UZBEK AND ENGLISH LANGUAGE COMPARISON ALGORITHM FOR TRANSLATOR SOFTWARE

Usmonova Kamola
kamolausmonova1234@gmail.com

ABSTRACT

The need to learn more languages is becoming important factor for all fields. Linguists are trying to compare languages to each other by many corners including language structure, sentence structure and meaning of the lexicon. As it is difficult to do this task without special program, the ADIOS program is used at large extent to show the similarities and differences of languages in the world. The aim of this work is to explain the ADIOS program and how it can be used for language comparison.

Key words: *ADIOS algorithm, corpus data, path, patterns, graph.*

Introduction

Sentences in a language can be both simple and complex for example in English language. in oreder to analyze the sentences in large amounts requires corpora based studies and application so that we can deal with raw, and authentic data. This helps for both language learning and teaching as well as comparing two languages and what are the similarities and differences are at the same time. The statistical-structural algorithm, ADIOS(automatic Distillation of structure) was developed by Solan et al (2005). This program helps to summarize precise and productive grammars out of realistic, and raw corpus data in various languages. On the other hand, the ADIOS program cannot deal with grammatically complex structures which makes the researcher to split the data into small parts and analyze in the program.

Literature review

ADIOS program

One of the common structures used by researchers today is the program of ADIOS developed by Solan et al (2005) which is used for unsupervised language learning. This takes a corpus out of some languages as an input. This program is used for generating symbolic results that are context-free(Solan et al, 2002). ADIOS can be used to show the difference between statistical, symbolic and representations. The consequences are context-free grammar that show the structures of sentences at various degrees. This ADIOS algorithm can be applied to a number of problems. The ADIOS program was

inspired by EMILE and ABL algorithms by Zellig Harris (1954) and its notion of "distributional structure.

There are a number of elements for the ADIOS algorithm:

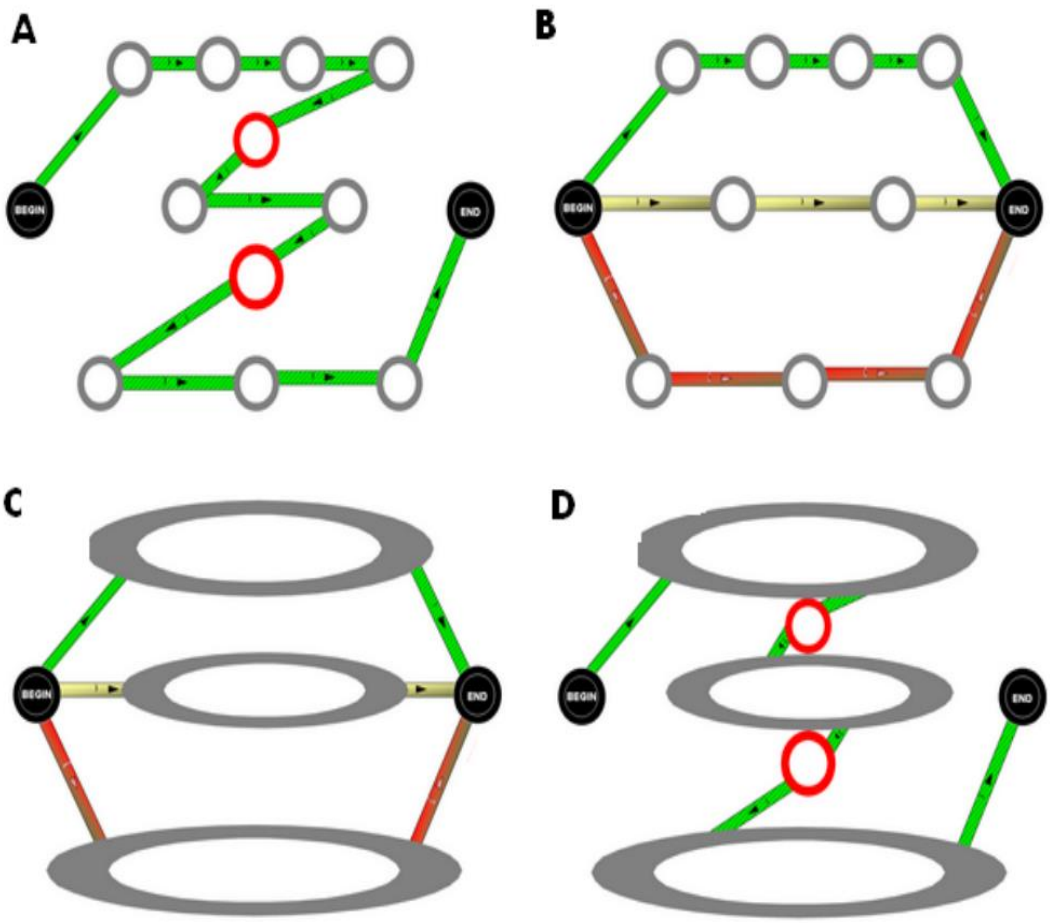
a) Graph: the ADIOS program shows a corpus as a directed multigraph, with the nodes of the graph originally showing the tiniest units within the corpus such as words, phonemes, part-of-speech tags or proteins on the basis of the problem. Each node is added by edges; as a result, the graph becomes a set of "bundles" of nodes and edges(Edelman et al, 2003). There are a BEGIN node and END node and all paths in the graph are joined by edges.

b) Paths: a sentence within the graph is displayed by a path. A path is a order of nodes, starting with the BEGIN node and finishing with the END node. The more the algorithm develops, the more nodes in the graph are reset by "patterns and equivalence" groups.

c) Patterns: they are orders of nodes. The node can be called as terminal nodes, equivalence classes r other patterns. ADIOS can be used to reduce the redundancy and replace the terminal nodes with the non-terminal nodes.

In case of cross-language syntactic comparison, the ADIOS algorithm can be used for more languages. For instance, Solan (2006) has used the program to translate the Bible translation for translating in English, French, Spanish, Swedish, and Chinese languages. The grammatical patterns were contrasted by gathering and making collocations of elements consisting resultant patterns. A pattern can contain elements of terminal nodes, equivalence classes and other patterns. Out of these six languages, Chinese was the most different one in sentence structure when analyzed in the case of ADIOS.

We use the algorithm, which consists of four phases. The first, every sentence are decomposed into simple sentences in corpus and those are split by conjunctions so the conjunctions are left out altogether from sentences. The second, the graph should be done in a simpler corpus and ADIOS algorithm is implemented. The third, the simple sentences are reconstituted in their new generalized form on condition that learning concludes. The third, the ADIOS algorithm is appealed again to the recomposed corpus. There are two phases in training: during the first stage, The learning course of action is constricted to deriving patterns from simple sentences only. Accordingly, this phase concludes then the original complex sentences are made visible to the algorithm and learning continues. Hence, conjunctions are really helpful to split complex sentence to simpler ones. It is also noteworthy fact that the ADIOS procedure, function words are often a dispersion point.

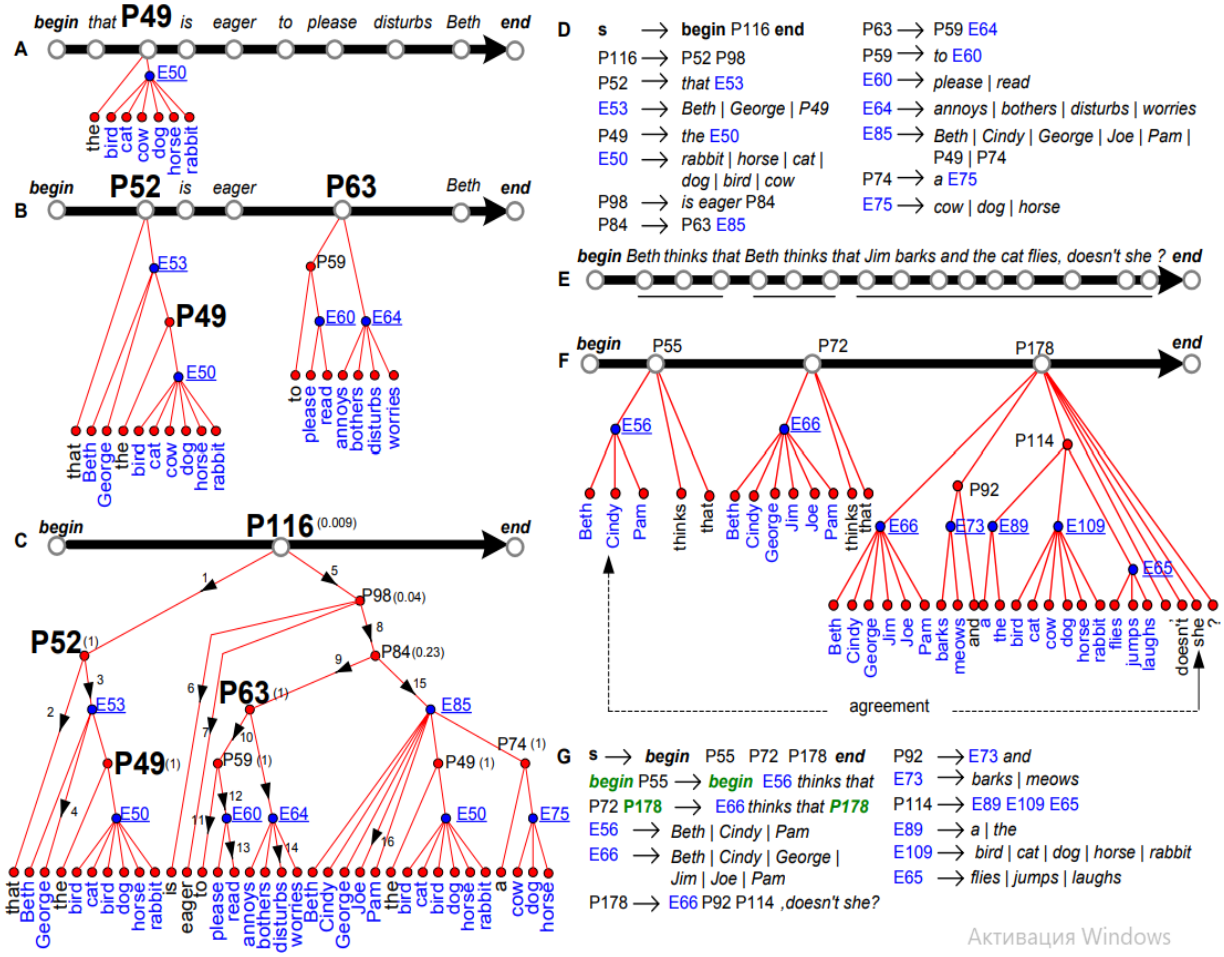


One path is shown in the graph in A. The red circle is conjunctions. In B the conjunctions nodes are split into several paths. We use ADIOS algorithm in order to carried out learning in simple paths. In C training is repeated and conjunctions are reinstated. As Diessel(2004) mentioned that we erected an artificial context- free grammar that generates three types of complex sentences.

A) Complementation - “he believes that John runs.”- u ishonadiki Jon yuguradi

(B) Relativization — “I like the books that you buy.” – Mengayoqadi, sensotibolgankitob

(c) Coordination — “After I run, I drink.” – Yuguribbo’lganimdankeyin men ichaman.



In Figure There are construction of forest trees rooted in vertices of graph (context- free grammar was generated by training data, TA1, with 50 terminals and 28 rules). A indicates pattern P49 equivalence and terminal class E50= {bird, cat, rabbit, cow, dog and horse} is executed. B) E=64 indicates verbs. C) P116 pattern adduced 896 novel sentences and eight out of that shows in the training corpus, novel sentences given like someone is eager to read disturbs someone. This pattern is root and a unit is on a final path. D sets indicate context free productions. E gives information about Initial path, which ADISON was used in context –sensitive mode. F) Patterns P55, P72 and P 178 are executed. G) there are two context- sensitive rules which is given in example are [begin P55-> Begin 56 thinks that] and [P 72 P178-> E66 thinks that P 178].

Men kuchugimni oldim va dostim bn aylanishga chiqdik

2. Garchi imtihon oson bo'lishiga qaramasdan Murod imtihondan o'ta olmadi.

3. Sayr qilib kelganimdan keyin ovqat tayyorlashni boshlayman

4. Beth va Candy universitetga qaytishganida George endigina chiqib ketayotgan edi.

5. Murodni uyida mushuk, kuchuk va ot borde uning o'rtog;inikida esa quyon va qushlar.

In order to apply linguistics into corpora it is necessary to look at information provided below:

1. Lexicograhic and lexical studies. As Hunson (2002) summarized that there are five "emphases" that change brought by corpora to dictionaries and other reference books.

Emphasis on

a) frequency
b) collocation and phraseology
c) variation
d) lexis in grammar
e) authenticity

2. Grammar studies
3. Register variation and genre analysis
4. Dialect distinction and language variety
5. Contrastive and translation studies
6. Diachronic study and language change
7. Language learning and teaching
8. Semantics
9. Pragmatics
10. Sociolinguistics
11. Discourse analysis
12. Stylistics and Literary studies
13. Forensic linguistics
14. What corpora can not tell us

Conclusion

Taking all things into consideration, the ADIOS algorithm promises the most effective results for the researchers when they work in combination with corpus data. This program can be applied in large extents including the grammatical comparison of world languages. By this way, the differences and similarities within languages can be discovered and can give ready made manual for language learners. This can prioritize which aspects of languages can be learned and which are not important to learn as it is similar to their mother tongue.

REFERENCES

1. Edelman, S., Solan, Z., Horn, D. & Ruppin, E. (2003). Rich Syntax from a raw corpus:unsupervised does it; a position paper presented at Syntax, Semantics and Statistics; aNIPS-2003 workshop, Whistler, BC, Dec. 2003.
2. Edelman, S., Solan, Z., Horn, D. & Ruppin, E.(n.d) Learning Syntactic Constructions from Raw Corpora. Tel aviv University.
3. Harris, Z. S. (1954). Distributional structure. Word 10: 146-162
4. Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge: Cambridge University Press
5. Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2005). Unsupervised learning of natural languages. Proceedings of the National Academy of Science, 102:11629–11634.
6. Solan, Z., Ruppin, E., Horn, D., Edelman, S. (2002). Automatic acquisition and e±cientrepresentation of syntactic structures. NIPS-2002.