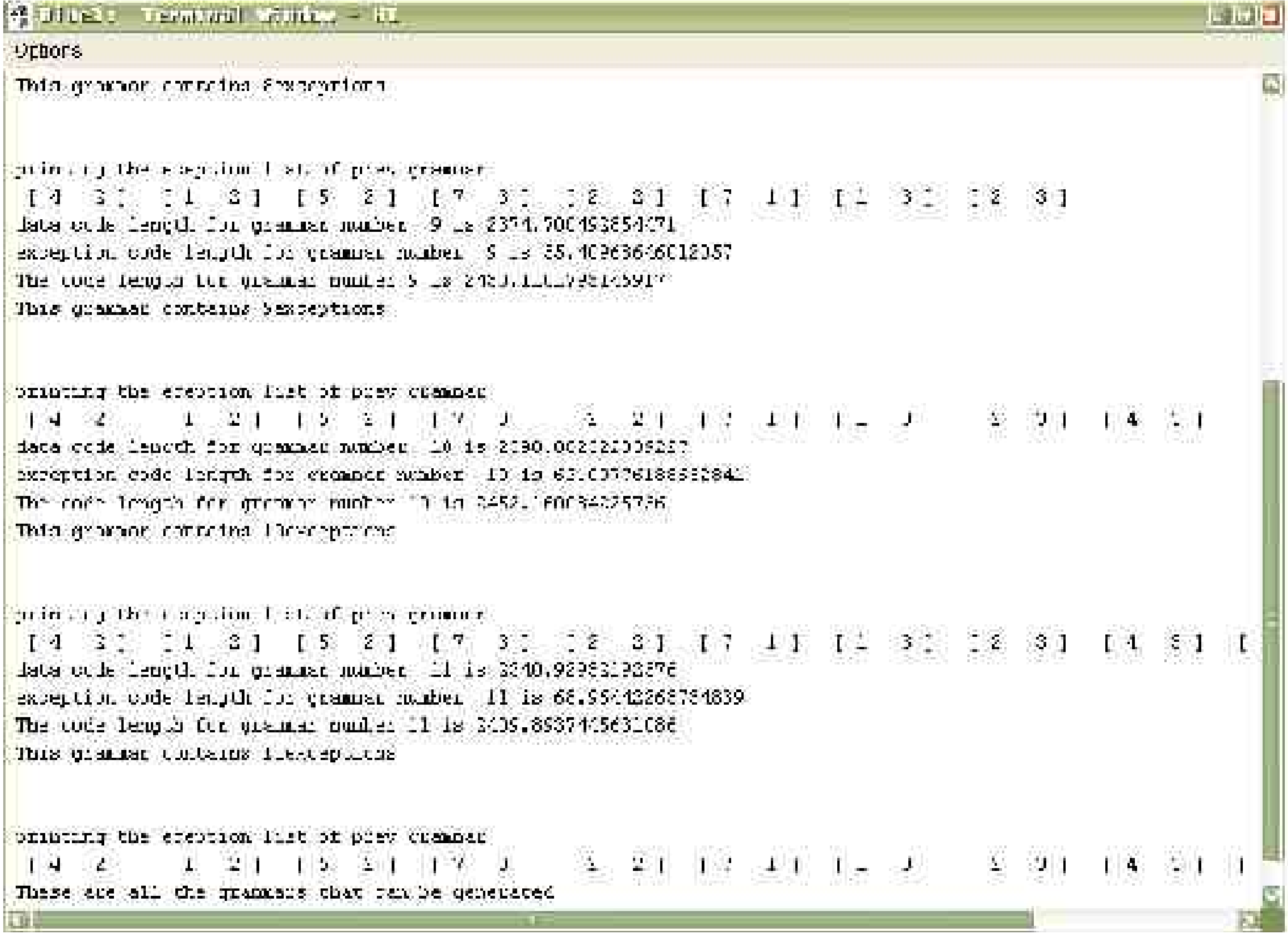
With the increased sample size as well, the grammar which had 11 exceptions was the one which had the least codelength(at times there were very minor variations). Again, all those 11 exceptions present in the grammar were found to be there in the grammar with the shortest codelength. Further grammars generated had longer codelegths.

2)Second Approach:

This approach which tried to model the transmission of the parents' exception list to the child and the acquisition of new exceptions by the child listening to utterances from the parents. This model fails and produces no converging results. Several grammars are produced for both the corpus sizes of 32 and 128 possible sentences.

This happens because, once the first learner posits an exception, he speaks keeping that exception in mind, so that sentence will never be spoken by him. He teaches his child the exception he knew (modeled by the fact that the exception element is carried forward across generations) and the child posits a new exception on the basis of the parent's utterance. Hence once an exception is posited it can never be accounted in the data to come, so it can never be removed. Hence the model isn't a good learning model.

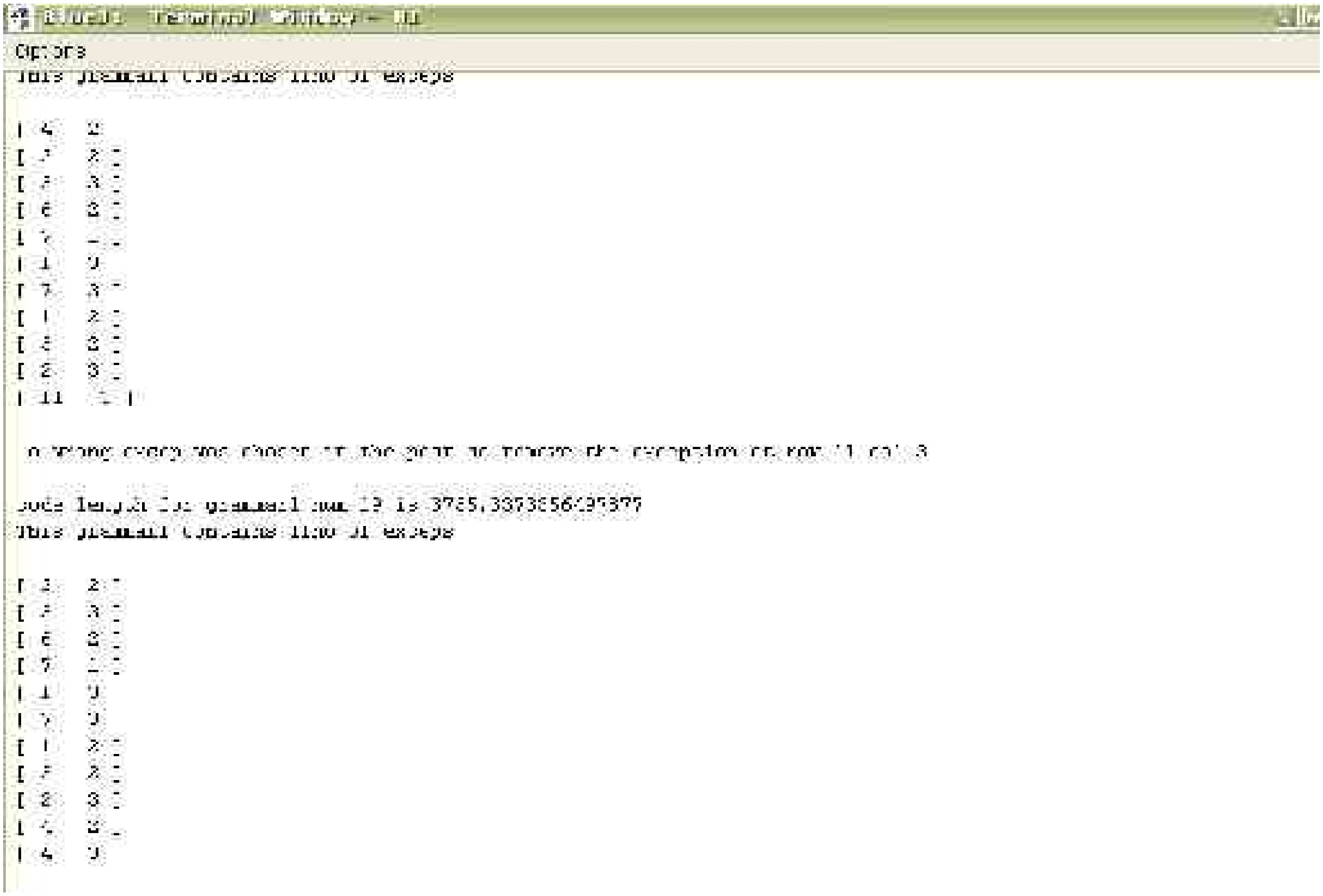
Results from the first approach with sample size = 32



```
This grammar contains 8exceptions


printing the exception list of prev grammar
 [ 4  3 ]  [ 1  3 ]  [ 5  2 ]  [ 7  3 ]  [ 2  3 ]  [ 7  1 ]  [ 1  3 ]  [ 2  3 ]
data code length for grammar number  9 is 2374.700193854071
exception code length for grammar number  9 is 55.40963646012057
The code length for grammar number 9 is 2430.1100795145917
This grammar contains 9exceptions


printing the exception list of prev grammar
 [ 4  3 ]  [ 1  3 ]  [ 5  2 ]  [ 7  3 ]  [ 2  3 ]  [ 7  1 ]  [ 1  3 ]  [ 2  3 ]  [ 4  3 ]
data code length for grammar number  10 is 2390.0022220095227
exception code length for grammar number  10 is 62.03776186532842
The code length for grammar number 10 is 2452.160034025736
This grammar contains 10exceptions


printing the exception list of prev grammar
 [ 4  3 ]  [ 1  3 ]  [ 5  2 ]  [ 7  3 ]  [ 2  3 ]  [ 7  1 ]  [ 1  3 ]  [ 2  3 ]  [ 4  3 ]  [
data code length for grammar number  11 is 3340.929532935?6
exception code length for grammar number  11 is 68.95412268734839
The code length for grammar number 11 is 3405.853741563108
This grammar contains 11exceptions


printing the exception list of prev grammar
 [ 4  3 ]  [ 1  3 ]  [ 5  2 ]  [ 7  3 ]  [ 2  3 ]  [ 7  1 ]  [ 1  3 ]  [ 2  3 ]  [ 4  3 ]  [
These are all the grammars that can be generated
```

The smallest codelength was observed for the grammar containing 11 exceptions and those exceptions were the same as in the speaker's language

Results from the first approach with a larger corpus size of 128



```
This grammar contains no of exceps

[ 4  2 ]
[ 3  2 ]
[ 3  3 ]
[ 6  2 ]
[ 7  1 ]
[ 1  0 ]
[ 7  3 ]
[ 1  2 ]
[ 5  2 ]
[ 2  3 ]
[ 11  1 ]

[illegible]

Code length for grammar num 19 is 3785.3373856495377
This grammar contains no of exceps

[ 3  2 ]
[ 3  3 ]
[ 6  2 ]
[ 7  1 ]
[ 1  0 ]
[ 7  0 ]
[ 1  2 ]
[ 3  2 ]
[ 2  3 ]
[ 4  2 ]
[ 4  0 ]
```

The smallest codelength was observed for the grammar containing 11 exceptions and those exceptions were the same as in the speaker's language

Conclusion

The problem enveloping language acquisition is that learning a language from experience alone cannot be possible because linguistic experience is too limited, quasi regular and even contradictory. Since we have finite lifetimes, and we hear only finitely many sentences, the mere absence of a construction cannot guarantee that it is an exception. The pertinent problem is that we must gauge and be able to distinguish between grammatical and ungrammatical constructions on the basis of this limited exposure. This problem becomes even more complex in case of quasi regular syntactic rules because some sentences in the space of allowable constructions are ungrammatical. This drives us to conjecture the presence of some innate learning principles: Universal Grammar However even this isn't a satisfactory explanation as the irregularities across languages are highly idiosyncratic and specific to languages. So these cannot be derived on the basis of universal principles and they necessarily must be acquired through experience. Also, the innate constraints for language acquisition posed by universal principles will cause all languages to develop perfectly compressible grammars, which isn't true for natural languages.

In the simulation above, the learner received no feedback on his learning other than more sample sentences. The sentences which have a low probability have a meager chance of getting selected. Hence these conditions weakly mirror the poverty of stimulus condition where children do not hear all possible sentences and they do not receive a negative feedback from their parents. A learner will always prefer to learn simpler grammars as opposed to more complex ones. The simplicity approach above does provide a metric for the quantitative measurement of simplicity. Also in such an approach, the poverty of stimulus, instead of becoming a drawback for language acquisition, manifests itself as a very major reason for emergence of exceptions and language evolution.

The fact that the grammar, which was the same as the original grammar, was the simplest (one with the shortest codelength) implies that the problem of language acquisition can be solved by including a bias towards the simplicity.

However, it would be highly erroneous on our part to conclude that simplicity is indeed the learning mechanism for acquiring grammars. In the first approach, each time, for each listener, a perfect set of sample sentences is being spoken. In reality this doesn't happen: one gets to listen to sentences which themselves contain exceptions. However, if these exceptions are transmitted and never removed, this will never lead to the acquisition of the proper grammar(as was shown in the second approach). We are not taking into consideration any communication functions ie. whatever is being spoken is being accurately understood without any ambiguity. However, real life communication is strongly influenced by the accent of the speaker, the stresses laid, emotions infused and also the comprehensive ability of the listener. Everytime a perfect communication does not occur. Since this language has no semantics and no communicative function, and it doesn't model the relation between meaning-signal-referents it cannot be proposed as a general learning mechanism.

X-------------X