Experiments on human incremental parsing

Leonid Mityushin
Institute for Information
Transmission Problems
Russian Academy of Sciences, Russia
lmityushin@gmail.com

Leonid Iomdin
Institute for Information
Transmission Problems
Russian Academy of Sciences, Russia
iomdin@gmail.com

Abstract

Experiments have been conducted in which the subjects incrementally constructed dependency trees of Russian sentences. The subject was successively presented with growing initial segments of a sentence, and had to draw syntactic links between the last word of the segment and the previous words. The subject was also shown a limited right context – a fixed number of words following the last word of the segment. The results of the experiments show that the right context of 1 or 2 words is sufficient for confident incremental parsing of Russian narrative sentences.

1 Introduction

The concept of incremental text comprehension implies that at any moment the reader/listener has a complete or almost complete linguistic and pragmatic interpretation of the part of the text perceived up to that moment, and that this interpretation, as a rule, does not change after new parts of the text have been perceived. Usually, this concept is used with regard to language learning (especially reading learning), literary studies, nontrivial semantic and pragmatic comprehension, and logical inference, which requires full understanding of subtle context; see e.g. a recent paper by E. Fischer et al. (2019). The aim of this work is to evaluate whether this is true for human comprehension of the syntactic structure of a text (as a matter of fact, of an individual sentence).

We have conducted experiments on incremental construction of dependency trees for Russian sentences. The subjects in the experiments were linguists with considerable experience of syntactic annotation. In a single experiment, the subject was successively presented with growing initial segments of a certain sentence, and had to draw syntactic links between the last word of the segment (**the active word**) and the previous words (**the left context**); the syntactic links created up to a certain moment form a partial syntactic structure of the sentence. At each step, the subject was also shown a limited **right context** – a fixed number of words following the active word. Three series of experiments have been conducted for the lengths of the right context 0, 1 and 2, with 100 sentences processed in each series.

2 ETAP syntactic model

We use the representation of syntactic structures of sentences in the formalism of dependency trees adopted in the ETAP multilingual multifunctional linguistic processor (Iomdin et al., 2012) and originally introduced by I. Mel'čuk (1974, 1988). The nodes of a dependency tree are the words of the sentence; punctuation marks are not included and constitute a kind of additional data – unlike, for example, the practice of the Universal Dependencies approach (https://universaldependencies.org). The nodes are connected by directed arcs called syntactic links, which are labelled with names of syntactic relations. The lists of syntactic relations for Russian and English include about 70 and 60 relations respectively.

Based on the ETAP syntactic formalism, a treebank named SynTagRus has been created which at present contains about 1.1 million words of Russian text (Dyachenko et al., 2015; Inshakova et al., 2019). Due to the complexity of the ETAP syntactic model, the developers of SynTagRus have always paid special attention to the reduction of the number of human errors. As a rule, each new sentence in SynTagRus is processed twice, by two different people: the annotator, who creates the complete

syntactic structure of the sentence (using the raw results produced by the ETAP linguistic processor), and the editor, whose role is to check the structures created by the annotator.

The syntactic link $a \rightarrow b$, where a and b are words of the sentence, is called projective if all the words between its head node a and dependent node b are directly or indirectly dominated by the word a, and non-projective otherwise. About 8% of syntactic links in SynTagRus are non-projective.

In SynTagRus, dependency trees for sentences with ellipsis contain additional "phantom" nodes that represent omitted words. Although ellipsis is not very frequent in Russian texts, it appears quite regularly; the proportion of elliptical sentences in SynTagRus is about 2%.

3 Modifications to the syntactic model

To facilitate the incremental construction of Russian dependency trees, certain modifications were made to the representation of subtrees containing prepositions and conjunctions; we will describe these changes using similar English examples. In the ETAP syntax, prepositions/conjunctions dominate the noun/verb groups that follow them. For example, the sentences

(1) *He arrived at work* and

(2) *He arrived at noon*

have the following dependency trees:

Figure 1. Dependency trees for the sentences beginning with *He arrived at ...*

Here for syntactic links entering the words, the abbreviated names of the assigned syntactic relations are shown; for full names and descriptions of English syntactic relations see (Apresjan et al. 1989). Being presented with the initial segment *He arrived at ...*, the subject cannot confidently decide which type of link connects *arrived* and *at*. The sentences

(3) *He saw Mary and Kate* and

(4) *He saw Mary and smiled*

have the following dependency trees:

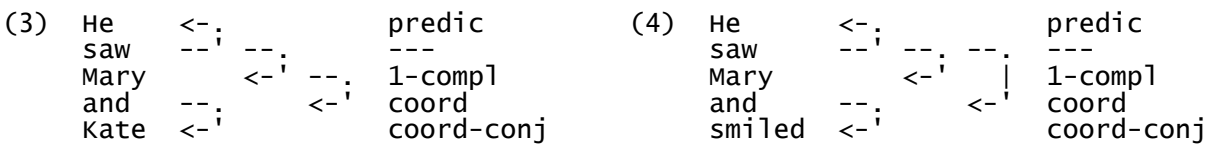

Figure 2. Dependency trees for the sentences beginning with *He saw Mary and ...*

Being presented with the initial segment *He saw Mary and ...*, the subject cannot decide which word is the head of the coordinating link: *saw* or *Mary*.

To avoid these difficulties, it was decided to invert the direction of the left-to-right links "preposition → X" and "conjunction → X" so that the links are directed from the word X to the function word; the names of the links remain unchanged. The links that entered a preposition or conjunction will now enter the word X, which in the new situation dominates the preposition/conjunction; again the names of the links remain unchanged. These modifications are purely technical and allow automatic transformation from the old form to the new and vice versa. It is worth noting that the new form of these constructions agrees with the principles of the Universal Dependencies approach; see the discussion in (Osborne and Gerdes, 2019).

The transformation described is not used for prepositions homonymous with adverbs, such as *naprotiv* ('opposite'), *poperek* ('across'), *posle* ('after'), *szadi* ('behind'), *vnutri* ('inside'), *vozle* ('near')

etc. Instead, such words are always considered as adverbs, and the dependent of the preposition formally becomes the dependent of the adverb (with the 1st completive syntactic relation instead of prepositive).

4 Tentative links

As shown by garden-path sentences (such as *The horse raced past the barn fell)*, which were first discussed by H.W. Fowler (1926) who actually introduced the incremental approach to syntax, a 100 percent confident incremental parsing of a sentence is impossible. There inevitably arise situations where it is necessary to revise decisions made earlier. We distinguish two types of such situations: those where the necessity of revision is surprising to the subject, and those where the possibility of revision was planned in advance. This "conscious uncertainty" is realized in the experiments in the form of tentative links.

Consider the sentence

(5) *I met her sister yesterday*,

and suppose the subject is given the first three words: *I met her ...* . The subject understands that in this segment the syntactic link *met → her* (1-compl) is possible and quite probable, and at the same time understands that the dependency tree of the complete sentence need not contain this link (as is indeed the case in this example, where the correct links are *met → sister* (1-compl) and *sister → her* (determ)). In this situation the subject inserts the link into the syntactic structure but marks it as "tentative" (the other links are called "final"). It is also allowed to create tentative links and keep them in reserve, without immediate insertion into the structure, – for example, when there is an alternative which seems more probable. While processing a sentence, the subject has the right to freely insert existing tentative links into the structure or remove them from the structure, on the condition that at any moment the syntactic structure should remain a well-formed directed tree or a union of disjoint well-formed trees.

Normally, the process of incremental construction of the syntactic structure consists in augmenting the structure by new syntactic links (final or tentative) that connect the active word and the words of the left context. It is also allowed to make "corrections", that is to insert into the structure or remove from it final links whose both ends belong to the left context. We always presume that processing a given sentence results in producing its correct complete dependency tree. The subject's performance on a sentence is measured with two indicators: the number of corrections and the number of created tentative links. In an ideal situation, both these numbers are equal to zero; in reality the subjects are instructed to avoid making corrections as much as possible and to keep the number of tentative links to a minimum. Accordingly, tentative links should only be created when the use of final links is associated with a significant risk of error.

In principle, the experiments might be conducted in a more straightforward way without an additional type of link. At each moment the subject would create a syntactic structure which is plausible enough for the known part of the sentence, for example, would include the link *met → her* (1-compl) in the structure for the segment *I met her...* If at a later stage certain links turn out to be incorrect, they are simply removed from the structure; similarly, missing links are added to the structure. In this case we have only one indicator of performance: the number of corrections. However, with this metric we cannot distinguish between changing the structure in situations of genuine ambiguity and correcting ordinary human errors such as those caused by carelessness.

5 Setup of the experiment

The experiment is conducted as a dialogue supported by a special program. The program takes as input a sentence in the form of a string of characters and splits it into words. The dialogue consists of N–1 steps numbered 2, 3, ... , N, where N is the number of words in the sentence. At step K the subject is presented with a text file which shows the first K words of the sentence (with the adjacent punctuation) plus the right context, that is, a fixed number of words following the word K. The syntactic links are also shown that were created at previous steps between the words of the left context (1, ... , K–1). When the last word of the sentence is shown, it is accompanied by the message [end of sentence]; until this message appears, the subject has no information about the length of the sentence.

The task of the subject is to create, if needed, new syntactic links between the active word K and the words of the left context. To create a link, the subject writes the name of syntactic relation and, if necessary, the number of the head (and in some cases dependent) in the appropriate field of the file.

Consider, for example, the English sentence

(6) *London Orbital is a 117 mile long motorway, encircling almost all of Greater London,*

and let the size of the right context be set to 1. If everything goes correctly, at step 8 the subject will be presented with the text file shown in Figure 3.

```
 London Orbital is a 117 mile long motorway,
 encircling ......

  1   London     <-.      compos
  2   Orbital    --' <-.  predic
  3   is             --'  ---
  4   a                   ---
  5   117        <-.      quantit
  6   mile       --' <-.  restr
  7   long           --'  ---
  8   motorway,
      encircling

------------------------------------------------------------
|  * -->  8        |
|  8 -->  3  is    |
|  8 -->  4  a     |
|  8 -->  7  long  |
------------------------------------------------------------

   TENTATIVE LINKS
------------------------------------------------------------
|  create and insert into the tree  |    -->
|  create                           |    -->
|  insert into the tree             |    -->
|  remove from the tree             |    -->
------------------------------------------------------------

   CORRECTION OF FINAL LINKS
------------------------------------------------------------
|  insert into the tree  |    -->
|  remove from the tree  |    -->
------------------------------------------------------------
```

Figure 3. The dialogue file at step 8.

The subject should create links between the active word 8 *motorway* and the previous words. In this case tentative links are not needed, and the subject only deals with final links, writing information about them in the first of the three frames (Figure 4).

```
------------------------------------------------------------
|  * -->  8        |  3 copulat
|  8 -->  3  is    |
|  8 -->  4  a     |  determ
|  8 -->  7  long  |  modif
------------------------------------------------------------
```

Figure 4. Creating new links between the active word 8 and the left context.

At step 9 these links are inserted into the structure, and word 9 becomes active (Figure 5). At this point the subject creates the link $8 \rightarrow 9$ (modif), and so on. The program keeps a complete record of the subject's actions at all steps of sentence processing.

```
London Orbital is a 117 mile long motorway,
encircling almost ......

 1   London     <-.              compos
 2   Orbital    --' <-.          predic
 3   is             --'     --.  ---
 4   a                  <-.   |  determ
 5   117        <-.       |   |  quantit
 6   mile       --' <-.   |   |  restr
 7   long       <-. --'   |   |  modif
 8   motorway,  --'     --' <-'  copulat
 9   encircling
     almost

------------------------------------------------------------
|  * -->  9      |
|  9 -->  3  is  |
------------------------------------------------------------

   TENTATIVE LINKS
   ......

   CORRECTION OF FINAL LINKS
   ......
```

Figure 5. The dialogue file at step 9.

6 Experimental dataset

The sentences for the experiments were taken from the two sets of sentences *dev.csv* and *train.csv* offered as training material for the competition "Automatic Gapping Resolution for Russian" held in association with the conference Dialogue 2019 (Dialogue Evaluation / AGRR-2019). These sets contain over 20,000 sentences of various genres, about one third of which are marked as elliptical. For our experiments, non-elliptical sentences were selected that satisfied the following additional requirements:

(1) the number of words does not exceed 30;
(2) the first alphanumeric character is a Russian capital letter;
(3) the last character is a small Russian letter or full stop;
(4) the proportion of small Russian letters among all alphanumeric characters is at least 90%.

The aim of these requirements was to restrict experimental material to "ordinary narrative Russian sentences of average length". As a result, a set of about 7,700 sentences was formed; the sentences for the experiments were taken from it without replacement using pseudorandom numbers. The distribution of sentence length in this set is rather "flat" on the segment from 7 to 30, with a mean of 17.4 and a standard deviation of 6.4. Hence, in a random sample of 100 sentences the average length has the same mean and a standard deviation of 0.64.

Size of the right context	Total number of links in the trees	Number of tentative links in the trees	Total number of created tentative links	Number of corrections
0	1627	34 (2.23%)	75	3
1	1741	21 (1.21%)	34	0
2	1607	8 (0.50%)	13	0

Table 1. The results of the experiments.

7 Results and future work

Three series of experiments were conducted for the sizes of the right context 0, 1 and 2, with 100 sentences processed in each series. The role of the subjects was played by the authors of this paper. They have considerable practical experience of developing the SynTagRus treebank, each having tagged not less than ten thousand sentences. The results of the experiments are given in Table 1.The figures in the table show that the right context of 1–2 words is sufficient for error-free and confident incremental parsing of Russian narrative sentences.

In the future, we plan to conduct experiments on incremental parsing of Russian elliptical sentences. Processing of a sentence is supposed to be similar to the procedure described in Section 5, but in addition to creating syntactic links, the subject will be able to create new nodes of the syntactic structure representing omitted lexical items.

Another possible area of future work is incremental parsing of English sentences. Generally, the results for English are expected to be more modest than for Russian, partly because the English inflectional system is not as rich as the Russian. However, preliminary experiments did not show a great difference in performance.

8 Conclusion

We believe that the experiments described in this paper characterize certain general features of human text comprehension. It could be argued that in fact we studied a much narrower phenomenon: text comprehension in people who are experts in linguistics. In our opinion, however, text comprehension is a highly automatic subconscious process which, in the case of native speakers, is not influenced by special linguistic training. But linguists, in contrast to ordinary speakers, have the tools which enable them to externalize their understanding of the text – for example, they can assign morphological features to wordforms or identify syntactic dependencies between words – and this is exactly what is required of the subjects in our experiments.

The results of the experiments, namely almost complete absence of errors (i.e. corrections of final links) and a small number of tentative links created, may be regarded as arguments in favour of the following general model of text comprehension. Suppose that while processing a sentence, only final links have been used. This means that the syntactic structure of the sentence was built in a strictly incremental way: links were added to the structure but never removed from it. In this case we can imagine the comprehension process to develop like this: for each new word, the reader/listener adds to the structure the links containing this word that satisfy the syntactic and semantic requirements, and later never returns to them. It may be assumed that this strategy of immediately adding plausible links to the structure is used universally, while relatively infrequent collisions (incompatibility of new potential links with those already in the structure) are successfully resolved on the basis of information available at the moment of collision. For this strategy to be efficient, natural language texts should be specially adapted to it. We assume that this adaptation is provided by their authors, who are interested in successful communication.

Acknowledgements

The authors express their gratitude to the Russian Foundation for Basic Research for their partial support of this work (grant No 19-07-00842).

References

Juri Apresjan, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Nikolai Pertsov, Vladimir Sannikov, and Leonid Tsinman. 1989. *Lingvisticheskoe Obespechenie Sistemy ETAP-2* [The linguistics of the ETAP-2 system]. Nauka, Moscow. (in Russian)

Dialogue Evaluation / AGRR-2019. https://github.com/dialogue-evaluation/AGRR-2019

Pavel Dyachenko, Leonid Iomdin, Alexander Lazursky, Leonid Mityushin, Olga Podlesskaya, Viktor Sizov, Tatiana Frolova, and Leonid Tsinman. 2015. Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (SynTagRus) [A deeply annotated corpus of Russian texts (SynTagRus):

contemporary state of affairs]. *Natsional'nyi korpus russkogo yazyka: 10 let proektu. Trudy Instituta russkogo yazyka im. V.V. Vinogradova. Vyp. 6* [The Russian National Corpus: 10 Years of the Project. Proc. of the V.V. Vinogradov Russian Language Institute. Issue 6]. Moscow. 272–299. (in Russian)

Eugen Fischer, Paul E. Engelhardt, Joachim Horvath, and Hiroshi Ohtani. 2019. Experimental ordinary language philosophy: a cross-linguistic study of defeasible default inferences. Preprint of paper forthcoming in Synthese. https://philarchive.org/archive/PHISEOL2.

Henry Watson Fowler. 1926. *A Dictionary of Modern English Usage.* Oxford University Press.

Evgeniya Inshakova, Leonid Iomdin, Leonid Mityushin, Viktor Sizov, Tatiana Frolova, and Leonid Tsinman. 2019. SynTagRus segodnya [SynTagRus today]. *Trudy Instituta russkogo yazyka im. V.V. Vinogradova* [Proc. of the V.V. Vinogradov Russian Language Institute]. Moscow. (in Russian) (to appear)

Leonid Iomdin, Vadim Petrochenkov, Viktor Sizov, and Leonid Tsinman. 2012. ETAP parser: state of the art. *Computational Linguistics and Intellectual Technologies. International Conference (Dialog'2012).* RGGU Publishers, 2012. Issue 11(18). Moscow. 830–843.

Igor Mel'čuk. 1974. *Opyt Teorii Lingvisticheskikh Modelei "Smysl ⇔ Tekst"* [Towards a theory of Meaning – Text linguistic models]. Nauka, Moscow. (in Russian)

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics,* 4(1):17, 1–28.