Creating a Tier 3 Word List for Computer Science

Bayan Shayeb, Emma McDonald, Carrie Demmans Epp
Department of Computer Science, University of Alberta

Background

A Tier 3 vocabulary refers to words that are used within a specific area while a Tier 2 vocabulary refers to words that are common across academic domains. Tier 1 refers to commonly used words. Corpus refers to a collection of texts. Word lists can be useful because they provide learners with a method of knowing which words are worth studying and spending their time on.

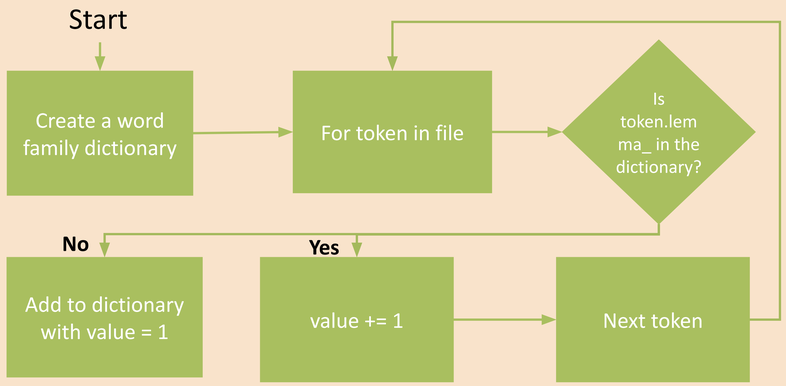
1. **Which word families occur frequently in a Computer Science Corpus that are domain-specific?**
2. **What percentage of the words in the Computer Science Corpus are covered by the Computer Science Word List?**

Methods

- The corpus had 4,506,546 individual words and 110,588 unique words.
- Averil Coxhead's method in *A New Academic Word List* (2000) was used as a guide for the development of the word list.

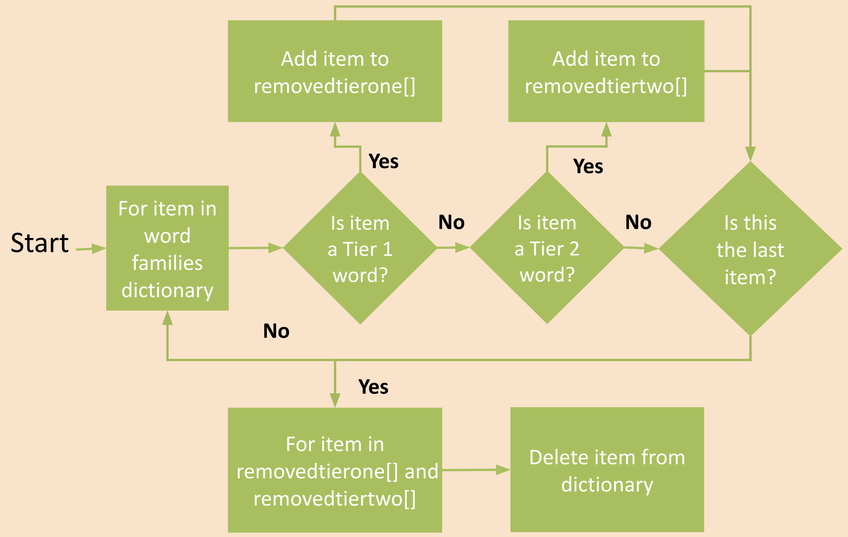
- The creation of the word list was automated
  - All punctuation, capitalization, numbers, symbols, and stop words were removed.
  - The below were counted:
    - total number of words
    - unique words
    - word families - the roots (lemmas)
  - Word families were only kept if:
    - They occurred 100+ times
    - They weren't in the Tier 1 or Tier 2 word lists

Word Family Dictionary Creation



Creation of a dictionary to hold all the word families in the corpus with no repetition

Tier 1 and 2 Word Removal



Creation of the final word list by removing all Tier 1 and 2 words

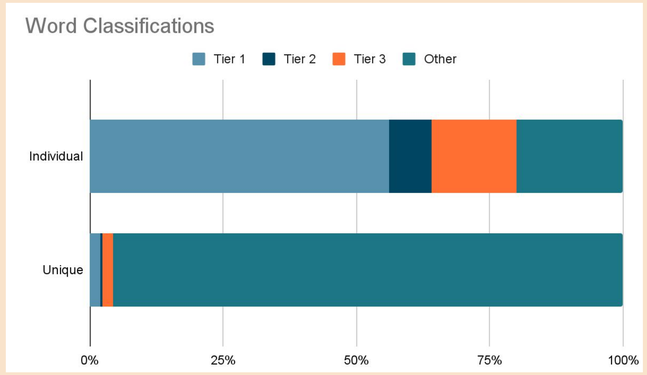
Final Word List

The Tier 3 word list produced consisted of 2,110 word families, representing 732,201 words. Below is a sample of the word list.

Word List Sample

Word Family	No. of Occurrences
algorithm	7,633
information	7,614
user	7,468
analysis	4,768
variable	3,885
computer	3,801
node	3,331
matrix	2,920
communication	2,629
cell	2,575
distribution	2,500

Corpus Coverage



Percentage of individual and unique words that are Tier 1, 2, 3 and other

- The word list accounted for 15.9% of individual words and 1.9% of unique words.

Discussion and Conclusion

- In comparison to Coxhead's Tier 2 percentage, this word list had a lower coverage of the corpus used.
- Further research could involve comparisons to other word lists or evaluations of different corpora.
- Many words in the final word list were not completely specific to Computer Science.
  - Additional manual data cleaning or increasing the minimum frequency per word could address this issue
- The SpaCy library used for the modification of the file did not lemmatize certain words correctly, which could have allowed certain Tier 1 and 2 words to remain.

References

- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951
- West, M. (1953). A general service list of English words. London: Longman, Green.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing Words to Life: Robust Vocabulary Instruction* (2nd ed.). The Guilford Press.

Acknowledgments

UNIVERSITY OF ALBERTA
EdTeKLA
Canada
MOTOROLA SOLUTIONS FOUNDATION