

2.7d Successor Stemmers

- The process determines the successor varieties for a word and uses this information to divide a word into segments and selects one of the segments as the stem.
- The [successor variety](#) of a segment of a word in a set of words is the number of distinct letters that occupy the segment length plus one character.

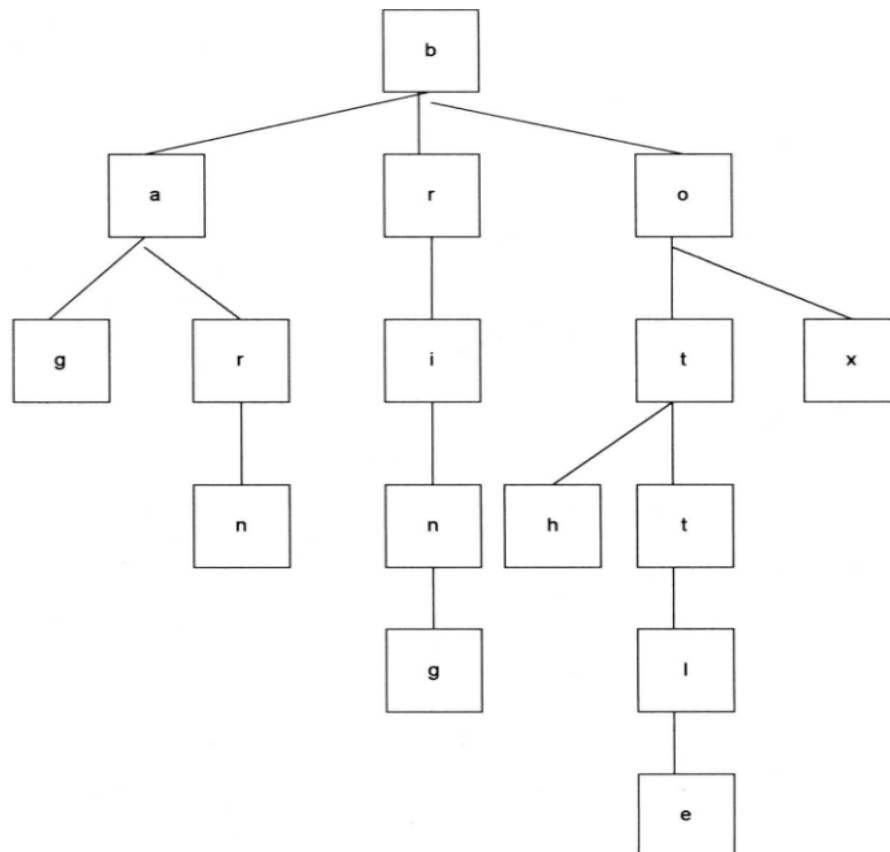


Figure 4.2 Symbol Tree for terms bag, barn, bring, box, bottle , both

- A graphical representation of successor variety is shown in a symbol tree.
- Figure shows the symbol tree for the terms bag, barn, bring, both, box, and bottle.
- The successor variety for any prefix of a word is the number of children that are associated with the node in the symbol tree representing that prefix.
- For example, the successor variety for the first letter “b” is three. The successor variety for the prefix “ba” is two.

— The successor varieties of a word are used to segment a word by applying one of the following four methods :

1. Cutoff method: a cutoff value is selected to define stem length. The value varies for each possible set of words.
2. Peak and Plateau: a segment break is made after a character whose successor variety exceeds that of the character immediately preceding it and the character immediately following it.
3. Complete word method: break on boundaries of complete words
4. Entropy method: uses the distribution of successor variety letters.

— Let $|D_{ak}|$ be the number of words beginning with the k length sequence of letters a.

— Let $|D_{akj}|$ be the number of words in D_{ak} with successor j.

— The entropy (Average Information) of $|D_{ak}|$ is:

$$H_{ak} = \sum_{j=1}^{26} -(|D_{akj}|/|D_{ak}|) (\log_2(|D_{akj}|/|D_{ak}|))$$

— Using this formula a set of entropy measures can be calculated for a word and its predecessors.

- _ To illustrate the use of successor variety stemming, consider the example below where the task is to determine the stem of the word READABLE.

Test Word: READABLE

Corpus: ABLE, APE, BEATABLE, FIXABLE, READ, READABLE
READING, READS, RED, ROPE, RIPE.

Prefix	Successor Variety	Letters

R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	BLANK

- _ Using the complete word segmentation method, the test word "READABLE" will be segmented into "READ" and "ABLE,".
- _ After a word has been segmented, the segment to be used as the stem must be selected.
- _ Hafer and Weiss used the following rule:

```

if(first segment occurs in <= 12 words in database)
    first segment is stem
else (second segment is stem)

```

- _ Selecting the first or second segment in general determines the appropriate stem.