

Dependency Parsing Based On Uzbek Corpus

Nilufar Abdurakhmonova

PhD, associate professor of Information technology department at Tashkent State University of the Uzbek language and literature

anilufar@navoiy-uni.uz

Abstract

Syntactic parsing is crucial stage among existing different types of parsing methods in the field of NLP. Syntactic parsing assists to identify the type sentence and word combinations that represented grammatical relations of the words. However, there are various grammatical features of the languages, almost all languages follow common linguistic rules. The Uzbek language belongs to agglutinative language family based on free constituent order language in syntax. Our investigations show that morphological aspect of word forms plays an essential role to identify and compose syntactic relations for the Uzbek language. Given morphological and lexical information can solve the some problems which connecting with syntactic parsing as well. Our article represents some main point of views the stages of parsing on CoNLLU format based on Uzbek corpus analysis.

Tabiiy tilni qayta ishlashda turli tahlil qilish metodlari orasida sintaktik analiz qilish muhim sanaladi. Sintaktik analiz tilning grammatik munosabatlari aks etgan soʻz birikmalari va gap turlarini aniqlashga xizmat qiladi. Tillarning turli grammatik xususiyatlari boʻlishiga qaramay, barcha tillar deyarli bir-biriga yaqin umumiy lingvistik qoidalariga boʻysunadi. Oʻzbek tili agglyutinativ tillar oilasiga mansub boʻlib, uning sintaksisi ancha erkin komponentlardan iborat. Bizning tadqiqotimizda sintaktik tahlil uchun soʻzshakllarni morfologik jihatdan sintaktik munosabatlarni tuzish va turlarini aniqlashda muhim ekanligi oʻz isbotini topgan. Morfologik va leksik maʼlumotlarning berilishi sintaktik tahlildagi lingvistik muammolarni aniqlashga ham yordam beradi. Maqolamizda oʻzbek tili korpusiga asoslangan CoNLLU formatida ifodalangan sintaktik tahlil bosqichlari tahlil qilingan.

Key words: CoNLLU, corpus, syntactic parsing, the Uzbek language

INTRODUCTION

The Uzbek language has rich paper versions of lexical resources. Currently gathering and selecting different types of Uzbek texts as a corpus implemented by Computational linguistics lab at Tashkent State university of Uzbek language and literature. One of general conception of composing computational models of corpus providing the texts is morphological analysis and syntactic parsing. Nevertheless, our corpus is not open available platform for users due to testing still the results of our project.

One is crucial issue for construction of corpus is to create the model that is ready for analyze the text morphologically and syntactically. Computational point of view grammar is more important for corpus driven language analysis. Parsing is a fundamental process in any natural language processing pipeline, since obtaining the syntactic structure of sentences provides us with information that can be used to extract meaning from them: constituents correspond to units of meaning, and dependency relations describe the ways in which they interact, such as who performed the action described in a sentence or which object is receiving the action (Carlos Gyzmez-Rodrigue, 2010)

As of early September 2018, there are 132 treebanks for 74 languages publicly available at <http://universalddependencies.org/>,¹ with 15 upcoming treebanks for a further 13 languages. New UD treebanks are often the result of converting corpora adhering to other annotation schemes—not only dependency-based, but also constituency-based (Adam Przepiorkowski , 2016).

THE NATURE OF UZBEK GRAMMAR

Grammar consist of two parts: morphology and syntax. Both of them are important layer of linguistics for NLP. The Uzbek language morphemes derived a number of combinations of word forms by concatenation root and affixes in most cases. Morphotactics of the language not every time follows the rules owing to some exceptions though there are an exact order of word combination.

Usual order of the words:

Root+DerAff+Pl+Poss+Case+Particle (kutub-xona-lar-i-ga-mi);

Root+ DerAff+Voice+Neg+Tense+Particle (oq-la-t-tir -ma-di-mi)

The general structure of the Uzbek language of the sentence follows SOV order. It is free constituent order because of all parts of speech nearly depends on the verb, therefore, there is not difference the meaning though changing places of parts of speech. Due to complex structure of the sentences with grammatical morphemes as a sequence of inflectional elements of parts of speech. We can see one example for this:

Men muzeyga sen bilan ertaga boraman.

$S^0=[a+b+c+d+f]$

Men sen bilan ertaga muzeyga boraman.

$S^1=[a+c+d+b+f]$

Men ertaga muzeyga sen bilan boraman. $S^2=[a+d+b+c+f]$

Even some inflectional groups give opportunity to identify the function of words by morphological markers. [Noun+Case] model as syntactic marker helps to identify the function of the words in the text, but not every time.

Noun+Gen=> Attributive (bolaning – child’s)

Noun+Acc=>Object (bolaga-to the child)

Conducting our research dedicated more on syntactic annotation in order to create the model of corpus analysis.

Our aim is to build Treebank as an example universal dependency structures so that to use them for NLP. The Uzbek language is more specialized morphological features sequences by order adding morphemes. Hence, representation morphological features is crucial for syntactic parsing as well.

Subsequently, we should clarify the forms and POS in the Uzbek language. The grammatical structure of words in Uzbek can be in following forms:

- 1) morphemes (affix and stem) – *chorvachilik*
- 2) morphological variations which composed different functions of parts of speech – *kitob + lar + im + ga (to my books)*
- 3) word combinations as syntactic level – *tug'ilgan kun uchun sotib olmoq*
- 4) compound words – *mashq qilmoq, sotib olmoq*
- 5) phrasal units represented as unique components – *yuragi dov bermaslik*

For accuracy of our parser needs morphological analysis. In Uzbek, being ambiguous word structures confuses the type of the grammatical features as example of verb:

[V]+ib+V=>*Sotib olmoq*-**compound verb**

[V]+ib +V=>*Gulib gapirmoq*-**adverbial clause**

[V]+ib +V=>*Kulib yubormoq*-**collaction**

Giving lexicon and rules for each lexical unit allow us to establish their lexical features and combine above pointed types of word forms though they are alike formally. Authors pointed out Word-based Model and IG-based model for choosing parsing units according to grammatical features (Gülşen Cebiroğlu Eryigit; Kemal Oflazer; Joakim Nivre, 2013).

Hence, we use FST technology to analyze at the first morphological analyzing, thanks to the Helsinki finite state technology it builds the amount of combination of morphemes (Abdurakhmonova N.; Tuliyeu U., 2018).

Furthermore, not only morphological categories but also syntactic relations between the words are important to classify the set of sentences in Treebank for Uzbek. Hence, we use tag set to identify each sentence type and word combination through morphological word forms.

WC	Word combination
COL C	Collocation
FP\F COL C	Free phrase\ Free collocation
NP	Noun Phrase
NA	Noun Adjoinment
NG	Noun Government
NCS	Noun Collateral subordination
VP	Verb Phrase
VA	Verb Adjoinment
VG	Verb Government
AG RM	Agreement
SLP	Singular personal pronouns
PPL	Plural personal pronouns

ICN\ CPC T	Interconnectedness\Complicity
S	Simple sentence
Sub	Subject
Obj	Object
Attr	Attributive
Mod	Modifier
Pre	Predicate

Probability of syntactic structures by pure grammatical approach is more complex than statistical approach by corpus. Consequently, it is necessary to be exist the corpus in order to construct dependency tagging.

METHODOLOGY OF PARSING

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. Our general workflow of parsing stages represented the following pic. 1.

The corpus consists of hand built selection of Uzbek fiction annotation with metadata respectively by genres. Here grammatical categories are crucial to give representativeness of features of parts of speech.

A special program intersecting composition was developed in order to facilitate the combining of the lexicon transducer and the two-level rule transducers (TWOLC-two-level compiler) and to avoid excessively large intermediate results (Alexandr Rosen, 19).

In order to morphological analysis there are three components of the Uzbek language: alphabet (Latin and Cyrillic), grammatical rules and Lexicon. In Uzbek the following morphotactics of words as example of Noun:

```

LEXICON NumC
+SG: Poss1;
+PL:lar Poss2;
LEXICON NumV
+SG: Poss2;
+PL:lar Poss2;
LEXICON Poss1
+PP1+PSG:m Case;
+PP2+PSG:ng Case;
+PP3+PSG:si Case;
+PP1+PPL:miz Case;
+PP2+PPL:ngiz Case;
+PP3+PPL:i Case;
0:0 Case;
LEXICON CyrPrePrefinal1
0:0 Final;
+PART:ми Final;
+PART:ку Final;
+PART:-^Ya Final;
+PART:-да Final;
+PART:-чи Final;

```

The algorithm of analysis represented Fig. 1.

We apply Turkish model to analyze the texts for CoNLLU format, hence there have been the sharp distinction between Turkish and Uzbek structures, but thank to by human correction, grammatical features tagging improved according to

newpar

```

# sent_id = 266
# text = Shoir yigitga dil-dildan achinarkan , uni ilk bor
uchratgan paytini esladi .
1      Shoir  Shoir  NOUN  Noun
      Case=Nom|Number=Sing|Person=3      2
      nmod      _      SpacesAfter=\r\n
2      yigitga  yigitga  NOUN  Noun
      Case=Nom|Number=Sing|Person=3      3
      nmod      _      SpacesAfter=\r\n
3      dil      dil      NOUN  Noun
      Case=Nom|Number=Sing|Person=3      13
      nsubj     _      SpaceAfter=No
4      -      -      PUNCT  Punc  _      13
      punct     _      SpaceAfter=No
5      dildan  dil      NOUN  Noun
      Case=Abl|Number=Sing|Person=3      6      obl
      _      SpacesAfter=\r\n
6      achinarkan  achin  VERB  Verb
      Aspect=Perf|Mood=Ind|Polarity=Pos|Tense=Pres|
VerbForm=Part  13      acl      _
      SpacesAfter=\r\n
7      ,      ,      PUNCT  Punc  _      13
      punct     _      SpacesAfter=\r\n

8      uni      u      NOUN  Noun
      Case=Acc|Number=Sing|Person=3      11      obj
      _      SpacesAfter=\r\n
9      ilk      ilk      ADJ   Adj   _      10
      amod     _      SpacesAfter=\r\n
10     bor      bor      NOUN  Noun
      Case=Nom|Number=Sing|Person=3      11
      obl      _      SpacesAfter=\r\n
11     uchratgan  uchrat  VERB  Verb
      Aspect=Perf|Mood=Ind|Polarity=Pos|Tense=Pres|
VerbForm=Part  12      acl      _
      SpacesAfter=\r\n
12     paytini  payt  NOUN  Noun
      Case=Acc|Number=Sing|Number[psor]=Sing|Per
son=3|Person[psor]=3      13      obj      _
      SpacesAfter=\r\n
13     esladi   esla   VERB  Verb
      Aspect=Perf|Mood=Ind|Number=Sing|Person=3|
Polarity=Pos|Tense=Past  0      root      _
      SpacesAfter=\r\n
14     .      .      PUNCT  Punc  _      13
      punct     _      SpacesAfter=\r\n\r\n

```

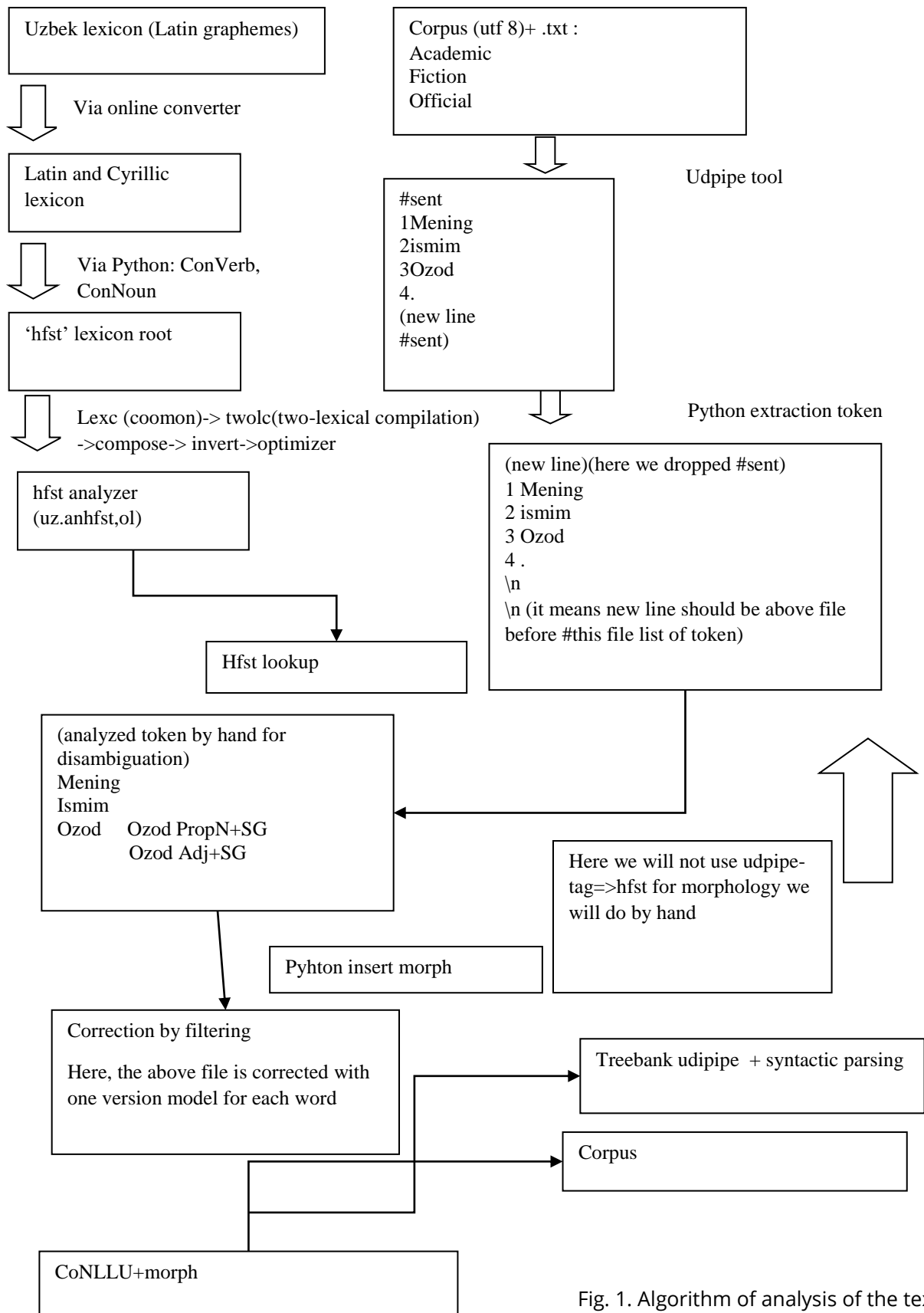
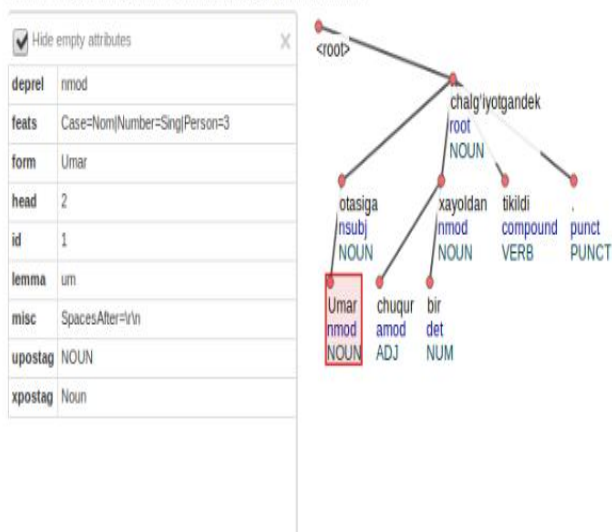


Fig. 1. Algorithm of analysis of the text

Applying CONLLU tool of universal dependency it will be represented by the following graph:

Umar otasiga chuqur bir xayoldan chalg'iyotgandek tikildi .



Oflazer, Kemal. (1994) Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Conclusion

Universal dependency is productive tool to analyze syntactic structures of the text for relative languages. Considering the importance of syntactic parsing in corpus analysis, give good opportunity to model a number of syntactic structures of the text. One of our conclusion is manual improvement given grammatical features of each sentence of corpus can provide for disambiguation through no grammar but morphological component of parts of speech.

Acknowledgement

I would like to express my sincere gratitude to Doctor Loic Boizou for his expert advice and encouragement throughout this project by implemented “El-yurt umidi” foundation (2018, at Vytautas Magnus University in Lithuania), as well as the head of Computational linguistics center Dr. Andrius Utka for his brilliant advices.

References

- Adam Przepiórkowski, Agnieszka Patejuk (2016) From Lexical Functional Grammar to enhanced Universal Dependencies The UD-LFG treebank of Polish Lang Resources & Evaluation <https://doi.org/10.1007/s10579-018-9433-z>
- Alexandr Rosen Morphological Tags in Parallel Corpora file:///C:/Users/user/Downloads/Morphological_Tags_in_Parallel_Corpora.pdf
- Carlos Gymez-Rodriguez (2010) Parsing Schemata for Practical Text Analysis Imperial College Press 5
- Gülşen Cebiroğlu Eryiğit, Kemal Oflazer, Joakim Nivre (2008) *Computational Linguistics* · December
- Abdurakhmonova N., Tuliyeu U. (2018), Morphological analysis by finite state transducer for Uzbek -English machine translation / *Foreign Philology: Language, Literature, Education*. №3 (68), - P. 59-66