ReCALL **27**(3): 304–320. 2014 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344014000408
First published online 30 December 2014

Using key part-of-speech analysis to examine spoken discourse by Taiwanese EFL learners

YEN-LIANG LIN

Department of English, National Taipei University of Technology 1, Sec. 3, Zhongxiao E. Rd., Taipei 10608 Taiwan (R.O.C)
(email: ericlin@ntut.edu.tw)

Abstract

This study reports on a corpus analysis of samples of spoken discourse between a group of British and Taiwanese adolescents, with the aim of exploring the statistically significant differences in the use of grammatical categories between the two groups of participants. The key word method extended to a part-of-speech level using the web-based corpus analytical tool, *Wmatrix*, highlights those linguistic domains which deserve particular attention. Specifically, it reveals the lexical and grammatical categories that occur unusually frequently or unusually infrequently in the English learners' discourse when compared with the language used by the native speakers of English in the sample. The research findings delineate the pedagogical merit of key domain analysis and thus help to inform English as a foreign language teachers and materials developers in the design of courses emphasising spoken interaction.

Keywords: key part-of-speech domain, spoken discourse, EFL learners, Wmatrix

1 Introduction

Key words are those items that occur unusually frequently (positive key words) or unusually infrequently (negative key words) in a target corpus as compared with some norm (Scott, 2010). They are identified on the basis of statistical comparisons of words in a text or collection of texts against a reference set of words. Thus any word that is found to be exceptional in its frequency in the comparison is considered key. Key word analysis therefore provides the analyst with lexical evidence that might be obscured from view, serving as an effective means of identifying salient features that warrant further exploration in context (Harvey, 2012; Lin, 2013; Scott, 2010; Scott & Tribble, 2006). Key word analysis has recently been widely used in many areas of applied linguistics research, particularly with regard to the identification of language variation, genre and discourse characterisation. For example, Baker (2006) compared the Freiburg-LOB corpus of British English (F-LOB)[1] with the Freiburg-Brown corpus of American English (FROWN)[2] and

[1] Further details can be found at www.helsinki.fi/varieng/CoRD/corpora/FLOB
[2] Further details can be found at www.helsinki.fi/varieng/CoRD/corpora/FROWN

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

found a number of underlying cultural differences between the two language varieties. Harvey (2012) compared a corpus of adolescent online health communication with a corpus of general spoken and written English and identified a set of key words which were then examined in their original discourse contexts to gain a deeper understanding both of teenagers' experiences with physical and mental health problems and of the linguistic particularities of online health communication.

A traditional way to explore particular linguistic features across texts is often attributed to the work conducted by Biber (1988, 1995). He employed a multi-dimensional methodology, specifically factor analysis, to analyse the distribution of linguistic features across texts and the systematic co-occurrence of patterns among linguistic features. Although this analysis method offers a reasonably robust basis for the linguistic specification of a genre, Lee (2001) has criticised the approach, raising many questions surrounding the statistical validity, empirical stability and linguistic usefulness of the linguistic dimensions from which Biber derives these text types, or clusters of texts sharing internal linguistic characteristics (p. 40). Moreover, Tribble (2000) identifies difficulties for the practical application of Biber's method in language teaching research, claiming that the approach will not give learners a direct insight into the ways in which native English writers draw on (often highly patterned and conventional) lexical resources, and that it requires the existence of a part-of-speech (POS) marked-up version of the research corpus prior to analysis, which is generally not available to classroom teachers. Xiao and McEnery (2005) compared Biber's multi-dimensional approaches with the key word analysis method in a case study of casual conversation, speech and academic prose in modern American English. They found that key word analysis could capture the same important genre features revealed by multi-dimensional analysis, but determined that the former constituted a better representation of a text and was less demanding to use, since Biber's approach requires sophisticated data extraction and statistical analysis. As such, key word analysis has been used in an increasing number of applied linguistics studies.

While previous studies have extensively employed key word analysis in the investigation of different contexts of language use, few, if any, have applied this method to understand intercultural discourse and to inform English teaching and learning. In addition, studies of intercultural communication or spoken discourse have typically selected particular linguistic features to study prior to the start of research (e.g. Hellermann & Vergun, 2007; Liaw & Master, 2010; Norrick, 2009). This current research, however, proposes a different approach to understanding intercultural discourse based on the methodology of corpus linguistics. It is a "data-driven" approach, which allows "macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focusing on the use of a particular linguistic feature)" (Rayson, 2008: 39). That is, decisions on the linguistic features to be studied are made on the basis of information extracted from the data itself. In this way, the specific linguistic features (the microscopic level) highlighted for further investigation are informed by macroscopic analysis. This macroscopic analysis is based on identifying significant differences (i.e. keyness) between the usage frequencies of lexical and grammatical features in two groups of texts, such as the parts-of-speech components used by Taiwanese learners and native English speakers in the present case. Once identified, these differences are then subjected to further (microscopic) analysis. For example, a key POS analysis might identify a statistically significant difference in the use of grammatical categories by different groups of participants.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

2 Key words and key POS domains

2.1 Extending key word analysis to key POS domains

Although key word analysis highlights those lexical items that are salient in a target corpus when compared with a reference corpus, there are a number of practical limitations. One is identified by Baker (2004), who claims that key word analysis only focuses on lexical differences, rather than semantic, grammatical or functional differences (p. 354). Baker (2006) further points out that it is sometimes the case that some words do not occur often enough to make a sufficient impact, and they are therefore not included in the key word list and tend to be overlooked. For example, Culpeper (2009) analysed the speech of six characters in Shakespeare's *Romeo and Juliet* by comparing each character's speech against the speech of all the other characters. Such an analysis helps to produce key items that reflect the distinctive styles of each character compared with the other characters in the same play. Although some of the results of the keyness analysis of POS categories are already apparent in the key word analysis, a number of important features about a character's style were overlooked when relying solely on the list of key words. Ooi, Tan and Chiang (2007) also applied key POS analysis to the examination of Singaporean English on personal weblogs and compared the text with large corpora of spoken and written English. A deeper understanding of the various cultural identities, gender differences and other sociolinguistic variables was obtained by extending the key word analysis to POS categories.

Furthermore, key word analysis usually generates far more key words than it is possible for the researcher to analyse (Berber Sardinha, 1999: 5; see also Rayson, 2008). Key POS domain analysis, therefore, would reduce the number of key items that the researcher should examine and highlight the most significant categories. This can be seen clearly in previous research on key domain analysis (Culpeper, 2009; Ooi *et al*., 2007; Rayson, 2008). As a result, extending the key word method to the POS level could provide valuable information on the understanding of particular language use, and thereby avoid the limitations of the key word analysis method.

2.2 Tests of statistical significance

Keyness refers to "a quality words may have in a given text or set of texts, suggesting that they are important … [and] reflect what the text is really about, avoiding trivia and insignificant detail" (Scott & Tribble, 2006: 55–56). Although there are a number of ways to calculate statistical significance for keyness, the Chi-square and Log-likelihood (LL) methods are probably the most common (Dunning, 1993; Oakes, 1998). They both compare the differences between the observed values and the expected values: the greater the difference between the two values, the more likely it is that the relationship between the two items is not due to chance, but that other factors influence their relationship (Adolphs, 2006; McEnery, Xiao and Tono, 2006). In this way, the items that have unusual characteristics in a target corpus can be generated by applying the Chi-square and LL tests. However, the LL test is preferred in this study as it does not assume that data are normally distributed (McEnery *et al.*, 2006). Dunning (1993: 65) notes that "using the normal distribution overestimates the significance"; consequently, the use of likelihood ratios leads to very much improved statistical results, particularly when analysing small volumes of text.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Another issue that arises with regard to the comparison of data in two corpora is that one might argue that in any comparison involving corpora of markedly different sizes, frequencies often need to be normalised to a common base. Rayson (2008) claims that there is no need to normalise the figures before conducting LL analyses as the calculation for the expected values takes the size of the two corpora into account. This is very important because claiming statistical significance is based on "not only the *magnitude* of the result but also the *size of the sample* investigated" (Dörnyei, 2007: 210). Rayson further presents the formula for calculating the LL statistic, based on the contingency table shown in Table 1.

In Table 1, the values *a* and *b* are the observed values (O), namely the raw frequencies of a word in two corpora; the values *c* and *d* correspond to the total numbers of words in the two corpora respectively. The expected values (E) are first calculated based on the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

At this point, according to Table 1, N1 = c, and N2 = d. So, for this item, the expected values (E) would be:

$$E_1 = \frac{c\ (a+b)}{(c+d)}, \text{ and } E_2 = \frac{d\ (a+b)}{(c+d)}$$

The log-likelihood value is then generated according to the following formula:

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

More precisely, in this case, the formula can be expanded as follows:

$$LL = 2\times\left(\left(a\times\ \ln\left(\frac{a}{E_1}\right)+b\times\ \ln\left(\frac{b}{E_2}\right)\right)\right)$$

$$\Rightarrow LL = 2\times((a\times(\ln(a)-\ln(E_1))+b\times(\ln(b)-\ln(E_2))))$$

To exemplify these formulae, for example, assuming the frequencies of a word in Corpus One and Corpus Two are 10 and 20 respectively, and the size of the two corpora is 400 and 600 words respectively, this case can be illustrated with the following numerical values, where a = 10, b = 20, c = 400 and d = 600.

Firstly, the expected values (E) can be calculated as follows:

$$E_1 = \frac{400\ (10+20)}{(400+600)} = 12, \text{ and } E_2 = \frac{600\ (10+20)}{(400+600)} = 18$$

Then the LL value can be generated as follows:

$$LL = 2\times((10\times(\ln(10)-\ln(12))+20\times(\ln(20)-\ln(18))))$$
$$\Rightarrow LL = 2\times((10\times(2.3025-2.4849)+20\times(2.9957-2.8903)))$$
$$\Rightarrow LL = 5.68$$

In such a case, the LL value in this keyness analysis is 5.68. In corpus analytical programmes such as *Wmatrix*, the procedure is automatically applied to each single item in the two frequency lists retrieved from the two corpora, while in POS or semantic domain

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Table 1 *Contingency Table for Log-likelihood Calculation*

	Corpus one	Corpus two	Total
Frequency of a word	a	b	a + b
Frequency of other words	c – a	d – b	c + d – a – b
TOTAL	c	d	c + d

analysis, the comparison is done with tag frequencies rather than word frequencies (Rayson, 2008). The most significant items or tags in Corpus One as compared to Corpus Two would then emerge, and they are sorted by the resulting LL values. In this case, the larger the LL value, the more significant the relative frequency difference between the two corpora. Nevertheless, the question remains how large the value needs to be in order to be considered statistically significant, and how confident one can be to make this claim. The latter is the so-called *p* value, which normally ranges from 0 to 1. In the LL test, the critical values with 1 degree of freedom (or d.f.) are 3.83, 6.64, 10.83 and 15.13 for the significance levels of 0.05, 0.01, 0.001 and 0.0001 respectively (McEnery *et al.*, 2006; Rayson, 2008).

3 The present study

This paper demonstrates the pedagogical merit of key POS analysis as a useful tool for EFL learners to obtain direct insight into the ways in which they and native speakers draw on lexical resources, and also as a means to inform teachers and materials developers in designing speaking courses. The key word approach extended to a POS level is applied in a case study of spoken discourse, investigating the lexical and grammatical items used by a group of adolescent Taiwanese learners of English interacting face-to-face with British adolescents. Specifically, the following research question is addressed:

What are the statistically significant differences in the use of grammatical categories between the Taiwanese and British participants?

The key grammatical categories that the Taiwanese participants used significantly differently from the British students are identified by applying key POS analysis to the comparison of the discourse between the two groups of participants.

3.1 Participants

The participants recruited for the study were 40 Taiwanese learners from Taiwan and 40 English secondary school students from the UK, between thirteen and fourteen years of age, all participating in the Connecting Classrooms Project. Most of the Taiwanese participants are English learners at a lower-intermediate level, having learned English for an average of five to six years, and nearly all (97.5%) had never had the experience of interacting with students with a different linguistic background, particularly from an English-speaking country. Similarly, few of the English participants had a friend or an online pen pal from an Asian country.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

The project began with an asynchronous online exchange, in which the participants interacted on an electronic discussion board. After a one-year period of online interaction, the British participants travelled to Taiwan and stayed with the Taiwanese participants for one week. They were divided into five groups with equal numbers of English-speaking natives and non-natives. Oral group discussions in a number of face-to-face meetings were audio-recorded and then co-transcribed by the researcher and the Taiwanese participants' English teachers. The transcriptions were then checked against the original recordings for accuracy.

3.2 Corpus

Since this study aims to examine the particular linguistic features of spoken communication, an analysis based on naturally occurring samples of language data is necessary. The British and Taiwanese Teenage Intercultural Communication Corpus (BATTICC), a newly developed specialised corpus, can therefore be of value as it represents the language use of specific people in specific contexts. The data that forms the basis of BATTICC was collected from casual face-to-face conversation in an intercultural exchange project involving British and Taiwanese teenage participants between thirteen and fifteen years of age. This spoken data collection resulted in approximately four hours of recorded chats, amounting to a total of 34,089 words, transcribed in accordance with standard orthographic practices in order to facilitate analysis using currently available corpus analytical tools. In the transcription, the speakers are generally numbered, and speaker codes (e.g. <TW01>, <BT01>, etc.) represent speakers from the two different countries, where TW and BT refer to the Taiwanese and British participants respectively. Extralinguistic information is also included in square brackets, such as [laughter], [coughing] and [inaudible speech]. Moreover, a sequence of two dots ".." indicates a brief break in speech rhythm, and a longer pause is marked as a sequence of three dots "…". In the Taiwanese scripts in particular, all repetitions of words and phrases (including self-interruptions and false starts) were also transcribed.

With regard to the construction of a corpus, *representativeness* and *sampling* are two essential criteria. McEnery *et al.* (2006) note that these are the features typically used to distinguish a corpus from an archive. That is, an archive is simply a random collection of texts whereas a corpus is designed to provide insight into a particular genre. In this study, BATTICC was constructed to present the informal nature of intercultural communication by adolescent learners in a face-to-face setting. As such, all the samples collected represent that genre. Biber (1995) defines representativeness as the extent to which a sample includes the full range of variability in a population. This suggests that one should strive to collect samples from all the possible situations within a certain genre to completely present the language being studied. However, in collecting spoken discourse data, it does not seem possible to record all of the spoken interactions in the participants' daily lives. Koester (2010) suggests that what is important is to ensure that the samples are collected from a range of fairly typical situations. In this regard, as the aims of the intercultural exchange project were to build relationships between participants in casual settings, spoken data for BATTICC was collected from a range of informal chats between the Taiwanese and British participants during the intercultural exchange programme, in a wide variety of locations

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

such as schools, homes, restaurants, tourist spots, public parks and social gatherings, with the entire speech events recorded whenever and wherever possible. Nevertheless, it needs to be noted that this paper has attempted to demonstrate the particular linguistic patterns in a case study of the intercultural communication project, and consequently the results from this sample should be interpreted with caution before generalising to intercultural interaction as a whole.

3.3 Data analysis

The analysis was begun by producing word frequency lists based on the Taiwanese and British participants' datasets, which offer an immediate snapshot of the characteristics of a particular language variety (Lin, 2013; Harvey, 2012). However, a word frequency list sometimes fails to capture crucial differences between texts, as has been noted by a number of corpus linguists (e.g. McCarthy & Handford, 2004). The key word analysis method was then applied; this works on the basis of statistical comparisons of frequency lists derived from the target corpus (i.e. the Taiwanese dataset) and the reference corpus (i.e. the British dataset). Furthermore, this study extends the key word method to the POS level, generating a list of key POS domains whose frequencies are unusually high or low in the Taiwanese learners' discourse as compared with the British dataset. In this case, the key grammatical categories identified are the ones that the Taiwanese participants used significantly differently from those the British participants used.

In order to incorporate grammar explicitly into a keyness approach, the data needs to be annotated, namely each word needs interpretative grammatical information. This study used *Wmatrix*, a web-based suite of tools provided by Rayson (2008; see also http://ucrel.lancs.ac.uk/Wmatrix), to annotate the datasets for grammatical categories, and then to identify the categories that were key. *Wmatrix* employs the CLAWS (Constituent Likelihood Automatic Word-tagging System) tagger for POS annotation, which automatically assigns POS fields respectively to each word in a corpus (Rayson, 2008).[3] Although *Wmatrix* provides users with a web-based tool to create and process the tagging automatically, Rayson (2008) reports that the accuracy rates of the POS tagging are approximately 96 to 97%. As a result, while we can be largely confident in the tagging accuracy, the interpretation of the results should take into account that possible tagging errors and manual checks of concordances are therefore necessary (Culpeper, 2009; Rayson, 2008).

4 Findings and discussion

The key POS analysis compared the relative use of grammatical categories of the two sets of texts from the Taiwanese and British participants' spoken communication. At the 99.999% level ($p < .0001$), the analysis generated ten significant POS tags. Table 2 displays these top ten tags (with the largest LL values) that occur unusually frequently (marked with a +) and unusually infrequently (marked with a –) in the Taiwanese students' data compared with the British participants' data.

[3] Further details of the software and tagsets employed can be found at http://ucrel.lancs.ac.uk/claws and http://ucrel.lancs.ac.uk/usas

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Table 2 *Ten Most significant differences at part-of-speech level between the Taiwanese and British discourse in BATTICC*

		Taiwanese participants		British participants		Overuse or		
Rank	POS code	Freq.	%	Freq.	%	underuse	LL	POS
1	VBDZ	9	0.13	128	1.07	–	67.36	was
2	UH	381	5.58	368	3.07	+	66.02	interjections
3	NN1	755	11.05	950	7.93	+	45.52	singular common nouns
4	VVN	19	0.28	119	0.99	–	35.23	past participle of lexical verb
5	PPH1	100	1.47	333	2.78	–	35.01	neuter personal pronoun: it
6	RR21	4	0.06	62	0.53	–	34.62	general adverbs – ditto tags
7	VVD	38	0.56	142	1.19	–	19.53	past tense of lexical verbs
8	VV0	268	3.93	338	2.82	+	16.05	base form of lexical verb
9	VBDR	2	0.03	29	0.24	–	15.76	were
10	VBZ	207	3.03	265	2.21	+	15.38	is

The table shows that the Taiwanese learners use significantly fewer words in the categories of *was* (VBDZ), past participle of lexical verb (VVN), third-person singular neuter personal pronoun *it* (PPH1), general adverbs – ditto tags (RR21), past tense of lexical verbs (VVD) and *were* (VBDR), while using words in the categories of interjections (UH), *is* (VBZ) and singular common nouns (NN1) significantly more frequently than the British participants. The list shows quite clearly that the majority of the categories with unusual frequencies are related to verb tenses. This paper will now focus particular attention on the use of tenses in BATTICC.

4.1 Tenses

As has been seen from Table 2, the categories related to verb tenses have unusual frequencies in the Taiwanese participants' discourse as compared with the British one. This raises the question of whether the Taiwanese learners have difficulties gaining full control of the appropriate use of verb tenses in their speech, particularly with regard to present simple and past simple tenses. Further examination of the scripts found many instances of the Taiwanese learners' mis-selections of the forms of appropriate tenses. For example:

1. <TW03>: That Treasure Hunt ***is*** pretty fun, right?
   <BT04>: Yeah,
   <BT05>: We tried to run to get back in time and we were ... we looked like
2. <TW05>: Confusing why? I think the question ***is*** boring.
   <BT06>: They were a bit confusing.
3. <TW07>: How about ... er ... remember today we ***are*** climbing the mountain right?
   <BT09>: Oh, that was really funny because I nearly fell over.

In examples (1) and (2) the Taiwanese students TW03 and TW05 used *is* to describe past activities, but their British peers responded in the past tense. This type of error made by the

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Taiwanese students was found in twelve instances when the researcher examined all of the concordance lines of *is* in BATTICC. Example (3) also shows the speaker's mis-selection of the appropriate verb form. Errors like this were found in five instances. Therefore, it seems that the Taiwanese speakers might have difficulty selecting the appropriate present simple and past simple verb tense forms. In addition, the overuse of *is* (VBZ) by the Taiwanese learners can be found in the following sentences.

4. <TW01>: Like our schools. Our schools ***is*** have many clubs and …
5. <TW16>: The bacon and the sausage ***is*** ... it tastes like ... erm ... I just
6. <TW09>: My parents ***is*** … My parents ***is*** very nice. My mom and dad let me do everything.

In example (4) speaker TW01 used an additional *is*, which was not necessary; in (5) and (6) speakers TW16 and TW09 used *is* instead of the plural form *are*. A notable occurrence of tense-aspect marking present in the Taiwanese speakers' speech is the use of the past tense of lexical verbs (VVD) and the base forms of lexical verbs (VV0). A closer look at the use of the lexical verbs in speech shows that in many cases it involves the non-marking of the past tense. For example:

7. <TW03> : this morning .. er: the first we ***go*** to the trail … the trail … Pretty tired and we ***climb*** [climbed] on rocks or something but we ***know*** we ... how to help each others and we ***have*** fun in this activity, yeah ...
   <BT04>: And we had to be sort of careful .. going up the rock face where people might fall.
   <BT05>: I did fall off ... [laughter]
   ……
   <TW03>: And some difficult words ... we ... just ***listen*** ... and ***hear*** you ... have ... we ***are*** confused.
   <BT04>: Yeah, like on the Treasure Hunt we went on later, there was the ... erm ... rhymes things and there was like ... there was confusing words on there that sometimes people wouldn't understand.

In this example the past time adverbial *this morning* by speaker TW03 sets the utterance in the past, namely the activities – trail walking and the Treasure Hunt – which they had just done that morning. Nevertheless, it can be noted that a number of lexical verbs used by the Taiwanese participant TW03 were in their base form, such as *go, climb, know, have, listen* and *hear*, where one would expect *went, climbed, knew, had, listened* and *heard* respectively. This feature can be considered as an example of adjacent default tense (ADT), which means that "if the overall tense of an utterance is marked in the context of the utterance, then, the 'adjacent' finite verbs in the utterance can (but may not necessarily) be set in their 'default' forms" (Xu, 2010: 69). Such "non-standard forms" (NSF) of tense-aspect marking were also identified by Kirkpatrick (2007: 157), particularly in Asian EFL learners' speech, in which a large amount of non-marking of tense by using the base form of the lexical verbs is present. As Ellis and Larsen-Freeman (2006) claim, a prominent characteristic of interlanguage in English as a second or foreign language acquisition is the lack of tense marking, and it constitutes a significant hurdle for English learners to overcome. One of the most important reasons for the Taiwanese learners' difficulties in the use of appropriate lexical verb tenses may well result from the influence of their first language, Mandarin Chinese,

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

which does not have a grammatical category of tense, and such a lack of tense inflection in Chinese may therefore have an impact on the frequency of verbal non-marking in Chinese-speaking learners' interlanguage (Brown, 2007; Ellis, 2008; Xu, 2010). Chen (2010), Xu (2010), and some other Chinese-speaking scholars have claimed that this distinctive use of English has constituted one of the most significant syntactic features of Chinese English.

Although the Taiwanese learners appear to have some difficulties with using appropriate tenses in their speech, their pre-test on writing shows the opposite. On examination of 700 sentences in which *is* was used by the Taiwanese participants, only four errors concerning the use of tenses were found. That is to say, they used the present simple form *is* properly most of the time in writing, while in speaking they frequently made a tense error by using the present simple tense to describe past events. This can be partially explained by Krashen's (1981: 52) Monitor Hypothesis:

> When we speak a second language, the forms we use come "first" from our subconsciously acquired competence. We then attempt to apply conscious rules, sometimes before we speak and sometimes not, sometimes successfully and sometimes not.

In this regard, "conscious learning acts as an editor, as a Monitor" when learners speak a foreign language (Krashen, 1982: 16), and they tend to utilise conscious rules to increase their grammatical accuracy in speaking when it does not interfere with communication. To that end, EFL teachers are advised to implement a variety of speaking activities to promote their students' self-monitoring strategies in the classroom. Self-monitoring can begin as gentle consciousness-raising exercises with the goal of helping students develop speech awareness and self-observation skills. This is most effective when students are actively involved in their own learning, instead of passively and detachedly repeating words and phrases in drills.

4.2 Interjections

In addition to verb tenses, another grammatical category that occurred with a significantly unusual frequency in the Taiwanese learners' discourse was the use of interjections (e.g. *yeah, oh, ah*), which normally refer to "exclamative utterances consisting of single words that do not easily fit into the major word classes (noun, verb, adjective, adverb)" (Carter & McCarthy, 2006: 224). They are particularly common in spoken discourse in that they provide speakers with useful interactional and organisational resources and serve to mark the boundaries of discourse units.

Yeah is by far the most frequent interjection overall in BATTICC. Jucker and Smith (1998) claim that *yeah* is most frequently used to acknowledge the receipt of information that is new to the discourse but consistent with the current active information. In this regard, *yeah* serves as a "continuer response token" in that it helps maintain the flow of the discourse and "encourage[s] the current speaker to continue" (O'Keeffe & Adolphs, 2008: 84). Many cases of *yeah* in BATTICC occurred alone as a continuer response token, for example:

8. <BT07>: And your recorder.
   <TW07>: ***Yeah***.
   <BT08>: That was really good. I was like how does .. I thought it was a recording.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Yeah may also function as a direct positive response to a question (Norrick, 2009), as can be seen in examples (9) and (10):

9. <TW07>: and er er in your country I know some ... some food is famous, right?
   <BT07>: ***Yeah***, like fish and chips.
   <TW07>: Fish and chips.
10. <TW15>: Er ... we usually play basketball and volleyball in our school.
    <BT17>: Ah .. volleyball?
    <TW15>: ***Yeah*** – but I don't really like it.

Additionally, *yeah* may signal agreement with a statement in the foregoing turn (Norrick, 2009). As in the excerpt below, *yeah* simply means *I agree with you.*

11. <BT09>: Our one was rubbish, but like the Taiwanese Dragon Boats are beautiful.
    <BT07>: ***Yeah***, they really are. But ours are like several canoes tied together basically.

Yeah also occurred as an initial transition word with no obvious positive response or agreeing function (Norrick, 2009), and as such it was used to express "a general acknowledgment of the previous interactive unit'' (Jucker & Smith, 1998: 181).

12. <BT10>: I'm not atheist, agnostic ... I don't … not quite sure if there is a God or not.
    <BT11>: ***Yeah***, if God appeared then I'd say hi God, I'd believe. We're explaining agnostic.

It appears that *yeah* fulfils a number of different discourse functions. This confirms Fung and Carter's (2007) observation that it functions to acknowledge, agree, affirm and mark continuation, and fits with Evison's (2008) views on its versatility in academic talk. However, based on the key POS analysis in *Wmatrix*, the Taiwanese learners used the interjection *yeah* significantly more frequently than the British pupils, with 2.29% and 1.51% respectively. Some of the instances in BATTICC show that the Taiwanese participants frequently and sometimes continuously used *yeah* in response to a question or just to acknowledge the previous utterance as a back-channel token, as in (13) to (15):

13. <BT09>: Have you lived in Haulien all your life?
    <TW07>: ***Yeah, yeah.***
    <BT07>: You know the dragon boat racing?
    <TW07>: ***Yeah, yeah, yeah.***
    <BT07>: What's it about?
    <TW07>: … You mean … story, or …
14. <BT02>: I feel very happy because I won.
    <TW02>: ***Yeah, yeah, yeah.***
15. <BT18>:Do you like the fruit that we have here?
    <TW16>: ***Yeah.***
    <BT18>: The fruit, like strawberries.
    <TW16>: ***Yeah.***

These extracts demonstrate that *yeah* is pervasive in the Taiwanese participants' utterances. Examples (13) and (14) even display the repetitive use of *yeah*. Such reduplications

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

were also observed in O'Keeffe and Adolphs's (2008) comparative study of response tokens in casual conversation in British and Irish English.[4] Evison (2008) also reports that repetitions of interjections are frequent, remarking that *yeah yeah* is the most frequent turn-initiator cluster in academic talk. In BATTICC, there are more reduplications in the Taiwanese than in the British English discourse. Moreover, the overwhelming tendency for *yeah* to occur alone can be found in the Taiwanese participants' data, as in (15). In some extreme examples, the Taiwanese participants even said *yeah* right after each sentence said by their British interlocutors. These two phenomena can result in a greater use of interjections in the Taiwanese dataset than in the British one. A possible explanation for this might be that the Taiwanese learners probably had no idea how to continue the conversation, or possibly that they did not even understand clearly the previous utterance (i.e. the matter of language proficiency), and thus the reduplication and stand-alone use of *yeah* are strategies they employed frequently to buy time for discourse planning.

Some other interjections that occur with extremely high frequencies in the Taiwanese participants' dataset include *ah*, *wow* and *oh*, occurring roughly two to three times more frequently in the speech of the Taiwanese pupils than in that of the British participants. The *ah* in (16) probably means *why didn't you try it?*, indicating the speaker's surprise since *stinky tofu* is one of the most famous and must-eat foods in Taiwan. The *ah* in (17) signals agreement with the statement, which simply means *I see/I understand*.

16. <TW17>: Do you like the stinky tofu?
    <BT19>: I didn't try it.
    <TW17>: ***Ah*** ... it's very good.
17. <BT19>: No. We've had one earthquake but it was very small, it was just
    <TW17>: ***Ah*** ... but our earthquake is always very big.

Oh is also one of the commonest interjections used in speech, and it is generally used to signal a change in cognitive state (Norrick, 2009) or an expression of surprise (Aijmer, 2011). In BATTICC, the Taiwanese learners use it appreciably more frequently than the British participants, but its usage neither signals a cognitive change nor a surprised expression. As in (18), speaker TW15 constantly used five *oh*s in a rather short conversation, and some *oh*s appear to have been used simply as a general acknowledgment of the previous utterance, which may indicate the speaker's interest and involvement. Moreover, reduplications can also be found in the use of *wow*, as in (19). Although speaker TW01 used four continuous *wows* in his/her response, the meaning of these *wow*s is not at all clear within the context of the discourse.

18. <BT17>: What did you do in Australia?
    <TW15>: ***Oh*** .. we went to school ... like local school and just stay in school. yeah
    <BT17>: Did you visit anywhere?
    <TW15>: ***Oh*** .. we visit the the Golden ... The Gold Coast.
    ……
    <BT17>: … we are all going when we are 20 and when we have a job.
    <TW15>: ***Oh*** ... you can tell me when you're coming to Taiwan.
    <BT17>: Yeah, I think I will tell everyone I have met.

[4] Their study notes that there is more reduplication in Irish English than in British English.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

<TW15>: And what's your school's name?
<BT17>: <School name01> .
<TW15>: ***Oh*** ... pictures everywhere.
<BT17>: Yeah. Are you coming, are you going to any more schools in England?
<TW15>: ***Oh*** we're just going to …

19. <BT01>: What email address do you use?
<TW01>: ***Wow, wow, wow, wow***.
<BT01>: Doing Facebook now.
<TW01>: Yes, that's good.

This section has presented some examples of the interjections that were used significantly frequently by the Taiwanese participants. Although the interjections usage did not result in misunderstandings or serious face-threatening situations in the samples of BATTICC, the meaning of these interjections is not always clear (and some of them are not appropriate). Reber (2010) suggests that EFL teachers should teach the sound patterns and usages of interjections on the basis of naturally occurring discourse rather than referring to invented conversation examples. In this way, learners would gain some knowledge about when and how to display affectivity and appropriate responses in spoken interaction.

As has been discussed, interjections and other interactional features of spoken grammar have very important discourse functions, such as organising the utterances by "breaking up utterances into smaller 'meaning chunks'" that may actually aid comprehension (Gilmore, 2004: 369). They can also indicate turn-taking, helping speakers keep the floor while formulating their next utterance, or in some cases indicating that they are ready to relinquish the floor. Moreover, they serve important interpersonal functions, which are highly relevant to successful interaction in an informal communication setting. As a result, it would be helpful for teachers to include these important commonly occurring spoken features in their EFL course syllabuses.

4.3 Ditto tags: Multi-word expressions

With regard to the domain of general adverbs – ditto tags (RR21), as shown in Table 2, far more instances are found in the British students' texts (63 tokens/0.53%) than the Taiwanese ones (4 tokens/0.06%). Ditto tags encode the notion that a token is not an individual unit, but rather is a (somewhat non-compositional) part of a larger "idiom" (Dickinson, 2005: 46). The ditto tags, RR21 here, indicate the (RR – general adverb) part of speech, the total number of elements in the idiom (two in this case) and the position of a certain word within the idiom (one in this case). Examples (with frequent occurrences) within this category include *sort of* (19), *a lot* (14), *a bit* (12), *as well* (7), *kind of* (3), *at all* (3), *of course* (2) and *by far* (2) in the British students' discourse, while only *a lot* (2), *as well* (1) and *of course* (1) were found in the Taiwanese texts. As can be seen in the following extract involving the British speaker BT18 and the Taiwanese speaker TW16, who were discussing the differences between Taiwanese and British food, *sort of* is used three times in one utterance.

20. <BT18>: Okay. Yeah, your food generally is a lot more ***sort of*** ... erm ... traditional and special than ours. Ours is just ***sort of*** simple, ***sort of***, ...

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

<TW16>: It's okay, I like it ... it's your culture actually.
<BT18>: Yeah.
<TW16>: But I am .. I don't like the traditional breakfast because it's too salty and the flavour is too strong.

Speaker BT18 is likely trying to lessen the strength of his assertion by frequently using *sort of* when giving comments about Taiwanese food so that the statement sounds less direct. This is perhaps explained by the uncertainty of speaker BT18 about his own assumption, and he thus intends to be less assertive; on the other hand, speaker TW16's response, *I don't like the traditional breakfast because it's too salty and the flavour is too strong*, seems much more direct compared to speaker BT18's statement. Miskovic-Lukovic (2009) calls such types of expressions "positive politeness strategies" (p. 622). These help to "*downtone* the force of the utterance" and to "*mitigate* against any potential threat to face" (O'Keeffe, McCarthy & Carter, 2007: 174).

With respect to the acquisition of multi-word expressions, it is widely accepted that mastering the formulaic expressions commonly used by native English speakers is difficult for young learners of English as a foreign language (Lin, 2014; O'Keeffe *et al.*, 2007; Schmitt, 2010; Wray, 2000). Wray (2000) indicates a number of possible reasons. One is the poor quality of the learner's learning experience in the classroom; it is also easy for the wrong sequences to be taught. One way around these problems, therefore, is to "provide learners with language experience which offers exposure to the most useful patterns of the language" and to help them "to notice patternings and to speculate about them" (Willis, 1990: 38). As such, when learners are engaged in meaningful activities that involve manipulating language, they learn more and retain that information longer (Reppen, 2010). Take, for example, the multi-word expression *sort of*, which may be difficult for many young Taiwanese English learners to use. A study of selected concordance lines taken from the British participants' transcripts would allow the Taiwanese students to be more aware of the appropriate use of multi-word expressions.

5 CONCLUSION

This study reports on a corpus analysis of samples of spoken discourse between a group of British and Taiwanese adolescents with a focus on exploring the statistically significant differences in the use of grammatical categories between the two groups. The key word analysis method extended to a part-of-speech level using *Wmatrix*, a web-based corpus analytical tool, highlights the POS domains which deserve particular attention. Comparisons of keyness at the POS level reduce the number of key items that the researcher needs to examine, as noted by Rayson (2008), thus addressing the most significant linguistic features that are key in discourse between adolescent British and Taiwanese learners.

Most significantly, the key POS analysis reveals that the Taiwanese participants have difficulty with the appropriate use of verb tenses in their speaking, and the category of general adverbs – ditto tags (e.g. *sort of, a bit, as well*) which was rarely used. On the other hand, significantly larger numbers of the verb to be *is* and interjections (e.g. *wow*, *yeah*) were found in the Taiwanese participants' discourse as compared with the British. The significance of the analysis of POS categories lies in its ability to demonstrate that the Taiwanese learners' data reveal the extent to which grammatical categories are used

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

significantly differently, as compared with the discourse of the native English speakers. Such findings pertaining to grammatical categories are not likely to be found when only word frequencies and key words are examined.

The result of this key domain analysis contributes to researchers' and educators' understanding of learner grammar and lexis, and thus has important pedagogical implications for EFL teaching and learning. First, it reveals a number of cultural and social differences in the use of words and grammar, and EFL teachers can therefore not only encourage their learners to observe these culturally relevant features but also further develop their intercultural awareness and spoken communication skills (Lin, 2013, 2014; O'Dowd, 2007). The spoken language features (e.g. interjections, multi-word expressions) have important interpersonal functions that serve to "create and maintain good relations between the speaker and hearer" (O'Keeffe *et al.*, 2007: 159). For EFL learners intending to maintain a good relationship in face-to-face conversations, it would therefore be very helpful for students to be aware of and to learn these interactional features. Second, by using automatic tagging software that assigns POS tags, macroscopic analysis can help to identify those linguistic features that deserve further attention, providing a reasoned basis for drawing learners' awareness to linguistic features specific to their own and the native-speaker discourse (Rayson, 2008; Tribble, 2000). The quantitative data derived from the statistically robust frequency of POS categories can be considered more objective in the sense that decisions about the linguistic features to be studied can be made on the basis of information mechanically extracted from the data itself. Last but not least, key domain analysis using the supporting *Wmatrix* software is accessible and practically useful for EFL classroom teachers due to its ease and rapidity of use without requiring extensive training. It allows users to simply upload their own data to the system, which automatically annotates the data and displays the results within a web browser.

To conclude, although keyness analysis has been applied in a growing number of diverse studies, few have applied this approach from an EFL pedagogical perspective. While illustrating that manual checking of concordances and interpretation of results remains useful in addressing occasional errors in automated keyness analysis (Culpeper, 2009; Rayson, 2008), this paper demonstrates the pedagogical merits of this type of analysis through its application in a case study of spoken discourse. In doing so, it has been argued that key domain analysis can provide researchers and educators with a better understanding of intercultural spoken discourse, giving English learners a direct insight into the ways in which they and native English speakers draw on lexical resources, and helping to inform EFL teachers and materials developers in their design of courses for adolescent intercultural interaction.

References

Adolphs, S. (2006) *Introducing electronic text analysis: A practical guide for language and literary studies*. New York: Routledge.

Aijmer, K. (2011) "Well I'm not sure I think…" The use of "well" by non-native speakers. *International Journal of Corpus Linguistics*, **16**(2): 231–254.

Baker, P. (2004) Querying key words: Questions of difference, frequency and sense in key words analysis. *Journal of English Linguistics*, **32**(4): 346–359.

Baker, P. (2006) *Using corpora in discourse analysis*. London: Continuum.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Berber Sardinha, T. (1999) Using key words in text analysis: Practical aspects. *DIRECT Working Papers*, **42**: n.p. http://www2.lael.pucsp.br/direct/DirectPapers42.pdf

Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1995) *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Brown, H. D. (2007) *Principles of language learning and teaching* (5th ed.), White Plains, NY: Pearson Longman.

Carter, R. and McCarthy, M. (2006) *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

Chen, X. (2010) *Discourse-grammatical features in L2 speech: A corpus-based contrastive study of Chinese advanced learners and native speakers of English*. City University of Hong Kong, unpublished PhD.

Culpeper, J. (2009) Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's 'Romeo and Juliet'. *International Journal of Corpus Linguistics*, **14**(1): 29–59.

Dickinson, M. (2005) Error detection and correction in annotated corpora. http://jones.ling.indiana.edu/~mdickinson/papers/diss/dickinson05-alt.pdf.gz

Dörnyei, Z. (2007) *Research methods in applied linguistics*. Oxford: Oxford University Press.

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistic*, **19**(1): 61–74.

Ellis, N. C. and Larsen-Freeman, D. (2006) Language emergence: Implications for applied linguistics. *Applied Linguistics*, **27**(4): 558–589.

Ellis, R. (2008) *The study of second language acquisition* (2nd ed.), Oxford: Oxford University Press.

Evison, J. (2008) *Turn-openers in academic talk: An exploration of discourse responsibility*. University of Nottingham, unpublished PhD thesis.

Fung, L. and Carter, R. (2007) Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, **28**(3): 410–439.

Gilmore, A. (2004) A comparison of textbook and authentic interactions. *ELT Journal*, **58**(4): 363–374.

Harvey, K. (2012) Disclosures of depression: Using corpus linguistics methods to interrogate young people's online health concerns. *International Journal of Corpus Linguistics*, **17**(3): 349–379.

Hellermann, J. and Vergun, A. (2007) Language which is not taught: The discourse marker use of beginning adult learners of English. *Journal of Pragmatics*, **39**: 157–179.

Jucker, A. H. and Smith, S. W. (1998) "And people just you know like wow": Discourse markers as negotiating strategies. In: Jucker, A. H. and Ziv, Y. (eds.), *Discourse markers. Descriptions and theory*. Amsterdam: John Benjamins, 171–201.

Kirkpatrick, A. (2007) *World Englishes: Implications for international communication and English language teaching*. Cambridge: Cambridge University Press.

Koester, A. (2010) Building small specialised corpora. In: McCarthy, M. and O'Keeffe, A. (eds.), *The Routledge handbook of corpus linguistics*. London: Routledge, 66–79.

Krashen, S. (1981) The "fundamental pedagogical principle" in second language teaching. *Studia Linguistica*, **35**(1): 50–70.

Krashen, S. (1982) *Principles and practice in second language acquisition*. Oxford: Pergamo.

Lee, D. Y. W. (2001) Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, **5**(3): 37–72. http://llt.msu.edu/vol5num3/pdf/lee.pdf

Liaw, M. L. and Master, S. B. (2010) Understanding telecollaboration through an analysis of intercultural discourse. *Computer Assisted Language Learning*, **23**(1): 21–40.

Lin, Y. L. (2013) Discourse functions of recurrent multi-word sequences in online and face-to-face intercultural communication. In: Romero-Trillo, J. (ed.), *Yearbook of corpus linguistics and pragmatics*. London: Springer, 105–129.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press

Lin, Y. L. (2014) Exploring recurrent multi-word sequences in EFL textbook dialogues and authentic discourse. *English Teaching & Learning*, **38**(2): 133–158.

McCarthy, M. and Handford, M. (2004) "Invisible to us": A preliminary corpus-based study of spoken business English. In: Connor, U. and Upton, T. A. (eds.), *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins, 167–201.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based language studies*. London: Routledge.

Miskovic-Lukovic, M. (2009) Is there a chance that I might kinda sort of take you out to dinner? The role of the pragmatic particles kind of and sort of in utterance interpretation. *Journal of Pragmatics*, **41**: 602–625.

Norrick, N. R. (2009) Interjections as pragmatic markers. *Journal of Pragmatics*, **41**: 66–891.

Oakes, M. (1998) *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

O'Dowd, R. (2007) Evaluating the outcomes of online intercultural exchange. *ELT Journal*, **61**(4): 144–152.

O'Keeffe, A. and Adolphs, S. (2008) Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. In: Schneider, K. P. and Barron, A. (eds.), *Variational Pragmatics*. Amsterdam: John Benjamins, 69–98.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Ooi, V., Tan, P. and Chiang, A. (2007) Analyzing personal weblogs in Singapore English: The Wmatrix approach *Studies in Variation, Contacts and Change in English*, **2**: n.p. http://www.helsinki.fi/varieng/journal/volumes/02/ooi_et_al

Rayson, P. (2008) From key words to key semantic domains. *International Journal of Corpus Linguistics*, **13**(4): 519–549.

Reber, E. (2010) Interjections in the EFL classroom: Teaching sounds and sequences. *ELT Journal*, **65**(4): 365–375.

Reppen, R. (2010) *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Schmitt, N. (2010) *Researching vocabulary: A vocabulary manual*. London: Palgrave Macmillan.

Scott, M. R. (2010) *WordSmith Tools help manual (Version 5.0)*. Liverpool: Lexical Analysis Software.

Scott, M. R. and Tribble, C. (2006) *Key words and corpus analysis in language education*. Amsterdam / Philadelphia: John Benjamins.

Tribble, C. (2000) Genres, key words, teaching: Towards a pedagogic account of the language of project proposals. In: Burnard, L. and McEnery, T. (eds.), *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang, 75–90.

Willis, D. (1990) *The lexical syllabus*. London: HarperCollins.

Wray, A. (2000) Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, **21**(4): 463–489.

Xiao, Z. and McEnery, T. (2005) Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics*, **33**(1): 62–82.

Xu, Z. (2010) *Chinese English: Features and implications*. Hong Kong: Open University of Hong Kong Press.

https://doi.org/10.1017/S0958344014000408 Published online by Cambridge University Press