Lexical Development in L2 English Learners' Speeches and Writings

Ishikawa, Shin'ichiro

(Citation)
Procedia - Social and Behavioral Science,198:202-210

(Issue Date)
2015-07-24

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90003297

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 202 – 210

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Lexical Development in L2 English Learners' Speeches and Writings

Shin'ichiro Ishikawa*

Kobe University, 1-2-1, Tsurukabuto, Nada-ku, Kobe, 657-8501, Japan

Abstract

In order to investigate how learners' L2 vocabulary develop in speeches and writings, the current study analyzed a new learner corpus including speeches and writings of varied Asian learners of English. The analyses revealed that as learners' L2 proficiency levels increased, lexical diversity decreased and then increased, lexical density remained unchanged, and lexical complexity steadily increased. Furthermore, lexical fundamentality increased and decreased in speeches and writings respectively, and the degree of noun orientation slightly decreased. It was also shown that learners' spoken and written vocabularies developed largely in a different way. Additional statistical analysis revealed that learners' L2 vocabulary use was influenced the least by L2 proficiency, and the most by L2 production mode, followed by nationality and L1.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Learner corpus; L2 production modes; L2 proficiency levels; lexical indices

1. Introduction

It is empirically known that novice and advanced learners use L2 vocabulary differently. However, it remains unclear how precisely L2 vocabulary develops as learners' overall L2 proficiency increases, how spoken and written vocabulary development may differ, and whether L2 learners exhibit a universal pattern in their lexical development. This is partly due to the limited availability of comparable written and spoken data from international learners of

* Corresponding author. Tel.: +81-78-881-1212.
E-mail address: iskwshin@gmail.com

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.437

differing L2 proficiencies. This study utilizes a new international learner corpus to quantitatively examine how the lexical diversity (LexDiv), lexical density (LexDen), lexical complexity (LexCom), lexical fundamentality (LexFun), and noun-orientation (NOr) of Chinese (CHN), Indonesian (IDN), Japanese (JPN), and Taiwanese (TWN) learners' speeches and writings change in proportion to their L2 proficiencies based on the Common European Framework of Reference (CEFR) proficiency band. Furthermore, English native speakers are used as a reference for comparison.

2. Literature Review

It is known that an individual's vocabulary develops according to his or her intellectual growth. Concerning the development of L1 vocabulary, Johansson (2008) examined the lexical diversity/density of spoken/written narratives, speeches (spoken expository), and essays (written expository) from monolingual Swedish speakers of differing ages. Her analysis revealed that while age does contribute to increased lexicon, such patterns are not necessarily evident unless long-term development is also considered. Moreover, these developmental trends are more easily identifiable by examining lexical diversity rather than lexical density. As for the development of L2 vocabulary, Meara and Milton's (2003) analysis of Greek and Hungarian EFL learners revealed that the receptive vocabulary sizes for learners at the B1, B2, C1, and C2 levels of the CEFR were 2500, 3250, 3750, and 4500 words respectively.

Vocabularies also change according to production modes (i.e., speeches and writings). Generally, spoken vocabularies are comparatively limited in size (Montgomery, 2007), and make use of more familiar, everyday words (Turk, 1985). Summarizing related studies, Schallert, Kleiman, and Rubin (1977) concluded that speeches are characterized by greater repetition, redundancy, and inefficient vocabulary use (i.e., the use of many words to convey an idea); in contrast, writings characteristically contain more diverse vocabulary (e.g., longer and less common words) and greater syntactic complexity. Crystal (1985) further noted that speech (which is time-bound, spontaneous, prosodic, phatic, ongoing, and usually face-to-face) is typified by the use of slang and longer coordinate sentences, in addition to grammatical informality. Comparatively, writing is space-bound, carefully organized, less context-dependent, and characterized by the occurrence of more elaborate syntactic structures, in addition to fewer errors and inadequacies. The differences between spoken and written vocabularies are not well defined, and therefore open to interpretation (Schallert, Kleiman, and Rubin, 1977). Consequently, any apparently clear-cut distinctions between them should be scrutinized, particularly in a modern electronic age, wherein such distinctions are becoming increasingly blurred (Crystal, 1985). As such, how can various types of vocabularies be reliably compared?

Here we briefly examine the five types of lexical measures adopted by this study. Lexical diversity (i.e., lexical variety/richness) is the first, and it is determined by dividing the number of types (i.e., different words) by the number of tokens (i.e., all words). This measure is generally used as an index of lexical sophistication, although it can also function as an index for weaker cohesion, since higher lexical diversity results in decrease of repetition evoking shared memories (Johnstone, 1987). The type/token ratio (TTR) is the simplest benchmark of lexical diversity, but since it is highly sensitive to token quantities, many modifications to it have been proposed, such as Guiraud's index, Herdan's C, and Malvern and Richard's (1997) D. Guiraud's index involves dividing the number of types by the square root of the number of tokens, while Herdan's C is determined by dividing the natural log of the number of types by the natural log of the number of tokens. Malvern and Richard's D is calculated by choosing a set of words from texts at random, and adjusting the raw TTR by referencing the best fitting curve (Šišková, 2012).

Lexical density is the second measure used in this study, which is based on the ratio of lexical and open-class words (i.e., content words), such as nouns, verbs, adjectives, and occasionally parts of adverbs, to the whole vocabulary. This measure can be interpreted in two different ways: an index of information-orientation and lexical easiness/immaturity. Generally, it is regarded as an index of the amount of information condensed into a text, since texts with a greater number of content words and fewer functional/grammatical words naturally include more information. Ure (1971) asserts that lexical density is less than and higher than 40% for speeches and writings respectively. Similarly, Eggins (1994) conducted a corpus analysis and determined that lexical density was 33% and

42% in speeches and writings respectively. Meanwhile, Halliday (1985) noted that children's language initially consists primarily of lexical words, and that grammatical words begin to appear in greater abundance later. Thus, lexical density is sometimes used as an index of lexical easiness and immaturity. In their comparison of English textbooks of varying difficulties and their lexical densities, To, Fan, and Thomas (2013) found that density consistently decreased from the pre-intermediate to intermediate and upper-intermediate levels, although it was not necessarily highest at the elementary level. These findings support the notion that lexical density can be indicative of easiness.

The third measure employed in this study is lexical complexity, which is determined by word length (i.e., the average number of letters in a word), and it is an index of a vocabulary's morphological complexity. That is, lexical complexity often leads to lexical difficulty, since longer words (e.g., *sophistication* and *comparability*, which are 14 and 13 letters respectively) are generally more difficult to use and understand than shorter words (e.g., *I* and *in*, which are 1 and 2 letters respectively).

Lexical fundamentality is the fourth measure utilized in this study, and is established based on the ratio of high-frequency fundamental words to the whole vocabulary. It is generally used as an index of dependence on basic vocabulary, and consequently non-native-learnerness as well. Hasselgren (1994) maintains that L2 learners, unlike English native speakers, often rely on a relatively limited set of familiar and comfortable words in order to convey various ideas. High-frequency fundamental words can be identified in many different ways, such as by extracting the top 1000-3000 words from existing word lists, or by selecting them directly from corpora. Nation (2012) chose the top 2000 words from a one-million-word corpus composed of speeches and essays in American and British English. The Range Software package calculates the ratios of the top 1000, 2000, and 3000 words from several sources.

The fifth measure employed by this study is noun orientation or nouniness (Biber, Conrad, and Reppen, 1982), which is usually determined by dividing the number of nouns by the number of verbs. The noun/verb ratio can be interpreted in different ways. In contrastive linguistics it is essentially an index of a particular language's syntactic uniqueness. For example, it has been established that the noun/verb ratio in some languages (e.g., Chintang) is approximately 1:0, although the ratio reaches nearly 2:0 in other languages (e.g., Sri Lankan Malay) (Seifart, Meyer, Zakharko, Bickel. Danielsen, Nordhoff *et al*., 2010). In studies concerning L1/L2 acquisition, the ratio is frequently used as an index of lexical immaturity, a feature characteristic of the language written and spoken by young children (Gentner, 1982) or non-native learners (Bates, Dale, and Thal, 1995), as it is known that children learn nouns earlier than verbs. In textual genre studies, the ratio serves as an index of the formality characterizing written texts. In their analysis of English corpora, Biber, Conrad, and Reppen (1984) found a noun/verb ratio of 2:2–2:5, 1:2–1:5, and 1:2–1:6 in academic prose, fiction, and speeches respectively. In essay writing the ratio functions as an index of dynamic and effective description. Indeed, many style guides advise writers to maintain a high verb/noun ratio in order to imbue their language with "a sense of vigor by eliminating unnecessary nouns and choosing powerful verbs" (Moxley, n.d.).

3. Research Design

3.1. Aim and Research Questions

The present study aims to quantitatively investigate lexical development in learners' L2 speeches and writings, and in doing so poses two research questions: How do five kinds of lexical aspects change in learners' speeches and writings in relation to increased L2 proficiency? (RQ1) and How is learners' overall lexical development influenced by L2 production modes, nationality, and L2 proficiency? (RQ2) Our comparison of speeches and writings will be limited to discussing their non-interactive forms. Although it is generally accepted that speeches and writings are interactive and non-interactive respectively, Crystal (1985) emphasizes that this is not always the case. For example, non-interactive speech might entail speaking to an answering machine; similarly, interactive writing could entail an electronic exchange of messages via e-mail or SMS. Concerning whether it is more appropriate to discuss linguistic

production modes (i.e., speeches/writings) and interactivity (i.e., interactive/non-interactive) separately, the current study primarily adopts the former approach.

3.2. Data and Methodology

The spoken (Baby Version 1.3) and written module (Version 2.1) of the International Corpus Network of Asian Learners of English (ICNALE) was used by this study (cf. Ishikawa, 2013, 2014). The ICNALE includes 60-second transcripts of impromptu speeches, and 200-300 word essays produced by college students from ten different Asian countries and territories. The prompts are commonplace ("It is important for college students to have a part-time-job" and "Smoking should be completely banned at all the restaurants in the country"), and the speaking/writing conditions are strictly controlled, which leads to greater reliability for use in contrastive studies. The ICNALE also contains detailed participant background data, including their age, gender, grade, major, L2 proficiency, learning motivation, and L2 learning history. Based on the scores of receptive vocabulary size tests and/or standard English proficiency tests (e.g., TOEFL, TOEIC, and IELTS), participants were classified into A2, B11, B12, and B2+ levels, in accordance with the CEFR proficiency band. An outline of the analyzed data is provided in Table 1.

Table 1. Outline of the data used for analysis

Country	Proficiency	Speeches			Writings		
		Ppt	Data	Token	Ppt	Data	Token
CHN	A2	14	56	34450	50	100	123768
	B11	48	192	117291	232	464	604695
	B12	78	312	196023	105	210	287540
	B2	10	40	24071	13	26	36935
IDN	A2	26	104	60459	32	64	83729
	B11	37	148	80187	82	164	208404
	B12	34	136	81200	83	166	221620
	B2	3	12	7358	3	6	8558
JPN	A2	14	56	18129	154	308	370714
	B11	34	136	51534	179	358	433018
	B12	27	108	40779	49	98	121325
	B2	25	100	44637	18	36	46452
TWN	A2	17	68	30701	29	58	69161
	B11	41	164	82317	87	174	217984
	B12	25	100	50383	61	122	157013
	B2	17	68	37049	23	46	61335
ENS	NA	75	300	249039	200	400	497533

Note: Ppt = number of participants. Each participant was required to produce four speeches (two about each prompt) and two essays about two different prompts. There were no overlaps between participants from the spoken and written modules.

To measure aspects of learners' L2 vocabularies, lexical diversity was calculated according to Herdan's C. Lexical density was determined according to the ratio of nouns, verbs, and adjectives to the whole vocabulary. Lexical fundamentality was established based on the ratio of the top 1000 most frequently occurring words, as identified by Nation (2012), to the whole vocabulary. Lexical complexity was calculated according to the number of letters per word, while noun orientation was determined based on the noun/verb ratio. The counting of nouns, verbs, and adjectives was based on the POS-tags supplied by Tree-Tagger. Nouns included proper use, verbs included grammatical use (e.g., *have* to), and adjective included all base, comparative, and superlative forms.

Concerning RQ1, three points are primarily addressed, namely the influence of (i) L2 production modes (i.e., speeches vs. writings), (ii) non-nativeness (i.e., native speakers vs. non-native learners), and (iii) L2 proficiency levels (i.e., A2, B11, B12, B2+) on vocabulary use. To answer RQ2, a principal component analysis (PCA) was performed, which is a statistical technique for transforming multiple variables into fewer principal components (e.g., PCA1, PCA2). The largest contributor, PCA1, was generally interpreted as a new integrative variable. The variables include the values of five different kinds of lexical indices, and the cases include thirty-two production data (4

learner nationalities, multiplied by 4 proficiency levels, multiplied by 2 production modes). After reordering a series of cases based on their PCA1 scores, we examined which contributed to their classification.

4. Results and Discussion

4.1. RQ1: Development of Learners' L2 Vocabulary

The results are shown in Table 2. When discussing learners in general, we pay attention to the average values summarizing vocabulary uses of four groups of learners at four different L2 proficiency levels, which are shown in italics.

Table 2. The development of five types of lexical aspects

Indices	L2Prof	Speech						Writings					
		CHN	IDN	JPN	TWN	Av	ENS	CHN	IDN	JPN	TWN	Av	ENS
LexDiv	A2	0.75	0.74	0.73	0.73	0.74	0.73	0.75	0.78	0.72	0.77	0.76	0.75
	B11	0.73	0.73	0.71	0.71	0.72		0.72	0.74	0.72	0.75	0.73	
	B12	0.72	0.73	0.73	0.73	0.73		0.75	0.75	0.74	0.77	0.75	
	B2	0.76	0.79	0.73	0.75	0.76		0.81	0.85	0.77	0.79	0.80	
	Av	0.74	0.75	0.72	0.73	*0.73*		0.76	0.78	0.74	0.77	*0.76*	
LexDen	A2	42.50	45.50	51.60	47.00	46.65	47.50	50.50	53.10	50.80	51.10	51.38	48.90
	B11	46.00	46.70	49.60	46.70	47.25		50.30	51.60	50.80	51.00	50.93	
	B12	45.30	45.20	49.30	46.00	46.45		50.50	52.10	50.20	50.10	50.73	
	B2	45.50	46.00	48.70	45.30	46.38		51.40	50.40	49.80	49.80	50.35	
	Av	44.83	45.85	49.80	46.25	*46.68*		50.68	51.80	50.40	50.50	*50.84*	
LexCom	A2	3.85	4.27	4.14	4.14	4.10	4.29	4.42	4.56	4.33	4.28	4.40	4.44
	B11	4.06	4.32	4.19	4.13	4.18		4.41	4.48	4.37	4.39	4.41	
	B12	4.11	4.21	4.16	4.18	4.17		4.46	4.52	4.36	4.44	4.45	
	B2	4.16	4.35	4.12	4.20	4.21		4.56	4.48	4.38	4.50	4.48	
	Av	4.05	4.29	4.15	4.16	*4.16*		4.46	4.51	4.36	4.40	*4.43*	
LexFun	A2	65.40	82.44	79.17	84.79	77.95	86.17	87.77	85.64	88.27	89.20	87.72	85.84
	B11	72.30	82.65	76.66	84.36	78.99		86.77	87.21	88.47	88.38	87.71	
	B12	72.89	82.09	79.94	81.99	79.23		83.84	86.84	88.87	86.85	86.60	
	B2	81.63	85.34	85.59	83.83	84.10		83.84	85.83	88.60	86.14	86.10	
	Av	73.06	83.13	80.34	83.74	*80.07*		85.56	86.38	88.55	87.64	*87.03*	
NOr	A2	108.90	119.60	118.70	93.40	110.15	91.60	127.30	167.90	119.00	119.40	133.40	109.60
	B11	114.10	117.50	110.50	96.20	109.58		125.90	154.20	121.40	119.90	130.35	
	B12	118.40	115.60	115.80	94.00	110.95		128.00	148.60	117.90	117.00	127.88	
	B2	91.00	119.10	106.10	96.90	103.28		135.50	146.40	118.00	118.60	129.63	
	Av	108.10	117.95	112.78	95.13	*108.49*		129.18	154.28	119.08	118.73	*130.31*	

First, lexical diversity is generally expected to be (1) higher in writings than in speeches, (2) lower for learners than for English native speakers, and (3) higher for upper-level learners than for lower-level learners. Our analysis supported the first expectation and partially supported the third, although the second was not supported.

Lexical diversity tended to be higher in writings than in speeches for both learners (0.73/0.76) and English native speakers (0.73/0.75), suggesting that a greater range of vocabulary was used in writings even when discussing the same topic. Furthermore, compared to English native speakers, learners generally exhibited the same or somewhat higher levels of lexical diversity in speeches (0.73/0.73) and writings (0.76/0.75) respectively. Specifically, B2+ upper-intermediate learners almost always used a greater range of vocabulary (0.73 to 0.85), thus indicating that lexical diversity was not chiefly a characteristic of English native speakers. Also, lexical diversity tended to change as learners' L2 proficiency levels ascended, wherein it decreased between A2 and B11 and increased between B11 and B2+ for both speeches (0.74—0.72—0.73—0.76) and writings (0.76—0.73—0.75—0.80). Although the development of learners' lexical diversity was not necessarily linear, the overall trend was for upper-level learners to use a greater range of vocabulary in their speeches and writings.

Second, lexical density, if interpreted as an index of information orientation, is also expected to be (1) higher in writings, (2) lower for learners, and (3) higher for upper-level learners. Our analysis supported the first expectation, although the third was not supported. The second expectation was supported partially, at least in relation to speech production.

Lexical density tended to be higher in writings than in speeches for both learners (46.68/50.84) and English native speakers (47.50/48.90). This suggests that writings were more information-oriented than speeches, even when identical topics were discussed, though the 40% threshold hypothesis (Ure, 1971) was not supported. When compared to English native speakers, learners exhibited lower and higher lexical density in speeches (46.68/47.50) and writings (50.84/48.90) respectively, thus indicating that a qualitative discrepancy existed between learners' spoken and written vocabularies. It is inappropriate to assume that the language of the learners was always less information-oriented, since they were capable of condensing a sufficient amount of information into writings when given adequate time to do so. Furthermore, lexical density scarcely changed even as learners' L2 proficiency levels increased, both in speeches (46.65—47.25—46.45—46.38) and writings (51.38—50.93—50.73—50.35). Unlike lexical diversity, lexical density did not necessarily develop steadily in relation to increased L2 proficiency.

Third, lexical complexity, like the two previous indices, is generally expected to be (1) higher in writings, (2) lower for learners, and (3) higher for upper-level learners. Our analysis supported all these expectations.

Lexical complexity tended to be higher in the writings than in speeches of both learners (4.16/4.43) and English native speakers (4.29/4.44), thus suggesting that longer and more morphologically complex words were used with greater frequency in writings even when identical topics were discussed. When compared to English native speakers, learners generally exhibited somewhat lower lexical complexity in speeches (4.16/4.29) and in writings (4.43/4.44), although B12 and B2+ upper-intermediate learners used more complex words than their native-speaking counterparts in writings (4.45 to 4.48/4.44). Hence, while lexical complexity may be a distinguishing factor between the speeches of English native speakers and learners, this may not necessarily be true for their writings. Moreover, lexical complexity tended to increase as learners' L2 proficiency levels ascended, both for their speeches (4.10—4.18—4.17—4.21) and writings (4.40—4.41—4.45—4.48). Although this increase was not necessarily constant, it could be argued that upper-level learners used more morphologically complex words in their speeches and writings.

Forth, lexical fundamentality is generally expected to be (1) lower in writings, (2) higher for learners, and (3) lower for upper-level learners. Our analysis supported the first assumption only among English native speakers, and the second and third assumptions for writings alone. A need to control the occurrence of fillers in speeches was also demonstrated.

Lexical fundamentality tended to be higher in speeches than in writings for English native speakers (86.17/85.84), although it was intriguingly lower among learners (80.07/87.03). For Chinese learners in particular the gap between production modes was quite large (73.06/85.56). This seemingly strange tendency was presumably attributable to learners' frequent use of fillers (e.g., *uh*, *er*, and *um*), as in the following excerpt from a speech delivered by an A2 Chinese learner.

> College student is a – college students are the, uh, people who, uh, um, transform from, um, transform the – from, uh, high school, in a high school to, um, uh, from – from school to, uh, uh, a society. So, uh, I think that a part-time job is very, uh, is a very, uh, is a necessity for our college students.... (CHN_PTJ_140)

Since many of these fillers are not included in the top 1000 word list, lexical fundamentality in learners' speeches was consequently low. When compared to English native speakers, learners generally exhibited higher lexical fundamentality in writings (87.03/85.84), but lower lexical fundamentality in speeches (80.07/86.17), a disparity that is attributable to learners' frequent use of fillers. Furthermore, a relationship was evident between learners' L2 proficiencies and lexical fundamentality in which the latter increased in speeches (77.95—78.99—79.23—84.10) and consistently decreased in writings (87.72—87.71—86.60—86.10). Learners' increasing and decreasing lexical

fundamentality in speeches and writings reflect their decreasing use of fillers in speeches and their less dependency on a limited range of high frequency words in writings.

Finally, noun orientation as an index of formality/writtenness (Biber, Conrad, and Reppen, 1984) is generally expected to be (1) higher in writings, (2) lower for learners, and (3) higher for upper-level learners. Our analysis supported the first expectation, although the second and the third were not supported.

The noun/verb ratio tended to be higher for the writings of both learners (108.49/130.31) and English native speakers (91.60/109.60), hence suggesting that more nouns and fewer verbs were used in their writings, despite identical topics being discussed. Since this trend was observed among learners and English native speakers alike, it can be deduced that noun orientation does reflect not lexical immaturity (Gentner, 1982) but formality/writtenness or dynamism. Furthermore, native speakers used verbs more frequently than nouns in speeches (91.60), and nouns more frequently than verbs in writings (109.6). This contrast between speech and writing in terms of noun/verb orientation was a clear characteristic of English native speakers. When compared to English native speakers, learners generally exhibited a significantly higher noun/verb ratio for both their speeches (108.49/91.60) and writings (130.31/109.60). Indeed, a greater written and spoken noun orientation was characteristic of learners in general. In addition, the noun/verb ratio gradually decreased in relation to increased proficiency for both speeches (110.15—109.58—110.95—103.28) and writings (133.40—130.35—127.88—129.63). Thus, it seems feasible for one to distinguish between B2+ upper-intermediate and novice/ intermediate learners based on noun orientation, which suggests that upper-intermediate learners are relatively closer to English native speakers in their use of nouns and verbs.

4.2. RQ2: Influence of L2 Production Mode, Nationality, and L2 Proficiency On Learners' Overall Lexical Development

Although several factors were extracted, the sole factor whose eigenvalue exceeded 1.0 was PCA1. The PCA1 loadings for lexical diversity, lexical density, lexical complexity, lexical fundamentality, and noun orientation were 0.57, 0.86, 0.95, 0.75, and 0.79 respectively. Since all loads were positive, PCA1 (whose contribution reached 63.12%) can be interpreted as a new variable summarizing the five types of lexical indices. Therefore, thirty-two L2 production data were arranged in ascending order according to their PCA1 scores (Figure 1.).

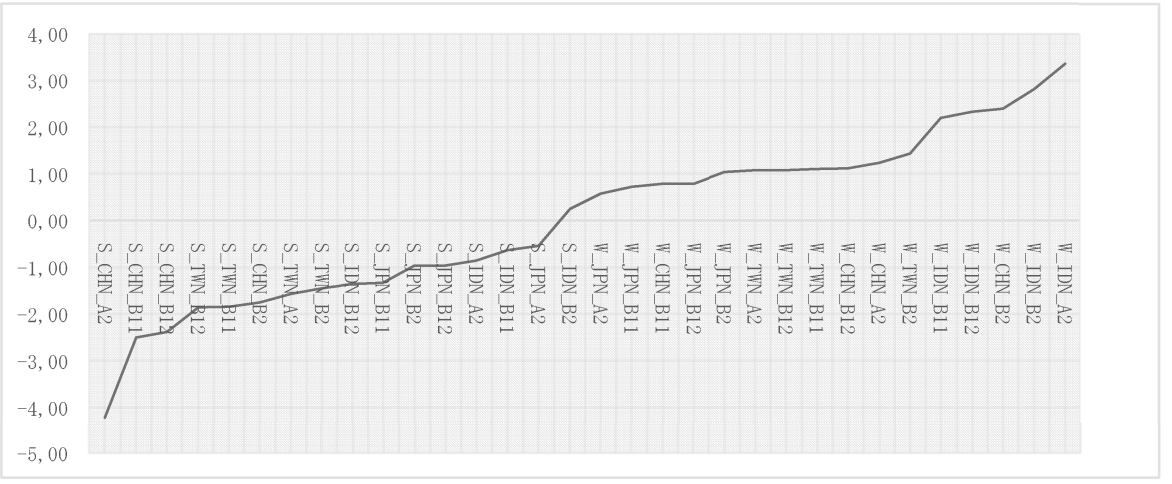

Fig. 1 Learners' spoken and written production data organized in ascending order according to PCA1 scores

Several findings are noteworthy here. First, as data samples with identical L2 production mode clustered together at the left and right halves of the graph respectively, it was proven that L2 production mode was the most significant factor in determining learners' L2 vocabulary use. This indicates that there exists a substantial gap in developments of learners' spoken and written vocabularies.

Second, the learners' nationalities influenced their L2 vocabularies, which is highlighted by the clear tendency for learners of identical nationalities to cluster together. For example, Indonesian learners' writings (W_IDN_A2, W_IDN_B2, W_IDN_B12, and W_IDN_B11) and Chinese learners' speeches (S_CHN_B2, S_CHN_B12, S_CHN_B11, and S_CHN_A2) cluster together on the left and right side of the graph. This demonstrates that learners of identical nationalities generally use vocabulary in a comparable fashion. Furthermore, clusters of Taiwanese and Chinese learners appear in close proximity for both speeches and writings, suggesting that learners of identical L1 backgrounds, even if they learn in different social contexts, might use L2 vocabulary similarly.

Finally, learners' L2 proficiencies, unlike L2 production mode or nationality, seemed to have a relatively limited effect on L2 vocabulary use. Data from learners with identical L2 proficiencies were not located in close proximity. For example, A2 learners' speeches (S_CHN_A2, S_IDN_A2, S_JPN_A2, and S_TWN_A2) appear in relative isolation.

Nevertheless, in some cases the effect of L2 proficiency can be observed more clearly as shown in Table 3.

Table 3. Four L2 proficiency levels in ascending order according to PCA1 scores

S_CHN		S_IDN		S_JPN		S_TWN		W_CHN		W_IDN		W_JPN		W_TWN	
A2	-4.22	B12	-1.36	B11	-1.33	B12	-1.86	B11	0.78	B11	2.19	A2	0.57	A2	1.08
B11	-2.50	A2	-0.86	B2	-0.97	B11	-1.85	B12	1.11	B12	2.32	B11	0.72	B12	1.08
B12	-2.38	B11	-0.63	B12	-0.96	A2	-1.57	A2	1.23	B2	2.82	B12	0.79	B11	1.10
B2	-1.75	B2	0.24	A2	-0.55	B2	-1.45	B2	2.39	A2	3.35	B2	1.04	B2	1.44

It is true that the positioning of proficiency levels is generally haphazard, such as for B12—A2—B11—B2 (S_IDN) and B11—B12—A2—B2 (W_CHN), but the order of speeches and writings of Chinese and Japanese learners seemed to reflect the order of proficiency development. In other words, the spoken and written vocabularies of Chinese and Japanese learners developed respectively steadily in unison with the development of L2 proficiency.

5. Conclusion

Using a large international learner corpus, this study examined lexical development in the speeches and writings of L2 learners. Our analyses revealed that as learners' L2 proficiency levels increased, lexical diversity decreased and then increased, lexical density remained unchanged, and lexical complexity steadily increased. Furthermore, lexical fundamentality, which was prone to the effect of filler use, increased and decreased in speeches and writings respectively, and the degree of noun orientation slightly decreased. It was also shown that learners' spoken and written vocabularies developed largely in a different way. Additional statistical analysis revealed that learners' L2 vocabulary use was influenced the least by L2 proficiency, and the most by L2 production mode, followed by nationality and L1.

Although this corpus-based study revealed many interesting facts concerning the development of L2 learners' spoken and written vocabularies, further research is nevertheless required to reach a generalizable conclusion. In a future study, we intend to expand the data range to be analyzed and to examine the replicability of the present study's findings.

References

Bates, E., Dale, P. S., and Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher, and B. MacWhinney (Eds), *Handbook of child language* (pp. 96-151). Oxford, UK: Basil Blackwell.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Crystal, D. (1985). Speaking of writing and writing of speaking. *Longman Language Review*, *1*, 5 - 8.

Eggins, S. (1994). *An introduction to systemic functional linguistics*. London, UK: Pinter Publishers.

Gentner, D. (1982). *Technical report No. 257: Why nouns are learned before verbs: Linguistic relativity versus natural partitioning.* Illinois, IL: Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Halliday, M. A. K. (1985). *Spoken and written language*. Oxford, UK: Oxford University Press.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*, 237 – 258.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian Learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91-118). Kobe, Japan: Kobe University.

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2* (pp. 63-76). Kobe, Japan: Kobe University.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers* (Department of Linguistics and Phonetics, Lund University), *53*, 61 - 79.

Johnstone, B. (1987). Perspectives on repetition: An introduction. *Text*, *7*, 205 - 214.

Malvern, D. D., and Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan, and A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon, UK: Multilingual Matters.

Meara, P., and Milton, J. (2003). *X_lex, The Swansea levels test*. Newbury, UK: Express.

Montgomery, M. (2007). *The discourse of broadcast news*. London, UK: Routledge.

Moxley, J. (n.d.). Maintain a high verb-to-noun ratio. Retrieved March 20, 2015 from http://www.collegewriting.org/

Nation, P. (2012). The BNC/COCA word family lists. Document bundled with Range program with BNC/COCA lists 25,000.

Schallert, D. L., Kleiman, G. M., and Rubin, A. D. (1977). *Technical Report No. 29: Analyses of differences between written and oral language*. Illinois, IL: Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Seifart, F., Meyer, R., Zakharko, T., Bickel. B., Danielsen, S., Nordhoff, S., and Witzlack-Makarevich, A. (2010). Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis. Paper presented at the DobeS Workshop "Advances in Documentary Linguistics" Nijmegen, 14-15 October 2010.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *Language Studies Working Papers*. *4*, 26 - 36.

To, V., Fan, S., and Thomas, D. (2013). Lexical density and readability: A case study of English textbooks. *Internet Journal of Language, Culture and Society*, *37*, 61 - 71.

Turk, C. (1985). *Effective speaking: Communicating in speech*. London, UK: Chapman & Hall.

Ure, J (1971). Lexical density and register differentiation. In G. Perren, and J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 443-452). Cambridge, UK: Cambridge University Press.