Grammar Frequency and Simplification: When Intuition Fails

David Kauchak, PhD
Computer Science Dept.
Pomona College
david.kauchak@pomona.edu

Gony Leroy, PhD
Management Information Systems
University of Arizona
gonyleroy@email.arizona.edu

Melissa Just, EdD
Rutgers University Libraries
Rutgers
melissa.just@rutgers.edu

Grammar frequency

5.4M sentences

Berkeley Parser

parse and extract 3-level tree

level 1 level 2 level 3

count and bin 3-level structures into 11 bins
(1%, 10%, 20%, ...)

Grammar frequency calculator

The heart is a muscular organ that pumps blood

[ [NP [DT MN]] [VP [VBZ NP]] ]

1 24,068

3-level structure bin 3-level frequency

Methods

Can a medical writer create simpler text by only changing the grammatical structure of a sentence?

Previous result: grammar frequency correlates with perceived difficulty (5-point Likert) and actual difficulty (Cloze test)

Data: 220 sentences, 10 from each grammar bin

Approach: A medical writer attempted to simplify sentences by only changing the grammatical structure:

However, due to its susceptibility to environmental stress cracking, its use is limited to low-stress applications.

However, its use is limited to low-stress applications due to its susceptibility to environmental stress cracking.

Evaluation:

Perceived difficulty: 5-point Likert

Actual difficulty: multiple choice Cloze test on 4 blanked nouns

30 participant from Amazon’s mechanical Turk for each sentence (before and after simplification)

Results

Grammar frequency:
The writer was effective at producing sentences with more frequent grammatical structures

- 52.7% increased in grammar frequency bin
- 22.3% stayed the same
- 25.0% reduced in grammar frequency bin

Perceived difficulty
More frequent structures were perceived to be easier

- 2.13 before simplification vs. 2.35 after

Actual difficulty
More frequent structures were not easier to understand

- The Cloze score did not change

Acknowledgments
Research reported was supported by the National Institutes of Health (NIH) NINDSNS011085. The content is solely the responsibility of the author and does not necessarily represent the official view of the NIH.