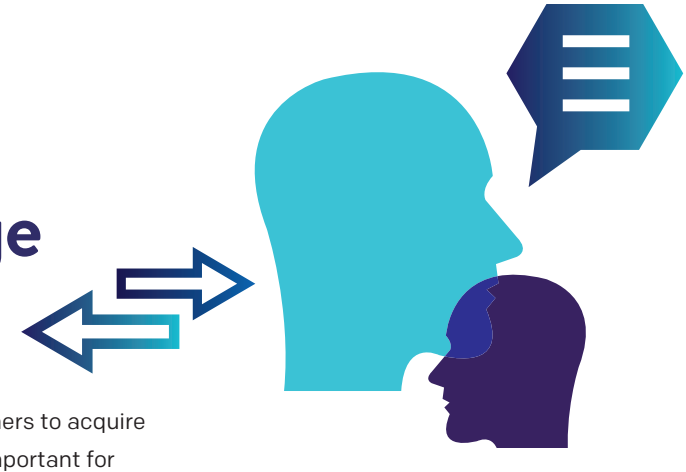
The word frequency effect on second language vocabulary learning



Research has shown that some words are harder for second language learners to acquire than others. Hence, estimating the difficulty level of an individual word is important for effective language instruction. In order to do so, it is necessary to identify the factors that make words difficult.

First language research has noticed a clear influence of word frequency on word difficulty. For instance, the word 'phone' is less difficult than the word 'floccinaucinihilipilification' (Yes, thats a real English word!) because we hear 'phone' more frequently than we hear 'floccinaucinihilipilification'. However, frequency is not the only factor that contributes to word difficulty. For instance, the fact that 'phone' is shorter in length than 'floccinaucinihilipilification' also makes it an easier word.

Two important questions arise from the above observations. (1) Does word frequency in the target language influence the second language vocabulary learning as well? (2) What other factors affect word difficulty in second language learning?

Voxy conducted original research to answer these questions. We investigated four factors that we hypothesized might contribute to the difficulty of a word as it pertains to second language learners: (1) Frequency of word usage (2) Word length in number of characters (3) Number of syllables in a word (4) Number of consonant clusters in a word. Word frequency, word length and number of syllables come from word difficulty studies in first language research. Word frequency is often treated as the quantifiable correlate of word familiarity, and word length and number of syllables measure structural complexity of a word. In the current study, we introduce consonant clusters as the measure of phonetic complexity. Phonetic complexity is a dimension of word difficulty that concerns perception and oral production of the word. Some languages have no (or very few) words with consonant clusters. As a result, speakers of those languages have difficulty perceiving and producing foreign words with consonant clusters.

Below, we present a brief description of Voxy's word difficulty experiment and discuss our key findings.

Experiment Design/Procedure:

We prepared a survey of 140 words, chosen randomly from a corpus of public domain books from Project Gutenberg, which we divided into four subgroups: words with varying frequencies, words with varying word lengths, words with varying counts of syllables, and words with varying counts of consonant clusters.The words in each subgroup were controlled for other variables, with equal number of words per condition within each subgroup. The subgroups and conditions are explained in more details below.

SUBGROUPS	CONDITIONS
1. Varying frequency bands	1-5, 5-50, 50-500, 500-5000
2. Varying word length	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
3. Varying number of syllables	1, 2, 3, 4
4. Varying number of consonant clusters	0, 1, 2, 3

Subgroup 1 consisted of 48 words belonging to four different frequency ranges - 1 to 5, 5 to 50, 50 to 500, and 500 to 5000. There were 12 words in each frequency range, and all the words were of length 5. Subgroup 2 consisted of 36 words of length 3 to 14. There were 3 words in each length condition. All the words were in the frequency range 50-500. Subgroup 3 consisted of 32 words with syllable counts 1 to 4. LIke subgroup 2, all 32 words belonged to the frequency range 50-500. Subgroup 4 consisted of 24 words divided equally among the four consonant cluster conditions - 0 clusters, 1 cluster, 2 clusters, and 3 clusters. All the words belonged to the frequency range 50-500.

The survey was sent to 217 Spanish and Portuguese Voxy users. Their task was to decide whether a word was 1. Easy to learn 2. Difficult to learn, or 3. Unknown word.

We used a three point scale (easy, difficult, and unknown) instead of two (easy and difficult) because we wanted to differentiate words that learners find difficult from the ones that they aren't familiar with. This distinction is especially relevant for subgroup 1 (words with varying frequencies). As mentioned above, word frequency is treated as the quantifiable correlate of word familiarity, and it does not make sense to measure familiarity of unknown words. However, for other measures of complexity (structural and phonetic) we treat unknown words as difficult words and report combined results.

RESULTS

(1) Frequency:

The results showed a negative correlation between word difficulty and word frequency; as frequency increased, difficulty decreased. This is similar to the relationship between word difficulty and word frequency in the first language. The correlation between word frequency and unknown words is also worth noticing. More words in lower frequency ranges were marked as unknowns than the words in higher frequency ranges.

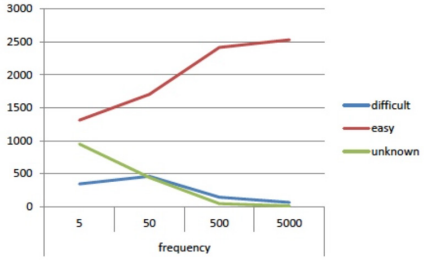


Figure 1: Effect of frequency

(2) Word length and number of syllables:

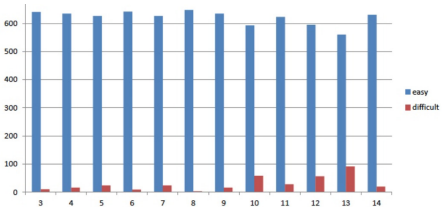


Figure 2: Effect of word length

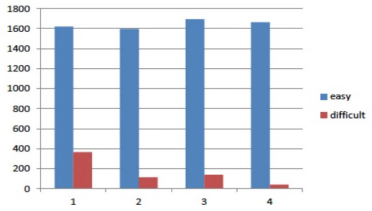


Figure 3: Effect of varying number of syllables

Unlike the frequency effect, the results did not show a clear trend for varying word length and varying counts of syllables. Most words in these two subgroups were rated as easy by most participants as shown in figure 2 and figure 3. Note that the words in these subgroups were controlled for frequency - they fall in the same frequency range. So, one reason for this result could be that frequency is a better predictor for word difficulty, and as these words fall in the same frequency band, they were rated to be almost equally difficult.

(3) Consonant Clusters:

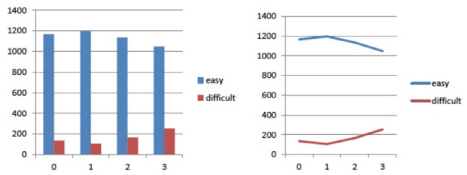


Figure 4: Effect of consonant clusters

Again, word-frequency seems to dominate participants' responses. As the words were in the same frequency band, they were rated similarly irrespective of varying number of consonant clusters. The second graph in figure 4 shows that difficulty increases with the increase in number of clusters, but the result is not significant.

CONCLUSION

The results show a correlation between English word frequency and perceived word difficulty of English words by Spanish and Portuguese speakers. Most participants rated low frequency words to be either difficult to learn or unknown words.

There were no clear results for factors other than word frequency. Most words in the other subgroups were categorized as easy to learn irrespective of their structural or phonetic complexities. As all those words belonged to the same frequency range (50-500), we have a strong reason to hypothesize that word frequency dominated other factors.

In order to examine the aforementioned hypothesis , a follow up experiment shall be conducted. In the follow up experiment, the words in subgroups 2, 3 and 4 will be replaced by 1) words in higher frequency range (500-5000), and 2) words in lower frequency range (1-5). Our hypothesis will be supported if most words in 1 are judged easy and most words in 2 are judged to be difficult irrespective of their structural and phonetic complexities.

"I've always been passionate about languages and that passion has driven me to Voxy where I have the opporutnity to work closely with a team of passionate educators and engineers from all over the world. Here, I'm able to put my love for computational modeling of linguisitic theories into practice by shaping our Natural Language Processing and Machine Learning tools.
In my opinion, Voxy is the perfect place to grow both personally and professionally."

CESAR KOIRALA