On the Structural Complexity of Natural Language Sentences

Dekang Lin*
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Rm 767, 545 Technology Square
Cambridge, Massachusetts, USA, 02139
E-mail: lindek@ai.mit.edu

Abstract

The objective of this paper is to formalize the intuition about the complexity of syntactic structures. We propose a definition of structural complexity such that sentences ranked by our definition as more complex are generally more difficult for humans to process. We justify the definition by showing how it is able to account for several seemingly unrelated phenomena in natural languages.

1 Introduction

Intuitively, certain syntactic structures are more difficult for humans to process than others. For example, compare the following to sentences:

(1) a. The cat that the dog that the man bought chased died.
    b. The man bought the dog that chased the cat that died.

It is obvious that sentence (1a) is much more difficult to understand than (1b). Since the two sentences are of the same length and involve the same set of semantic relationships, the difficulty in understanding (1a) can only be attributed to its syntactic structure.

The objective of this paper is to formalize the intuition about the complexity of syntactic structures. We propose a definition of **structural complexity (SC)** such that sentences ranked by our definition as more complex are generally more difficult for humans to process than otherwise similar sentences. In other words, suppose a pair of sentences A and B consist of the same set of words and have essentially the same meaning, then sentence A is more difficult to process than sentence B if SC(A)>SC(B). For example, the proposed definition of structural complexity correctly predicts that (1a) is much more difficult to process than (1b).

The notion of structural complexity proposed in this paper offers explanations for a set of seemingly unrelated phenomena:

- We will show that the definition of structural complexity explains why a Dutch sentence involving cross-serial dependencies is slightly easier to understand than a corresponding center-embedded German sentence.
- We will also show that extrapositions, such as heavy-NP shift and PP extractions are motivated by reducing syntactic complexity. The extraposition of an element is only warranted when the structural complexity of the sentence is reduced as a result.
- NP modifiers of a head tend to be closer to the head than its PP modifiers, which in turn tend to be closer than its CP (clausal) modifiers. In Generalized Phrase Strcuture Grammar (GPSG) (Gazdar et al., 1985), these linear order constraints are stated explicitly in the grammar. The notion of structural complexity provides an explanatory account.

There are several reasons why the notion of structural complexity is useful. Firstly, in natural language generation, a generator should generate the simplest sentence that conveys the intended meanings. Structural complexity can be used to choose the syntactic structures with the lowest structural complexity so that the resulting sentence is easier to understand than other alternatives.

Secondly, structural complexity is also needed in assessing the readability of documents. It is well known that the length of a sentence is not a reliable indicator of its readability. Yet, the readability of texts has up to now been measured by the lengths of sentences and familiarities of the words in the documents. Using structural complexity instead of sentence length allows the readability of documents to be measured more accurately.

Finally, we propose, in Section 4, that extrapositions are motivated by reduction of structural

*On leave from the University of Manitoba, Winnipeg, Manitoba, Canada. This research has been supported by NSERC Research Grant OGP121338. The author is very grateful to the reviewers who pointed out several mistakes in the draft.

complexity. In other words, extrapositions are only allowed if the structural complexity of the sentence is reduced as a result. This constraint is useful both in parsing sentences with extrapositions and in deciding where to use extraposition during generation.

The notion of structural complexity is defined in Section 2. We then justify the definition of structural complexity by demonstrating in Sections 3, 4, and 5 that sentences with lower structural complexity are easier to understand than otherwise similar sentences with higher structural complexity.

2 Structural Complexity

The definition of structural complexity presumes the notion of dependency relationships between words in a sentence. In dependency grammars (Hudson, 1984; Mel'čuk, 1987), a dependency relationship is a primitive relationship between two words, called the **head** and the **modifier**. In constituency grammars that contain the X-bar theory as a component, dependency relationships between words are implicitly specified in X-bar structures. The modifiers of a word *w* are the head words of the specifier, complements, and adjuncts of *w*. For example, Figure 1 is the X-bar structure of (2). The word "will" has two modifiers: the head word of its NP specifier ("Kim") and the head word of its VP complement ("bring"). The dependency relationships in the X-bar structure in Figure 1 are shown in Figure 2. Each directed link in Figure 2 represents a dependency relationship with the direction going from the head to the modifier.

(2) Kim will bring the wine in the evening.

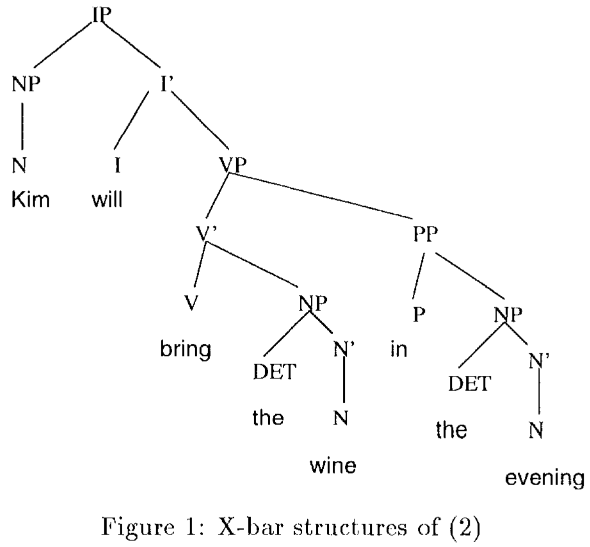


Figure 1: X-bar structures of (2)

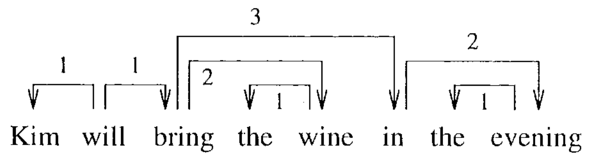


Figure 2: Dependency structure of (2)

In order to recognize the structure of a sentence, a parser must establish the dependency links between the words in the sentence. Structural complexity measures how easy or difficult it is to establish these dependency links. The definition of structural complexity is based on the assumption that the shorter dependency links are easier to establish than longer ones, where the length of a dependency link is one plus the number of words between the head and the modifier. For example, the lengths of the links in Figure 2 are shown by the numbers attached to the dependency links.

Definition 2.1 (Structural Complexity) *The structural complexity of a dependency structure is the total length of the dependency links in the structure.*

For example the structural complexity of the dependency structure in Figure 2 is 11.

In the next three sections, we will show that the definition of structural complexity does indeed reflect the difficulty in processing a sentence. We will present examples in which sentences with lower structural complexities are easier to process than similar sentences with higher structural complexities.

3 Center embedding

The difficulty in processing center embedding sentences, such as (1a), has been explained by its requirement on the size of the stack in a parsẽr. This explanation presumes that the human parser uses a push-down stack to store the partially built constituents. The notion of structural complexity provides an explanation of the difficulty of processing center embedding that makes much weaker commitment to the parsing model. Figure 3 shows the lengths of the dependency links in a center-embedding sentence (1a) and a non-center-embedding sentence (1b) with similar semantics. The structural complexity of the center-embedding sentence is 30, which is much higher than the structural complexity (=12) of the non-center-embedding sentence.

The presumption that human sentence processor uses a push-down stack is challenged by the contrast between cross-serial dependencies in Dutch (e.g., Figure 4a) and center-embedding sentences in German (e.g., Figure 4b.)

Since the cross serial dependencies are much more difficult to handle with push-down stacks

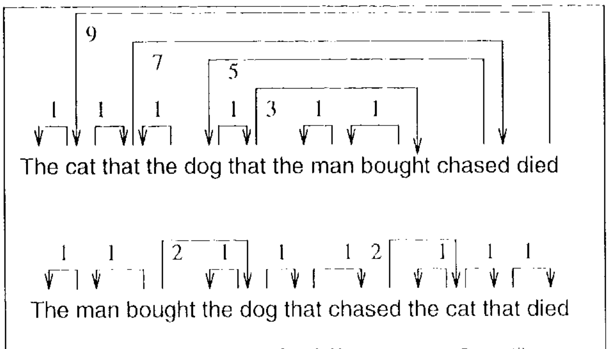
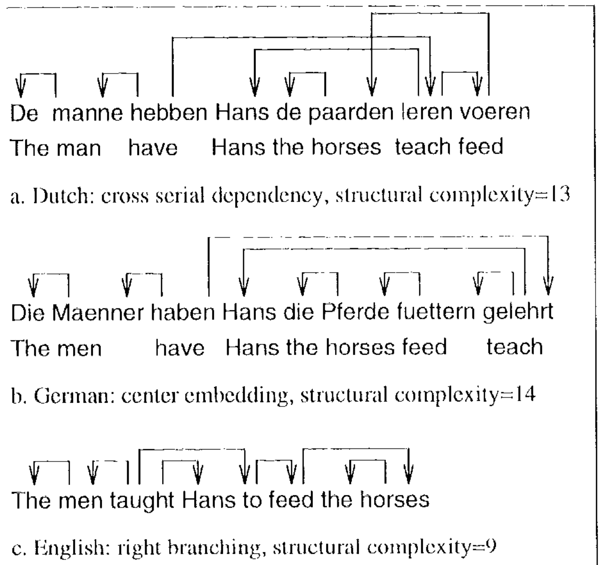

9
7
5
1 1 1 1 3 1 1
The cat that the dog that the man bought chased died
1 1 2 1 1 1 2 1 1 1
The man bought the dog that chased the cat that died

Figure 3: Center-Embedding vs. Non-Center-Embedding Sentences


De manne hebben Hans de paarden leren voeren
The man have Hans the horses teach feed
Die Maenner haben Hans die Pferde fuettern gelehrt
The men have Hans the horses feed teach
The men taught Hans to feed the horses

a. Dutch: cross serial dependency, structural complexity=13

b. German: center embedding, structural complexity=14

c. English: right branching, structural complexity=9

Figure 4: cross-serial dependency vs. center-embedding vs. right-branching

than nested dependencies, the hypothesis that human parser uses a push-down stack would predict that the Dutch sentences such as Figure 4a should be much more difficult to understand than the corresponding German sentences with nested dependencies (Figure 4b). However, data from psycho-linguistic experiments suggest that they are in fact slightly easier to process than the corresponding German sentences with nested dependencies (Bach et al., 1986). This observation can be accounted for by structural complexity, since the structural complexity of the Dutch sentence (Figure 4a) is 13, which is slightly lower than the structural complexity (=14) of the corresponding German sentence Figure 4b. It was also observed in (Bach et al., 1986) that "For someone with even a limited competence in English and either of the other languages, the patterns in Dutch and German seem to be more difficult to process and produce than their English counterparts" (p. 249). This is also consistent with the structural complexity account, since the structural complexity of Figure 4c is 9, which is significantly lower than its Dutch and German counterparts (Figure 4a and 4b).

4 Extrapositions

Extraposition refers to the movement of an element from its normal position to a position at or near the end of the sentence. Examples of extraposition in English include:

Heavy-NP shift

(3) a. Joe sent the book he found in Paris *to his pal*

b. Joe sent *to his pal* the book he found in Paris

Extraposed relative

(4) a. A man that no one knew *stood up*

b. A man *stood up* that no one knew

PP-extraposition

(5) a. I read a description of Hockney's latest picture *yesterday*

b. I read a description *yesterday* of Hockney's latest picture

Extraction from AP

(6) a. How certain that the Mets will win *are you*?

b. How certain *are you* that the Mets will win?

Mechanism for constraining extraposition is urgently needed in both parsing and generation. To the best of the author's knowledge, none of the broad-coverage parsers or generators handles extrapositions in a principled fashion. The reason for this is that extrapositions appear to be dependent upon certain aspects of contexts that are not captured by usual syntactic features. For example, compare the following pair of sentences

(7) a. I talked with a man *yesterday* with a mustache

b.*I talked with a man *one year and four months ago* with a mustache

The syntactic structures of (7a) and (7b) are the same, which is shown in Figure 5, except that the adverbial phrase AdvP is "yesterday" in (7a) and "one year and four months ago" in (7b). Although the two adverbial phrases are two different strings, they are identical in their syntactic features. Yet, extraposition is good in (7a) but bad in (7b).

We propose that the purpose of extraposition is to make a sentence easier to understand. Therefore, extraposition is only allowed when the structural complexity of the sentence is reduced as a result. Note that reduction of structural complexity is not the only constraint on extraposition. There

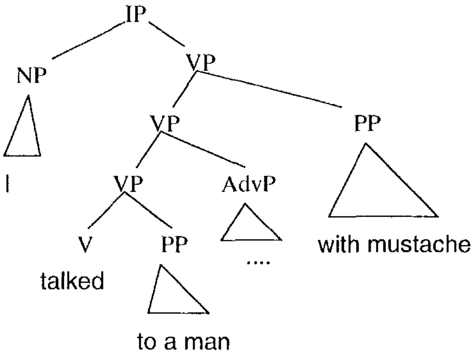


Figure 5: Parse tree of (7a) and (7b)

are also syntactic constrains such as Right Roof Condition (Ross, 1967) or Complement Principle (Rochemont and Culicover, 1990).

When a phrase is extraposed, the set of dependency relationships remains the same. However, the lengths of some of the dependency links will change. The structural complexity of the sentence may change as a result. Figure 6 illustrates how extrapositions affect the lengths of dependency links in (3), (4), (5), and (6). Only the dependency links whose lengths are changed are shown there. In all cases, structural complexity is reduced by the extraposition.

Consider the difference between (7a) and (7b). In (7a), the extraposition of [$_{\mathrm{PP}}$ with a mustache] increases the length of the dependency link between "man" and "with" by 1, but reduces the length of the dependency between "talked" and "yesterday" by 3. Therefore, the structural complexity is reduced by 2 as a result of the extraposition. In contrast, in (7b), the extraposition of [$_{\mathrm{PP}}$ with a mustache] increases the length of the dependency link between "man" and "with" by 6 and reduces the length of the dependency link between "talk" and "ago" by 3. Thus the structural complexity is increased when [$_{\mathrm{PP}}$ with a mustache] is extraposed.

The hypothesis that extraposition must reduce the structural complexity also explains why in heavy-NP shift, the extraposed NP must be heavy, i.e., consisting of many words. When the complement NP of a verb is 'shifted' to the right across an adjunct modifier of the verb, the length of the dependency link from the verb to the head of the NP is increased by length the adjunct modifier. On the other hand, the length of the dependency link from the verb to the adjunct modifier is reduced by the length of the NP. Therefore, the structural complexity of the sentence can only be reduced as a result of the extraposition when the NP is longer than the adjunct modifier,

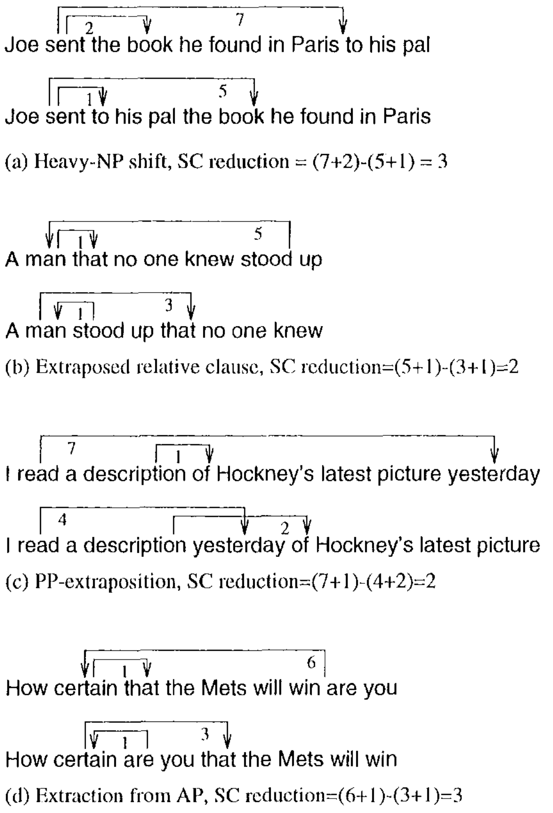

2
7
Joe sent the book he found in Paris to his pal
1
5
Joe sent to his pal the book he found in Paris
(a) Heavy-NP shift, SC reduction = (7+2)-(5+1) = 3
1
5
A man that no one knew stood up
1
3
A man stood up that no one knew
(b) Extraposed relative clause, SC reduction=(5+1)-(3+1)=2
7
1
I read a description of Hockney's latest picture yesterday
4
2
I read a description yesterday of Hockney's latest picture
(c) PP-extraposition, SC reduction=(7+1)-(4+2)=2
1
6
How certain that the Mets will win are you
1
3
How certain are you that the Mets will win
(d) Extraction from AP, SC reduction=(6+1)-(3+1)=3

Figure 6: Extraposition must reduce structural complexity

5 Linear Precedence

In most languages, the NP modifiers of a word tend to be closer to the word than its PP modifiers, which, in turn, tend to be closer to the word than its CP (clausal) modifiers. In GPSG (Gazdar et al., 1985), these linear order constraints are stated explicitly as the linear precedence rules. In this section, we show that the linear precedence rules in GPSG can be derived from the assumption that the linear order among different types of modifying phrases, such as NP, PP, and CP, should minimize the structural complexity so that the sentence is as easy to process as possible.

Suppose a word w has n modifiers XP_1, XP_2, ..., XP_n; the number of words in XP_i is l_i; and the head word of XP_i is w_i, which is the p_i'th word in XP_i. Without loss of generality, let us assume that w precedes its modifiers. If the order of the modifiers is XP_1, XP_2, ..., $\mathrm{XP}_{n'}$, then the length of the dependency link between w and the head of XP_i is $(p_i + \sum_{j=1}^{i-1} l_j)$ and the total length of dependency links within the maximal projection of w is:

$$\sum_{i=1}^{n}\left(p_i + \sum_{j=1}^{i-1} l_j\right)$$
$$= (n-1)l_1 + (n-2)l_2 + \ldots + l_{n-1} + \sum_{i=1}^{n} p_i$$

Among all permutations of XP_1, XP_2, ..., XP_n, the above sum is the minimal when $l_1 \leq l_2 \leq \ldots \leq l_n$. In other words, the total length of dependency links is minimal when the modifiers with fewer words are closer to the head. Generally speaking, PPs contain more words than NPs and CPs contain more words than PPs. Therefore, NP modifiers should be closer to the head word than PP modifiers and PP modifiers should be closer to the head word than CP modifiers if the structural complexity of the maximal projection of the head word w is to be minimized.

6 Discussion

We used the total length of the dependency links in the definition of structural complexity. The examples presented in the previous sections are also consistent with a definition that uses the maximum length of structural links. The reason we choose to use the sum is that the definition naturally incorporate the length into consideration.

The arguments presented in previous sections are preliminary. Our future work include backing up the hypothesis with empirical evidence and investigate the application of structural complexity in handling extraposition in parsing and generation.

7 Conclusion

We have proposed a notion of structural complexity. A sentence with higher structural complexity is more difficult to process than a similar sentences with lower structural complexity. Structural complexity is needed in both parsing and generation. It can also be used to assess the readability of documents. We support the definition of structural complexity with a set of seemingly unrelated phenomena: the contrast between center-embedding and right-branching sentences, extrapositions, and the linear order among modifying phrases. In all of these cases, sentences with lower structural complexity are easier to understand.

References

E. Bach, C. Brown, and W. Marslen-Wilson. 1986. Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, 1(4):249 262.

Gerald Gazdar, Ewan Klein, Geoffery Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Basil Blackwell Publisher Ltd, Oxford, UK.

Richard Hudson. 1984. *Word Grammar*. Basil Blackwell Publishers Limited., Oxford, England.

Igor A. Mel'čuk. 1987. *Dependency syntax: theory and practice*. State University of New York Press, Albany.

Michael S. Rochemont and Peter W. Culicover. 1990. *English Focus Constructions and the Theory of Grammar*. Cambridge Studies in Linguistics. Cambridge University Press.

J. Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, M.I.T., Cambridge, MA.