Experiments on human incremental parsing of English

Leonid Mityushin, Leonid Iomdin
A.A. Kharkevich Institute for Information Transmission Problems
Russian Academy of Sciences
mit@iitp.ru, iomdin@iitp.ru

Abstract
Experiments have been carried out in which human subjects incrementally constructed dependency trees of English sentences. The subjects were 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. They were also shown a fixed number of lookahead words following the last word of the segment. The results of the experiments show that lookahead of 1 or 2 words is sufficient for confident incremental parsing of English declarative sentences.

Keywords: incremental parsing; human parsing; dependency tree; English language
DOI: 10.28995/2075-7182-2021-20-505-513

1 Introduction
In this work experiments are described on incremental construction of dependency trees by humans. The subjects1 in the experiments were linguists well experienced in syntactic annotation. In each experiment, the subject was successively shown growing initial segments of an English sentence and had to draw syntactic links between the last word of the segment (the active word) and the previous words (the left context). In addition to the initial segment, at each step the subject could see a fixed number of words following the active word (lookahead).

This work is a repetition for English of the experiments on human incremental parsing of Russian carried out by the authors previously [Mityushin, Iomdin, 2019], and our motivation was to find out

---

1 Here and below the word subject is used in the meaning 'a person taking part in an experiment' (the equivalent of the Russian term испытуемый).
whether the results for English and Russian would be similar. We could not expect it a priori as these languages are not closely related; actually, their morphology and syntax are typologically very different.

In situations of uncertainty the subjects were supposed to use so-called tentative syntactic links instead of ordinary ones. An additional option in the English experiments was a temporary increase of the lookahead size. This proved to be very useful, especially in the case of zero default lookahead. Three series of experiments were carried out for default lookahead sizes 0, 1 and 2, with 100 sentences processed in each series. The results show that lookahead of 1 and especially 2 words is enough for quite confident incremental parsing (as is the case with Russian).

This leads to the conclusion that natural language texts have certain implicit properties that make incremental parsing effective. It is reasonable to assume that such properties may be rooted in specific features of human text generation, most probably associated with gradual, incremental deployment of information and meaning (the course of which the linguist reconstructing the parse tree tries to guess).

Although produced by the authors independently, this assumption seems to be much in line with the theory of dynamic syntax, which appeared in 1990s [see Blackburn, Meyer-Viol, 1994] and has been actively developed since, involving differently structured languages [see e.g. Kempton et al., 2001; Tugwell, 2006; Kempton et al., 2011; Kempton, Gregoromichelaki, 2017]. The ideas of dynamic syntax lie at the intersection of linguistics, cognition and brain science, and heavily rely on incrementality.

2 Syntactic model

As in the experiments with Russian, we use the representation of syntactic structures of sentences in the form of dependency trees introduced by I. Mel’čuk [1974, 1988] within the framework of the Meaning ⇔ Text theory. This representation is used in the ETAP multifunctional multilingual linguistic processor [Iomdin et al., 2012]. In this format, the nodes of a dependency tree are the words of the sentence, and the links are labelled with names of syntactic relations. In the current version of the ETAP English syntax, 64 relations are used [Apresjan et al., 1989]; a similar set of relations is described by Mel’čuk and Pertsov [1987].

To facilitate incremental construction of dependency trees, certain formal changes were made to subtrees for phrases containing prepositions or conjunctions. In the ETAP syntax, prepositions/conjunctions dominate the noun/verb groups that follow them. For example, the sentences She lives in Paris and She lives in luxury have the following dependency trees:

She <-. predicative                  She <-. predicative
lives --' --. --                     lives --' --. --
in --. <-' 1st completive           in --. <-' 1st completive
Paris <-' prepositional              luxury <-' prepositional

Here for each syntactic link entering a word, the corresponding name of syntactic relation is shown. Being presented with the initial segment She lives in ..., the subject cannot confidently decide which type of link connects the words lives and in. The sentences He drank tea and coffee and He drank tea and left have the following dependency trees:

He <-. predicative                   He <-. predicative
drank --' --.                        drank --' --. --
tea <-' --. 1st completive          tea <-' | 1st completive
and <-. <-' coordinative             and <-. <-' coordinative
coffee <-' coordinate-              left <-' coordinate-
    conjunctional                    conjunctional

Being presented with the initial segment He drank tea and ..., the subject cannot decide which word is the head of the coordinative link: tea or drank.

In order to avoid these difficulties, we used a different representation of prepositional/conjunctional constructions. The construction \( X \xrightarrow{a} P \xrightarrow{b} Y, \) where \( P \) is a preposition or a conjunction, \( a \) is an arbitrary syntactic relation, and \( b \) is one of the four relations: prepositional, coordinate-conjunctinal, subordinate-conjunctinal, or comparative-conjunctinal, is replaced with the
construction $X \overset{b}{\leftarrow} P \overset{a}{\rightarrow} Y$. This transformation is purely technical and can be performed automatically in both directions. The new representation allows us to postpone making decisions about preposition/conjunction attachment until the relevant content words appear. It is worth noting that the new form of these constructions is accepted in multiple syntactic models today. In particular, it conforms to the principles of the Universal Dependencies framework [see Osborne, Gerdes, 2019].

3 Experimental setup

Processing of a sentence is organized as a dialogue supported by a specially created computer program. The input to the program is a sentence in the form of a string of characters; the program splits it into words using blanks as separators. Isolated punctuation marks, such as dashes, are coupled with adjacent words, usually those on the left. 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 dialogue text file which shows the first $K$ words of the sentence with the adjacent punctuation plus a fixed number of words following the word $K$ (lookahead). 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.

At step $K$, the dialogue file also shows the partial syntactic structure (PSS) created at the previous steps on the words of the left context [1, ..., $K-1$] of word $K$. PSS is the main data structure supported by the program. For a given sentence, PSS may be any set of syntactic links between the words of the sentence with the additional condition that these links form either a single well-formed directed tree or a union of disjoint well-formed trees. At the beginning of the experiment, PSS is empty. The task of the subject is to add to PSS new syntactic links, so as to obtain a complete dependency tree of the sentence after step $N$.

At each step, the subject can create new syntactic links between the active word $K$ and the words of its left context. There are two types of link: ordinary and tentative. Ordinary links are "permanent", they are added to PSS at the moment of creation and are supposed to remain there. New tentative links are added to the so called "tentative pool" (another data set supported by the program, also empty at the beginning). The links of the tentative pool can be added to or removed from PSS at any moment; in particular, they can be added to PSS at the moment of creation.

If necessary, it is also allowed to add to PSS or remove from it ordinary links whose both ends belong to the left context [1, ..., $K-1$]. However, these actions are considered as "error correction", and the subject is instructed to prevent them as much as possible.

There is also a different sort of action: the subject can ask the program to show one additional word of lookahead. As a result, the subject gets the same dialogue file with one word added; the active word remains the same. If the command to show an additional word is given $M$ times in a row, $M$ additional words will be shown. At subsequent steps, the lookahead size returns to its default value.

We always presume that processing a given sentence results in producing its correct complete syntactic structure (the correspondence between sentences and their syntactic structures is discussed in Section 7). The subject's performance on a sentence is measured by three indicators: the number of corrected errors, the number of created tentative links and the number of commands to increase lookahead. In an ideal situation, all these numbers are equal to zero; in real practice the subjects are instructed by the experimenters to avoid making corrections as much as possible and to keep the number of tentative links and lookahead expansions to a minimum. Accordingly, unless there is a significant risk of error, it is preferable to use ordinary links and refrain from increasing lookahead.

4 Example

In this section we illustrate the use of ordinary links, the main building material for dependency trees. The input sentence is "That was how the words occurred in the old Latin poem.", and the lookahead size is 2 words. If everything goes correctly, at step 6 the subject is presented with the following dialogue file:
That was how the words occurred in the ......

1 That <-; predicative
2 was --' --
3 how --
4 the <-; determinative
5 words --' --
6 occurred
   in
   the

---------------------------------------
|  * --> 6         |
|  6 --> 2 was    |
|  6 --> 3 how    |
|  6 --> 5 words  |
---------------------------------------

TENTATIVE LINKS
---------------------------------------
create and add to PSS
create
add to PSS
remove from PSS
increase lookahead
---------------------------------------
ERROR CORRECTION
---------------------------------------
add to PSS
remove from PSS
---------------------------------------

At this point, syntactic links should be created between the active word 6 (occurred) and the words of the left context. The subject writes in the first field the abbreviated names of syntactic relations for the new links (and for the link going left to right to the active word, also the number of its head word):

|  * --> 6         |  2  copulat |
|  6 --> 2 was    |
|  6 --> 3 how    | adverb     |
|  6 --> 5 words  | predic     |

At the next step, these links are added to PSS, and the subject gets this dialogue file:

That was how the words occurred in the old ......

1 That <-; predicative
2 was --' --
3 how <-; adverbial
4 the <-; determinative
5 words --' --
6 occurred <-; <- predicative
7 in
   the
   old

---------------------------------------
|  * --> 7       |
|  7 --> 2 was   |
---------------------------------------

TENTATIVE LINKS
......
ERROR CORRECTION
......

Then the subject moves on without creating new links until the last word 11 (poem) becomes active:
That was how the words occurred in the old Latin poem.

1. That  <--.  predicative
2. was  --'  --.  --
3. how  <--.  adverbial
4. the  <--.  determinative
5. words  <--.  predicative
6. occurred  <--.  copulative
7. in  --
8. the  --
9. old  --
10. Latin  --
11. poem.  --'  --'  --'  --'  --'  2nd completive

[complete structure]

That was how the words occurred in the old Latin poem.

5 Increasing lookahead

Technically, the command to increase lookahead is given by typing any single character in the blank field on the right of the words "increase lookahead". It is incompatible with other commands, so no other changes should be made to the dialogue file. In the experiments with a zero default lookahead this command was used quite frequently – roughly speaking, “for every singular noun”. The reason is that we often cannot correctly parse noun phrases until we get to the end of them. For example, consider the sentence

*The temperature control device adjustment technique is described in the Appendix.*
Here we have a chain of nouns connected with the compositive relation: temperature ← control ← device ← adjustment ← technique, and the is dominated by the rightmost noun technique (with the determinative relation). We cannot correctly attach the until we identify the last word of the chain. So it would be quite natural, when we are at step 2 (that is, when we are shown the segment The temperature ...), to repeatedly increase lookahead until the noun phrase comes to an end, and only then establish the syntactic links within the phrase moving from step 3 to step 6.

6 Tentative links

Another way to deal with uncertainty is to use tentative links. For example, with the sentence in Section 5 we might move from active word 2 (temperature) to 6 (technique), and at each step create a tentative determinative link from the current active word to the article the. As these links are mutually incompatible, we would keep them in the tentative pool rather than add to PSS. Simultaneously, at steps 3 to 6, we would create ordinary compositive links between the active word and the previous one. Then, when we come to the active word 7 (is), we would add the tentative link the ← technique to the PSS, and the processing of the sentence would continue.

It is clear, however, that in this case increasing lookahead is much more convenient. Nevertheless, there are situations of "long distance non-confidence" where it is quite natural to use tentative links. Consider the sentence

Weather forecast: clouds with sunny spells and occasional showers.

Its syntactic structure contains an explicative link between the heads of the two parts divided by the colon: forecast → clouds. In contrast, this sentence with two words added

Weather forecast: clouds with sunny spells and occasional showers are expected.

has an explicative link forecast → are. When presented with the initial segment Weather forecast: clouds with clouds as the active word (a few words of lookahead can also be shown), the subject must decide what to do with the explicative link forecast → clouds. Obviously, it would be too risky to create it as an ordinary link. The subject can repeatedly increase lookahead, but the number of increases needed may be large (of the same order as the length of the sentence). In this case the advisable course of action is to create a tentative explicative link forecast → clouds and add it to PSS. If it later proves to be incorrect it will be replaced with the correct one "free of charge".

On the whole, to avoid errors (in particular with garden-path sentences such as The horse raced past the barn fell), the subject should use tentative links and lookahead increases whenever straightforward use of ordinary links carries a real risk of error, however small.

7 What is the correct structure?

The task of the subjects in our experiments was to create syntactic structures for given sentences. The input sentences had no structures assigned to them in advance, so there was no gold standard for comparison. As the subjects were linguists with considerable experience of syntactic annotation, it was agreed that they themselves should decide what dependency tree is the correct structure for the given sentence. If any ordinary links in the current tree are incorrect or missing, the tree is repaired using the "error correction" field in the dialogue file. This can be done at any step up to N, or when the complete structure is presented after step N.

Sometimes the input sentence cannot be assigned a dependency tree – for example, this is true of many elliptical sentences. Such sentences are excluded from the experiment as soon as the subject discovers that they are unparsable. On the other hand, a sentence can have more than one correct syntactic structure. There are two main types of such situations: semantic indeterminacy and genuine semantic ambiguity.

Semantic indeterminacy [Ziering, Van der Plas, 2015] can be illustrated by the sentence They mined the roads along the coast, where the phrase along the coast may be attached either to the verb or to its object without essentially changing the meaning. In such cases the alternative syntactic structures are considered equally correct, and the subject is free to choose any of them as the target one.
Genuine semantic ambiguity is represented by N. Chomsky's sentence *Flying planes can be dangerous*, which has two semantically different dependency trees:

```
  Flying    <-.      modificative           Flying    --. <-.  predicative
   planes    --' <-.  predicative            planes    <-'   |  1st completive
   can       --. --'  --                     can       --. --'  --
   be        <-' --.  1st completive         be        <-' --.  1st completive
     dangerous     <-'  copulative             dangerous     <-'  copulative
```

The subject has no reason to prefer one interpretation to the other (partly because sentences in the experiments are taken out of context). Moreover, at step 2 (and also at step 3 if the lookahead is 0 or 1) the subject doesn't know how long the sentence is going to be. The sentence may continue in a way that makes either of these readings the only correct one. Hence the right thing in this situation is to keep both options open by creating four tentative links: *planes → flying, can → planes, flying → planes* and *can → flying*. It should be said, however, that genuine ambiguity of this kind never really occurred in our experiments.

8 Experimental material

The sentences for the experiments were taken from the written subcorpus of the British National Corpus (BNC) [Burnard, 2007]. BNC covers British English of the late 20th century from a wide variety of genres. Its written subcorpus contains approximately 5 million sentences. For our experiments we selected sentences that satisfied the following additional requirements similar to those used in [Mityushin, Iomdin, 2019]:

1. the number of words in the sentence is between 6 and 30;
2. the first alphanumeric character is a capital letter;
3. the last character is a small letter or full stop;
4. the proportion of small letters among all alphanumeric characters is at least 90%.

The aim of these conditions was to restrict the experimental material to "ordinary declarative English sentences of moderate length". The number of sentences in the written part of BNC satisfying condition (1) is about 3.3 million, and the number of those satisfying all four conditions is about 2.6 million. The sentence length in this subset has the mean 17.0 and the standard deviation 6.7. Sentences for the experiments were selected from this subset using pseudorandom numbers.

9 Results and discussion

Three series of experiments were carried out for the default lookahead size equal to 0, 1 and 2, with 100 sentences processed in each series. The subjects in the experiments were the authors of this paper, who have an advanced command of English. The results are presented in Table 1.

<table>
<thead>
<tr>
<th>Lookahead size</th>
<th>Total number of links in the trees</th>
<th>Number of lookahead increases</th>
<th>Number of created tentative links</th>
<th>Number of corrections</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1621</td>
<td>373</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>1555</td>
<td>56</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1686</td>
<td>17</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. The results of the experiments.

The results demonstrate good improvement in performance for each additional word of lookahead. Although the number of corrections in the experiments was 8, the actual number of errors was 4, but each error needed two separate "correcting actions": removing the incorrect link from PSS and adding...
the correct one. For the two-word lookahead, only 1.5 percent of links were accompanied with "signs of doubt" (increasing lookahead or creating a tentative link instead of ordinary one).

It is interesting to compare the present results with those for Russian [Mityushin, Iomdin, 2019]. In the Russian experiments, the average sentence length over 300 sentences was 17.6 words; in the English it was 17.2 words. The general setup of the experiments was the same, with one important difference: for Russian, the option of temporary increase of lookahead was not available. This makes strict quantitative comparison impossible. Another obstacle to strict comparison is a "greater weight" of Russian words. Analysis of parallel English–Russian corpora shows that English texts contain on average 20–30 percent more tokens than their Russian counterparts [Russian National Corpus, 2020]. The reason is the massive use of function words such as articles (absent in Russian), auxiliary verbs (infrequent in Russian) and adverbial particles (absent in Russian). This means that k Russian words as lookahead are on average more informative than k English words. Yet another important distinction is the richer morphological system of Russian and much wider use of grammatical agreement, which helps to identify syntactic links with greater confidence.

Nevertheless, the results can be compared on a qualitative level. For Russian, error correction was used 3 times (for zero lookahead), and the number of created tentative links was 75, 34 and 13 for 0, 1 and 2 words of lookahead respectively. We may say that the level of performance for English with 2 words of lookahead was somewhere between the levels for Russian with 1 and 2 words of lookahead.

10 Conclusion

The results of the experiments described in this paper, as well as those for Russian, may be regarded as arguments supporting the following general picture of text comprehension. Suppose that only ordinary links have been used to build the syntactic structure of the input sentence. This means that the structure was constructed in a strictly incremental way: links were added to it but never removed. In this case we can imagine the parsing process to develop like this: the reader/listener adds to the partial syntactic structure the links between the active word and the left context that satisfy the syntactic and semantic requirements, and later never returns to them. This strategy of immediately adding plausible links to the structure may be assumed to be used universally, even in the cases of doubt, with infrequent collisions (incompatibility of new links to be added and those already in the structure) being successfully resolved on the basis of information available at the point of collision. To make this strategy effective, natural language texts should have certain implicit properties, possibly connected with the specific character of human text generation.

Acknowledgements

This research was supported by a grant from the Ministry of Science and Higher Education of Russia No. 075-15-2020-793.

References

Experiments on human incremental parsing of English