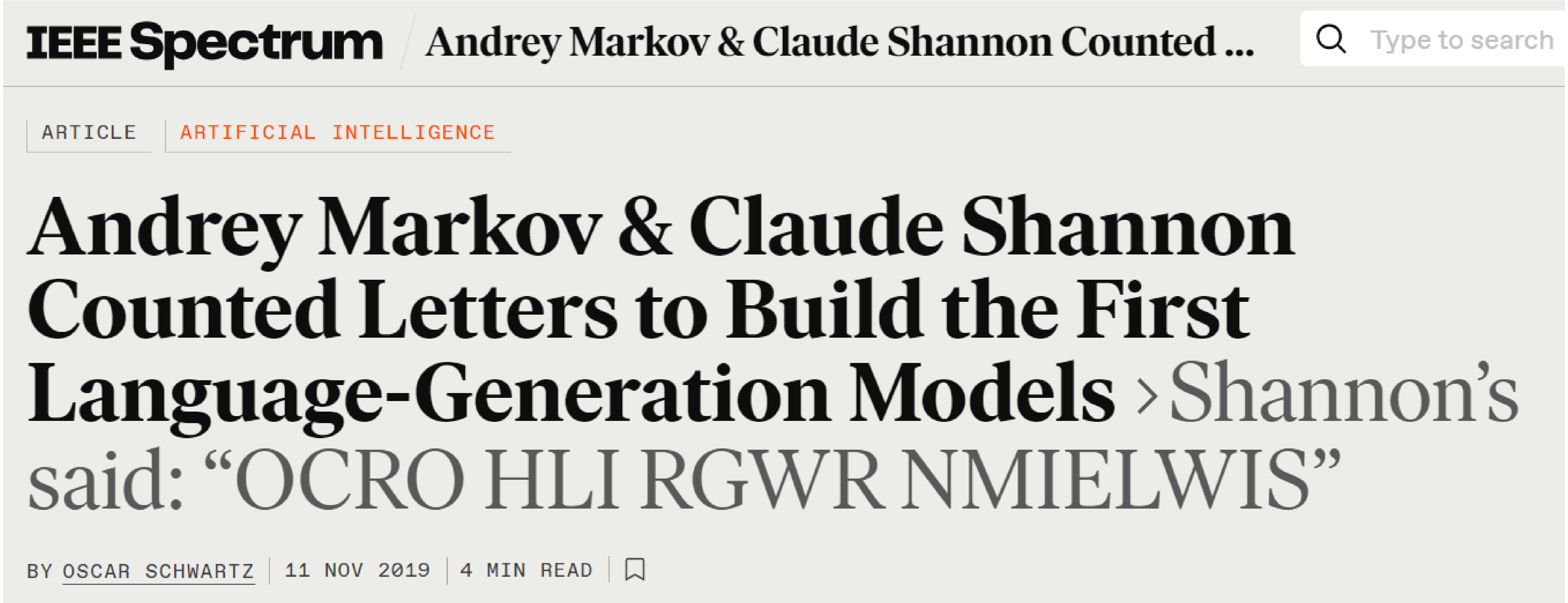
Language Models - Basics



ARTICLE ARTIFICIAL INTELLIGENCE

Andrey Markov & Claude Shannon Counted Letters to Build the First Language-Generation Models › Shannon's said: "OCRO HLI RGWR NMIELWIS"

BY OSCAR SCHWARTZ | 11 NOV 2019 | 4 MIN READ

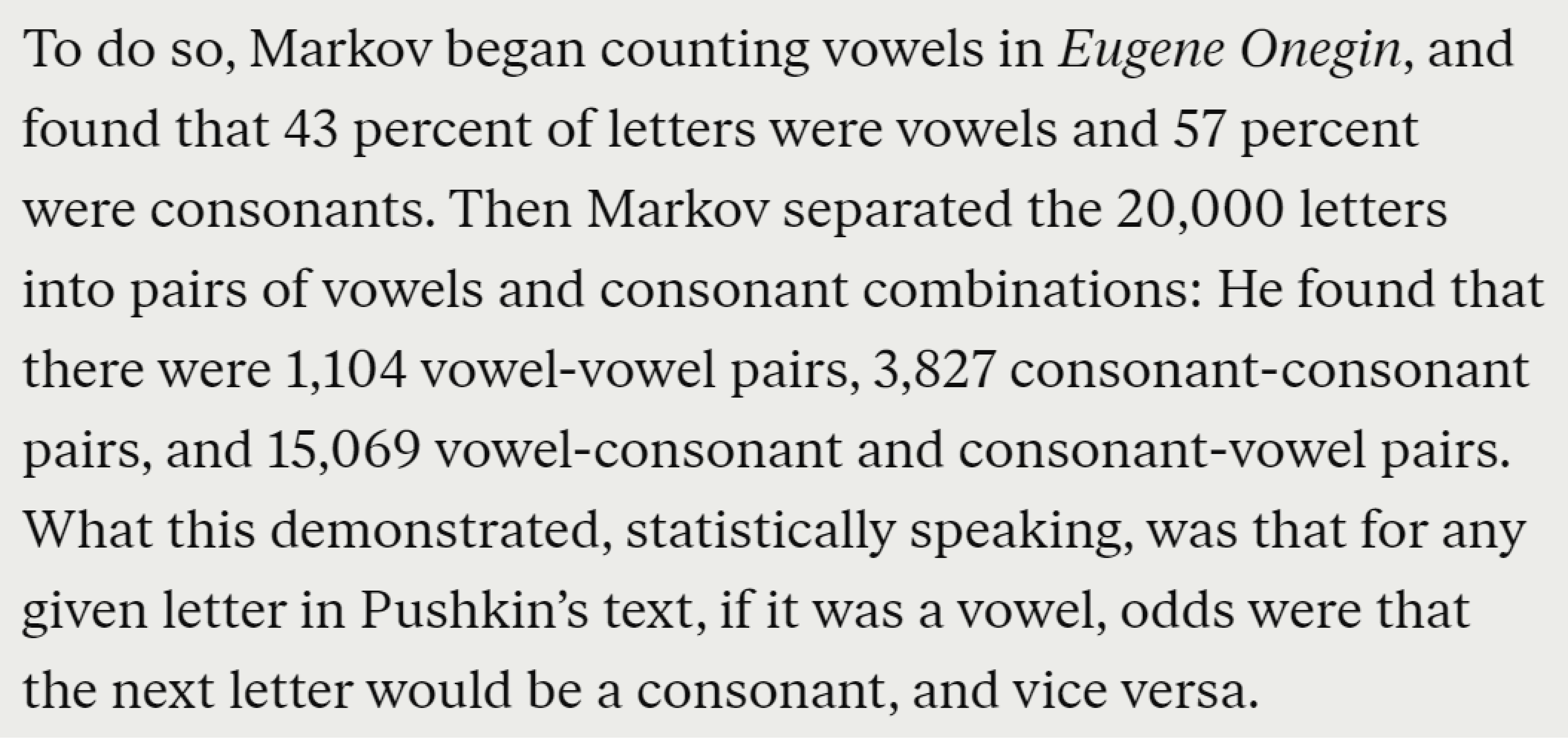
Science in Context **19**(4), 591–600 (2006). Copyright © Cambridge University Press
doi:10.1017/S0269889706001074 Printed in the United Kingdom

Classical Text in Translation

An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains

A. A. Markov

(Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 23 January 1913)[1]



To do so, Markov began counting vowels in *Eugene Onegin*, and found that 43 percent of letters were vowels and 57 percent were consonants. Then Markov separated the 20,000 letters into pairs of vowels and consonant combinations: He found that there were 1,104 vowel-vowel pairs, 3,827 consonant-consonant pairs, and 15,069 vowel-consonant and consonant-vowel pairs. What this demonstrated, statistically speaking, was that for any given letter in Pushkin's text, if it was a vowel, odds were that the next letter would be a consonant, and vice versa.

Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

``It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source.''

``A second method is to delete a certain fraction of the letters from a sample of English text and then let someone attempt to restore them. If they can be restored when 50% are deleted the redundancy must be greater than 50%.''

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

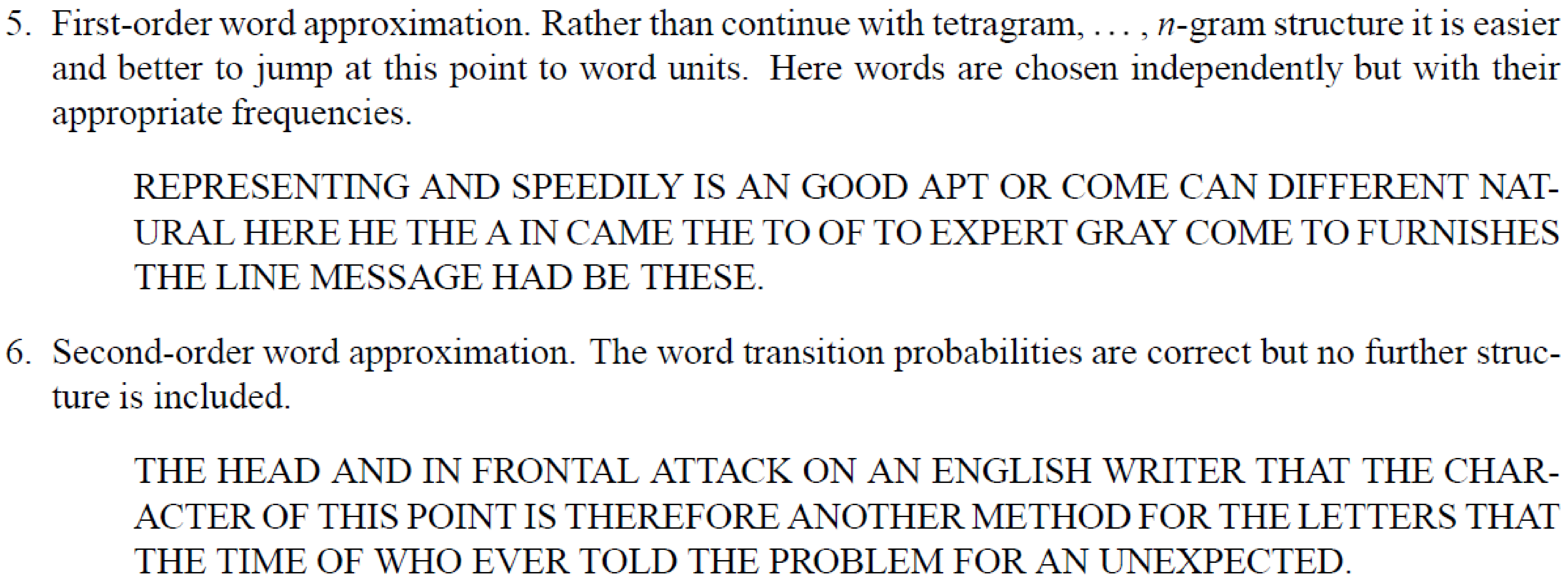
   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.



5. First-order word approximation. Rather than continue with tetragram, … , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The resemblance to ordinary English text increases quite noticeably at each of the above steps.

3. Second-order approximation (digram structure as in English).

> ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

To construct (3) for example, one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. A similar process was used for (4), (5) and (6). It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage.

Shannon's Method

Assigning probabilities to sentences is all well and good, but it's not terribly illuminating . A more interesting task is to turn the model around and use it to generate random sentences that are like the sentences from which the model was derived.

- Generally attributed to Claude Shannon.

1. *Sample a random bigram (<s>, w) according to its probability*
2. *Now sample a random bigram (w, x) according to its probability*
   - *Where the prefix w matches the suffix of the first.*
3. *And so on until we randomly choose a (y, </s>).Then string the words together.*

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>

Thank you !!