ENTRIES AND AVERAGE MUTUAL INFORMATIONS FOR A ‘CHOSEONG,’ A ‘JUNGESEONG,’ AND A ‘JONGSEONG’ IN MULTI-SYLLABLE KOREAN WORDS

Jae Hong Lee, Nonmember and Chaehag Yi, Associate Member

SUMMARY In this paper, we compute entropies and average mutual informations for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ in multi-syllable Korean words. In these computations, we consider the property that each Korean syllable is divided into a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ which are a consonant, a vowel, and a consonant, respectively, without exception. We compute entropies and average mutual informations for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ from the cumulative frequency of Korean syllables with respect to word length and syllable location in a word. We compute average mutual informations for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong,’ and a ‘jongseong’ between the two adjacent syllables in a word. The correlation between the two adjacent syllables in a word is not large on the average.

Key words: entropy, average mutual information, Korean words, syllable location

1. Introduction

The entropy of a language was defined as a statistical parameter which measures how much information is produced on the average for each letter of a text in a language [1]. Shannon found upper and lower bounds for the entropy of printed English based on the number of trials required for a subject to guess subsequent symbols in a given text [2]. The entropy of printed Portuguese was obtained by the statistical calculation [3].

A Korean word is composed of syllables. Without exception, a written Korean syllable is divided into a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ which are a consonant, a vowel, and a consonant, respectively. This property has not been found in other written languages. Considering this property of Korean syllables, entropies and average mutual informations were computed for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ in a syllable in written Korean [4].

In this paper, we compute various entropies and average mutual informations of a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ in a multi-syllable Korean word. We compute the length distribution of Korean words and the L-th order entropies for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ in an L syllable word. We compute entropies for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ in a syllable at various locations in a word. Also, we compute average mutual informations for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ between the two adjacent syllables in a word as measures of the correlation of occurrence.

2. Entropy of Korean

A Korean word is composed of syllables. A written Korean syllable is regarded as a random variable according to its probabilistic properties in occurrence. A Korean syllable is divided into a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’ which are also regarded as random variables.

We define the random variable $S$ to denote a syllable. Also, we define the random variables $X$, $Y$, and $Z$ to denote a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong,’ respectively. The sample spaces of these random variables $X$, $Y$, and $Z$ are given by

$$A_X = \{ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ\},$$

$$A_Y = \{ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ, ㅐ, ㅒ, ㅔ, ㅖ, ㅚ, ㅟ, 食品药品, ㅢ, ㅥ, ㅧ, ㅨ, ㅩ, ㅪ\},$$

$$A_Z = \{\text{blank}, ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ, ㅐ, ㅒ, ㅔ, ㅖ, ㅚ, ㅟ, 食品药品, ㅢ, ㅥ, ㅧ, ㅨ, ㅩ, ㅪ\}.$$  \hspace{1cm} (1)

The sizes of the sample spaces are $|A_X| = 19$, $|A_Y| = 21$, and $|A_Z| = 28$. The joint probability $p(x, y, z)$ is the probability that a specific syllable $s = (x, y, z)$ occurs.

![서울](image)

Word length = 2 syllables.

First syllable: $X_1 = 'a'$, $Y_1 = 'i'$, and $Z_1 = 'blank'$.

Second syllable: $X_2 = 'o'$, $Y_2 = 't'$, and $Z_2 = 'a'$.

Fig. 1 Structure of a Korean word for a ‘choseong,’ a ‘jungseong,’ and a ‘jongseong’.
There are $19 \times 21 \times 28 = 11,172$ possible syllables in Korean, although some of them are not used at all. An example is given in Fig. 1 to explain the structure of a Korean word for a 'choseong,' a 'jungseong,' and a 'jongseong.' The word '가이(Seoul)' is composed of two syllables. The first syllable '가' is divided into a 'choseong' '가,' a 'jungseong' 'ㅏ,' and a 'jongseong' 'ㅏ', 'ㅏ'. The second syllable '이' is divided into a 'choseong' 'ㅔ,' a 'jungseong' 'ㅔ,' and a 'jongseong' 'ㅔ.'

Entropy and average mutual information are used as the measure of information [1]. The entropy of a 'choseong' $X$ in bits is given by

$$ H(X) = -\sum_x p(x) \log_2 p(x) $$

where $p(x)$ is the probability of a 'choseong' $X$. Other entropies are given similarly: the entropy of a 'jungseong' $H(Y)$, the entropy of a 'jongseong' $H(Z)$, the joint entropy of a 'choseong' and a 'jungseong' $H(X, Y)$, the joint entropy of a 'choseong' and a 'jongseong' $H(X, Z)$, the entropy of a 'jungseong' and a 'jongseong' $H(Y, Z)$, and the joint entropy of a 'choseong,' a 'jungseong,' and a 'jongseong' $H(X, Y, Z)$. The average mutual information between a 'choseong' $X$ and a 'jungseong' $Y$ is given by

$$ I(X; Y) = H(X) + H(Y) - H(X, Y). $$

Other average mutual informations are given similarly: the average mutual information between a 'choseong' and a 'jungseong' $I(X; Z)$, the average mutual information between a 'jungseong' and a 'jongseong' $I(Y; Z)$, the average mutual information between a 'choseong' $X$ and a 'jungseong' and 'jongseong' $I(Y, Z)$, the average mutual information between a 'jungseong' and a 'jongseong' $I(Y; X, Z)$, and the average mutual information between a 'jungseong' and a 'jongseong' $I(Z; X, Y)$.

In an $L$ syllable word, the $L$-th order entropy is considered. We define a random vector $\vec{X} = (X_1, X_2, \ldots, X_L)$ to denote 'choseong's in a word where $X_i$ represents the 'choseong' of the $i$-th syllable in an $L$ syllable word. Similarly, we define a 'jungseong' vector $\vec{Y}$ and a 'jongseong' vector $\vec{Z}$. The $L$-th order entropy of a 'choseong' vector $\vec{X}$ is given by

$$ H(\vec{X}) = -\frac{1}{L \cdot x_1, x_2, \ldots, x_L} \sum_x p(x_1, x_2, \ldots, x_L) \log_2 p(x_1, x_2, \ldots, x_L) $$

where $p(x_1, x_2, \ldots, x_L)$ is the probability of a 'choseong' vector $\vec{X}$. Other average mutual informations are given similarly: the $L$-th order entropy of a 'jungseong' vector $H(\vec{Y})$, the $L$-th order entropy of a 'jongseong' vector $H(\vec{Z})$, the $L$-th order joint entropy of a 'choseong' and a 'jungseong' vectors $H(\vec{X}, \vec{Y})$, the $L$-th order joint entropy of a 'choseong' and a 'jongseong' vectors $H(\vec{X}, \vec{Z})$, the $L$-th order joint entropy of a 'jungseong' and a 'jongseong' vectors $H(\vec{Y}, \vec{Z})$, and the $L$-th order joint entropy of a 'choseong,' a 'jungseong,' and a 'jongseong' vectors $H(\vec{X}, \vec{Y}, \vec{Z})$.

We use average mutual informations for a 'choseong,' a 'jungseong,' and a 'jongseong' between the two adjacent syllables in a word as measures of the correlation of occurrence between the two adjacent syllables. The following average mutual informations are computed: the average mutual information between a 'choseong' in the $(n+1)$-th syllable and a 'choseong' in the $n$-th syllable $I(X_{n+1}, X_n)$, the average mutual information between a 'jungseong' in the $(n+1)$-th syllable and a 'jungseong' in the $n$-th syllable $I(Y_{n+1}, Y_n)$, the average mutual information between a 'jongseong' in the $(n+1)$-th syllable and a 'jongseong' in the $n$-th syllable $I(Z_{n+1}, Z_n)$, and the average mutual information between a 'choseong' in the $(n+1)$-th syllable and a 'jongseong' in the $n$-th syllable $I(X_{n+1}, Z_n)$.

3. Numerical Results

For the computation of entropies and average mutual informations, we use the frequency distribution of Korean word survey [5]. In the survey, 56,077 words were counted in written Korean. With the Korean word distribution, we compute joint and conditional probabilities for the random variables of a 'choseong,' a 'jungseong,' and a 'jongseong.' And then, with these probabilities we compute various joint and conditional entropies and average mutual informations for the three random variables.

The length distribution of Korean words is shown in Table 1. Korean words have average length of 1.82 syllables per word. Most of Korean words have length of one or two.

Entropy and average mutual informations for a 'choseong,' a 'jungseong,' and a 'jongseong' in a syllable are computed which results are given in Table 2. The entropy of a 'choseong' $H(X)$ is greater than that

![Fig. 2](image-url)
Fig. 3 Entropies for a 'choseong,' a 'jungseong,' and a 'jongseong' with respect to word length and syllable location in a word.
of a 'jungseong' $H(Y)$, even though the size of sample space of a 'choseong' $|A_X|$ is less than that of a 'jungseong' $|A_Y|$. The entropy of a 'jungseong' $H(Y)$ is greater than that of a 'jungseong' $H(Z)$, even though the size of sample space of a 'jungseong' $|A_Y|$ is less than that of a 'jungseong' $|A_Z|$. Although the number of all possible Korean syllables is very large, the entropy $H(X, Y, Z)$ is only 7.24 bits. The average mutual information $I(X; Y)$ is greater than $I(Y; Z)$ or $I(X; Z)$. The average mutual information $I(Y; X, Z)$ is greater than $I(X; Y, Z)$ or $I(Z; X, Y)$.

We compute seven types of the $L$-th order entropies for a 'choseong,' a 'jungseong,' and a 'jongseong' in an $L$ syllable word. In Fig. 2, we plot seven types of the $L$-th order entropies. There are only a few Korean words with $L \geq 8$, and hence each of words with $L \geq 8$ has different 'choseong' vectors, 'jungseong' vectors, and 'jongseong' vectors. This explains the reason why seven types of entropy are exactly same in a word with $L \geq 8$. As the word length $L$ increases, the $L$-th order entropies decrease.

In Fig. 3, we plot the seven types of entropies for a 'choseong,' a 'jungseong,' and a 'jongseong' with respect to word length and syllable location in a word: $H(X), H(Y), H(Z), H(X, Y), H(X, Z), H(Y, Z),$ and $H(X, Y, Z)$. In Figs. 3(a) to (g), seven types of entropy show a similar tendency in their values with respect to word length and syllable location in a word. It is shown that the seven types of entropy are relatively large commonly for the second and the third syllables in a word with the length of four to five. Also it is shown that the seven types of entropy are relatively small for the last syllable in a word of any length.

We compute the average mutual information between a syllable and its adjacent syllable in a word with respect to word length and syllable location. In Fig. 4, we plot the average mutual information with respect to word length and syllable location. In Fig. 4, it is shown that the average mutual information $I(S_{n+1}; S_n)$ has relatively large value in a word with the length of three to six, and that $I(S_{n+1}; S_n)$ increases as $L$ increases in a word with $L \geq 6$. In a word with $L \geq 7$, the correlation between the two adjacent syllables is not large on the average.

We compute the average mutual informations for a 'choseong,' a 'jungseong,' and a 'jongseong' between the two adjacent syllables in a word with respect to word length and syllable location. In Fig. 5, we plot the average mutual informations. The correlation between the two adjacent syllables in a word is not large on the average.

4. Conclusions

Korean words have the average length of 1.82 syllables per word. Most of Korean words have length of one or two. The entropy of a 'choseong' is greater than that
of a 'jungseong,' and the entropy of a 'jungseong' is greater than that of a 'jongseong' on the average in total syllable. We compute the $L$-th order entropies for a 'choseong,' a 'jungseong,' and a 'jongseong' in an $L$ syllable word. As the word length increases, the $L$-th order entropies decrease. We compute entropies for a 'choseong,' a 'jungseong,' and a 'jongseong' with respect to word length and syllable location in a word. Entropies are relatively large for the second and the third syllables in a word with the length of four to five syllables and are relatively small for the last syllable in a word. We compute average mutual information between a syllable and its adjacent syllable in a word as a measure of the correlation of occurrence between the two adjacent syllables. Also, we compute average mutual informations for a 'choseong,' a 'jungseong,' and a 'jongseong' between the two adjacent syllables in a word. The correlation between the two adjacent syllables in a word is not large on the average.

The results of this paper would have applications in the Korean speech recognition and synthesis, and cryptography.

References