**GTU Department of Computer Engineering**

**CSE 484/654 Natural Language Processing**

**Fall 2021 - Homework 2 Report**

**Akif KARTAL**

**171044098**

**1) Problem Definition**

The problem is to develop a statistical language model of Turkish that will use N-grams of Turkish syllables.

**2) Solution**

The homework was finished as expected in homework pdf file. Solution steps are following;

**2.1) Creating Corpus**

The given turkish-wikipedia-dump text was to big(441MB) to test therefore, I used small portion of it which is 6,39 MB of data and 0.366 MB test data (5% of the set).

**2.2) Dividing Turkish words into syllables**

First, I convert all the letters to small case letters.

As in hw1, I used following program to divide turkish words into syllables.

<https://github.com/MeteHanC/turkishnlp>

**Output:**

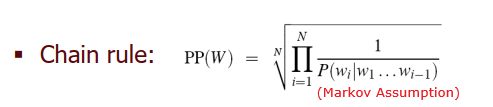
1-line syllabled corpus text;



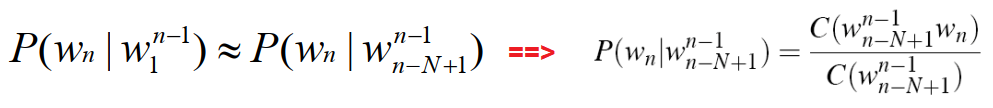
**2.3) Calculating N-grams**

**2.4) Calculating perplexity with the Markov assumption**

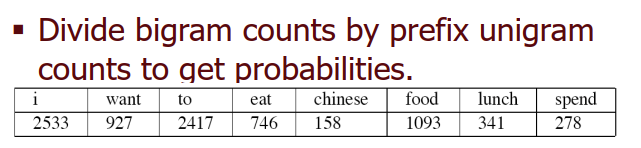
Perperlixty formula with the Markov assumption;

****

Markov assumption and calculating probabilities

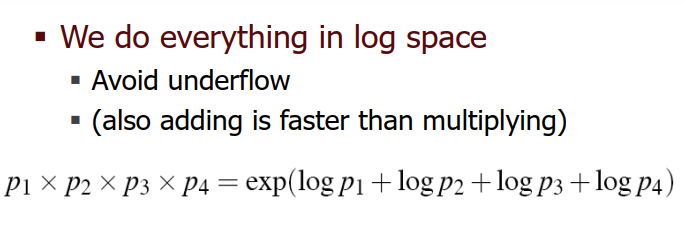


Calculating probabilities



**Using logarithm of the multiplication of the chain rule formula**

Following formula will be used while calculating probabilities



**Putting all of these together and getting result**