Sentence syntax trees should be made from morphemes. Semantically ordered trees.

Dinar Qurbanov
Contacts: qdinar in Gmail, Twitter, Vk.com, Facebook.
January 26 – April 15, 2015.

Abstract

Some critique of construction of sentence parse trees in modern linguistics. Two propositions on constructing trees, as mentioned in the title. Introduction of an English-to-Tatar translator program that is being developed by the author. Precedence by specificity. There are no new ideas for science, but, this paper is for popularisation of known ideas.

Preface

I, the author, do not know English well. Also, I do not know well how to write in scientific paper format. I have not cited some texts from previous works I found because I am not sure whether they are appropriate.

1. First proposition: Sentence syntax trees should be made from morphemes

Currently, people make English language sentence constituency trees with words at leaf nodes, and dependency trees with words at nodes, i.e. they use words as basic elements.

I think, morphemes should be used as basic elements of sentence syntax trees, in canonical form of trees, in grammar books, etc.

1.1 Demonstration of modern situation. How people draw trees.

1.1.a English language syntax tree usage (construction) examples:

1) Wikipedia: See fig. 1.

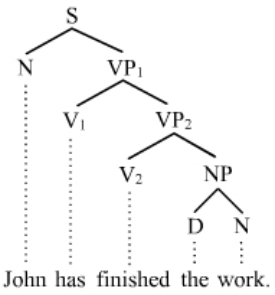

Figure 1. A tree from Verb phrase, Wikipedia.

2) Stanford Parser sample: Parse tree of "*The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today."* is:

```
(ROOT
 (S
  (S
   (NP
    (NP (DT The) (JJS strongest) (NN rain))
    (VP
     (ADVP (RB ever))
     (VBN recorded)
     (PP (IN in)
      (NP (NNP India)))))
   (VP
    (VP (VBD shut)
     (PRT (RP down))
     (NP
      (NP (DT the) (JJ financial) (NN hub))
      (PP (IN of)
       (NP (NNP Mumbai)))))
   (, ,)
   (VP (VBD snapped)
    (NP (NN communication) (NNS lines)))
   (, ,)
   (VP (VBD closed)
    (NP (NNS airports)))
   (CC and)
   (VP (VBD forced)
    (NP
     (NP (NNS thousands))
     (PP (IN of)
      (NP (NNS people))))
    (S
     (VP (TO to)
      (VP
       (VP (VB sleep)
        (PP (IN in)
         (NP (PRP$ their) (NNS offices))))
       (CC or)
       (VP (VB walk)
        (NP (NN home))
        (PP (IN during)
         (NP (DT the) (NN night))))))))))
  (, ,)
  (NP (NNS officials))
  (VP (VBD said)
   (NP-TMP (NN today)))
  (. .)))
```

(2)

3) Others: Same way is used in lots of other works; for example: Carnie, 2000, p.39; the "Parse tree", "Constituent (linguistics)", "Dependency grammar" Wikipedia articles; Kulick & Bies & Mott, 2012; Rosa et al, 2014.

4) But, usage of morphemes separately is also known in science, see section 1.5.

There are not many word forms in English language, and such trees are not very bad. If they used morphemes, their trees would be harder to read.

1.1.b Parse trees in other language linguistics:

1) Modern Tatar philology dependency scheme example, see fig. 3.

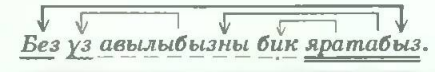

Figure 3. A tree from a Tatar language schoolbook (Miftakhov & Sungatov, 2002).

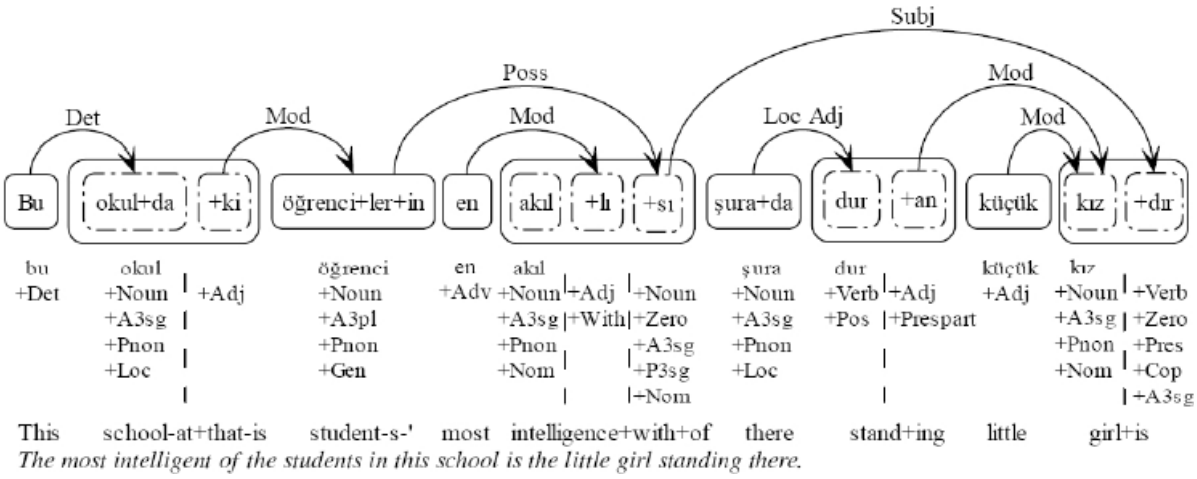

Figure 4. A tree from Oflazer, 2007.

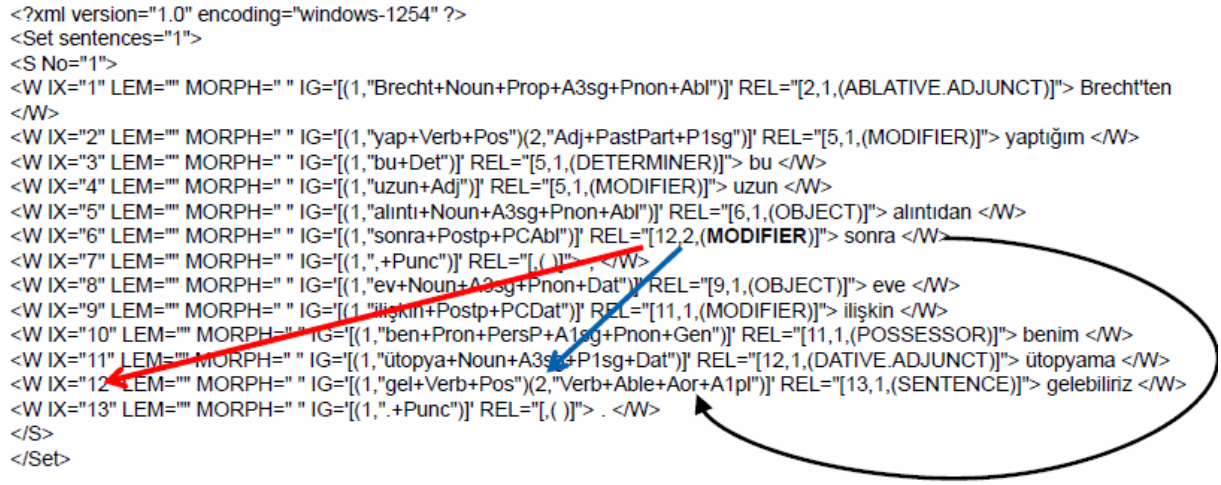

Figure 5. A tree from Oflazer, 2007.

2) Modern Turkish linguistics example, see fig. 4, 5. Words (*öğrencilerin*), parts of word (*okulda*), "inflectional groups" (*biliriz*, I think, it is just an archaic word), consisting of several morphemes, are in some of leaf nodes. But morphemes are separated and set in some nodes, examples are in "1.5 Previous works" section.

3) Modern Arabic linguistics example, see fig. 6. Though some morphemes, like "*bi*", are shown with different colour and regarded as separate element, case endings are not separated, and complex modifications of root are not shown as abstract morphemes, and "*al*"s are not set at tree nodes.

But, I think, what the author, Kais Dukes, has chosen, is appropriate for his purpose; if all morphemes were separated, the trees would not be beautiful and easy.

4) Japanese and English trees, that are made from words, can be seen in the papers referenced by this paper.

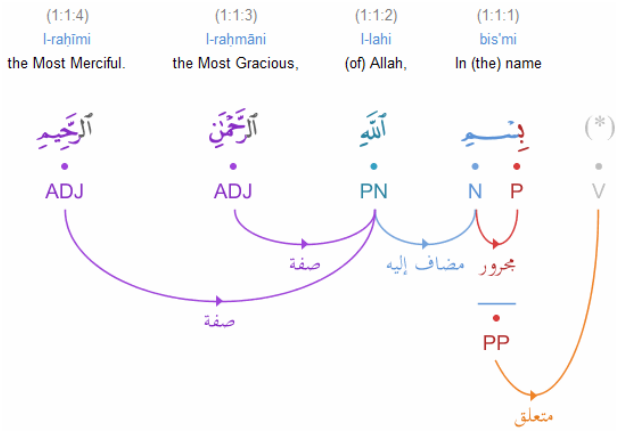

Figure 6. A tree from Dukes, 2010.

1.2 Demonstration of my first proposition. How do I suggest drawing parse trees.

Example tree for my proposition: see fig. 7. I have marked here phrases, which are made from

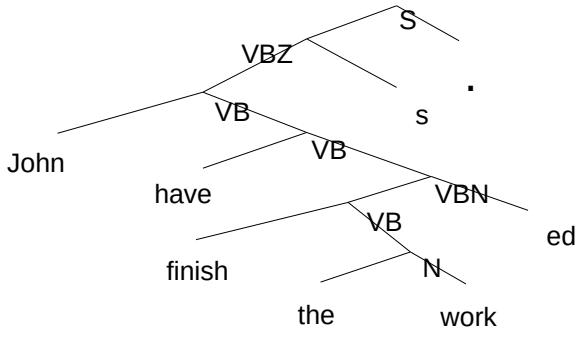

Figure 7. My proposition. Morphemes are at leaf nodes.

morphemes (and from inner phrases), (subtrees) with Penn Treebank part-of-speech tags. (Originally, in Penn Treebank, and in other works, that tags, *VB*, *N*, etc, are used to mark words, and other tags like *VP*, *PP*, *S* are used to mark phrases, which are made from words (and from inner phrases). For examples, see section 1.1.a of this article.)

1.3 Reasons

Though there are not many word forms in English language, there are reasons to use morphemes at sentence tree nodes:

1) In speech, words may be, sometimes, like joined ("for them") or, sometimes, like split ("underestimate"); written spaces and words are just a writing tradition, and, so, should not be blindly followed.

2) If the traditional word-based trees are used for translation, they would have to be transformed into this structure, for example:
There are several phrases like these 5, made from a phrase from Tree 2:

(VP (VBD snapped)
(NP (NN communication) (NNS lines)))

(VP (VB snap)
(NP (NN communication) (NNS lines)))

(VP (VBN snapped)
(NP (NN communication) (NNS lines)))

(VP (VBG snapping)
(NP (NN communication) (NNS lines)))

(VP (VBZ snaps)
(NP (NN communication) (NNS lines)))

But they all would be translated into another language by translating

(VP (VB snap)
(NP (NN communication) (NNS lines)))

with addition of something corresponding to past simple or past participle or other (*s*, *ing*) suffixes.

1.5 Previous works that use morphemes as tree nodes

This proposition is not result of a formal scientific research; I had seen this incorrectness when I saw how we drew word dependency connections in our Tatar (see fig. 3) and Russian language lessons, and later I saw that same way is used also in English linguistics. But, for I am writing in format of scientific paper, I should make some bibliography research.

1) Cite from Embick & Noyer, 2005:
In its essence the Distributed Morphology approach to morphology is syntactic. As a consequence of the architecture of the grammar, in the simplest case, morphological structure and syntactic structure are the same. Because there is no Lexicon in which complex objects are assembled according to rules distinct from the rules of syntax, the generation of all complex forms must be performed in the syntax.

2) Oflazer, 2007: See fig. 4 and 5. Morphemes are inconsistently separated and set at tree nodes, examples: *lı*, *dur*, *ki* etc in fig. 4.

3) Dukes, 2010: See fig. 6. Some morphemes, like "*bi*", are shown with different colour and regarded as separate element.

4) Hangstom, 2001; Carnie, 2000, p. 155: *-ed* suffix is used separately.

5) Previous publications by me: 1) Handwritten notes on linguistics in a notebook, I have written that nearly in 2000. I used morphemes at sentence tree leafs. 2) English-Tatar translator program, written in PHP, 2013-2015, it's not production software for now. 3) A blog post, 2013.

2. Second proposition: Semantically ordered trees

There is an idea of semantically ordered binary tree of syntax of sentence, consisting of morphemes, with head-final order of child branches in each node. Also similar tree with head-initial order exists. I think, usage of these forms of tree is better, I explain that in Reasons section below. Example of such tree:

```
(
  0:(
    0:(
      0:John
      1:(
        0:(
          0:(
            0:(
              0:the
              1:work
              )
            1:finish
            )
          1:ed
          )
          1:have
        )
      )
    1:s
    )
  1:.
)
```

(8)

By saying "semantically ordered", I mean the consistent ordering of main part (child branch) and

dependent part (child branch) in every constituent (in every node of tree).

2.1 Demonstration in form of tree

See fig. 9.

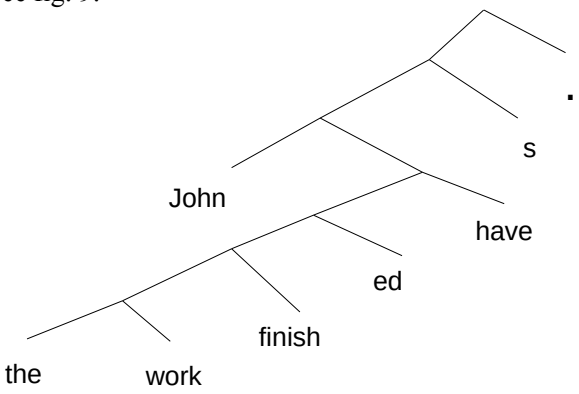

Figure 9. My proposition. Dependents (specifiers) are at left branches, heads (main parts) are at right branches. (and morphemes are at terminal nodes).

2.2 Reasons

1) Trees are intended to show semantic structure of sentence, at least, as I regard them, and, the phonetic phenomenon (phenomenon of level of 1-dimensional speech) of morpheme order should not be mixed into/with the semantic order (precedence, hierarchy, priority) representation.

2) Tatar and English languages become very similar in this form and easy to translate from one to another, and I believe, also all other natural languages mostly use semantically similar morphemes and similar trees of this form and they are easily translatable to each other. For example, how much semantic tree of the English sentence (Tree 8) is similar to corresponding semantic tree of Tatar language sentence (Tree 10):

```
(
   0:(
      0:(
         0:Джон (John)
         1:(
            0:(
               0:(
                  0:эш (work)
                  1:сен
                  )
               1:не
               )
            1:тәмамла (finish)
            )
         )
      1:ды (ed)
      )
   1:.
)
(10)
```

3) This approach is already used in science of computer programming languages and known as reverse Polish notation and postfix notation, (and reverse version as forward Polish notation and prefix notation).

2.3 Some ideas

1) The trees were semantically ordered almost in all previous works, because head and dependents are usually marked either by part-of-speech tags (like in fig. 1) or by arrows (like in fig. 6) or by higher position (in dependency trees). So, this my proposition is just about better visual representation of such order, and, in computing, about detecting and direct indexing/marking/sorting of the main (head) part (child branch) and dependent part (child branch) of each branching point.

2) You can see that main words has similar position in both languages in this example: (John (work (finish))). I think, this precedence of verb and its specifiers is same phenomenon as with order of noun with specifiers, when word, which is for more close feature of noun, takes more close position to it. Example: *big red book*, colour is usually more inseparable feature of thing than its size, and for that reason, its word (*red*) joins to head noun phrase with higher precedence (earlier) than the word of size (*big*). Same happens in Arabic, and corresponding words get reverse order in speech, because adjective should be located after the word it specifies (the word that is specified by it), in Arabic. By the "semantically ordered trees", I suppose, all languages will have similar structures for corresponding sentences (unlike the dissimilarity of word order, in speech level, between Arabic and English). (Same explains the order (John (work (finish))): *work* is more close feature of *finish* compared to *John*, so *work* takes precedence over *John* in joining order.)

A cite from a previous work: Halliday & Matthiessen, 2004: *So there is a progression in the nominal group from the kind of element that has the greatest specifying potential to that which has the least.*

2.4 Previous works

I have found some previous works, also for the sake of writing in scientific paper format, but also I introduce my other work:

1) Similar way of head-final ordering of English was proposed by Taijiro Tsutsumi (1986), Katsuhito Sudoh et al (2011), Isao Goto et al (2012).

2) I have used this idea in my prototype English-to-Tatar translator. Example from its output:

```
Array
(
   [0] => Array
      (
         [0] => i
         [1] => Array
            (
               [0] => Array
                  (
                     [0] => every
```

```
                [1] => day
            )
        [1] => Array
            (
                [0] => Array
                    (
                        [0] => school
                        [1] => to
                    )
                [1] => go
            )
    )
)
[1] => pr-si
)
```

(11)
I have come to it like the Japanese scientists, in process of planning translator program; Tatar language has morpheme order similar to Japanese', it's almost fully head-finally ordered, i.e. it has semantic order, and I just ordered English also in same way, in parsing stage.

2.5 Previous works regarding some other features of the proposed tree

1) Binary trees for syntax were proposed previously by Richard S. Kayne (1984).
2) Usage of trees for semantic structure, also, is not a new thing, for example, more semantically ordered trees are named “deep structures” in Chomsky, 1965.

3. General conclusion for the two propositions

Head-final or head-initial tree of morphemes, with usage of abstract morphemes for some languages, can be regarded as semantic layer of grammar of language. Children in each {pair of children of a node} comply rules of locating (taking their places) after or before each other, when they, the child morphemes (or phrases) become part of speech (in speech layer); and each morpheme or constituent has its rules (properties) of phonological joining with morphemes which become its closest neighbours in speech. For example, any adjective in Arabic takes place after head, but any adjective takes place before head in English, Turkish, Russian, Japanese. Another example: specifier object of morpheme “to” takes place after head i.e. after “to”, or corresponding morpheme of other language, in English, Arabic, Russian, (if cases are not counted; actually, there are a case morpheme, in tree, between them, in these languages), but takes place before head in Turkish and Japanese. Words appear just as phonetically joined blocks of morphemes at surface layer (speech layer), and, semantically, that morphemes (which compose a word) do not have so much close relation to each other.
(If dependent consist of many morphemes, it can get several copies of main part into itself, in some languages, for example, in Russian:
(krasivaya kniga) (abstract dative case morpheme) becomes
krasivoy knige
in speech level; i.e. it takes this form (surface structure):
(krasivaya dative) (kniga dative)
, before it comes to speech level; same happens in Arabic).
Both my propositions, and also binary trees, semantic trees, and idea of order by specificity are already presented in linguistics, and this my paper has at least intention of popularisation of these ideas.

References

Dukes, Kais. 2010. The Quranic Arabic Corpus. Syntactic Treebank - Dependency Graphs. http://corpus.quran.com/treebank.jsp , 2015-02-11.

Carnie, Andrew. 2000. Syntax: A generative introduction. Published by Blackwell Publishers in 2002, 2006, 2013.

Chomsky, Noam. 1965. Aspects of the Theory of Syntax, p. 16. MIT Press.

Embick, David; Noyer, Rolf. 2005. Distributed Morphology and the Syntax/Morphology Interface, p. 10. To appear in *The Oxford Handbook of Linguistic Interfaces (2007)*. http://babel.ucsc.edu/~hank/mrg.readings/E_N_DM_SM_Interface.pdf , 2015-02-02.

Hagstrom, Paul. 2001. CAS LX 522 Syntax I, Handouts, Week 3: X-bar Theory, p. 13. http://www.bu.edu/linguistics/UG/course/lx522-f01/ , 2015-02-19.

Halliday, M.A.K. & Matthiessen, C.M.I.M. 2004. An Introduction to Functional Grammar, p. 322. Hodder Arnold.

Isao Goto, Masao Utiyama, Eiichiro Sumita. 2012. Post-ordering by Parsing for Japanese-English Statistical Machine Translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island, Korea. Association for Computational Linguistics. http://www.aclweb.org/anthology/P12-2061 .

Katsuhito Sudoh, Xianchao Wu, Kevin Duh, HajimeTsukada, Masaaki Nagata. 2011. Post-ordering in Statistical Machine Translation. In *Proceedings of the 13th Machine Translation Summit*. http://www.mt-archive.info/MTS-2011-Sudoh.pdf .

Kayne, Richard S. 1984. Connectedness and Binary Branching, pp. 133-5. Foris Publications.

Kulick, Seth & Bies, Ann & Mott, Justin. 2012. Using Supertags and Encoded Annotation Principles for Improved Dependency to Phrase Structure Conversion. *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. http://papers.ldc.upenn.edu/NAACL2012/NAACL2012KulickBiesMottPaper.pdf .

Мифтахов, Б. М., Сөнгатов, Г. М. 2002. Татар теле (Tatar language, for 7th grade of school), p. 28. Мәгариф.

Oflazer, Kemal. 2007. The Turkish Treebank. *Treebank Workshop*. Hindi/Urdu Treebank Project at University of Washington. http://faculty.washington.edu/fxia/treebank/workshop07/agenda.htm , 2015-01-30.

Qurbanov, Dinar. 2003. Notes on linguistics, pp. 57, 58, 60. http://qdb.narod.ru/tattyazmaindex.htm .

Qurbanov, Dinar. 2013. Right (correct) analysis of phrase structure… http://qdb.wp.kukmara-rayon.ru/2013/11/26/right-correct-analysis-of-phrase-structure/ .

Qurbanov, Dinar. Tarjima, output of index2.php. *https://github.com/qdinar/tarjima/archive/812c17e3cfd340001a8ae5aa445bde94b4eb6b29.zip* .

Rosa, Rudolf & Mašek, Jan & Mareček, David , Popel, Martin & Zeman, Daniel & Žabokrtský, Zdeněk. 2014. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. http://www.lrec-conf.org/proceedings/lrec2014/summaries/915.html .

Taijiro Tsutsumi. 1986. A Prototype English-Japanese Machine Translation System for Translating IBM Computer Manuals. *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, pp. 646—648. http://www.aclweb.org/anthology/C86-1152 .

The Stanford Parser: A statistical parser. The Stanford NLP (Natural Language Processing) Group. http://nlp.stanford.edu/software/lex-parser.shtml#Sample , 2015-02-02.

Verb phrase - *Wikipedia, the free encyclopedia.* https://en.wikipedia.org/wiki/Verb_phrase , 2015-02-02.