Linguistic Research 30(2), 163-193

An error analysis of L2 English discourse reference through learner corpora analysis*

Peter Crosthwaite
(University of Cambridge)

Crosthwaite, Peter. 2013. An error analysis of L2 English discourse reference through learner corpora analysis. *Linguistic Research* 30(2), 163-193. This research investigates errors of referring expressions used in L2 narrative discourse through two learner corpora, namely the Cambridge Learner Corpus (CLC), and a subsequent controlled corpus created in CLAN. By adopting the cross-sectional approach to learner language used in the English Profile Programme, the research identifies the frequency and type of errors of reference made by Mandarin and Korean L2 English groups from A1 (beginner) to C2 (advanced) levels of the CEFR (Council of Europe, 2001), and asks if and when L2 learners cease making errors of reference, and whether L1 background is a factor in the frequency and type of errors made. The results suggest that L2 learners produce little to no syntactic marking of reference at lower proficiencies, gradually incorporating the appropriate markings of the L2 target at higher proficiencies. It was also found that Korean L2 English learners produce more errors compared to the Mandarin L2 English group at each CEFR level. The difference in the type and frequency of errors between the L2 groups is suggested to lie in the potential grammaticization of numerals and demonstratives in L1 Mandarin to sharing the same functions as the English indefinite and definite articles. (Li and Thompson, 1976, 1989, Hedberg, 1996, Chen, 2004), giving the Mandarin group an advantage in mapping syntactic form (articles) to pragmatic function (introducing and maintaining reference) in the L2. **(University of Cambridge)**

Keywords L2 Error analysis, L2 discourse reference, English as a foreign language, Korean L2 English, Mandarin L2 English, Learner Corpus Analysis, Cambridge Learner Corpus

1. Introduction

This study aims to show the difficulties that second language (L2) learners have

* This paper was presented at the 2012 KACL (Korean Association of Corpus Linguistics), Busan, South Korea, December 2012. I would like to extend my thanks to the participants of KACL who offered helpful commentary after the presentation, and to the two anonymous reviewers who provided invaluable input during the editing process.

when maintaining reference to animate entities in narrative discourse. There has been a renewed interest in the coherence of discourse reference in the light of syntax-pragmatic approaches to reference such as Accessibility Theory (Ariel, 1991, 2008, 2010), where a speaker's selection of referring expressions for the maintenance of coherent reference is claimed to be motivated by the pragmatic accessibility of the referent in the mental model of the discourse between speaker and listener. While the principle of referent accessibility is claimed to be universal to all languages, the form and distribution of referring expressions that signal varying degrees of referent accessibility may be different between languages, and thus present particular problems for the second language learner during language acquisition if the form and distribution of referring expressions in the first language (L1) is not congruent with that of the L2.

Authentic narrative data (often rich in long chains of referring expressions) is vital to any exploration of the principles governing reference. Therefore, learner corpora, or searchable collections of the written or spoken production of language learners, have become increasingly vital tools when researching how referent introduction and maintenance is achieved by second language learners. These tools allow the researcher to explore authentic learner texts parsed and tagged for syntactic, lexical and morphological criteria at a speed and level of ease that was not available to researchers in the past.

Therefore, two studies were conducted into the use of referring expressions by second language learners using learner corpora, focusing on errors made when maintaining reference. The first study uses the Cambridge Learner Corpus, which has been error-coded for a range of linguistic phenomena such as violations of syntactic structure, missing or incorrect use of determiners (including articles) with noun phrases, inappropriate use of discourse markers, and many more besides (Nicholls, 2003). A second controlled corpus was also built using data from a standardized narrative picture sequence, using the CLAN (Computerized Language ANalysis) program, following the principles of the CHILDES corpus (MacWhinney, 2000). The results of both studies point towards a clear developmental trajectory, from errors of reference such as missing anaphor and missing determiners at lower proficiency levels, to the development of appropriate L2-target-like use of reference at higher proficiencies. The following sections define the scope of this study, and explain the choice of source languages used in the investigation of L2 English referring

expressions.

2. Reference and the choice of referring expressions

Bach (2008) defines reference as "a four-place relation" [in which] "a speaker uses an expression to refer his [or her] audience to an individual" (Bach, 2008:17). In the performance of a narrative, speakers have to introduce new individuals into the story, and once introduced, maintain reference to those individuals throughout the story for the audience, in a cohesive and coherent fashion. This is achieved through the selection of particular referring expressions that are suitable for those functions, within the range of referring expressions available within their language. For example, in English, subsequent mentions of a referent will likely be a full noun phrase (with definite article) or a pronoun, as with the example below:

"A boy enters a shop. He goes to the counter"

The speaker's selection of subsequent-mention referring expressions should ensure that the audience is left in no doubt about who is being referred to, and is dependent on the *accessibility* (Ariel, 1990, 2008, 2010) of the referent at the moment of expression. Accessibility depends on a number of factors, such as *distance* between mentions, *competition* from other referents, degree of *salience* (such as the recurrent topicality of the referent), and *unity* of references within a single discourse sequence. The speaker's selection of referring expression represents the relative accessibility of a referent, with different referring expressions functioning as *low* or *high* accessibility markers. It is claimed that full noun phrases are used for referents with low accessibility, and shorter lexical forms (pronouns, zero anaphor) are used for highly accessible referents, according to the following scale (from low to high accessibility):

Full name > long definite description > short definite description > last name > first name > distal demonstrative > proximate demonstrative > stressed pronoun > unstressed pronoun > cliticized pronoun > verbal person inflections > zero

(taken from Ariel, 2008:44)

While this scale is taken to be a universal ranking of the accessibility of referential forms between languages, the configuration and relative frequencies of referring expressions used to signal relative statuses of accessibility are language-specific (Ariel, 2008:53). It is conceivable that L2 learners may have more difficulty using the L2 target-like system of referring expressions, if their L1 reference system is not congruent with that of the L2 target − experiencing either negative transfer from the L1 or a lack of positive transfer from the L1. Difficulty in using the target-like referential system may result in errors that may damage the overall coherence of the complex verbal task the speaker and audience are engaged in.

The studies in this paper investigate the use of syntactically inappropriate referential forms used for reference maintenance[1] in L2 narrative production. The next section defines how the first languages of English, Mandarin Chinese, and Korean maintain coherent reference, so that hypotheses might be drawn about the likelihood of errors in L2 English production found as the result of L1 transfer.

3. Language-specific strategies for reference maintenance

Subsequent or 'given' mentions (following Chafe, 1980) involve a wide range of referring expressions that represent different levels of *accessibility* (Ariel, 1990, 2008, 2010) of referents in the common ground. As mentioned above, high accessibility referring expressions (such as zero anaphors, overt pronouns) are used for highly activated referents (highly topical, typically co-referent with the previous subject/topic), while low accessibility referring expressions (definite article + noun constructions, bare nominals) are used for less topical, more distant referents.

3.1 English

In English, the most common high accessibility marker is the pronoun, with separate forms to indicate the gender of the referent:

---

[1] The study of referent introductions has been left out of this paper due to space considerations.

A boy went upstairs. *He* smiled.
A girl went upstairs. *She* smiled.

Zero anaphors may be used as a high accessibility marker in English, but only in clauses linked by conjunctions such as ‘and’ or ‘but’, and they may only be used to refer to the subject of the previous clause, or as part of lists that include a final conjunction:

A boy$_i$ went upstairs and ø $_I$ went to the bathroom.
A boy went upstairs, ø $_I$ went to the bathroom, ø $_I$ looked in the mirror and ø $_I$ smiled.
*A boy went upstairs. ø $_I$ went to the bathroom.

The most common low accessibility marker is the definite article + noun construction, or proper names are occasionally also used:

A boy went upstairs. Another boy followed. *The first boy* said hello.

3.2 Mandarin

In Mandarin, the most common high accessibility marker is the zero anaphor. Mandarin is known as a ‘topic drop’ language, where highly accessible referents are frequently omitted from the discourse in the context of a ‘topic chain’ when they are inferable to the audience (although the omission can also extend to the entire proposition, rather than just the referent alone):

那 輛 車$_i$ 价錢 太 貴, ø $_I$ 顏色 也 不好, ø $_j$我 不 喜歡 ø $_I$, ø $_j$不想 買ø $_I$
Nà liàng chē$_i$ jiàqian tài guì, ø $_i$ yánsè yě bùhǎo, ø $_j$ wǒ bù xǐhuan ø $_I$, ø $_j$ bùxiǎng mǎi ø $_I$
That CL car, price too high, color also not good, **I** not like, not want buy
ø $_j$ 昨天 去 看 了 一下ø $_I$, ø $_j$還 開 了 一會儿ø $_I$, ø $_j$還是不喜歡 ø $_I$

ø j Zuótiān qù kàn le yīxià ø I, ø j hái kāi le yīhuìr ø I, ø j háishi bù xǐhuan ø I
Yesterday go see ASP a-bit, also drive-ASP a while, still not like.
That car is too expensive, the color is not good either. I don't like (it) and (I) don't want to buy it. Yesterday, I went there to take a look, I even drove it for a while. But I still don't like it. (from Li, 2004: 25)

Mandarin also makes use of third-person pronouns, but these are most commonly used to signal a shift in reference from one referent to another after a topic chain has been established. In the written form, there are separate pronouns for male and female referents, but the spoken form of the language uses only a single third person pronoun form that is gender neutral:

他	非常	高興.
tā	fēi cháng	gāo xìng.
He/she	very	happy

The most common low-accessibility marker in Mandarin is the bare (no determiner) noun, although the use of the proximal demonstrative 这 – zhè - is also increasingly being used to signal definiteness, in a manner similar to the definite article in English (Li & Thompson, 1976, 1989; Hedberg, 1996; Chen, 2004), as with the following example where a competing referent (the black cat) is the topic of the second sentence, reducing the overall accessibility of the goldfish from the first sentence, and prompting the demonstrative NP (assuming the goldfish and black cat had already been introduced into the discourse in previous clauses):

zhei zhi jinyu qiaqiao shi zai zhoumian shang.
this CL goldfish happen be at table on
Keshi hei mao mei you faxian **zhe zhi jinyu**,
but black cat not have notice **this CL goldfish**
'this goldfish happened to be on [the] table, but [the] black cat didn't notice this fish' (Hedberg, 1996:12)

3.3 Korean

In Korean, a series of postpositional affixes are used to signal the passage of referents from ‘new’ to ‘given’ information, although these affixes signal grammatical role rather than givenness, unlike the function of definite/indefinite articles in English (see Kang, 2004). Referents are usually marked with the subject marker ‘이/가’- ‘ee/ga’ in the case of referent introduction or low subsequent-mention accessibility, through to ‘은/는’- ‘un/nun’ topic marking for mid-accessible referents, through to no marking (accompanying zero anaphora) to highly accessible referents:

한	남자**가**	왔어요.		남자**는**	문을
han	namja-ga	wasseoyo.		namjaneun	muneul
A	man-**SUB**	came-PAST-DEC-POL.		Man-**TOP**	door-OBJ
열었어요,			**O** 거울을	봤어요.	
yeoreosseoyo,			O geoureul	bwasseoyo.	
open-PAST-DEC-POL,			**O** mirror-OBJ	look-PAST-DEC-POL.	

‘A man came. (The) man opened the door, and looked in the mirror.’

In Korean, the most common high accessibility marker is the zero anaphor (as seen in the example above). Korean is similar to Mandarin in that referents are often referred to with zero anaphors when inferable. Korean has a system of third person pronouns marked for gender (그 – geu, ‘he’, 그녀 – geunyeo, ‘she’), but these are rarely used in Korean, with Koreans preferring titles and kinship terms for honorific purposes (Brown, 2011). Koreans, when referring to a referent of a particular social status (such as a teacher, or general, etc.) would refer to that referent by their title only (without other kinds of definite markings such as demonstratives) and would not use third person pronouns to refer to that referent, unlike more generic referents such as ‘a boy’, who may receive overt pronominal reference.

The most common low accessibility marker in Korean is the bare noun. Demonstrative + noun combinations are occasionally used, although to a lesser extent than in Mandarin, due again to the preference for titles and kinship terms described above.

3.4 Summary

The differences in strategy used to maintain reference between these L1s have caused Huang (2000) to label English as a 'syntactic' language, and Mandarin and Korean 'pragmatic' languages. Under this definition, Mandarin and Korean have a far higher rate of zero anaphor than English (the resolution of which has to be inferred), Chinese style 'topics' are linked to their associated comment pragmatically, and Mandarin and Korean make little use of morphology to signal (in)definiteness, at least compared to the L1 English article system.[2] Given the lack of congruence between these three first languages in terms of reference maintenance, it is now necessary to explain how this lack of congruence might affect the acquisition of referring expressions in a second language.

4. L2 development and reference management

As seen above, Mandarin and Korean use different strategies to introduce and maintain reference than L1 English. Mandarin and Korean speakers of English as a second language have to change their L1 referential strategy from a 'pragmatic' mode (categorised as the prevalence of bare and zero forms in L1 Korean and Mandarin) to the 'syntactic' mode required of English (Huang, 2000). To do so, they must acquire the appropriate syntactic features of English, in the form of grammatical articles on the noun, as well as the use of overt pronouns. It has been documented that learners from pro/topic drop languages may omit overt pronouns required in the L2 at lower proficiency levels (Gundel & Tarone, 1983; White, 1986). It is also claimed that L2 learners then become over-explicit in reference at intermediate proficiency levels, using full NPs where pronouns are expected (Hendriks, 2003; Kang, 2004). Articles, in particular, are claimed to be particularly difficult to learn, in that the English article system is claimed to be 'a complex set of abstract distinctions which are, to some extent, arbitrarily mapped onto surface forms' (Ekiert, 2007:1). Learners looking for a one-form to one-function mapping for new/given reference find it difficult to do so with English articles, which have a

[2] Although some researchers argue that these differences are, in fact, syntactic in nature, see Zribi-Hertz, 2009.

variety of form-function mappings (Ekiert, 2007:2). For example, L2 learners from L1s without an article system have trouble assigning features of definiteness and/or specificity with the article system, and overuse indefinite articles where definite articles are expected (Ionin, Ko & Wexler, 2004). While miscommunication is certainly possible in native L1 to native L1 communication (Ryan, 2012), it is more likely to occur in a second language context, given the potential for L1 transfer of referential strategies into the L2, and the time and effort it takes to acquire the appropriate L2 target-like syntactic markings for cohesive referent introduction and maintenance. This data on errors collected in the present study allows us to see how each factor affects L2 reference, and whether the effect of each factor is eventually overcome by these L2 learners.

By performing a cross-sectional analysis of the L2 learner's attempts to produce the appropriate referential forms in English, the existence of errors such as missing and/or incorrect anaphors and determiners should allow the researcher to pinpoint the stage of development where the learner eventually masters the 'syntactic' mode of reference required in the target-language. Therefore, the following research questions are proposed:

1) Will there be a developmental pattern in error types of reference in the L2 English production of L1 Mandarin and L1 Korean speakers?
2) When will L2 English learners from L1 Mandarin and L1 Korean backgrounds acquire the appropriate L2 target-like 'syntactic' use of particular referring expressions without error? Will either group eventually manage native-like proficiency in this respect?

It is predicted in this research that at earlier proficiency levels, the L1 'pragmatic' strategy employed for reference by Mandarin and Korean learners of English should result in a significant number of references coded as being errors of missing anaphor, where overt anaphoric expressions are expected. Once the learners are aware that obligatory reference is required, they may then struggle with the L2 'syntactic' article system, from omitting obligatory determiners, to supplying the incorrect form of the determiner, or overusing determiners where determiners are not required. As the English article system is so complex, the L2 learners are not expected to master this system until quite late in acquisition, and some errors may

persist even in very advanced learners, in certain cases.

Moreover, it is argued in Li & Thompson (1976, 1989), Hedberg (1996) and Chen (2004) that the proximal demonstrative这 – zhè - is also increasingly being used to signal definiteness, in a manner similar to the definite article in English. While demonstratives may be used to signal definiteness in Korean, it is not claimed that Korean is acquiring an English-like article system to the extent that the same claim is made for Mandarin. If this is the case, then it is conceivable that those from Mandarin-speaking backgrounds already have (at least part) of the functional features of the English article system marked in their L1, and therefore may have *less* difficulty mapping L2 syntactic form to pragmatic function. This leads to the final research question below:

3) Will L1 background be a factor in the frequency and type of errors made in L2 reference production?

5. Study 1 – Referential errors in the Cambridge Learner Corpus

The CLC contains data taken from learner production on a wide range of Cambridge ESOL exams, including IELTS, FCE, CAE, and CPE tests. The data covers levels A1-C2 of the Common European Framework for Languages (CEFR), which is the standard criterion-based framework of English proficiency levels for learning, teaching and assessment developed by the Council of Europe (2001a) (see Nicholls, 2003; Alexopolou, 2008; or Hawkins & Buttery, 2008; Salamoura & Saville, 2010, for a detailed description of how this was achieved). A summary of the CEFR levels is found below:

Table 1. Summary of CEFR levels

CEFR Level	Comparison with other standard assessments
A1 (Basic user)	IELTS 1-2.5, TOEIC 0-250, TOEFL(IBT) 9-29
A2 (Basic user)	IELTS 3, TOEIC 255-400, TOEFL(IBT) 30-40
B1 (Independent user)	IELTS 3.5-4.5, TOEIC 401-500, TOEFL(IBT) 41-52
B2 (Independent user)	IELTS 5-6, TOEIC 501-640, TOEFL(IBT) 53-78

C1 (Proficient user)	IELTS 6.5-7, TOEIC 640-780, TOEFL(IBT) 79-95
C2 (Proficient user)	IELTS 7.5-9, TOEIC 785-990, TOEFL(IBT) 96-120

This large collection of learner data is filtered according to different searchable criteria, including the proficiency level of the learner, task type (essay, narrative etc.) or the first language of the learners. Users of the corpus are able to see precisely what aspects of language are acquired at key stages of a learners' development, and the differences in how learners of different L1s structure their L2 English.

5.1 Sample

Narratives were chosen as the source of investigation as they are considered a 'basic' discourse form, 'acquired early in all cultures and integral to all ages' (McCabe & Bliss, 2003; Kang, 2005), and provide 'important information about the narrator's linguistic competence and pragmatic sensitivity in the target language' (Kang, 2005:260). The texts selected from the corpus were texts that included a narrative structure in the main answer. The texts selected were either clearly narrative tasks (in the case of B2/C2 texts) or were personal/business letters (B1/C1) that had within them a narrative structure. This led to the exclusion of low proficiency (A1-A2) data from the analysis, as narrative tasks were not found within those datasets. Only the narrative part of the text was analysed in the case of personal/business letters, discounting headers/footers/salutations or direct reference to the reader of the letter (such as 'I hope you are well') from the word count. B1 texts were taken from the PET exam suite, B2 texts were taken from the FCE exam suite, C1 texts from the CAE exam suite and C2 texts from the CPE exam suite. The number of words and the number of texts analysed for each proficiency level are given below:

Table 2. Word counts analysed in CLC data

Level	Word Count - Mandarin L2 English	Word Count - Korean L2 English
B1	2844 (20 texts)	3173 (20 texts)
B2	12570 (39 texts)	8807 (42 texts)
C1	3626 (11 texts)	3132 (11 texts)
C2	17500 (36 texts)	14733 (33 texts)

Only reference to animate discourse referents was analysed as these were more likely to be referred to again after their introduction. Reported speech between characters was not included in the word counts, nor was it included in the error analysis. References at the end of reported speech, such as “”…” said *John*” were included in the counts and the analysis. C1 level texts were severely underrepresented in the data, compared to B2 and C2 level texts. Due to the wording of the question often containing the referential form to be used for the first-mention of a referent, only subsequent-mention errors have been included here in the analysis.

5.2 Error coding of CLC data

The Cambridge Learner Corpus is coded for errors that learners produce in their texts. Of importance to reference management, the frequencies of each type of the following error were collected:

Table 3. Error codes used in CLC corpus

Kind of Error	Code	Example
Missing Anaphor (a word is needed for the construction to be complete)	ERMA	793215.0 In addition<NS type=”MA”>|**I**</NS>hope you can accept my application
Missing Determiner (a word is needed for the construction to be complete)	ERMD	563553.0 We especially recommend this hotel because you can find special offers there all through<NS type=”MD”>|**the**</NS>year
Replace Incorrect Anaphor (the form exists and is appropriate but the choice of word is wrong)	ERRA	4285.0 For example, I made one trip to Karachi City,<NS type=”RA”>**it|which**</NS> is a major city in Pakistan.
Replace Incorrect Determiner (the form exists and is appropriate but the choice of word is wrong)	ERRD	563567.0 Johnny always said that if any of us wanted to preserve<NS type=”RD”>**the|our**</NS><NS type=”RN”>relationship
Unnecessary Anaphor (A valid word has been used, but its presence makes the sentence incorrect)	ERUA	720907.0 I also wanted to know if I have to take some money to buy souvenirs and food or will that be paid for<NS type=”UA”>**me**</NS>, too?

Unnecessary Determiner (A valid word has been used, but its presence makes the sentence incorrect)	ERUD	563552.0 We decided to find<NS type="UD"> **a**</NS>proper accommodation

5.3 Analysis

As the data is not normally distributed (significant Levene's and Shapiro-Wilk tests) when comparing the total number of errors per text per level, the data was organised to look at the type of errors per individual reference as a binary set per level. In doing so, it was possible to compare the likelihood of an L2 group making a particular error type at a particular CEFR level with the likelihood of another L2 group doing so, through logistic regression analysis. Each analysis passed goodness-of-fit statistics.

5.4 Results

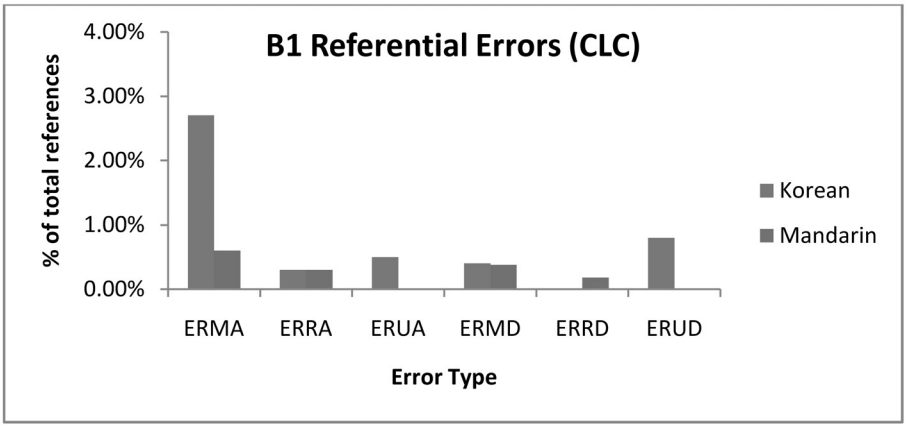

Figure 1. Referential errors found in CLC corpus at B1 level

At B1 level, the Koreans had a significantly higher percentage of missing anaphor (ERMA) in their narratives, being 1.3 times as likely than the Mandarin group to produce an error of this type under logistic regression analysis[3] (β=0.335, Wald

[3] β = *slope of regression,* Wald = *strength of regression,* Sig. = *significance of independent variable,* EXP(β) = *odds ratio.*

=4.790, Sig. = p<0.05, EXP(β=1.379), as well as a higher number of errors tagged as unnecessary anaphor (ERUA) and unnecessary determiners (ERUD) as the Mandarin group did not make any such errors. The average number of references per text between groups (Korean = 22, Mandarin = 26) was not significantly different according to *t*-test comparison.

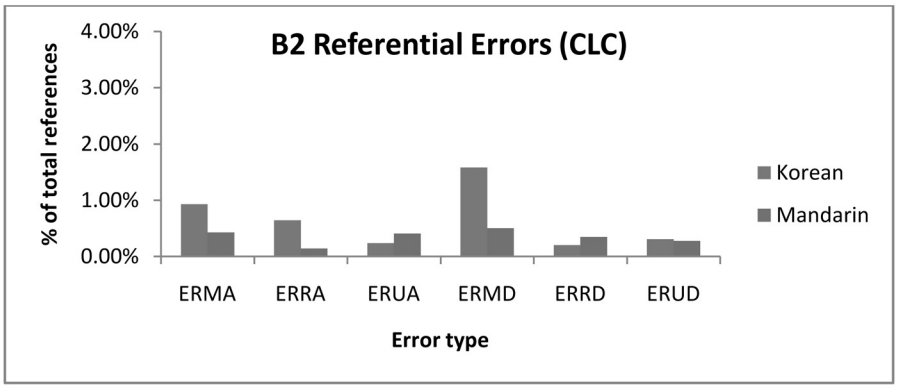

Figure 2. Referential errors found in CLC corpus at B2 level

At B2 level, errors of missing determiners (ERMD) are the most common type in both groups, rather than the errors of missing anaphor found at B1. The Korean group has a significantly higher number of referring expressions tagged for ERMD, and are 1.5 times more likely to produce this kind of error than the Mandarin group according to a logistic regression analysis (β=0.276, Wald=4.441, = p<0.05, EXP(β =1.317). They are also 1.6 times more likely to make errors of incorrect anaphor (ERRA) (β=0.519, Wald=4.447, p<0.05, EXP(β=1.681). The average number of references per text between groups (Korean = 30.4, Mandarin = 34.7) was not significantly different according to *t*-test comparison.

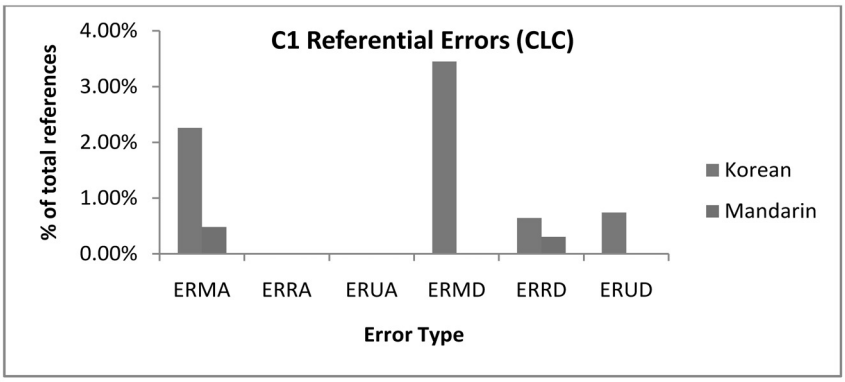

Figure 3. Referential errors found in CLC corpus at C1 level

At C1 level, errors of missing determiners (ERMD) are common in the Korean group, along with a higher number of missing anaphor (ERMA) errors within the Korean group from B2 level (although this may be a symptom of the low number of texts for C1, rather than evidence of a developmental rise of this error type with proficiency). The Korean group has a higher number of referring expressions tagged for ERMD and ERMA than the Mandarin group, with the Mandarin group making very few errors of any type within this dataset. However, the results of a linear regression analysis do not show any significant differences between the two groups, due to the low number of narrative texts found at this proficiency level (n=11 for both groups). The average number of references per text between groups (Korean = 21.2, Mandarin = 21.8) was not significantly different according to t-test comparison.

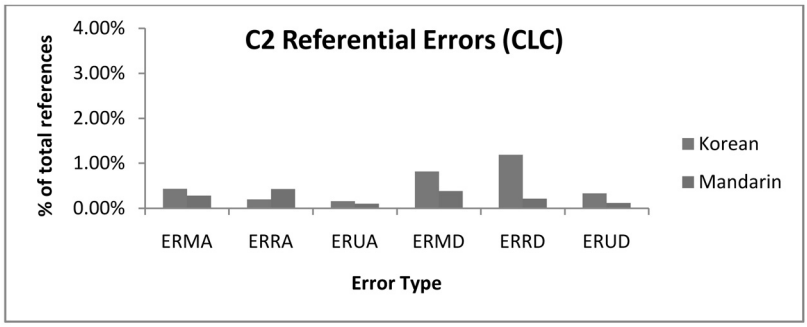

Figure 4. Referential errors found in CLC corpus at C2 level

At C2 level, errors of incorrect determiners are still found in the Korean group, while the Mandarin group maintain a negligible number of missing determiner errors. The Korean group has a significantly higher number of referring expressions tagged as having incorrect determiners (ERRD), and were 1.8 times more likely to produce this kind of error according to a logistic regression analysis (β=*.608,* Wald=4.964, p<0.05, EXP(β=1.837) than the Mandarin group, although again this a very small amount of errors overall. They are also 1.5 times more likely to make references tagged as having missing determiners (ERMD) (β=0.431, Wald=4.578, p<0.05, EXP(β=1.538). The average number of references per text between groups (Korean = 54.71, Mandarin = 62.28) was not significantly different, after a t-test comparison (t = -1.438, df = 31, p=.160).

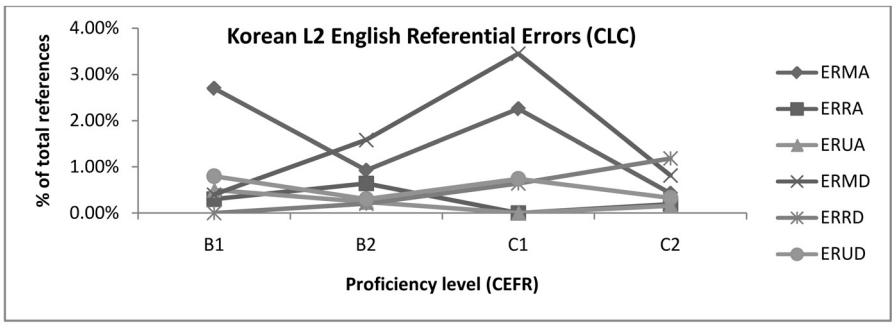

Figure 5. Referential errors found in CLC corpus across CEFR levels by Korean L2 English learners

The Korean L2 English data seems to follow a developmental trajectory of errors, with particular error types found more commonly at certain CEFR proficiencies than other error types. At B1 level, errors of missing anaphor are common, then as learners learn not to freely omit referents in English from B2 level, the number of references tagged as having missing determiners increases, peaking at C1 level. The high figure for missing anaphor at C1 may be explained by the low text count at that proficiency level, as the expected trend was downward from B2 level onwards. At the highest level of proficiency (C2), the Korean L2 English learners are managing to use overt references when required, with an accompanying determiner, but the number of referring expressions tagged as having incorrect determiners (ERRD) is at its highest at this proficiency.

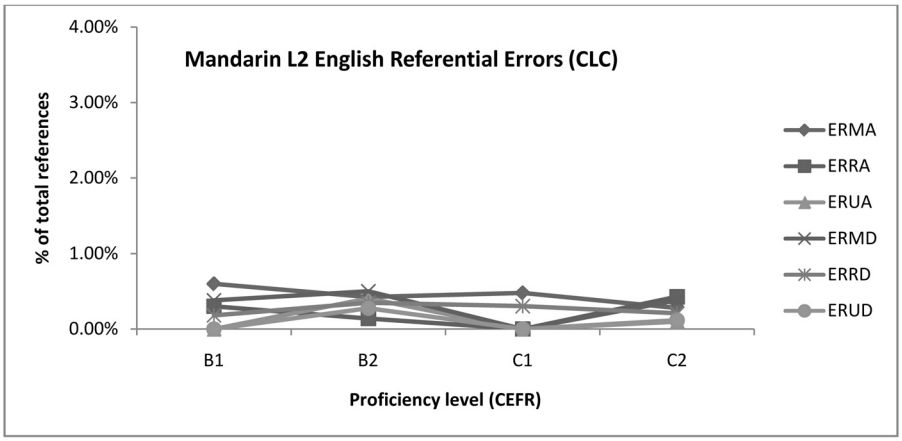

Figure 6. Referential errors found in CLC corpus across CEFR levels by Mandarin L2 English learners

The Mandarin L2 English data also seems to follow a developmental trajectory of errors linked to CEFR proficiency, although the figures are too small to state this clearly. At B1 level, errors of missing anaphor are found, as with the Korean L2 English data (although significantly fewer). At B2 level, a very small (<1%) number of references are tagged as having missing or incorrect determiners, and this error type is largely eliminated by C1 level. The only error type that remains in the Mandarin data at C2 level are references tagged as having the incorrect anaphor form ('he' instead of 'she', for example). In Mandarin, L1 transfer may account for the late prevalence of this error type, as third person pronouns in spoken Mandarin are gender neutral, while there are separate pronominal forms in the Korean data. However, the number of errors of this kind are still very small (<1%), and are not significantly higher in frequency than the same kind of error produced by the Korean L2 English group.

5.5 Discussion

The CLC data points towards a developmental trajectory of error frequency and type found in both L2 groups, in support of research question one. However, the number of errors overall is very small, particularly at C2 level, and as such it is difficult to really determine whether this is due to development or simply is indicative of errors

that a native speaker might make for written texts. The pattern of errors starts from a number of missing anaphor present in the texts, then to a number of missing determiners, then to the incorrect form of determiners being used for referring expressions. This suggests a process of gradual L2-target-like syntacticization from the L1 referential strategies found in Mandarin and Korean. The unnecessary use of anaphor or determiners was limited even at low proficiency levels.

For research question three, the data appears to point towards the Mandarin group generally making fewer referential errors that their Korean counterparts, with some significant differences between the two groups in the form of missing anaphor errors, and missing or incorrect determiner errors at different stages of language acquisition.

5.6 Complications with CLC analysis

Given that A1-A2 level narrative data was unavailable in the CLC, and that C1 narrative data is underrepresented in the corpus, the analysis of the CLC data is perhaps inconclusive with regards to research question 2, about *when* learners acquire L2 reference.

Another main issue with the analysis of the narrative texts in the CLC was one of perspective. The B1 and C1 texts were only told from a 1st person perspective, while B2 and C2 texts had a mix of 1st and third person perspectives, which has implications for the number of pronouns that a text might receive compared to the number of full noun phrases. The word counts for each perspective are shown below (*n*= number of narrative texts analysed at that level for each language).

Table 4 and 5. Word counts by perspective in CLC narratives

1st person perspective CLC narratives

Level	Korean	Chinese
B1	2844 (n=20)	3173 (n=20)
B2	2812 (n=12)	2192 (n=9)
C1	3132 (n=11)	3626 (n=11)
C2	2693 (n=6)	3616 (n=6)

3rd person perspective CLC narratives

Level	Korean	Chinese
B2	5995 (n=30)	7534 (n=30)
C2	12040 (n=27)	13884 (n=30)

The choice of perspective was influenced by the explicit wording of the question

offered to the learners before they began to write their narratives. For example, one of the questions for the B1 level texts was phrased as follows:

> "Write a letter to a friend with the first line 'I must tell you what happened to me recently"

Von Stutterheim & Klein (1989), in the *quaestio* (or 'implicit question') model, propose that the structure of a text 'is constrained on both global and local levels by the nature of the question which the text in its entirety is produced to answer' (1989:41). This has important implications for reference, and different *quaestiae* posed (e.g. what happened? vs. what happened to X?) have been shown to make speakers produce different referential forms depending on what the person asking the question wanted to know (Campbell, Brookes & Tomasello, 2000; Matthews & Lieven, 2005). As the question posed in the exam prompt is explicit in the perspective or referential form to be used, it is difficult to compare the referential forms produced as the result of one *quaestio* with the forms produced in an alternative *quaestio*, if concrete generalizations are to be made from the results.

The final complication with the CLC analysis was that the data was entirely composed of *written* narratives produced under exam conditions. Under such conditions, the writer has a chance to correct their production, and is likely more careful to do so. If one is to understand the cognitive difficulties faced by the second language user in producing the appropriate L2 referential forms, it is necessary for *spoken* narrative data to be analysed.

While the CLC data was useful for hypothesis testing, a new set of learner data was collected using a narrative picture sequence, collecting oral narrative data, and where the *quaestio* was controlled for, allowing for a greater degree of generalization to be drawn from the results.

6. Study 2 – Controlled learner narrative corpus

6.1 Method

Narrative data was collected from L2 English learners from Mandarin and Korean

backgrounds using a narrative picture sequence as the elicitation device. The materials were designed to capture a variety of first and subsequent-mention phenomena while following a standardized narrative structure, using an approach known as the 'story grammar model', taken from Stein & Glenn (1979). Within this structure, a main character tries to perform an action (such as playing with a ball), with the story focusing on how the character manages to perform that action in the face of some difficulty to overcome, and finishing with the character finally able to perform the action. The materials are shown below:

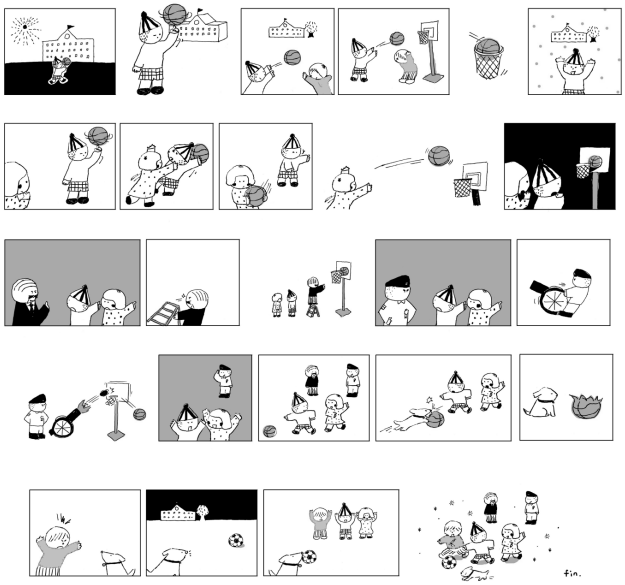

Figure 7. The elicitation materials for experiment 2

For the *quaestio*, the participants were told that they should try to make a story with a 'beginning, middle and end', and were also told to make the story 'as interesting as possible' so that the listener would enjoy it. This meant that the speaker was free to produce referential forms without any referential form explicitly given in the *quaestio*.

6.2 Sample

Sixty native Koreans and fifty-eight native Mandarin Chinese adults provided the L2 English narratives used in the L2 study. These participants were all pre-tested for English proficiency using the Oxford Quick Placement Test, a 30 minute multiple choice standardized language test, which can be used to ascertain a participant's CEFR level from A1-C2. Ten participants from the Korean group for each CEFR level (n=60) joined the study, with ten participants per level in the Mandarin group from A1-B2 levels, and nine each for C1 and C2 levels (n=58). The total number of words in the narratives is shown below:

Table 6. Word counts in corpus for experiment 2

Level	Word count - Mandarin L2 English	Word Count - Korean L2 English
A1	1884 (10 texts)	1526 (10 texts)
A2	2516 (10 texts)	1808 (10 texts)
B1	3197 (10 texts)	2633 (10 texts)
B2	3286 (10 texts)	3148 (10 texts)
C1	4016 (9 texts)	3440 (10 texts)
C2	3559 (9 texts)	4154 (10 texts)

The participant's narratives were digitally recorded and later transcribed in the CLAN program using a simplified version of the CHAT transcription system from the CHILDES database (MacWhinney, 2000). This program provides basic corpus analysis tools such as frequency counts and concordances. The analysis procedure followed the CLC study (reference to animate referents, no reported speech etc). As with the CLC data, the data is not normally distributed (significant Levene's and Shapiro-Wilk tests) when comparing the total number of errors per text per level, and as such the data was organised to look at the type of errors per individual reference as a binary set per level, allowing for logistic regression analysis. Each analysis passed goodness-of-fit statistics.

6.3 Error coding of controlled corpus data

The error coding followed the forms found in the Cambridge Learner Corpus. The following additional codes were also included to capture a wider range of potential

errors that were not covered in the CLC data:

Table 7. Additional error codes in study 2

INF	Problem with plural inflection	Grammatically unacceptable	**Two boy** (s) arrived.
REFAMB	Ambiguous reference	Ambiguity	A boy arrived, then another boy also arrived. A girl then arrived. **The boy** said hello.
REPAIRR	Repair mid-stream	Hesitation	A boy arrived. **She - - he** said hello.

While these codes are not directly comparable with the CLC data, it was considered useful additional evidence in support of research question 1.

6.4 Results

In general, there is a developmental trajectory of error forms found in subsequent-mention referring expressions in both L2 groups. The following figure shows the distribution of error types by proficiency level for the Mandarin L2 English group:

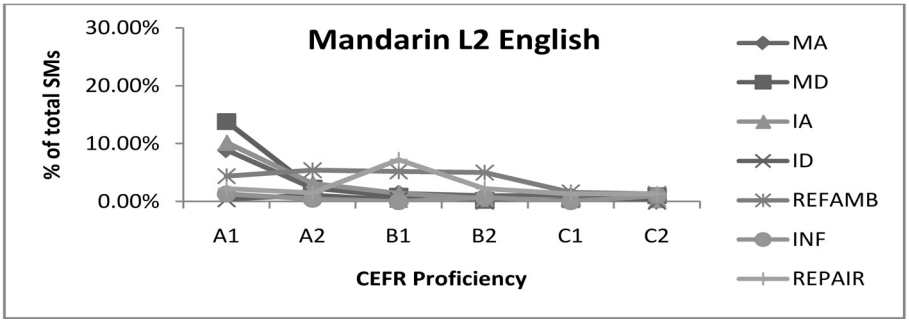

Figure 8. Distribution of inappropriate subsequent-mention NPs in Mandarin data (as % of total subsequent-mentions)

At A1 level, problems with missing anaphor (MA), missing determiners (MD), and incorrect anaphor selection (IA - such as 'he' where a female referent is being referred to) account for an average 10% respectively of the total number of

subsequent-mentions made. By A2 level, these types of inappropriate NP forms are largely gone from the learner's production. Replacing the IA error type at A1/A2, there is a spike of self-repairs (REPAIR) at B1 level, almost all of which are where personal pronouns are to be used, with the learner hesitating between 'he'/'she' but eventually settling on the appropriate form for the particular referent they are referring to. By B2 level, the self-repairs have stopped as the learners become more confident and more automatic in their selection of the appropriate pronominal form. However, from A2-B2 level around 5% of the total subsequent-mention references were ambiguous and the identity of the referent could not be resolved (REFAMB). By C1 level, inappropriate NP forms of any kind were rare. Problems with incorrect determiners (ID), or inappropriate pluralisation (INF) are rare even at the lowest levels in subsequent-mention reference in this L2 group, and errors of unnecessary anaphor and determiners (UA/UD) are limited to one or two throughout the entire dataset.

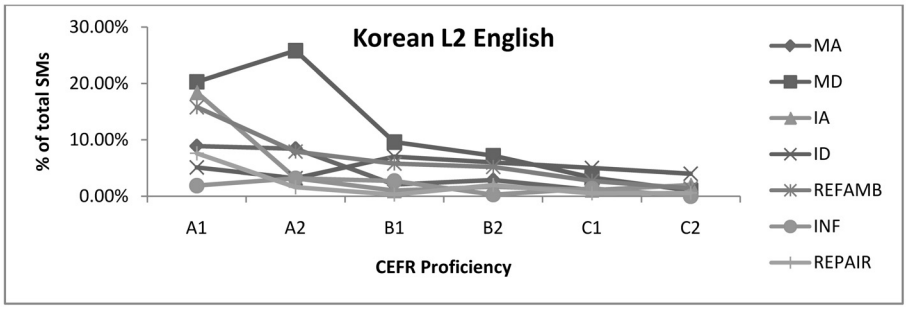

Figure 9. Distribution of inappropriate subsequent-mention NPs in Korean data (as % of total subsequent mentions)

In the Korean L2 English dataset, missing determiners (MD) are a major feature of the subsequent-mentions made by this group, and this type of error persists until C1 level. At A1-A2 levels, issues with missing anaphor (MA), incorrect anaphor (IA) and ambiguous reference (REFAMB) are common, accounting for around 10% each of the total number of subsequent-mentions made. By B1-B2 level, as the number of missing determiners falls, the number of incorrect determiners (ID) rises to just under 10% of all subsequent-mentions made at those levels. The number of self-repairs (REPAIR) is small overall, as is the number of inappropriately pluralised

references (INF).

6.5 Summary of subsequent-mention error forms comparing the two L2 groups

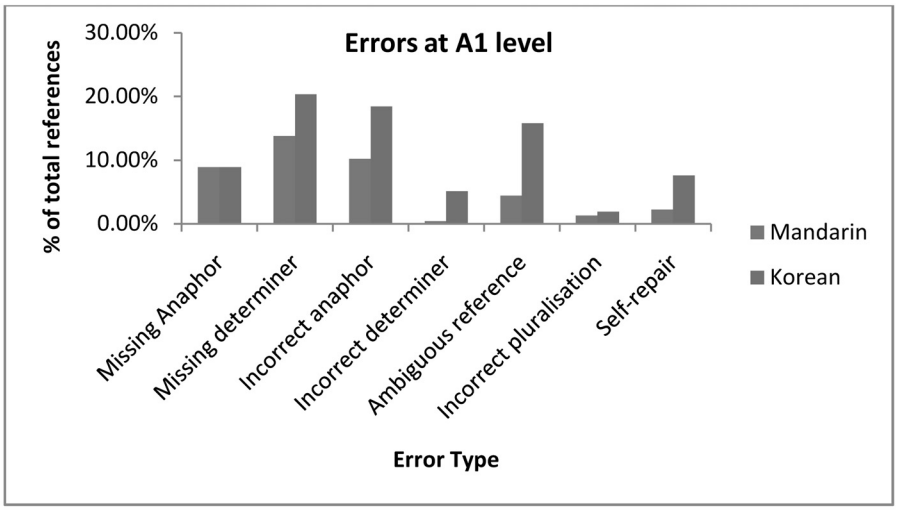

Figure 10. Distribution of errors at A1 level between L2 groups (as % of total referring expressions)

Figure 10 represents the frequencies of different error types found at A1 level, which is where the majority of errors were made by both L2 groups. A series of logistic regression analyses with L1 background as the predictor of errors by type found that the Korean L2 English group were twice as likely to make errors of incorrect anaphor choice (IA) than their Mandarin counterparts at A1 level (β=*.680,* Wald=5.104, p<0.05, EXP(β=1.974). The Korean group were almost 12 times as likely to use the incorrect determiner when producing reference than their Mandarin counterparts (β=2.480, Wald=5.415, p<0.05, EXP(β=11.947), and they were four times more likely to produce an ambiguous reference than their Mandarin counterparts (β=1.397, Wald=12.817, p<0.01, EXP(β=4.041).

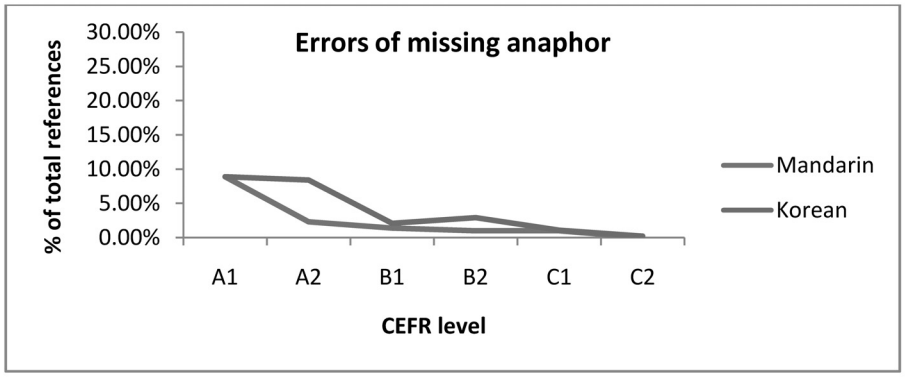

Figure 11. Distribution of missing anaphor errors between L2 groups across CEFR (as % of total referring expressions)

Figure 11 shows the distribution of errors of missing anaphor between Mandarin and Korean L2 English learners across each CEFR proficiency level. The Korean L2 English group has a higher number of errors of missing anaphor across the CEFR proficiencies, and are almost four times more likely to produce such an error than their Mandarin counterparts at A2 level, where the gap between the two L2 groups is the greatest (β=1.136, Wald=7.779, p<0.05, EXP(β=3.908).

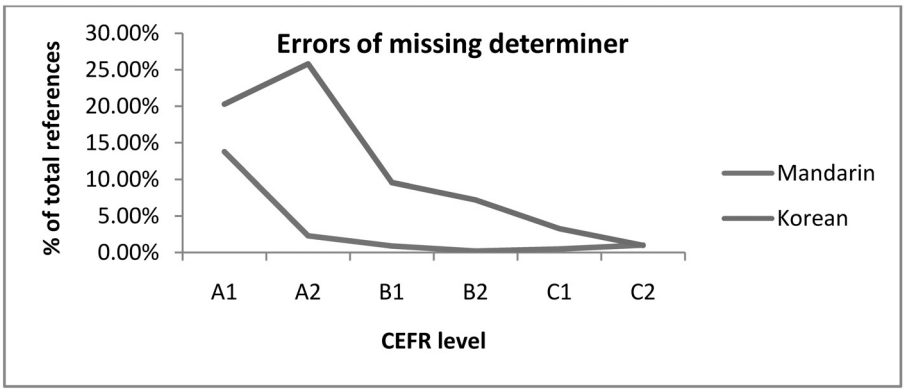

Figure 12. Distribution of missing determiner errors between L2 groups across CEFR (as % of total referring expressions)

Figure 12 shows the distribution of errors of missing determiners between Mandarin and Korean L2 English learners across each CEFR proficiency level. The Korean L2

English group has a higher number of errors of missing determiners across the CEFR proficiencies. They are fourteen times more likely to produce such an error than their Mandarin counterparts at A2 level (β=2.963, Wald=36.599, p<0.01, EXP (β=14.770), twelve times more likely to do so at B1 level (β=*2.499,* Wald=16.622, p<0.01, EXP(β=12.172), thirty times more likely to do so at B2 level (β=3.343, Wald=11.270, p<0.01, EXP(β=30.960), and six times more likely to do so at C1 level (β=1.866, Wald=5.924, p<0.05, EXP(β=6.472).

6.6 Discussion

From both studies, a clear developmental trajectory of errors was found in support of research question one, from a L1 like 'pragmatic' use of reference at lower proficiencies where referents were frequently omitted when obligatory in an English context, to the learner struggling to assign determiners to these obligatory references, to errors assigning the correct determiner at higher proficiency levels. For research question two, errors of reference are largely absent from the learner data by C1 level, although a small number of errors are still found even at the highest level (C2). Finally, for research question three, it was shown that at every proficiency level of the CEFR from A1 to C2, those from L1 Mandarin-speaking backgrounds produced fewer referential errors than their L1 Korean counterparts.

As a potential explanation for the findings of research question two, the fact that even high-level (C1-C2) language users still show some errors in reference production may be explained by theories of language acquisition that claim that the interaction of *language-internal interface* properties and *language-external interface* constraints make conditions difficult for language acquisition. Reference may be particularly vulnerable to the internal/external division of labour, given the claims of Accessibility Theory as well as the effects of the *Quaestio* model. Interface theories of language acquisition include the Interface Hypothesis (Sorace & Felice, 2006), and Interlanguage Pragmatics (Kasper & Rose, 2002), that suggests that elements of language acquisition that rely on both language-internal and language-external factors (such as the management of coherent reference) are more difficult to acquire that language-internal-only features of language (such as mapping phonology to semantic meaning). The method of investigation of interface difficulties in language acquisition has primarily been conducted through traditional data collection methods

such as grammaticality judgement tasks (Slabakova, Rothman, Mendez, Campos & Kempchinsky, 2011) or context sentence matching tasks (Iverson & Rothman, 2008), but so far investigation has not been fully extended to the field of corpus linguistics, and this could therefore be a potentially interesting approach to language acquisition for corpus linguists.

In terms of research question three, the better performance of the Mandarin L2 English group over the Korean L2 English group at each proficiency level requires an explanation, given that both languages employ markedly different strategies for referent introduction and maintenance from English, but are quite similar to each other in terms of the referring expressions used to maintain reference (zeros, bare nominals). The marked success of the Mandarin group in assigning determiners to subsequent-mention references over their Korean counterparts, and across CEFR proficiencies, appears to be firm evidence that these learners have less difficulty in mapping target-like syntactic form (articles) to pragmatic function (reference management). The Korean learners are unable to assign obligatory determiners to noun phrases in both first and subsequent- mentions until C2 level in the corpus for experiment two, while the Mandarin group are able to do this roughly by B1 level. The Mandarin group are aware that determiners have to be used to signal discourse-newness or discourse-oldness, and if Mandarin is shifting diachronically towards an English-like article system, as suggested in Li & Thompson, (1976, 1989), Hedberg (1996), and Chen (2004), then this would be the likely reason for their success. Further research is needed to see whether a correlation can be found between the increased use of numerals and demonstratives as indefiniteness/ definiteness markers in Mandarin narratives, and an increase in L2 referential proficiency in terms of the appropriate use of English articles for first- and subsequent-mention NPs.

7. Conclusion

Learners at different CEFR levels make different kinds of referential errors as they produce narrative discourse, which may damage the cohesion and coherence of that discourse. While errors of reference are largely overcome by L2 learners by C2 level of the CEFR, there is also a clear effect of L1 background on the frequency of

errors found at each CEFR level between Mandarin and Korean L2 English learners. While the existence of these errors is not the whole story when considering the cohesion and coherence of L2 discourse, the data provided in these studies should be of use to researchers working on these languages and to corpus linguists interested in the study of reference in the L2. Future research needs to consider reference that is syntactically/semantically appropriate but is pragmatically infelicitous, such as over-explicit full NP reference where pronominal reference is required in the L2.

References

Alexopoulou, T. 2008. Building new corpora for English Profile. *Research Notes* 33: 15-19.

Ariel, M. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics* 16(5): 443-463.

Ariel, M. 2008. *Pragmatics and grammar*. Cambridge: Cambridge University Press.

Ariel, M. 2010. *Defining pragmatics*. Cambridge: Cambridge University Press.

Bach, K. 2008. On referring and not referring. In Gundel, J. K. and Hedberg, N. (eds.), *Reference: Interdisciplinary perspectives,* 13-58. New York: Oxford University Press, USA.

Brown, L. 2011. *Korean honorifics and politeness in second language learning*. Amsterdam: John Benjamins.

Campbell, A. L., Brooks, P., and Tomasello, M. 2000. Factors affecting young children's use of pronouns as referring expressions. *Journal of Speech, Language And Hearing Research* 43(6): 13-37.

Chafe, W. L. 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production* (Advances in Discourse Processes Vol. 3). Norwood, NJ: Ablex Publishing Corporation.

Chen, P. 2004. Identifiability and definiteness in Chinese. *Linguistics* 42(6): 1129-1184.

Clancy, P. M. 1980. Referential choice in English and Japanese narrative discourse. In Chafe, W. L. (ed.), *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production* (Advances in Discourse Processes Vol. 3), 127-201. Norwood, NJ: Ablex Publishing Corporation.

Clark, H. H. 1975. Bridging. In Schank, R.C. and Nash-Webber, B. L. (eds.), *Theoretical issues in natural language processing*, 169-174. New York: Association for Computing Machinery.

Council of Europe. 2001. *Common European framework of reference for languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

Ekiert, M. 2007. The acquisition of grammatical marking of indefiniteness with the indefinite article *a* in L2 English. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics* 7(1): 1-43.

Gundel, J. K., Hedberg, N., and Zacharski, R. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2): 274-307.

Gundel, J. K., and Tarone, E. 1983. Language transfer and the acquisition of pronominal anaphora. In Gass, S. M. and Selinker, L. (eds.), *Language transfer in language learning*, 87-101. Amsterdam: John Benjamins.

Hawkins, J., and Buttery, P. 2009. Using learner language from corpora to profile levels of proficiency - insights from the English profile programme. In Taylor, L. and Weir, C. (eds), Language testing matters: *Investigating the wider social and educational impact of assessment − proceedings of the ALTE Cambridge conference, April 2008*, 158-176. Cambridge: Cambridge University Press.

Hedberg, N. 1996. Word order and cognitive status in Mandarin. In Fretheim, T., and Gundel, J. K. (eds.), *Reference and referent accessibility*, 179-192. Amsterdam: John Benjamins.

Hendriks, H. 2003. Using nouns for reference maintenance: A seeming contradiction in L2 discourse. In Ramat, G. (ed.), *Typology and second language acquisition*, 291-326. Berlin: Mouton de Gruyter.

Huang, Y. 2000. *Anaphora: a cross-linguistic approach*. New York: Oxford University Press, USA.

Ionin, T., Ko, H., and Wexler, K. 2004. Article semantics in L2 acquisition: The role of specificity. *Language Acquisition* 12(1): 3-69.

Iverson, M., and Rothman, J. 2008. The syntax-semantics interface in L2 acquisition: Genericity and inflected infinitive complements in non-native Portuguese. In Bruhn de Garavito, J., and Valenzuela, E. (eds.), *Selected Proceedings of the 10th Hispanic Linguistic Symposium,* 78-92.

Kang, J. Y. 2004. Telling a coherent story in a foreign language: analysis of Korean EFL learners' referential strategies in oral narrative discourse. *Journal of Pragmatics* 36(11): 1975-1990.

Kasper, G., and Rose, K. R. 2002. *Pragmatic development in a second language*. Oxford/ Malden, MS: Blackwell.

Li, C. N., and Thompson, S. A. 1976. Subject and topic: A new typology of language in subject and topic. In Li, C. N. (ed.), *Subject and topic*, 457-489. London/New York: Academic Press.

Li, C. N., and Thompson, S. A. 1989. *Mandarin Chinese: A functional reference grammar*. California: University of California Press.

Li, W. 2004. Topic chains in Chinese discourse. *Discourse Processes* 37(1): 25-45.

MacWhinney, B. 2000. *The CHILDES project: tools for analyzing talk, Volume II: The data-*

base (Vol. 2). Mahwah, NJ: Lawrence Erlbaum.

Matthews, D. E., and Lieven, E. 2005. *Constructivist investigation into the development of word order and reference*. Manchester, UK: University of Manchester Press.

Nicholls, D. 2003. The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT, in Archer, D., P. Rayson, A. Wilson, and T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference, UCREL technical paper number 16*. UCREL: Lancaster University, UK.

Prince, E. F. 1981. Toward a taxonomy of given-new information. In Cole, P. (ed), *Radical pragmatics*, 223-256. New York: Academic Press.

Ryan, J. 2012. *Acts of reference and the miscommunication of referents by first and second language speakers of English*. Doctoral dissertation, University of Waikato. http://researchcommons.waikato.ac.nz/handle/10289/6599 (accessed 1/12/2012).

Salamoura, A., and Saville, N. 2010. Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. In Bartning, I., Maisa, M., and Vedder, I. (eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*. EuroSLA Monographs Series (1), 101-132.

Slabakova, R., Rothman, J., Mendez, T. L., Campos, G., and Kempchinsky, P. 2011. Pragmatic features at the L2 syntax-discourse interface. In *Proceedings of the 34th Annual Boston University Conference on Language Development*.

Sorace, A., and Filiaci, F. 2006. Anaphora resolution in near-native speakers of Italian. *Second Language Research* 22(3): 339-368.

Stalnaker, R. 1974. Pragmatic presuppositions. In Munitz, M. K., and Unger, P. K. (eds.), *Semantics and philosophy*, 197-213. New York: New York University Press.

Stein, N. L., and Glenn, C. G. 1979. An analysis of story comprehension in elementary school children. In Freedle, R. O. (ed.), *New directions in discourse processing (volume 2)*, 53-120. Norwood, NJ: Ablex.

Von Stutterheim, C., and Klein, W. 1989. Referential movement in descriptive and narrative discourse. In Dietrich, R., and Graumann, C. F. (eds.), *Language processing in social context*, 39-76. North Holland: Elsevier Science Publishers B. V.

White, L. 1986. Implications of parametric variation for adult second language acquisition: an investigation of the pro-drop parameter. In Cook, V. J. (ed.), *Experimental approaches to second language acquisition*, 55-72. Oxford: Pergamon.

White, L. 2011. Second language acquisition at the interfaces. *Lingua* 121(4): 577-590.

Peter Crosthwaite
Department of Theoretical and Applied Linguistics
University of Cambridge
E-mail: prc34@cam.ac.uk

Received: 2013. 03. 20
Revised: 2013. 07. 23
Accepted: 2013. 07. 23