English Usage Comparison between Native and non-Native English Speakers in Academic Writing

Bei Yu *(beiyu@uiuc.edu)*
University of Illinois at Urbana-Champaign
Qiaozhu Mei *(qmei2@uiuc.edu)*
University of Illinois at Urbana-Champaign
Chengxiang Zhai *(czhai@cs.uiuc.edu)*
University of Illinois at Urbana-Champaign

Introduction

Discovering the differences in the language usage of native and non-native speakers contributes to contrastive linguistics research and second language acquisition. Unlike grammatical errors, the language usage differences are grammatically correct, but somehow do not conform to the native expressions. For example, if an English phrase is used popularly in a native English speaker group (G1) but not in a Chinese speaker group (G2), or vice versa, one possible reason may be first language (L1) impact on second language (L2). Such an impact may be due to grammar differences or even culture differences.

Isolating the language usage differences is never an easy task. There are generally two approaches to detect such differences: controlled experiments and corpus-based analysis. Many English as a Second Language (ESL) studies were conducted in a controlled environment, for example, through manually comparing the short in-class writing samples by a small number of college students in G1 and G2. The small data sets and the small subject groups undermined the result generalization because conflicting results were sometimes produced regarding the same language usage.

In contrast, corpus-based approaches (for example, computational stylistics) facilitate analysis of larger sample sets written by many subjects. Such approaches involve three major steps: 1) building a corpus of two comparable subsets; 2) automatically extracting a language usage feature set; 3) selecting a subset of the features that distinguishes G1 and G2. Given a corpus and a feature set, we can transform the above task into a text categorization problem. The goal is to categorize the texts by the authors' language background. Picking the subset is then transformed into the feature selection problem in text categorization.

Oakes used Chi-square test to find vocabulary subsets more typical of British English or American English. Oakes' work focused on identifying two feature categories: (1) the features common in G1 but not in G2; and (2) the features common in G2 but not in G1, while we hypothesize the existence of a third discriminative feature category (3): features common in both G1 and G2, but with different usage frequencies. All the other features are considered irrelevant. Distinguishing these different categories of features allows us to discover subtle differences in the language usage.

In this paper, we propose a simple approach of comparative feature analysis to compare the language usage between native English speakers and Chinese speakers in academic writing. We first select candidate features that are common in at least one group and then categorize them into the above three feature categories. Features in category (1) and (2) are ranked by their differences in document frequency, and features in category (3) are ranked by their difference indices as defined in the next few paragraphs.

We use two heuristic constraints to select a "good" candidate discriminative feature:

- Constraint (1): The feature should be common within at least one of the groups.
- Constraint (2): The feature should be contrastive across the groups.

We use the normalized document frequency (DF) to measure the feature commonality within a group. DF means the number of documents containing this feature. Denote DF_1 and DF_2 as the document frequencies in G1 and G2, respectively, and T=0.3 as a DF threshold. A feature falls into

- category (1) if (DF_1-DF_2)>=T;
- category (2) if (DF_2-DF_1)>=T;
- category (3) if |DF_2-DF_1|<T and (DF_1 >= T) and (DF_2 >= T).

Intuitively, the features in category (3) are those that are sufficiently popular in both G1 and G2 and have comparable DFs in G1 and G2.

After categorizing the features, we then define a Difference Index (DI) to measure the feature discriminating power and rank the features in category (3) by DI. We define TF as the number of occurrences of a feature in a document. Let m_1 and m_2 be the mean of the TF value of a feature in G1 and G2, respectively, we define DI as DI =sig(m_1 - m_2)* max(m_1,

m_2)/min(m_1, m_2). sig(m_1 - m_2) is 1 if m_1>m_2 and is -1 otherwise. A positive DI means the feature is more heavily used in G1 and a negative DI means it is more popular in G2. The larger the |DI| is, the bigger the difference is.

Corpus Construction

We propose the following fairness constraints for corpus construction: 1) the subjects in each group should have similar English proficiency; 2) the text genre and the topic should not interfere with the language usage analysis.

We collect two datasets that satisfy the constraints above. The first has 40 selected electronic theses and dissertations (ETD) from the ETD database at Virginia Tech., in which 20 are from Chinese students and 20 from American students, all from computer science and electronic and engineering departments to avoid genre interference. The second has 40 selected research articles downloaded from Microsoft Research (MSR), in which 20 are contributed by Chinese researchers in Beijing, China and 20 by their British colleagues in Cambridge, UK. The documents in ETD collection are long and each strictly attributed to one author, while the documents in MSR collection are much shorter and many are co-authored. We restrict the co-authors to the same language background. The biographies in theses and the resumes on MSR website help us identify the authors' first and second language.

Feature Extraction

We extracted plain text from the original pdf and ps files. We limit our search for features at the lexical level because syntactic parsing does not perform well in such a technical writing corpus due to formulas, tables and figure captions, etc. In order to avoid topic interference, we choose a feature set popular in computational stylistic analysis (Koppel et. al.), the n-gram (we set 0<n<4) common word sequences (CWS). A common word list consisting of 626 functional words and some common content words for technical writing (e.g. "problem") is used to generate CWS. For example, in "SW4 is among the few known wireless system tools for in-building network design.", the following 3-gram CWS features are extracted: "is among the", "among the few", "the few known".

Experiments and Results

We used the aforementioned procedure to analyze both ETD and MSR corpora. The results show that most "differences" found in one corpus do not repeat in the other one, but there are some "stable" features across the corpora. Figure 1 lists the 1-gram, 2-gram and 3-gram CWS features in all the three categories.

Category	1-gram CWS	2-gram CWS	3-gram CWS
1	fact, might, placed, appropriate, course, seem, indeed, unfortunately, mean, allow, particularly, yet, could, look, never	this would, until the, to an, we would, fact that, note that, these are, to give, the fact, such that, would be, with this, could be, given the, along with, as to, and so, for which, which the, and not, to allow, it to, the appropriate, not the, a particular, can then, the form, then be, if a , of which, this provides, that all, where there	the fact that, can then be, this is a, to give a
2	according		
3 (positive)	specifying, being, would, itself, must, able, produce, useful, rather, question, were, allows, particular	a way, in which, form of , a given, must be, the use, in any, use of, a more, rather than	in which the, the use of
3 (negative)	novel, respectively, build	of different, show that, we may, and more, according to, and it, than that, we use, see that, is still, or the, is very, but also	used in the, show that the

Figure 1: 1-gram, 2-gram and 3-gram CWS in category 1, 2, and 3.

- *G1: native English speakers (British/American); G2: Chinese speakers.*
- *Category 1: CWS common in G1 but not G2.*
- *Category 2: CWS common in G2 but not G1.*
- *Category 3: CWS common in both G1 and G2 but the frequencies differ.*
- *Category 3 (positive): CWS with mean frequency in G1 at least twice as that in G2.*
- *Category 3 (negative): CWS with mean frequency in G2 at least twice as that in G1.*

An example in category (1) is the word "never". It appears in 75% American students theses in ETD but just in 20% Chinese student theses, and its frequency mean in the American group is six times as that in the Chinese group. Similarly, "never" appears in 45% MSR papers from UK, but just in 10% MSR papers from China, and British researchers use it five times more often than their Chinese colleagues.

Category (2) is surprisingly small with only one stable feature "according".

Category (3) includes features common in both groups. It has a positive subset consisting of features more heavily used in the

British/American groups, and a negative subset consisting of features more popular in the Chinese groups. Examples in the positive subset are "specifying", "must", "were", "rather than", "the use of", etc. Examples in the negative subset include "novel", "respectively", "build", "show that", "according to", "used in the", etc.

We noticed that some feature groups are worth further study, such as the negation words, modals, personal pronouns, and parallelism indicators as listed in figure 2.

language usage aspects	representative features
negation	"nothing", "cannot", "not", "none", "no", "nobody", "nothing", "nowhere", "neither", "never"
modal	"can", "could", "may", "might", "will", "would", "shall", "should"
personal pronoun	"I", "my", "me", "you", "your", "us", "our", "we"
parallel	"and", "or", "nor", "but"

Figure 2: special groups of interesting features

As shown in figure 3, the native English speakers always use more negation words than the Chinese. It is probably due to culture difference rather than grammar impact.

Negation	ETD	MSR
cannot	1.53	MAX
not	1.77	1.09
never	MAX	MAX
is not	1.42	1.16
do not	1.88	MAX
not the	MAX	MAX
and not	MAX	MAX

Figure 3: comparison of negation words usage

Note: In figure 3, 4, and 5, a number (positive or negative) in a cell means the corresponding feature belongs to category (3) and the number is the feature's difference index (DI) value as defined in the paper. "MAX" instead of a number in a cell means this feature belongs to category (1) and thus it does not have a DI value. Similarly, "-MAX" means the feature belongs to category (2).

As shown in figure 4, the native speakers also use more modals such as "might", "would" and "could". The Chinese use "will" more often. There is no big difference between their uses of "can".

Modal	ETD	MSR
will	-1.41	-1.31
it will	-1.97	-1.50
we will	-1.72	-2.62
will be	-2.16	-1.61
we may	-4.04	-2.21
can	-1.08	-1.00
we can	-1.08	-1.16
which can	-1.24	-1.11
this can be	1.10	MAX
can be used	1.39	MAX
can be used to	1.60	MAX
can then	MAX	MAX
can then be	MAX	MAX
might	MAX	MAX
could	MAX	MAX
would	5.60	MAX
this would	MAX	MAX
we would	MAX	MAX
would be	MAX	MAX
could be	MAX	MAX

Figure 4: Comparison of Modal Usage

As shown in figure 5, "us", "our" and "we" are the three mostly used personal pronouns for both groups, but the native speakers use "us" more often while the Chinese use "our" and "we" more often.

Personal Pronoun	ETD	MSR
us	1.55	1.61
we would	MAX	MAX
we are	1.29	MAX
that we	1.16	1.10
our	-1.05	-1.87
we	-1.16	-1.13
if we	-1.44	-1.36
we may	-4.04	-2.21
we can	-1.08	-1.16
we use	-3.19	-MAX
in our	-1.81	-2.96
we also	-1.04	-2.29
we will	-1.72	-2.62
we use the	-1.86	-MAX

Figure 5: Comparison of personal pronoun usage

"And" and "but" are two parallel structure indicators commonly used by both groups. The Chinese group use "and" slightly more than the American/British group, but they use "but" almost twice as less than the native English speakers.

Conclusion and Future Work

We use a simple comparative feature analysis method to compare the differences in the English common word usage between the native British/American English speakers and the Chinese speakers in their academic writing. The proposed method helps us find some interesting or even surprising differences between these two groups. We also see that common words are a very limited feature set. We shall explore more meaningful linguistic features to find more useful differences.

Bibliography

Koppel, M., S. Argamon, and A.R. Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Literary and Linguistic Computing* 17.4 (2003): 401-412.

Oakes, M. "Text Categorization: Automatic Discrimination between US and UK English using the Chi-square Text and High Ratio Pairs." *Research in Language* 1 (2003): 143-156.