# Part-of-speech Sequences and Distribution in a Learner Corpus of English

Rebecca H. Shih[*], John Y. Chiang[+] and F. Tien[+]

[*]Department of Foreign Languages and Literature

[+]Department of Computer Science and Engineering,

National Sun Yat-sen University

Page 171 ~ 177
Proceedings of Research on Computational Linguistics
Conference XIII (ROCLING XIII)
Taipei, Taiwan
2000-08-24/2000-08-25

# Part-of-speech Sequences and Distribution in a Learner Corpus of English

Rebecca H. Shih[*], John Y. Chiang[+] and F. Tien[+]

[*]Department of Foreign Languages and Literature
[+]Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung,Taiwan, R.O.C.
E-mail: hsuehueh@mail.nsysu.edu.tw

## Abstract

Computer learner corpora have been widely used by SLA/EFL specialists since mid 1990s to gain better insights into authentic learner language. The work presented in this paper examines the inter-language of Taiwanese learners of English from a part-of-speech sequence perspective. Two pre-tagged corpora (one learner corpus and one native corpus) are involved in this work. The experimental results indicate that there are more than one third of eligible POS trigrams that are never practiced by the Taiwanese learners in their writing and the learners have stronger preference than native speakers in using pronouns, especially right after punctuations, verbs and conjunctions.

## 1. Introduction

With the recognition of its theoretical and practical potential, computer learner corpora (CLC) have been subsequently built up around the world since early 1990s.[1] CLC research aims to gain a better insight into learners' inter-language from the authentic data. The research often involves comparisons between inter-language that learners possess and native language on various linguistic features. For instance, the frequency distributions of most commonly-used words in a native and seven eastern European learner corpora are compared on various parts-of-speech categories[2]; the use of complement clauses in terms of their frequencies in four learner corpora as contrasted with their native counterparts [3] is studied; the use of adverbial connectors by Swedish learners in comparison with the natives' is examined [4]. The quantitative information as such often guides the researchers to carry out insightful qualitative analysis. And this kind of cross-language approach helps SLA and EFL specialists find out what linguistic features the language learners are apt to overuse/underuse, what particular areas of language behavior that are shared by learners with different backgrounds, and to what extent these phenomena appear in learner English.

The aim of the work in this paper is to discover distinctive inter-language features of Taiwanese learners of English in terms of part-of-speech sequences and distribution. It is based on two corpora: Taiwanese Learner corpus of English (TLCE) and British National Corpus (BNC). Both corpora are tagged by TOSCA tagger, using the TOSCA-ICLE tagset. The details of the corpora and the tagger will be stated subsequently in Section 2, which is followed by a series of experiments in Section 3. Conclusions are drawn in Section 4 with future work.

## 2. Methodology

### 2.1 Corpora: TLCE and BNC

As stated in the introduction, CLC-research often compares non-native data with native data in order to reveal the overuse and/or underuse phenomena in a learner corpus. In this work, the Taiwanese Learner Corpus of English (TLCE) is under investigation and the British National Corpus (BNC) is used for comparison. TLCE of 455,000 words is a growing corpus of English compositions and weekly journals written mainly by college English majors( freshmen, sophomores and juniors) from Sun Yat-sen and Chi-nan universities in Taiwan. The BNC contains modern British English and is a unique collaboration between three major U.K. dictionary publishers, two universities, and the British Library [5]. The work here utilizes mainly its subset of 1 million words (from BNC Sampler written text).

### 2.2 Tagger: TOSCA

The corpora are lemmatized and part-of-speech tagged with the TOSCA tagger [6]. TOSCA is a stochastic tagger, supplemented with a rule-based component which tries to correct observed systematic errors of the statistical components. TOSCA also gives each word form its lemma (basic form). For instance, word forms such as *takes*, *took*, *taken*, and *taking* have the same lemma *take*. This function facilitates the collocation analysis under the same lemma. TOSCA operates with a lexicon, which currently contains about 160,000 lemma-tag pairs, covering about 90,000 lemmas. The TOSCA-ICLE tagset contains 270 different tags within 18 major word classes. For simplicity, only the major word classes are considered in the current study (see Appendix A)

## 3. Experiments and Results

### 3.1 Corpus Perplexity in Bigram and Trigram models

Perplexity, in speech recognition community, is often referred to as the number of equi-probable choices at each step of word prediction in a language model such as a bigram/trigram model under the assumption that a word depends merely on the previous one/two words. In this work, given a corpus *L*, the perplexity of the corpus, *S(L)*, can be viewed as a measure of diversity for the next POS in a language model, and it is defined as:

$$S(L) = 2^{H(L)}$$

$$H(L) = \frac{1}{N}\sum_{c} H_c(k)$$

$$H_c(k) = -\sum_{k} P(k \mid c)\log_2 P(k \mid c)$$

where *H(L)* is the entropy of the corpus *L*, *N* is the size of part-of-speech set, and *P(k|c)* is the probability that *k* will be the next POS when the current POS is *c*.

.

In this experiment, the perplexities of BNC and TLCE corpora are calculated using both bigram and trigram models, and the results are shown in Table 1:

| | S(BNC) | S(TLCE) |
|---|---|---|
| Bigram | 4.91 | 4.36 |
| Trigram | 3.01 | 2.15 |

Table 1: Corpus perplexity

As can be seen in Table 1, the perplexities of BNC corpus in the two language models are both greater than those of TLCE, especially in the trigram model where the degree of POS diversity in the learner corpus is only 2/3 of BNC's. The above phenomena can be explained by the limiting sentence structure varieties the learners possess.

### 3.2 Structure Variety

In order to further understand the limit of structure variety in learners' writing, the numbers of POS trigrams, i.e. sequences of three POSs, used in the two corpora are compared and shown in Table 2. As seen in the table, there are 2531 trigram patterns in BNC, 1649 in TLCE, and 1574 in both. If those appearing in BNC can be viewed as the only eligible patterns for English, then the learners merely use 62% of correct trigram structures in their writing, and leave 38% in tact.

| BNC | TLCE | overlap |
|---|---|---|
| 2531 | 1649 | 1574 |

Table 2: the number of POS trigrams in the corpora

Under the same assumption, Figure 1 depicts the divergence of learners' use of trigrams from BNC, the optimum indicated by the square curve, on the scale of top-ranking trigrams in use. The diamond curve denotes the number of the learners' trigrams that overlap with BNC at the same rank. As illustrated, the learners' curve moves away from the optimum when the scope of the rank enlarges, especially after the rank of 1000.

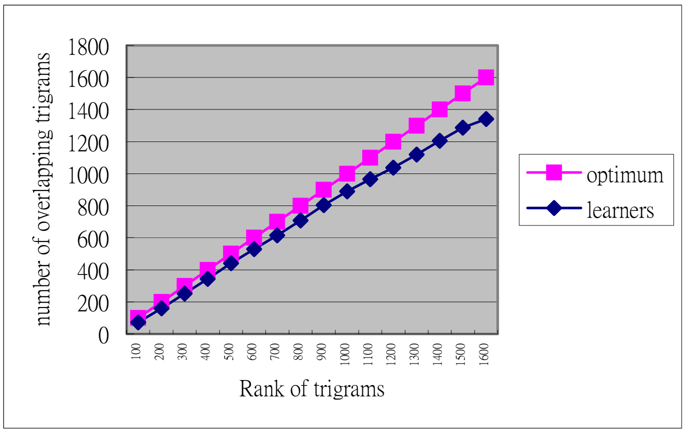

Figure 1: The divergence of the use of POS trigrams

### 3.3 POS Distribution

As the learners have preference in using certain POS trigrams it is then desirable to understand the learners' preference in using POSs themselves as well. Figure 2 shows the POS distribution in each corpus, and only those taking up at east 5% of the corpus are indicated. Two significant phenomena are observed from the figure. Firstly, although N(Noun) and VB(Verb) are the first two leading POSs in both corpora, there exists a distinct discrepancy of the percentage difference between the two. The difference in distribution percentage between N and VB in BNC reaches 9%, whereas merely 1% difference in TLCE. Secondly, PRON(pronoun), the 3rd highest distribution in the learner corpus but the 7th in BNC, apparently is overused the learners.

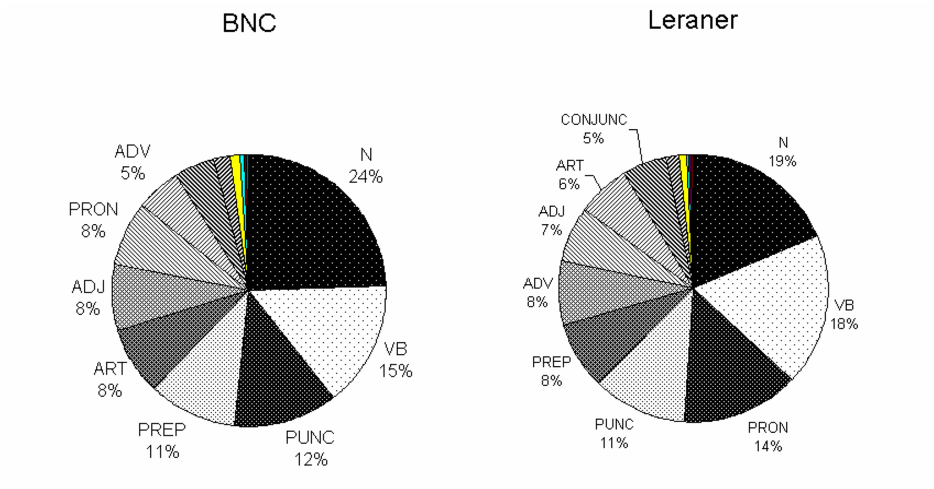

Figure 2: POS distribution

### 3.4 Distribution of Preceding POSs in PRON bigrams

As the previous figure indicates the excessive use of PRON in the learner corpus, the phenomenon is further analyzed by examining the likelihood of each POS preceding PRON in the bigrams. Figure 3 shows the distribution of preceding POSs of PRON in each corpus. As seen, PUNC(punctuation), VB and CONJ(conjuction) are the three most likely POSs in TLCE to be followed by PRON, and the learners also have stronger preference in using these bigrams than the native speakers. By contrast, the bigrams, PREP(preposition)+PRON and N+PRON, are used more frequently by the native speakers than the learners.

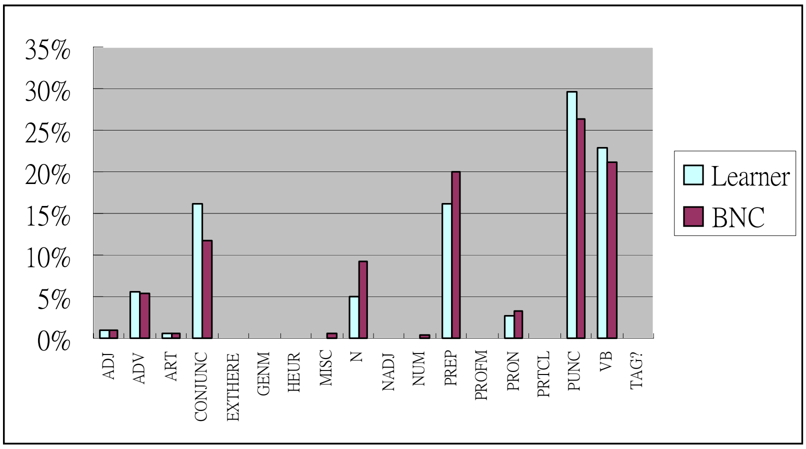

Figure 3: Distribution of Preceding POSs in PRON bigrams

## 4. Discussions and future work

The results of the preliminary experiments above show that there are more than one third of BNC trigrams that the learners never practice in their writing, whereas there are 4.5% of TLCE trigrams which do not appear in the BNC's. It is intended to believe that this small proportion of TLCE trigrams is contributed from the learner's writing errors. However, increasing the size of the native speaker corpus to observe any changes in the distribution of the trigrams will clarify the findings. It is also worth looking into those BNC trigrams that the learners do not know or are not aware of, and then isolating those with high frequency for the pedagogical purpose.

The experimental results also suggest that the learners use pronouns excessively in their writing and that they have stronger preference than native speakers in using pronouns right after punctuations, verbs and conjunctions but less preference after prepositions and nouns. Pronouns often appear in the informal register, and as the corpus is composed of college students' compositions as well as their weekly journals, the informality of the journals may contribute partly to their excessive use of pronouns. So, it is desirable in the next stage of the work to divide the learner corpus in terms of its different registers and compare their POS distributions with the native speaker corpus.

## Acknowledgements

The authors would like to thank the National Science Council, Taiwan, for supporting this project, and Prof. Ching-Yuan Tsai for his insightful comment.

## Appendix A

| Label | Major word class |
|---|---|
| ADJ | Adjective |
| ADV | Adverb |
| ART | Article |
| CONJUNC | Conjection |
| EXTHERE | Existential there |
| GENM | Genitive marker |
| HEUR | (unknown) |
| MISC | Miscellaneous |
| N | Noun |
| NADJ | Nominal adjective |
| NUM | Numeral |
| PREP | Preposition |
| PROFM | Proform |
| PRON | Pronoun |
| PRTCL | Particle |
| PUNC | Punctuation |
| TAG? | Word unable to tag |
| VB | verb |

## References

1. Granger, S., *The International Corpus of Learner English*, in *English Language Corpora: Design, Analysis and Exploitation*, J. Aarts, P.d. Haan, and N. Oostdijk, Editors. 1993, Rodopi: Amsterdam. p. 57-69.
2. Lorenz, G., *Overstatement in advanced learners' writing: stylistic aspects of adjective intensification*, in *Learner English on Computer*, S. Granger, Editor. 1998, Addison Wesley Longman Limited. p. 53-66.
3. Biber, D. and R. Reppen, *Comparing native and learner perspectives on English grammar: a study of complement clauses*, in *Learner English on Computer*, S. Granger, Editor. 1998, Addison Wesley Longman Limited. p. 145-158.
4. Tapper, M., *The use of adverbial connectors in advanced Swedish learners' written English*, in *Learner English on Computer*, S. Granger, Editor. 1998, Addison Wesley Longman Limited. p. 80-93.
5. Aston, G. and L. Burnard, *The BNC Handbook*. 1998: Edinburgh University Press.
6. Aarts, J., H. Barkema, and N. Oostdijk, *The TOSCA-ICLE Tagset Software and Tagging Manual*, . 1997, The Department of Language and Speech, University of Nijmegen, The Netherlands.