

Does high frequency lead to automaticity? A corpus study.

Bybee (2002) proposes that articulatory reduction is observed in high-frequency words because the high frequency of use makes the production of such word more automatic. A prediction of this hypothesis is that the production of high-frequency words should be harder to interrupt than production of low-frequency words because automatic behavior is less subject to conscious inhibition. Some evidence supporting this prediction has been provided by Logan (1982) who found that interrupting typing in response to a signal while typing the highly frequent word 'the' is harder than while typing some other, less frequent word.

The present corpus study tested this hypothesis by examining instances of repetition and replacement repair found in the Sitchboard Corpus (Godfrey et al. 1992). Repetition repair, illustrated in (1)-(2), occurs when a speaker repeats a word s/he has just produced or started producing. Replacement repair, illustrated in (3)-(4), occurs when the speaker replaces a word that has just been (partially) produced with another one. As examples (1)-(4) show, there is variability in whether word production is interrupted or goes to completion before the repair.

The results show that for both repetitions and replacements, words that are interrupted tend to have lower token frequency than words that are not interrupted. Figure 1 shows the effect for replacement repairs. Longer words are more likely to be interrupted but the frequency effect holds when length of the word, in terms of number of segments, number of syllables, or mean duration is controlled.

While these results support the automatization hypothesis, a number of alternative interpretations need to be considered. First of all, since high-frequency words tend to be shorter, the speaker has fewer opportunities to decide to interrupt speech production while producing a high-frequency word. However, if the location of interruption is determined by whether the speaker has time to decide that s/he should interrupt the flow of speech during the production of a word, what remains of an interrupted word should be as long as the entire duration of an uninterrupted word, which is not true for our data.

An alternative explanation for the replacement repair results would appeal to the frequency of the replacement word. Since high-frequency words are more accessible, the replacement would be more likely to become available during the production of the to-be-replaced word if the replacement word has high frequency. If frequent words tended to replace rare words or vice versa, that is, if there were a negative correlation between the frequency of the replaced and the frequency of the replacement, rare words would be interrupted more often simply because alternatives would come to mind more quickly. However, the correlation between the frequency of the replaced and that of the replacement is actually positive, making this explanation not viable.

Finally, when a word is interrupted, the coder needs to make a guess about its lexical identity, which provides a source of error. However, an observer faced with incomplete information about lexical identity tends to be biased in favor of coming up with high-frequency words (e.g., Goldiamond and Hawkins 1958). Thus, the observer is likely to err by coming up with interrupted words that are higher in frequency than the actual words intended by the speaker. It is therefore a bias against the automatization hypothesis, which states that interrupted words should have lower frequency than words that are completed.

Thus, the study provides strong evidence for the hypothesis that words differ in how automatic their production is, such that production of high-frequency words is more automatic than the production of low-frequency words, supporting Bybee's (2002) explanation for the link between articulatory reduction and high frequency of use.

Data (Switchboard, Godfrey et al. 1992):

- (1) *We've gotten, gotten pulled into these superfund deals*
- (2) *He is living now in Maryland but he li-, lived in Grapevine for a long time*
- (3) *We were surprised to find 'Toyota' written, I mean, imprinted on the engine*
- (4) *It was pathe-, it was horrible.*

Figure 1. Replacement repair (* $p < .05$, ** $p < .01$, ***, $p < .001$)

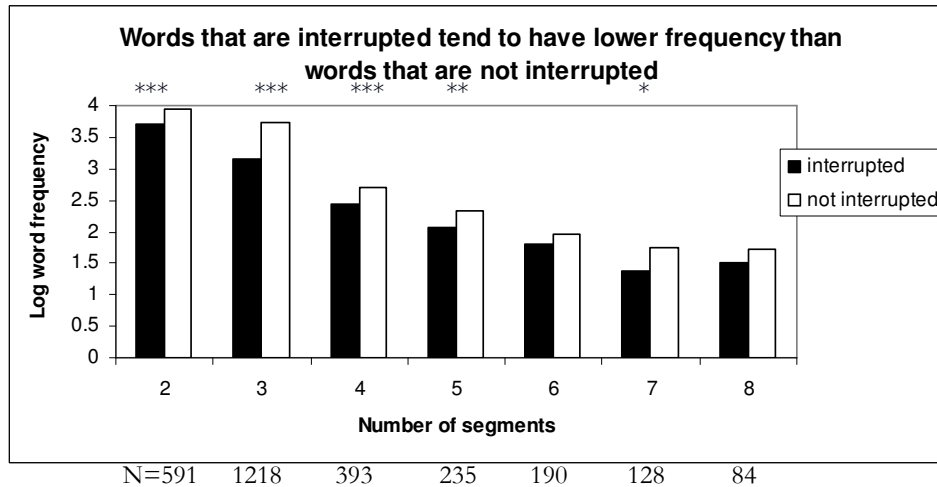
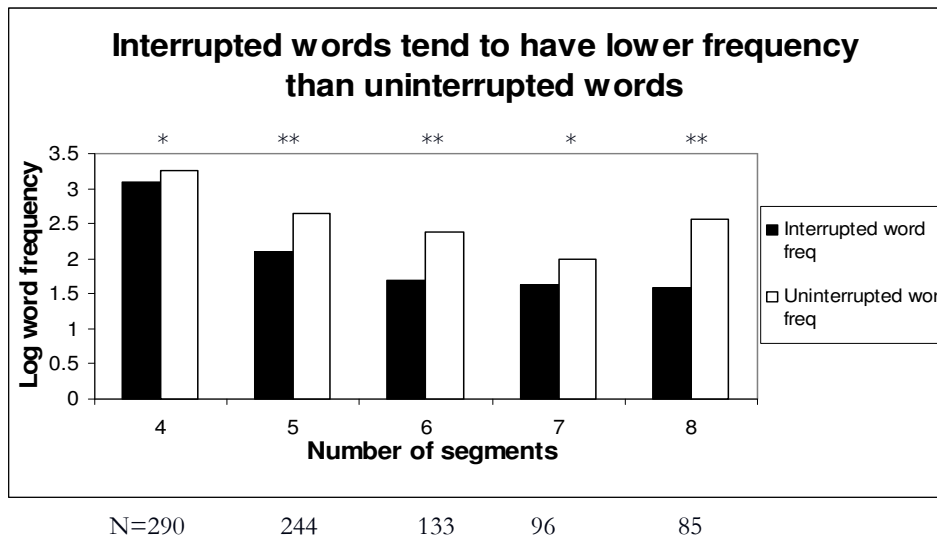


Figure 2. Repetition repair



References:

Bybee, J. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14, 261-90.

Godfrey, J. J., E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. *IEEE ICASSP*: I517-15.

Goldiamond, I., and W. F. Hawkins. 1958. Vexiersversuch: the log relationship between word-frequency and recognition obtained in the absence of stimulus words. *JEP*, 6, 457-63.

Logan, G. D. 1982. On the ability to inhibit complex movements: A stop-signal study of typewriting. *JEP: Human Perception and Performance*, 8, 778-92.