Sentence Complexity

Sentence complexity: Definition

Complexity generally refers to a sentence containing a subordinate clause or clauses. A complex sentence has one independent clause and at least one dependent clause. An independent clause (unlike a dependent clause) can stand alone as a sentence. Sentence complexity has been shown to vary with age (older people produce less complex sentences) while educational level does not appear to be related to complexity, only to vocabulary. Sentence complexity is linked to how easily sentences are understood (text readability) and how accurately they are recalled.

Sentence complexity: Measures

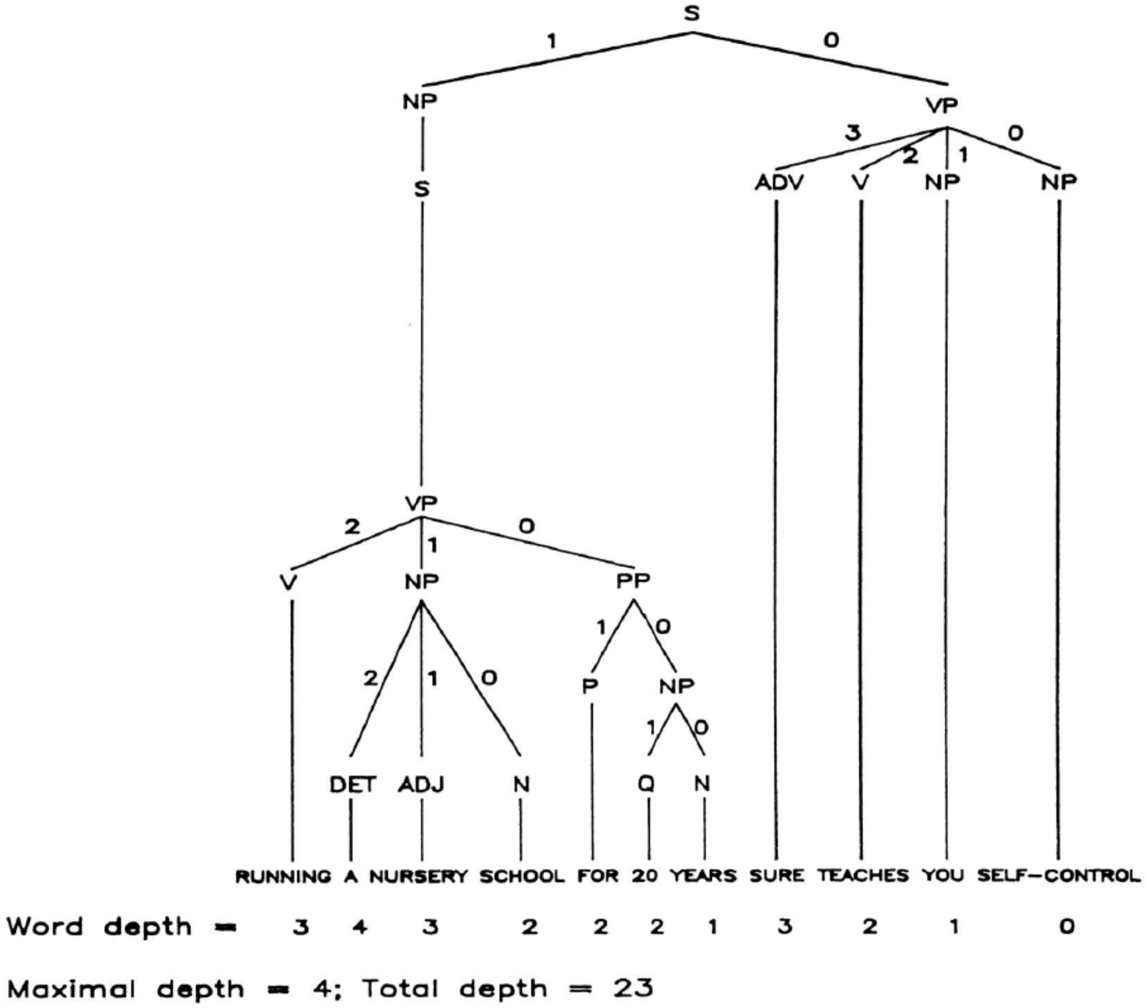
The NLP Suite implementation of sentence complexity provides five different measures of complexity: *Yngve* (score and sum) and *Frazier* (score and sum). The Yngve and Frazier measures are two of the most commonly used measures of sentence complexity.

The script also provides the *sentence length* in number of words as a measure. Needless to say more complex sentences result in longer sentences, characterized by more than one clause; but not necessarily longer sentences are more complex; think of a list of nouns followed by a verb and perhaps a direct object.

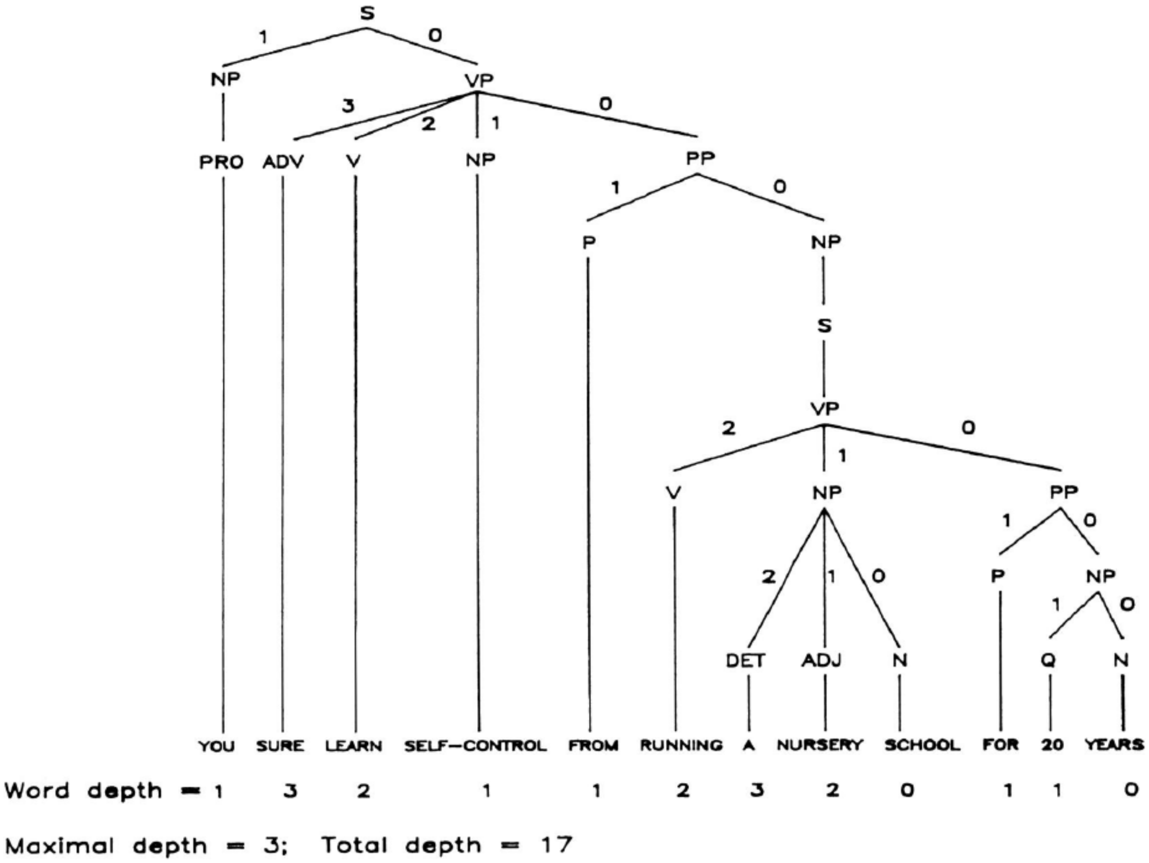
The computation of sentence complexity requires parsing the sentence to obtain the sentence dependency tree. The current implementation relies on the output of the Stanford PCFG Parser (http://nlp.stanford.edu/software/lex-parser.shtml).

The way the Yngve and Frazier measures of sentence complexity are computed is well illustrated in Cheung and Kemper (1992) and we reproduce here both their figures and explanations.

“Two Yngve depth measures were determined for each sentence: (a) Maximal Yngve depth is the largest number associated with any word in the sentence, and (b) Total Yngve depth is the sum of all depth counts for each word in the sentence. Maximal Yngve depth was, therefore, a “local” measure that was independent of sentence length, whereas Total Yngve depth was confounded with the number of words in the sentence.”

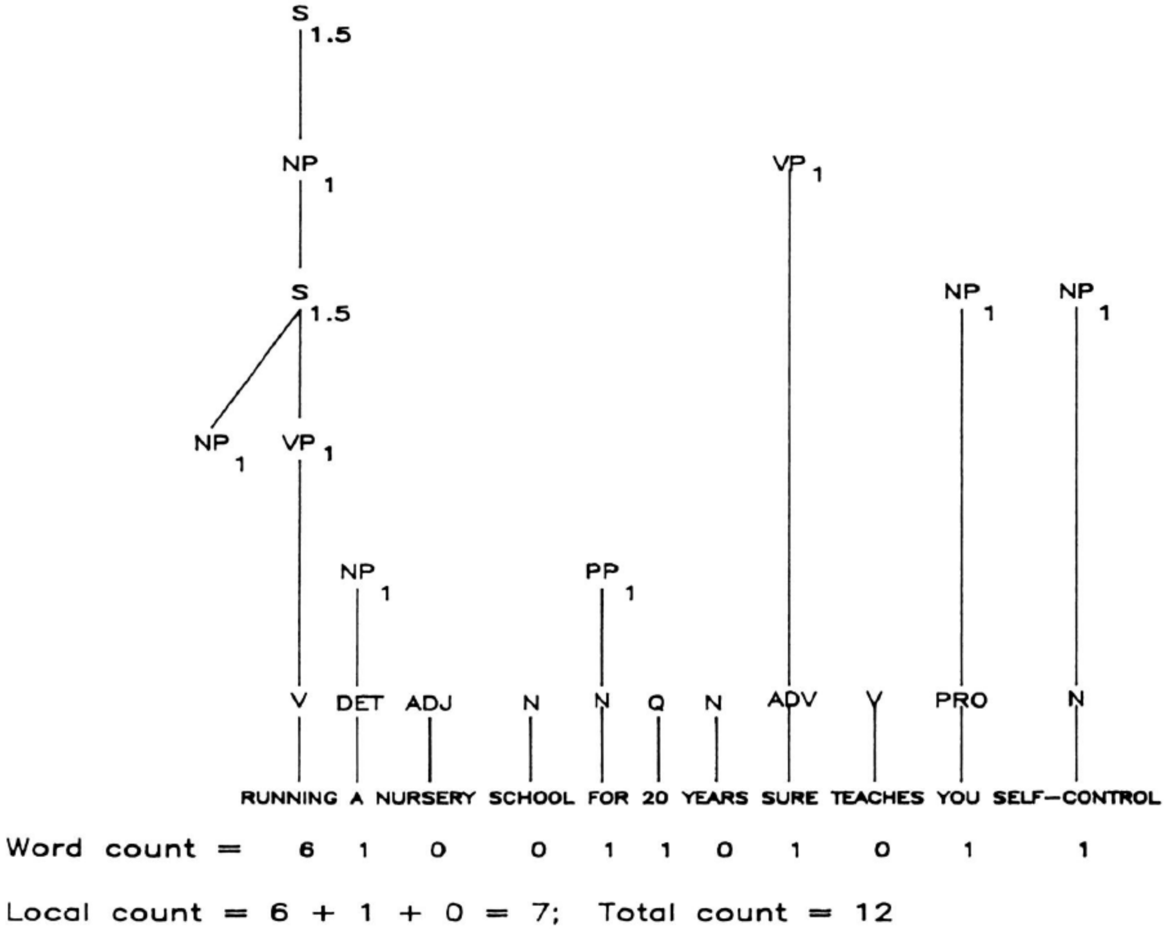

S
1
0
NP
VP
3
2
1
0
ADV
V
NP
NP
S
VP
2
1
0
V
NP
PP
1
0
2
1
0
P
NP
1
0
DET
ADJ
N
Q
N
RUNNING A NURSERY SCHOOL FOR 20 YEARS SURE TEACHES YOU SELF–CONTROL
Word depth = 3 4 3 2 2 2 1 3 2 1 0
Maximal depth = 4; Total depth = 23

Yngve measures illustrated in Cheung and Kemper (1992).

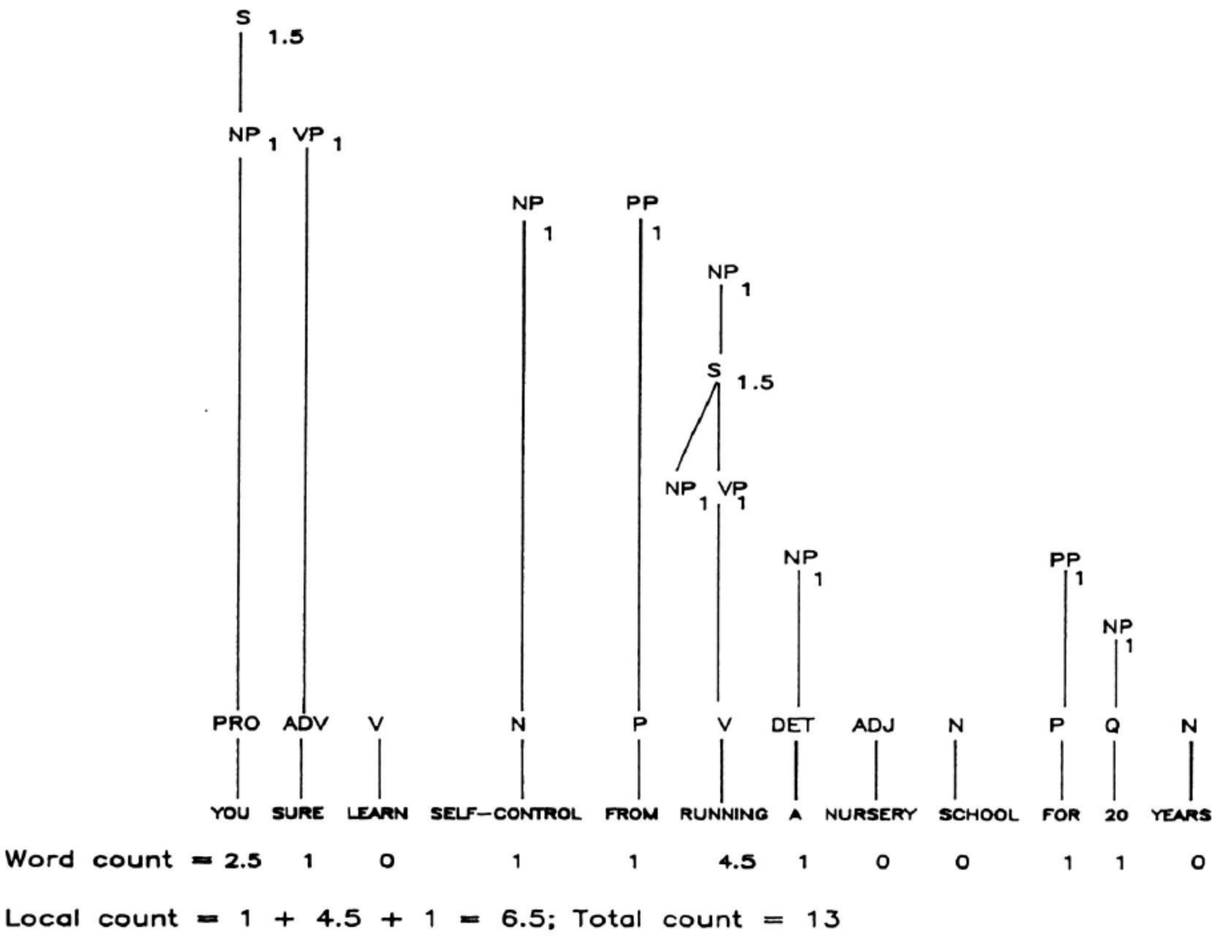


Yngve measures illustrated in Cheung and Kemper (1992).

“Total Frazier node count, were derived from the rules given by Frazier (1985). Figures 3 and 4 illustrate the calculation of the Local Frazier and Total Frazier counts. The Frazier counts were based on a surface phrase structure analysis in which all (nonterminal) nodes in the phrase structure of the sentence were assigned a point value of 1 except for sentence nodes and sentence-complement nodes, which were assigned a point value of 1.5. Counts for each word were then determined by summing up the points assigned to all the nodes dominating each word in the sentence.”


S 1.5
NP 1
VP 1
S 1.5
NP 1
NP 1
NP 1
VP 1
NP 1
PP 1
V
DET
ADJ
N
N
Q
N
ADV
V
PRO
N
RUNNING A NURSERY SCHOOL FOR 20 YEARS SURE TEACHES YOU SELF—CONTROL
Word count = 6 1 0 0 1 1 0 1 0 1 1
Local count = 6 + 1 + 0 = 7; Total count = 12

Frazier measures illustrated in Cheung and Kemper (1992).


S 1.5
NP 1 VP 1
NP 1
PP 1
NP 1
S 1.5
NP 1 VP 1
NP 1
PP 1
NP 1
PRO ADV V N P V DET ADJ N P Q N
YOU SURE LEARN SELF—CONTROL FROM RUNNING A NURSERY SCHOOL FOR 20 YEARS
Word count = 2.5 1 0 1 1 4.5 1 0 0 1 1 0
Local count = 1 + 4.5 + 1 = 6.5; Total count = 13

Frazier measures illustrated in Cheung and Kemper (1992).

The mean for example for the Frazier depth is calculated by summing the depths of each token in the parse tree and dividing the result by the number of tokens, where tokens are the words in each sentence and all the punctuation marks. For example, the total Frazier depth for the sentence "Miracles thicker than fog on flight of No 10607." is 11.5: the sentence has 10 tokens (8 words, one number (10607) and one period. So, 11.5/10 = 1.15. For the sentence "We had taken off below take-off speed, we had survived a crashing bounce back to earth, we had smashed through trees, we had nearly rolled over on our back, and then we had survived a crash landing through fire.", the total Frazier depth is 44; the sentence has 46 tokens (the word "take-off" is split into 3 tokens by the parser "take" "-" "off"). 44/46 = 0.9565, and so on.

Comparing sentence complexity across documents

To compare syntactic complexity scores across different sets of documents (e.g., different books), Pakhomov et al. recommend using one-way ANOVA with subsequent pairwise post-hoc t-tests using Tukey's Honestly Significant Differences (HSD) approach to adjust for multiple comparisons.

Sentence complexity: Visualization

The NLP Suite sentence complexity algorithm provides several Excel charts of complexity:

1. Line chart by Sentence Index of Yngve and Frazier scores;
2. Line chart by Sentence Index of Yngve and Frazier sums;

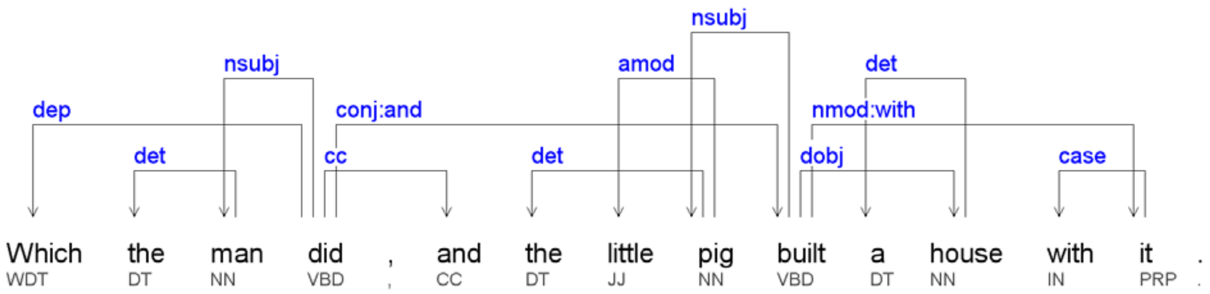
3. Line chart by Sentence Index of sentence length.

Other useful tools

We recommend several other useful tools for understanding how complex a text is.

Visualizing the parse tree

Do try visualizing the parse tree using the java script dependenSee.jar.



Clausal analysis

Sentence complexity fundamentally depends upon how a sentence is structure in different clauses. Run the Clausal analysis script to see how a text organizes its sentences.

Text readability

The text readability script provides several measures of readability, in terms of how many years of formal education it takes to understand a text, starting from grade 1 (of the American education system).

Unusual words

Unusual words (and not just because a word is misspelled) do not contribute to sentence complexity at the syntactical level but they do contribute to text readability at the semantic level. Run the NLTK algorithm to get a list of unusual words in your documents.

References

Cheung, H. and S. Kemper (1992). “Competing complexity metrics and adults’ production of complex sentences.” *Applied Psycholinguistics*. 13:53-76.

Frazier, L. (1985). “Syntactic complexity.” In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computation, and theoretical perspectives* (pp.129-189). Cambridge: Cambridge University Press.

Pakhomov, S., Chacon, D., Wicklund, M., Gundel, J. (2010). “Computerized Assessment of Syntactic Complexity in Alzheimer's Disease: A Case Study of Iris Murdoch’s Writing.” *Behavioral Research Methods*. 43(1):136-144.

Yngve, V. (1960). “A model and a hypothesis for language structure.” *Proceedings of the American Philosophical Society*, 104, 444-466.