How native language and L2 proficiency affect EFL learners' capitalisation abilities: a large-scale corpus study

Itamar Shatz[1]

Abstract

Capitalisation is a salient orthographic feature, which plays an important role in linguistic processing during reading, and in writing assessment. Learners' second language (L2) capitalisation skills are influenced by their native language (L1), but earlier studies of L1 influence did not focus on learners' capitalisation, and examined primarily 'narrow' samples. This study examines capitalisation error patterns in a large-scale corpus of over 133,000 texts, composed by nearly 38,000 EFL learners, who represent seven different L1s and a wide range of English proficiency levels. The findings show that speakers of all L1s made a large number of capitalisation errors, in terms of errors per word and error proportion (out of all errors), especially at lower L2 proficiency levels. Under-capitalisation was more common than over-capitalisation, though this gap narrowed over time. Interestingly, L1s which share English's Latin script had higher error rates, suggesting that (assumed) perceived similarity between the L1 and the L2 increases interference, though this interference could not be explained only through direct negative transfer. There was also an interaction between L1 influence and L2 proficiency, so that differences between speakers of different L1s became smaller as their L2 proficiency improved.

Keywords: capitalisation, crosslinguistic influence, English as a foreign language, L2 writing, native language L1 transfer.

[1] Department of Theoretical and Applied Linguistics, University of Cambridge, 9 West Road, University of Cambridge, Cambridge, CB3 9DP, United Kingdom.
Correspondence to: Itamar Shatz, *e-mail*: is442@cam.ac.uk

Corpora 2019 Vol. 14 (2): 173–202
DOI: 10.3366/cor.2019.0168

www.euppublishing.com/cor

1. Introduction

Capitalisation is a salient orthographic feature, in which certain words in the written form of the language start with an upper-case letter (e.g., *Friday*). Languages that have this feature, which relies on a case distinction between upper-case and lower-case letters, can have different capitalisation schemes. These schemes are based on aspects such as the word's position in the sentence, its context and its part of speech. For example, both English and German capitalise sentence-initial words, but German capitalises all nouns, while English capitalises only proper nouns, such as names of individuals, places or days of the week (Beaufays and Strope, 2013; and Hohenstein and Kliegl, 2013).[2]

Capitalisation serves an important role in linguistic processing during reading. It can, for example, aid syntactic parsing, facilitate lexical access and provide semantic preview benefits (Peressotti *et al*., 2003; Farghaly and Shaalan, 2009; Spitkovsky and Jurafsky, 2012; Beaufays and Strope, 2013; Hohenstein and Kliegl, 2013; and Rayner and Schotter, 2014). Furthermore, capitalisation is also important from a pedagogical perspective; this stems from its aforementioned influence on linguistic processing, and from its frequent use as a criterion in second language (L2) writing assessment. In addition, in the case of English, learners of English as a foreign language (EFL) often struggle with learning the language's capitalisation rules, and experience difficulties with implementing them correctly in their writing (Morris, 1998; Al-Jarf, 2004; Lee, 2006; Darus and Ching, 2009; Nezakatgoo, 2011; Gustilo and Magno, 2012; and Sawalmeh, 2013).

Learners' capitalisation abilities in the L2 are influenced, among other factors, by their native language (L1). This cross-linguistic influence affects various domains in the language, and occurs primarily due to transfer of structures (or lack thereof) from the L1 to the L2. As such, L1 influence depends on the structural similarity, as well on the perception of similarity between the languages, which is labelled by Kellerman (1978, 1983) as 'psychotypology'. Essentially, when a structure is perceived to be similar between the L1 and the L2, due to some form of linguistic similarity between the languages, learners transfer the structure from the L1 to the L2. When this cross-linguistic transfer results in target-like forms in the L2, which facilitates acquisition, it is referred to as 'positive transfer'. Conversely, when this cross-linguistic transfer results in L2 errors, which hinders acquisition, it is referred to as 'negative transfer' (Odlin, 1989, 2003; Ringbom, 1992, 1987; Benson, 2002; Koda, 2005; O'Sullivan and Chambers, 2006; Bennui, 2008; R. Ellis, 2008; Jarvis and Pavlenko, 2008; Sersen, 2011; Tolentino and Tokowicz, 2014; and Jarvis, 2015).

---

[2] Note that while days of the week are generally treated as proper nouns in practice, onomastic studies do not treat them as such, due to the fact that they are not mono-referential (Van Langendonck, 2008).

A better understanding of how learners' L1 influences their errors during the L2 acquisition process has significant pedagogical implications. Primarily, this is because such understanding can be used to optimise teaching materials, educational tools and curricula, in order to better account for learners' L1 influence (Chang and Chang, 2004; Chang *et al*., 2008; and Ge, 2015). For example, Chang *et al*. (2008) developed an online collocation aid for Taiwanese EFL learners, which helps them cope with the frequent miscollocations in L2 writing that stem from negative L1 transfer. In addition, a better understanding of L1 influence also helps teachers to raise awareness of the subject among learners, which in turn facilitates L2 acquisition (Benson, 2002; Bennui, 2008; Sersen, 2011; and Shekhzadeh and Gheichi, 2011). For example, Sersen (2011) showed that making Thai EFL learners consciously aware of L1 interference led to a reduction in errors stemming from negative transfer, which improved their English writing skills.

Several previous studies examined the effects of L1 influence on the writing of EFL learners (e.g., Bhela, 1999; Darus and Ching, 2009; Darus and Subramaniam, 2009; Sersen, 2011; Sawalmeh, 2013; and Sönmez and Griffiths, 2015). However, while these studies often show that EFL learners frequently make capitalisation errors, the role of L1 influence is not clear in these studies, especially from a developmental perspective. This can be attributed to three primary factors. First, capitalisation errors are often not the focus of such studies, as they reflect a strictly orthographical feature of the language, which does not manifest in speaking. Second, capitalisation errors are generally prevalent in English writing, even among native speakers, which makes it difficult to discern the role of L1 influence on learners' capitalisation skills (Lunsford and Lunsford, 2008; and Wilcox *et al*., 2014). Most importantly, these studies focussed for the most part on narrow samples, by looking at a small number of L1s or a small range of L2 proficiency levels. This is exemplified by Darus and Ching (2009: 252), who examined the English writing of Chinese-speaking EFL learners, in the discussion of their study's limitations:

> This study is limited to the written work of 70 essays of Form One Chinese students who are studying in one public school. The written essays collected are from a specific topic only. Therefore, the study will not be able to give conclusive evidence regarding other Form One Chinese students from other proficiency levels [. . .].

These limitations necessitate a direct investigation of learners' capitalisation errors during the second language acquisition (SLA) process. Such investigation should compare the capitalisation abilities of learners with different L1s and with a wide range of L2 proficiency levels. This large-scale analysis can be accomplished through the use of a learner corpus, which consists of texts produced by learners during the language acquisition process (Granger, 1994; Biber *et al*., 1998; Ghadessy *et al*., 2001; Reder *et al*., 2003;

Myles, 2005; McEnery *et al*., 2006; Granger and Leech, 2014; and Krummes and Ensslin, 2014). Learner corpora already serve an important role in research on L1 influence (e.g., Borin and Prütz, 2004; Ene, 2008; Crompton, 2011; Laufer and Waldman, 2011; and Murakami and Alexopoulou, 2016), but there has not yet been a large-scale investigation of L1 influence on EFL learners' capitalisation errors.

This study attempts to overcome the limitations of previous studies by focussing specifically on capitalisation errors and by using a large-scale learner corpus called the EF-Cambridge Open Language Database (EFCAMDAT), which contains English texts composed by EFL and ESL learners with a wide variety of linguistic backgrounds. The main motivation in examining capitalisation, in addition to the considerations mentioned earlier, is that it serves as a structure which enables us to examine L1 influence during L2 acquisition from a new perspective. Specifically, it allows us to examine whether similarity in capitalisation rules between learners' L1 and the target L2 increases transfer from the L1, and whether this transfer facilitates L2 acquisition or hinders it. Furthermore, it allows us to examine L1 influence from a psychotypological perspective, by researching how differences in script between learners' L1 and the target L2 affect the rate of L1 transfer. That is, the research will first show us how L1–L2 linguistic similarity in capitalisation rules affects acquisition of capitalisation in the L2. Then, it will show us whether the assumed perception of dissimilarity (i.e., psychotypological distance), due to the difference in script between the L1 and the L2, reduces transfer of capitalisation rules from the L1. If this is indeed the case, then we will see that for learners whose L1 shares the same script as the target L2, there will be increased L1 transfer compared to learners whose L1 uses a different script.[3]

In addition to outlining the general patterns of capitalisation errors in EFL learners' writing, this study seeks to answer the following questions:

(*i*) How does similarity in capitalisation rules between learners' L1 and the target L2 affect their acquisition of capitalisation rules in the L2?
(*ii*) If L1 influence does affect capitalisation, is this effect moderated by differences in script between learners' L1 and the target L2?
(*iii*) Does learners' L2 proficiency moderate the effect of L1 influence on the acquisition of L2 capitalisation?

[3] Note that learners' perception of L1–L2 similarity is not measured directly in the study. Rather, we analyse the assumed perception of similarity, based on whether learners' L1 script is the same as the script used in the target L2.

2. Methodology

2.1 The corpus

The EFCAMDAT was developed at the Department of Theoretical and Applied Linguistics at the University of Cambridge in a collaboration with EF Education First (EF), an international, private education company (Geertzen *et al*., 2014). EF's outreach allowed for the creation of a large-scale database, which consists of 551,036 texts, written by 84,864 learners from 172 nationalities (Geertzen *et al*., 2014).[4]

The EFCAMDAT is composed of essays on a wide variety of topics, submitted by English learners to EF's online English school. The school spans sixteen proficiency levels, aligned with common language proficiency standards such as the TOEFL, the IELTS, and the Common European Framework of Reference for languages (CEFR). When learners start the course, they first undertake a placement test to determine their initial proficiency level. Each level contains eight units; at the end of each unit, the learner writes, in response to an essay prompt, a short text which is graded by a language teacher. If the learner receives a passing grade, they advance to the next unit, but if they receive a failing grade, they are required to repeat the current one. The average text in the sub-corpus targeted in this study contains 54.6 words (Standard Deviation = 34.4), and because of the course's requirements and learners' growing abilities, texts become significantly longer as learners' English proficiency grows (correlation between proficiency and word count is $r = 0.7648$, $p < 0.001$). Two sample texts are shown in Figure 1.

In addition to a grade, teachers also provide feedback to learners using a standardised set of twenty-three error tags, which cover concepts such as grammar, spelling and punctuation. At present, the EFCAMDAT contains error annotations for approximately 36 percent of texts, distributed evenly across texts from all nationalities and proficiency levels, based on each nationality's prevalence in the corpus. Only texts containing error annotations were included in this analysis. This study looks specifically at capitalisation errors, which are marked using the 'capitalisation' tag in the two following scenarios:

(*i*) Cases where the learner under-capitalised, meaning that they failed to capitalise a word when necessary (e.g., '. . . the mail will arrive on sunday').
(*ii*) Cases where the learner over-capitalised, meaning that they capitalised a word when they should not have (e.g., '. . . the mail will arrive Tomorrow').

[4] Unfortunately, while the EFCAMDAT contains information regarding learners' nationality and English proficiency level, it does not contain further metadata such as learners' gender, age or additional spoken L2s.

LEVEL 3, UNIT 5. BRAZILIAN. GRADE = 98.

TOPIC: GIVING SUGGESTIONS ABOUT CLOTHING.

Hi Nina! The purple top is nice, but it's very expensive. Why don't you buy that orange summer skirt? It's very nice and cheap. How about the red hat? I love it.

LEVEL 12, UNIT 1. JAPANESE. GRADE = 96.

TOPIC: TURNING DOWN AN INVITATION.

Dear Graham, Thank you ever so much for the invitation, but I'm afraid we can't make it. We have already got an invitation for the birthday party of my mother-in-law. I'm afraid that that's the only occasion I can catch up with her. I'd be utterly grateful if you and Lucy could come over to our house to have dinner with us on Thursday night in next week. Finally,I'd like to thank you again for your kind invitation. I'm looking forward to your response. Kind regards, Chris

Figure 1: Sample texts from the corpus.

LEVEL 5, UNIT 4. ITALIAN.

TOPIC: GIVING INSTRUCTIONS TO A HOUSE-SITTER.

ORIGINAL SENTENCE:

In the afternoon On monday, Wednesday and Friday I feed the fish.

ANNOTATED VERSION:

In the afternoon <change><selection>On </selection><tag><symbol>C</symbol>

<correct> on </correct></tag></change><change><selection>monday</selection><tag>

<symbol>C </symbol><correct>Monday</correct></tag></change>, Wednesday and

Friday I feed the fish.

Figure 2: Annotation of a sample sentence in the corpus. Note that the words containing capitalisation errors are underlined in the original sentence, and that in the error annotations, these errors are denoted using the 'C' symbol.

Note that the EFCAMDAT annotations do not distinguish between the two scenarios, and mark both under-capitalisation as well as over-capitalisation errors using the same error tag ('capitalisation'). This is demonstrated in the annotation shown in Figure 2.

Native language	Percent incidence as an L1 in nationality	No. of learners (percent of sample)		Number of texts (percent of sample)	
French	93 [a]	2,680	(7.1)	8,376	(6.3)
German	90 [a]	3,312	(8.7)	10,793	(8.1)
Italian	95 [a]	2,464	(6.5)	7,928	(6.0)
Japanese	99 [b]	1,204	(3.2)	4,038	(3.0)
Portuguese	95 [b]	19,286	(50.8)	70,185	(52.7)
Russian	95 [b]	4,881	(12.8)	16,148	(12.1)
Spanish	93 [c]	4,165	(11.0)	15,632	(11.7)
Total	–	37,992		133,100	

Table 1: The distribution of learners and texts in the sample, grouped by native language. (Note: L1 Spanish refers to Mexican learners, while L1 Portuguese refers to Brazilian learners. The large portion of Portuguese texts is potentially as a result of EF's role in the preparation for the 2014 World Cup in Brazil (EF Education First, 2011). The number of learners and texts listed here represents the sample after excluding texts at the highest (16th) CEFR level, due to the insufficient number of texts at that level (as explained in the next section).)
[a] Data from the Special Eurobarometer 243 (2006).
[b] This was calculated based on Lewis *et al*. (2015).
[c] Data from The World Factbook (2015).

2.2 Sample selection

2.2.1 Native language

Initially, texts of learners from the ten nationalities with the most texts were selected, in order to ensure that there is a sufficiently large sample for each nationality; these nationalities account for 93.9 percent of all the texts in the database, and other nationalities did not have enough texts for analysis, especially at the higher proficiency levels. Given that the EFCAMDAT does not contain information regarding learners' L1 directly, nationality was used to approximate it, as in previous studies on the database (e.g., Geertzen *et al*., 2014; Jiang *et al*., 2014; and Murakami, 2016). Three nationalities where under 90 percent of the population speaks a single L1 were excluded from the initial sample.[5] The seven remaining nationalities are listed in Table 1.

[5] These nationalities were: Taiwanese (Klöter, 2004), Saudi Arabian (Lewis *et al*., 2015) and Chinese (Lewis *et al*., 2013).

2.2.2 English proficiency

Of the sixteen proficiency levels in the database, Levels 1 to 15 were used in the sample here; Level 16 was excluded because it had an insufficient number of texts for analysis.[6] Based on the EFCAMDAT's background information, these levels correspond to different CEFR levels (Geertzen *et al*., 2014). CEFR was chosen as the metric for L2 proficiency in the analysis, as it is a commonly used standard for assessing language proficiency (Lehtonen and Karjalainen, 2008; Little, 2011; Nagai and Dwyer, 2011; and Geertzen *et al*., 2014).[7] There was some minor variation in the average L2 proficiency between learners from different L1s; this was accounted for at the data analysis stage. Table 2 shows the original levels in the data and the corresponding CEFR level, together with the number of texts in the sample at each ranking. These texts represent all the texts from Levels 1 to 15 in the seven L1s examined.

Proficiency levels	Equivalent CEFR ranking	No. of texts (percent of sample)	
1–3	A1	65,424	(49.2)
4–6	A2	39,637	(29.8)
7–9	B1	19,335	(14.5)
10–12	B2	6,992	(5.3)
13–15	C1	1,712	(1.3)

Table 2: Proficiency levels in the sample.

2.3 Assessing L1–L2 similarity

The similarity between learners' L1 and English, the target L2, was assessed in two ways. First, each language's script was considered (e.g., Latin or Cyrillic), because capitalisation is an orthographic feature, and is therefore likely to be associated with the language's script. Second, each language's capitalisation rules were considered. The focus was on summarising the overall tendency to capitalise in the language, and how this tendency compares with the capitalisation scheme in English. The assessment of L1–L2 similarity is therefore qualitative in nature, due to the difficulty of accurately quantifying the degree of similarity between the languages, which

[6] Level 16 could not be grouped with Levels 13 to 15, as it has a higher CEFR ranking (C2 *versus* C1).
[7] For a detailed explanation of the CEFR, see the original report by the Council of Europe (2002).

stems from wide variability in the nature of the capitalisation rules and scripts of the different languages.

2.4 Data analysis

The data used in the analysis included the number of capitalisation errors in each text, as well as the total number of errors in the text, and the number of words. In addition, a further distinction between under-capitalisation and over-capitalisation errors was achieved using an additional analysis of the texts.[8] The analysis examined two commonly used metrics of learners' error rates:

(*i*) *Errors per words* – how many capitalisation errors a learner makes for every 100 words of writing in the text. For example, a certain text could contain an average of five capitalisation errors for every 100 words.

(*ii*) *Error proportion* – the percentage, out of the total errors in the text, that capitalisation errors account for. For example, capitalisation errors could account for 10 percent of the total errors which appear in a certain text.

Both metrics have been validated in previous SLA studies, and are well-attested in existing literature. (For studies which examined errors-per-words, see, for example, Chandler, 2003; Kovac, 2011; and Polio, 1997. For studies which examined error proportions see, for example, Chaudron, 1988; Darus and Ching, 2009; and Yang and Li, 2012). Their combined usage here is intended to be complementary, in order to give a comprehensive account of learners' error patterns.

The analysis took place in several stages. First, two analyses of covariance (ANCOVA) were used in order to determine whether learners' L1 affects the rate of their capitalisation errors, in terms of errors per words and error proportion. Learners' English proficiency served as the covariate, in order to control for the minor difference in average L2 proficiency between the different L1s, and error rates were transformed using a square-root transformation, in order to linearise the relationship between the error rates and the L2 proficiency covariate.[9] Next, estimated marginal means of

[8] As mentioned, the 'capitalization' error tag in the EFCAMDAT does not distinguish between instances of over-capitalisation or under-capitalisation. This distinction was achieved using an analysis of learners' texts in Python, using the 'xml.etree.ElementTree' module (Python Version 3.5.2, xml.etree.ElementTree version 20.5). This module was used to extract texts marked with the 'capitalization' error tag, and identifies whether the first letter of the target word is uppercase or lowercase.

[9] The use of analysis of covariance in non-randomised, observational studies is a methodologically acceptable way to reduce biases which result from systematic differences

the two error rates (which control for the L2 proficiency covariate) were calculated for each L1, and pairwise comparisons between the L1s were performed. Then, the Spearman's rho correlations were calculated, between learners' proficiency level and their capitalisation error rates. This was followed by significance testing (using Fisher's transformation), in order to determine whether the correlations were different for speakers of different L1s. Finally, error patterns for each error rate were plotted across learners' CEFR proficiency levels.

In addition, an analysis was run in order to examine learners' under-capitalisation and over-capitalisation patterns. This included a logistic regression, with learners' L1 as the predictor, their L2 proficiency as a covariate, and the type of capitalisation error (i.e., under-capitalisation or over-capitalisation) as the dependent variable. This was followed by pairwise comparisons of the different error rates between the L1s, together with correlation calculations, which examined how these error rates changed over time, and finally with the plotting of errors patterns.

The statistical tests were all two-tailed and Bonferroni-adjusted, and treated each text as a separate case. In all calculations, when L2 proficiency was used as a variable, the original, ungrouped variable was used (on a scale of 1 to 15).[10]

3. Results

3.1 Scripts and capitalisation rules

The following is a description of each language's script and capitalisation rules. Table 3 (p. 185) contains a summary of this information.

English is a language with a Latin script, and it capitalises a wide variety of word types. These consist primarily of proper nouns, such as personal names (e.g., *John*), names of days (e.g., *Sunday*), names of months (e.g., *January*) and names of places, including the accompanying geographical terms (e.g., *Mount Everest*). In addition, the first-person pronoun *I* is also capitalised, as are proper adjectives, which are adjectives that are derived from proper nouns (e.g., *American*). English also capitalises sentence-initial words, a trait shared by the other languages in the sample

---

between the groups. This remains the case even when there is an association between the predictor and the covariate, unlike in randomised experimental studies, where this represents a potential problem (Schafer and Kang, 2008; and Huitema, 2011).

[10] Data analysis was conducted using SPSS 20. Preliminary data wrangling and correlation calculations were performed with R, using the following packages: plyr, data.table, reshape2, psych, and mada (Wickham, 2007, 2011; Doebler, 2015; Dowle *et al*., 2015; R Core Team, 2016; and Revelle, 2016). Figures were created using ggplot2 and gridExtra (Wickham, 2009; and Auguie, 2016).

which have a case distinction (McCaskill, 1998; and The Chicago Manual of Style, 2010).[11]

French, Italian, Spanish and Portuguese also have a Latin script, and capitalisation schemes which are relatively similar to English, and strongly similar to one another: all four languages capitalise proper nouns, though fewer word types are capitalised than in English. As such, names of individuals are capitalised, for example, but not names of days or months. Names of places are capitalised, but not the accompanying geographical terms (e.g., *mont Everest* in French). Proper adjectives are not capitalised either. The first-person pronoun is also not capitalised, though the formal second-person pronoun is capitalised in Spanish (Hutchinson and Lloyd, 2003; Jenkins, 2004; Kattán-Ibarra and Pountain, 2003; Proudfoot and Cardo, 2005; and The Chicago Manual of Style, 2010).

German also shares English's Latin script, and has some similarities in capitalisation rules to the previous languages. These include no capitalisation of proper adjectives or of the first-person pronoun, though the formal second-person pronoun is capitalised. However, a significant difference is that all nouns are capitalised in German, rather than just proper nouns (Schenke and Seago, 2004; The Chicago Manual of Style, 2010; and Hohenstein and Kliegl, 2013).

Russian uses a Cyrillic script, and capitalises words to a lesser extent than English, in a manner similar to French, Italian, Spanish and Portuguese. It capitalises only some proper nouns (e.g., names of individuals but not names of days or months, or geographical terms), and does not capitalise proper adjectives or the first-person pronoun, though the second-person pronoun is sometimes capitalised in formal situations (Kaufman *et al*., 2006; and The Chicago Manual of Style, 2010).

Japanese uses a script which consists of a combination of hiragana and katakana (Japanese scripts) together with kanji (Chinese characters), which do not have a case distinction, and therefore capitalisation does not appear in the language (Daniels and Bright, 1996; and Hadamitzky and Spahn, 1997).[12]

---

[11] In English and in the other languages, there is sometimes variation in the capitalisation schemes when writing titles of works, as opposed to main texts. However, such distinction is difficult to control for, especially as languages may have more than convention regarding capitalisation of titles. This is not crucial in this analysis, as the vast majority of learners' texts did not contain titles.

[12] It is important to note that while hiragana and katakana do *not* contain a case distinction, most Japanese speakers are taught to read and write the Rōmaji script, which consists of Latin characters, and which is used for the transliteration of Japanese. Capitalisation in Rōmaji is supposed to be similar to English, though this varies in practice, so that sometimes words in Rōmaji are not capitalised at all, or are written in all upper-case letters (Daniels and Bright, 1996; and Hadamitzky and Spahn, 1997). This means that Japanese speakers are exposed, to some degree, to English rules of capitalisation, and prior research shows that this exposure affects Japanese speakers' ability to process words in English (Yamada *et al*., 1988), though overall the usage of Rōmaji in daily life is relatively rare, which mitigates its influence (Ida *et al*., 2015).

3.2 Capitalisation error rates

The estimated marginal mean of capitalisation errors per word overall was 2.815 (Standard Error = 0.036, 95 percent Confidence Interval [2.744, 2.885]), and the mean for capitalisation error proportion was 0.117 (SE = 0.001, 95 percent CI [0.116, 0.119]). The ANCOVAs showed that learners' L1 significantly affected the rates of their capitalisation errors, both in terms of errors per word ($F(6, 133092) = 442.74$, $p < 0.001$, $\eta^2 = 0.020$) as well as in terms of error proportion ($F(6, 133092) = 323.70$, $p < 0.001$, $\eta^2 = 0.014$).[13]

The ANCOVAs also showed that learners' L2 proficiency covariate (which was controlled for in the analysis) significantly affected the rates of their capitalisation errors, both in terms of errors per word ($F(1, 133092) = 4955.27$, $p < 0.001$, $\eta^2 = 0.036$) as well as in terms of error proportion ($F(1, 133092) = 2458.01$, $p < 0.001$, $\eta^2 = 0.018$). The correlation analysis showed that the correlation between learners' L2 proficiency and capitalisation errors per word was $r_s = -0.153$ ($p < 0.001$, 95 percent CI [–0.158, –0.148]), while the correlation between proficiency and capitalisation errors proportion was $r_s = -0.120$ ($p < 0.001$, 95 percent CI [–0.125, –0.115]).

Mean capitalisation error rates for each L1 are shown in Table 4. Correlations between L2 proficiency and error rates for each L1 are shown in Table 5. Plots of capitalisation error rates per L1 and across L2 proficiency are shown in Figures 3 and 4.

The pairwise comparisons show that all differences in errors per words were statistically significant at $p < 0.001$ following Bonferroni adjustment, except for German–Japanese, German–Portuguese, Japanese–Portuguese and Japanese–Russian, which were not statistically

[13] Note that these effect sizes are relatively small compared to those typically found in SLA research (Oswald and Plonsky, 2010), though they still fall within a range that is considered to be meaningful in cognitive research (Cohen, 1988). This could be attributed to several factors. First, typological similarity between several of the languages (particularly French, Italian, Spanish and Portuguese), means that there are relatively small differences in error rates between speakers of those L1s, and this reduces the overall effect size of L1 influence. Second, the fact that differences in error rates between speakers of different L1s decrease significantly over time, as we will see below, also reduces the overall effect size of L1 influence, since the majority of transfer appears to take place during the initial stages of acquisition. Third, the fact that individual texts were examined in the analysis means that there were relatively large measurement errors, which leads to a large amount of noise, and thus to a smaller effect size. In addition, error rates are in general considered to be a relatively crude measure of accuracy, which cannot perfectly isolate and capture learners' capitalisation ability – a trait which further increases measurement noise, and decreases effect sizes (Schachter and Celce-Murcia, 1977). Finally, it's important to remember that learners' L2 capitalisation abilities are, similarly to other L2 abilities, influenced by a multitude of factors beyond their L1, so that it is not surprising to find a relatively small effect size when focussing on the influence of a single factor.

	English	French, Italian, Spanish, Portuguese	German	Russian	Japanese
Script	Latin	Latin	Latin	Cyrillic	Mixed (hiragana and katakana)
Capitalisation scheme	Large number of word types	Smaller number of word types	Capitalises all nouns	Smaller number of word types	No

Table 3: Summary of script and capitalisation rules for each language. (Note: this represents the generalisations of the capitalisation rules in each language as they pertain to the present analysis. For a more detailed description of the rules, see the earlier text in this section.)

L1	Errors per 100 words		Error proportion	
	Mean (SE)	95 percent CI	Mean (SE)	95 percent CI
French	3.778 (0.101)	[3.579, 3.977]	0.145 (0.002)	[0.141, 0.149]
German	2.416 (0.090)	[2.240, 2.592]	0.123 (0.002)	[0.119, 0.127]
Italian	3.177 (0.104)	[2.972, 3.382]	0.130 (0.002)	[0.125, 0.134]
Japanese	1.993 (0.146)	[1.706, 2.279]	0.092 (0.003)	[0.086, 0.099]
Portuguese	2.436 (0.035)	[2.367, 2.505]	0.114 (0.001)	[0.112, 0.115]
Russian	1.558 (0.073)	[1.415, 1.702]	0.074 (0.002)	[0.071, 0.077]
Spanish	4.344 (0.074)	[4.199, 4.490]	0.142 (0.002)	[0.139, 0.145]

Table 4: Capitalisation error rates.

significant ($p > 0.05$). All differences in error proportion were significant except for French–Spanish and German–Italian.

The pairwise comparisons show that all the differences in correlations of errors per word and L2 proficiency between the different L1s were statistically significant at $p < 0.05$ following Bonferroni adjustment, except for German–Italian, German–Spanish, Italian–Spanish and Japanese–Portuguese. In the correlations between error proportion and proficiency, the differences were significant at $p < 0.05$ except for German–Italian, German–Spanish, Italian–Japanese, Italian–Portuguese,

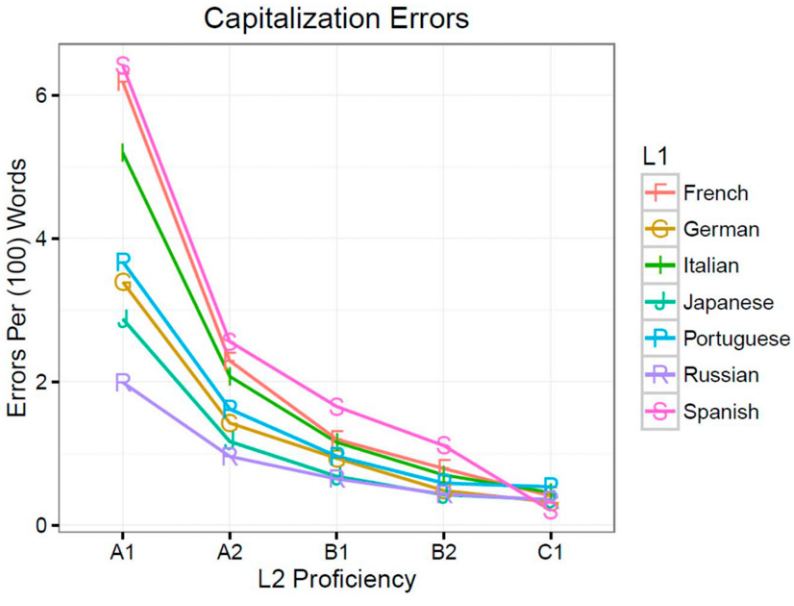

Figure 3: Average capitalisation errors per word, for speakers of different L1s, across (CEFR) L2 proficiency levels.

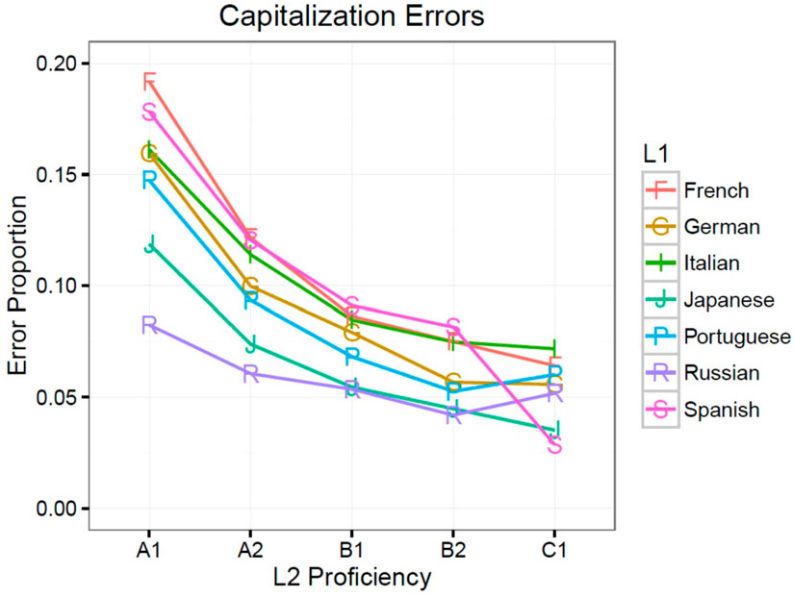

Figure 4: Average capitalisation errors proportion (out of total errors), for speakers of different L1s, across (CEFR) L2 proficiency levels.

L1	Errors per 100 words		Error proportion	
	rs	95 percent CI	rs	95 percent CI
French	–0.244	[–0.264, –0.224]	–0.199	[–0.219, –0.178]
German	–0.191	[–0.21, –0.173]	–0.154	[–0.173, –0.136]
Italian	–0.175	[–0.196, –0.153]	–0.132	[–0.153, –0.11]
Japanese	–0.113	[–0.144, –0.083]	–0.088	[–0.118, –0.057]
Portuguese	–0.139	[–0.147, –0.132]	–0.109	[–0.117, –0.102]
Russian	–0.013	[–0.029, 0.002]	0.003	[–0.013, 0.018]
Spanish	–0.173	[–0.188, –0.158]	–0.127	[–0.142, –0.111]

Table 5: Correlation between L2 proficiency and capitalisation error rates. (Note: all the correlations were significant at $p < 0.001$, following Bonferroni adjustment, except for Russian, which was not statistically significant for either errors per word or for error proportion.)

Italian–Spanish, Japanese–Portuguese, Japanese–Spanish and Portuguese–Spanish.

Note that the sharp decrease in error rates for Spanish speakers at the final proficiency level appears to result from an idiosyncrasy in the data, as it is inconsistent with the previous rate of decrease, and does not appear in any of the other languages.

3.3 Under-capitalisation and over-capitalisation rates

The logistic regression showed that learners' L1 significantly affected their under-capitalisation and over-capitalisation rates ($\chi^2(7) = 3026.34$, $p < 0.001$, *Nagelkerke* $R^2 = 0.043$), when controlling for learners' L2 proficiency covariate, which significantly affected their under-capitalisation rates ($B = –0.106$, $SE = 0.003$, $p < 0.001$). The estimated marginal mean of under-capitalisation error proportion (out of all capitalisation errors) was 0.765 (SE = 0.003, 95 percent CI [0.760, 0.770]). Information on the under-capitalisation and over-capitalisation rates for each L1 are shown in Table 6. Correlations between L2 proficiency and under-capitalisation/over-capitalisation rates for each L1 are shown in Table 7. Plots of under-capitalisation rates per L1 and across L2 proficiency are shown in Figure 5.

All pairwise comparisons in under-capitalisation/over-capitalisation rates between the L1 were statistically significant at $p < 0.001$ following

L1	Under-capitalisation error proportion	
	Mean (SE)	95 percent CI
French	0.850 (0.006)	[0.838, 0.863]
German	0.769 (0.006)	[0.756, 0.781]
Italian	0.852 (0.007)	[0.838, 0.865]
Japanese	0.643 (0.011)	[0.621, 0.666]
Portuguese	0.744 (0.002)	[0.740, 0.749]
Russian	0.726 (0.006)	[0.714, 0.738]
Spanish	0.772 (0.004)	[0.764, 0.781]

Table 6: Proportion of under-capitalisation errors (out of total capitalisation errors) for speakers of different L1s. The remaining proportion of errors represents cases of over-capitalisation.

L1	Correlation between under-capitalisation and L2 proficiency	
	rs	95 percent CI
French	–0.153	[–0.183, –0.121]
German	–0.093	[–0.124, –0.063]
Italian	–0.132	[–0.165, –0.098]
Japanese	–0.157	[–0.211, –0.101]
Portuguese	–0.101	[–0.113, –0.09]
Russian	–0.155	[–0.184, –0.125]
Spanish	–0.125	[–0.146, –0.103]

Table 7: Correlations between learners' under-capitalisation error proportion and their L2 proficiency. Over-capitalisation rates are a direct inverse of these correlations. (Note: all the correlations were significant at $p < 0.001$, following Bonferroni adjustment. The overall correlation between L2 proficiency level and under-capitalisation rates was rs = –0.112 ($p < 0.001$, 95 percent CI [–0.121, –0.104]).)

Bonferroni adjustment, except for French–Italian, German–Spanish and Portuguese–Russian, which were not statistically significant ($p > 0.05$). German–Portuguese was significant at $p = 0.006$.

All differences in correlations between under-capitalisation/over-capitalisation and L2 proficiency were *not* statistically significant ($p > 0.05$ following Bonferroni adjustment), except for French–Portuguese, German–Russian and Portuguese–Russian.

Note that the minor increase of under-capitalisation error rates at the final proficiency level was not statistically significant for any of the L1s, based on Bonferroni-adjusted pairwise comparisons between the under-capitalisation rates for each L1 at the C1 and B2 proficiency levels, except

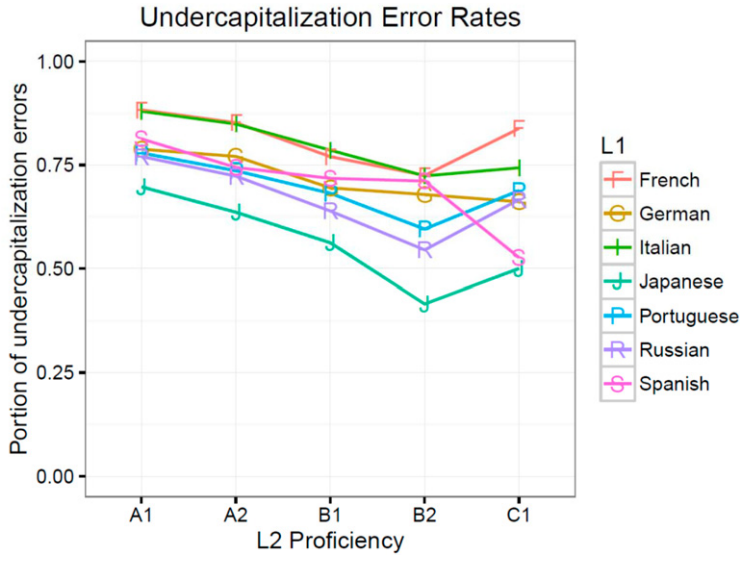

Figure 5: Mean portion of under-capitalisation error rates (out of total capitalisation errors), for speakers of different L1s, across (CEFR) L2 proficiency levels. Since over-capitalisation rates are the direct inverse of under-capitalisation rates, only the under-capitalisation rates are shown in the figure.

for Portuguese ($p = 0.04$). The decrease for Spanish was not statistically significant either, and appears to result from the potential idiosyncrasy in the data, as mentioned in the previous section.

4. Discussion

This paper examines how EFL learners' L1 affects their acquisition of L2 capitalisation rules. We will now discuss the overall patterns that arise from the results, followed by a discussion of the specific error patterns for speakers of different L1s. We will conclude by discussing the findings in aggregate, together with their implications and with some suggestions for future research.

In general, capitalisation errors were common among speakers of all L1s, and accounted for a relatively large proportion of the total errors, especially at the early L2 proficiency levels. However, the error rates decreased as learners' proficiency level improved, both in terms of errors per word and in terms of error proportion. This drop-off was high at first, and tended to level out over time. At the highest measured proficiency level (C1), these errors occurred infrequently, though they still accounted for a non-negligible portion of the total errors (~2.5 percent to 7.5 percent). In addition, under-capitalisation errors were more common than over-capitalisation

errors, though the difference between the two decreased slightly over time. There were significant differences in the error patterns of the different L1s, which indicates that L1 influence played a role in learners' L2 capitalisation abilities.[14]

French, Italian and Spanish all share English's Latin script, and share similar capitalisation rules with one another, as they all capitalise words similarly to English, but to a smaller degree. Consequently, these languages all had a similar pattern of capitalisation errors, which included the highest rate of capitalisation errors out of the L1s in the study, as well as a high drop-off in error rates as learners' proficiency increased. The difference in error rates between these languages and the other L1s became smaller over time. In addition, these three languages had the highest rates of under-capitalisation, which decreased slightly over time (compared to over-capitalisation errors), but at a rate matching the decrease in under-capitalisation in the other L1s.

In common with the previous languages, German also shares English's Latin script, but it has distinctly different capitalisation rules, as it capitalises all nouns. German therefore had a lower rate of capitalisation errors, both in terms of errors per word and in terms of error proportion. German speakers also had a lower rate of under-capitalisation (in comparison with over-capitalisation) than speakers of French, Italian and Spanish. However, the pattern of capitalisation errors over time was similar between these languages, and there was a significant decrease in capitalisation errors among German speakers as their L2 proficiency increased, as well as a decrease in the proportion of under-capitalisation in comparison with over-capitalisation.

Russian has a Cyrillic script and capitalisation rules that are similar to French, Italian and Spanish, so that it capitalises words similarly to English, but to a smaller degree. Conversely, Japanese utilises a mixed Japanese and Chinese script which contains no capitalisation at all, though Japanese speakers are exposed to Latin characters and capitalisation rules through the Rōmaji script. Both Russian and Japanese had similar error patterns, as they had the lowest rates of capitalisation errors among the L1s in the study, in terms of errors per word and in terms of error proportion, with Russian having the lower rates of the two. Japanese experienced a drop in capitalisation error rates over time, similarly to the other languages, though at a more moderate pace, while Russian did not experience a statistically significant decrease. Both had the lowest rates of under-capitalisation overall, with Japanese having the lower rates of the two. The rates of under-capitalisation decreased over time here as well, but at an equal rate to the

[14] It is important to note that the fact that the errors patterns were generally different for groups of speakers of different L1s, does not automatically mean that L1 influence played a role here. This consideration, based on the criteria for identifying L1 influence which were suggested by Jarvis (2000), is discussed more in depth in the 'limitations and future research' section.

decrease in the other languages, so that the difference between the languages remained consistent over time.

Taken together, these results provide important insights regarding the effects of L1–L2 linguistic similarity, as well as regarding the effects of the assumed perception of similarity between the L1 and the L2 (i.e., assumed psychotypological distance). Specifically, perceived similarity between the L1 and the L2, as a result of similarity in script, appears to have led learners to erroneously transfer their L1 capitalisation rules to the L2, despite the differences in capitalisation rules between the two languages. Conversely, when learners' target L2 was perceived to be different from their L1 due to the use of a different script, learners did not appear to transfer their L1 capitalisation rules to the L2. This could explain why speakers of languages which share English's Latin script (i.e., French, Italian, German, Portuguese and Spanish), had higher capitalisation error rates than speakers of Russian, which has a different script from English, but similar capitalisation rules to most other Latin-script using languages. This also explains why speakers of Russian had similar error rates to speakers of Japanese, whose language uses a different script from English, while also lacking capitalisation.

The error pattern of Japanese is slightly surprising given the prior results, as the linguistic difference in script and capitalisation rules between Japanese and English is greater than the difference between Russian and English, and therefore it would be expected that Japanese would have lower error rates than Russian. A possible explanation is that Japanese speakers experienced minor interference from their experience with the Rōmaji script, which utilises Latin characters, but whose usage may involve some deviations from English capitalisation rules, though as noted earlier, Rōmaji plays a relatively peripheral role in the language. Nevertheless, Japanese speakers did have lower error rates than speakers of the other L1s, which does fit with the general trend in the results. Overall, this indicates that speakers are more strongly influenced by their L1's capitalisation rules when the L1 shares its script with English, as perceived dissimilarity due to difference in script appears to reduce L1 interference, in terms of capitalisation ability.

The error pattern of Portuguese was also somewhat unexpected. Typologically, Portuguese is similar to French, Italian and Spanish, as it shares English's Latin script, while capitalising fewer word types. Hence, we predicted that Portuguese speakers will have similar error patterns to speakers of these languages. However, while Portuguese exhibited a similar decrease in capitalisation error rates over time as did these languages, it had significantly lower error rates, at a level similar to German. In addition, its under-capitalisation rates were also lower than in these languages, and were on a par with the under-capitalisation rates of Russian. Overall, the results for Portuguese are difficult to explain, and future research on the topic is necessary in order to investigate them, though they do not necessarily contradict the other findings in this study. Such research could examine, for example, the difference in acquisition of English capitalisation between speakers of Brazilian Portuguese, which does not capitalise days and months,

and speakers of European Portuguese, which capitalises those word classes similarly to English.[15,16]

There was also variation between speakers of different L1s in terms of under-capitalisation rates (in comparison with over-capitalisation rates), as speakers of L1s with higher rates of capitalisation errors tended to under-capitalise more than speakers of languages with low error rates. This could be attributed to L1 interference: the greater the interference in the languages which tend to capitalise less than English, the more learners tended to under-capitalise, and consequently, the more errors they made in their English writing. Interestingly, this was also true in the case of German, which capitalises more words than English, but whose speakers nonetheless under-capitalised more often than over-capitalised. This suggests that interference is associated, at least partially, with difficulty of implementing the capitalisation rules of English, beyond direct negative transfer of the capitalisation rules from the L1 to the L2. That is, similarity in capitalisation rules between learners' L1 and the target L2 makes it more difficult for learners to successfully implement the capitalisation rules in the L2, but this does not necessarily occur because the capitalisation rules from the L1 are transferred directly into the L2. Furthermore, this also provides evidence against the possibility that hypercorrection plays the primary role in explaining these error patterns. This is because hypercorrection is likely to lead to increased rates of over-capitalisation, rather than under-capitalisation, since English capitalises more word classes than the other language which were examined in the study (except for German).

There was also an interaction between the effects of learners' L1 and their L2 proficiency, so that in general, the greater the rate of learners' capitalisation errors was, the faster the decrease in error rates occurred over time. At the same time, the reduction in error rates became smaller as learners' L2 proficiency increased, meaning that learners' improvement rate (in terms of capitalisation ability) gradually decreased as their L2 proficiency increased. Because of this, the differences in error rates between the L1s became smaller as learners' L2 proficiency improved. This pattern of acquisition has a rough power-law distribution, which appears in the acquisition of other cognitive skills, including in the domain of SLA (Newell and Rosenbloom, 1981; Dekeyser, 1997; and N.C. Ellis, 2002). It is similar, for example, to the pattern found by Dekeyser (1997), who showed that

[15] Note however that the Brazilian–Portuguese form of capitalisation rules was preserved in the Portuguese Language Orthographic Agreement of 1990, whose implementation began in 2009 (Guia do Acordo Ortográfico, 2008).

[16] Exposure to European Portuguese was initially considered as a factor which could explain the acquisition patterns among the Brazilian–Portuguese speakers in the study. While this possibility was not ruled out, and may still be confirmed by future research on the topic, it is important to note that for the most part, speakers of Brazilian Portuguese have little exposure to literature which is written using the European–Portuguese capitalisation conventions (Baxter, 1992).

participants' performance in L2 comprehension and production tasks follows a power-law learning curve, where the rate of improvement decreases as participants' proficiency in the task improves. Overall, this indicates that, in the case of capitalisation, L1 influence plays a role primarily in the initial stages of acquisition, and its effects decrease over time, as learners' capitalisation skills improve, though it continues to play a role throughout the acquisition process.

4.1 Limitations and future research

This study examined L1 influence during the acquisition of English capitalisation by looking at two primary metrics: the frequency of capitalisation errors, in terms of errors per word and error proportion, and the type of capitalisation errors, in terms of whether learners capitalised a target word unnecessarily (over-capitalisation) or failed to capitalise a word when necessary (under-capitalisation). Examining these error patterns provides compelling evidence with regards to the effect of L1 influence. Specifically, Jarvis (2000) suggests the following three criteria for identifying L1 influence in speakers' productions: inter-L1-group differences, intra-L1-group similarities, and L1–IL performance similarities. These are all found in this study, to varying degrees. First are inter-L1-group differences, which are evident in the variation in error patterns between speakers of different languages. Second are intra-L1 group similarities, which are evident in the fact that the order of the different L1s, in terms of error rates, remained roughly consistent across the L2 proficiency spectrum (as we saw in Figures 3, 4 and 5), which suggests that there are intra-L1 similarities within groups. Finally, there are L1-interlanguage (IL) performance similarities, which are evident in how L1–L2 similarities and perceived similarities affected learners' use of L2 capitalisation. Specifically, we saw that the IL performance differences between the L1 groups could be explained by the differences in learners' L1s, since perceived similarity between the L1 and the L2 led to the transfer of capitalisation rules from learners' L1 to the target L2.

However, future studies could expand on this work, by examining additional aspects of L2 capitalisation patterns, in order to strengthen the evidence found in this study with regard to the criteria proposed by Jarvis (2000), and expand our understanding of this influence. Such studies could examine, for instance, systematicity and individual variation in the learners' productions, using mixed-effects models (Baayen *et al*., 2008; and Murakami, 2016), in order to confirm the presence of intra-L1 group similarities. Furthermore, such studies could also examine whether there are different patterns of acquisition for different word types (e.g., months *versus* names) or for different parts of speech (e.g., nouns *versus* adjectives), which could strengthen the support for evidence of L1–IL performance similarities, and answer open questions, such as why German speakers tend to under-capitalise in English, similarly to speakers of other languages.

5. Conclusion

Overall, this study reveals several key insights into how EFL learners with various linguistic backgrounds cope with the task of learning how to capitalise in English. First and foremost, capitalisation errors were found to be common for speakers of all L1s, especially in the early L2 proficiency levels. In addition, under-capitalisation was significantly more common than over-capitalisation, though the difference between the two decreased over time. There was also significant variation in capitalisation error patterns between speakers of different L1s, which indicates that cross-linguistic influence from learners' L1 affected their English capitalisation abilities. This influence depends on the assumed perception of dissimilarity between the L1 and the L2 (i.e., assumed psychotypological distance), which depends on the similarity in script between the languages. Specifically, perceived similarity between the L1 and the L2, due to the use of a similar script, led learners to erroneously transfer their L1 capitalisation rules to the L2, despite the differences in capitalisation rules, while the use of different scripts mitigated this transfer. In accordance with this, speakers of languages which share English's Latin script (i.e., French, Italian, German, Portuguese and Spanish), had higher capitalisation error rates than speakers of Russian, which has a non-Latin script, but similar capitalisation rules to most of the other Latin-using languages. This also explains why speakers of Russian had similar error rates to speakers of Japanese, whose language uses a different script from English, while also lacking capitalisation.

Furthermore, this interference does not appear to be explained *only* through direct negative transfer of the capitalisation rules from the L1, as speakers of German tended to under-capitalise more than to over-capitalise, similarly to speakers of other languages, despite the fact that more words are capitalised in German than in English and in the other L1s. This does not entirely rule out the possibility that direct transfer from the L1 plays a role here, but rather suggests that if such transfer does play a role, it is only a partial one, as learners' L2 capitalisation abilities are likely to be influenced by other factors. In addition, the study shows that an increase in learners' English proficiency is associated with a decrease in their capitalisation errors, both in terms of errors per words and also in terms of error proportion (out of total errors), and that this decrease occurs at a greater rate when there is a greater degree of interference from the L1. Taken together, this suggests that in the case of capitalisation, perceived similarity between the L1 and the L2, as a result of similarity in script, increases interference and inhibits learning. However, the effects of interference lessen over time and, consequently, the differences in capitalisation ability between speakers of different L1s also decrease. Furthermore, most learners successfully learn how to capitalise in English as their L2 proficiency improves, regardless of their L1.

In addition to their theoretical implications, the findings are also important from a pedagogical perspective, since an improved understanding of L1 influence will allow educators to raise awareness of the topic among

learners and optimise educational tools and curricula. Specifically, the findings show which learners are likely to struggle when learning how to capitalise in English, based on their L1 and on their English L2 proficiency, while also giving an indication regarding which type of capitalisation error they are more likely to make (in terms of under-capitalisation *versus* over-capitalisation). Future research, which investigates direct implementation of the findings in educational processes, could help discover how this knowledge, together with other findings on the role of L1 influence in L2 writing, can be utilised most effectively.

References

Al-Jarf, R.S. 2004. 'The effects of web-based learning on struggling EFL college writers', Foreign Language Annals 37 (1), pp. 49–57.

Auguie, B. 2016. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.2.1. Available online at: https://cran.r-project.org/package=gridExtra

Baayen, R.H., D.J. Davidson and D.M. Bates. 2008. 'Mixed-effects modeling with crossed random effects for subjects and items', Journal of Memory and Language 59 (4), 390–412.

Baxter, A.N. 1992. 'Portuguese as a pluricentric language' in M. Clyne (ed.) Pluricentric Languages: Differing Norms in Different Nations, pp. 11–43. Berlin & New York: Mouton de Gruyter.

Beaufays, F. and B. Strope. 2013. 'Language model capitalization' in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, pp. 6749–52. 26–31 May 2013. Vancouver, British Columbia.

Bennui, P. 2008. 'A study of L1 interference in the writing of Thai EFL students', Malaysian Journal of ELT Research 4, pp. 72–104.

Benson, C. 2002. 'Transfer / Cross-linguistic influence', ELT Journal 56 (1), 68–70.

Bhela, B. 1999. 'Native language interference in learning a second language: exploratory case studies of native language interference with target language usage', International Education Journal 1 (1), pp. 22–31.

Biber, D., S. Conrad and R. Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Borin, L. and K. Prütz. 2004. 'New wine in old skins: a corpus investigation of L1 syntactic transfer in learner language' in G. Aston, S. Bernardini and D. Stewart (eds) Corpora and Language Learners, pp. 67–87. (Volume 17.) Philadelphia, Pennsylvania: John Benjamins.

Chandler, J. 2003. 'The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing', Journal of Second Language Writing 12 (3), pp. 267–96.

Chang, J.S. and Y.-C. Chang. 2004. 'Computer assisted language learning based on corpora and natural language processing: the experience of project CANDLE' in Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning, pp. 15–23. 10 December 2004. Tokyo, Japan.

Chang, Y.C., J.S. Chang, H.J. Chen and H.C. Liou. 2008. 'An automatic collocation writing assistant for Taiwanese EFL learners: a case of corpus-based NLP technology', Computer Assisted Language Learning 21 (3), pp. 283–99.

Chicago Manual of Style, The. 2010. (Sixteenth edition) Chicago: University Of Chicago Press.

Chaudron, C. 1988. Second Language Classrooms: Research on Teaching and Learning. Cambridge: Cambridge University Press.

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum.

Council of Europe. 2002. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Crompton, P. 2011. 'Article errors in the English writing of advanced L1 Arabic learners: the role of transfer', Asian EFL Journal 50 (1), pp. 4–35.

Daniels, P.T. and W. Bright. 1996. The World's Writing Systems. New York: Oxford University Press.

Darus, S. and K.H. Ching. 2009. 'Common errors in written English essays of Form One Chinese students: a case study', European Journal of Social Sciences 10 (2), pp. 242–53.

Darus, S. and K. Subramaniam. 2009. 'Error analysis of the written English essays of secondary school students in Malaysia: a case study', European Journal of Social Sciences 8 (3), pp. 483–95.

Dekeyser, R.M. 1997. 'Beyond explicit rule learning: automatizing second language morphosyntax', Studies in Second Language Acquisition 19 (2), pp. 195–221.

Doebler, P. 2015. mada: Meta-Analysis of Diagnostic Accuracy. Available online at: https://cran.r-project.org/package=mada.

Dowle, M., A. Srinivasan, T. Short, S. Lianoglou, R. Saporta and E. Antonyan. 2015. data.table: Extension of Data.frame. R package version 1.9.6. Available online at: https://cran.r-project.org/package=data.table.

EF Education First. 2011. Cambridge University funds unique EF Education First research project. Hong-Kong: [Press Release]. Available online at: http://www.ef.com/sitecore/__~/media/efcom/efc/pdf/Press-release/PR110706Camb.pdf. (Archived at: http://www.webcitation.org/6b56cuVKy.)

Ellis, N.C. 2002. 'Frequency effects in language processing: a review with implications for explicit language acquisition', Studies in Second Language Acquisition 24, pp. 143–88.

Ellis, R. 2008. The Study of Second Language Acquisition. (Second edition.) Oxford: Oxford University Press.

Ene, E. 2008. 'Developmental stages in advanced SLA: a corpus-based analysis of academic writing by ESL graduate students', International Journal of Applied Linguistics 156 (1), pp. 53–86.

Farghaly, A. and K. Shaalan. 2009. 'Arabic natural language processing: challenges and solutions', ACM Transactions on Asian Language Information Processing 8 (4), pp. 1–22.

Ge, Z.G. 2015. 'Enhancing vocabulary retention by embedding L2 target words in L1 stories: an experiment with Chinese adult e-learners', Journal of Educational Technology and Society 18 (3), pp. 254–65.

Geertzen, J., T. Alexopoulou and A. Korhonen. 2014. 'Automatic linguistic annotation of large scale L2 databases: the EF-Cambridge Open Language Database (EFCamDat)' in R.T. Millar, K.I. Martin, C.M. Eddington, A. Henery, N.M. Miguel and A. Tseng (eds) Selected Proceedings of the 2012 Second Language Research Forum, pp. 240–54. Somerville, Massachusetts: Cascadilla Proceedings Project.

Ghadessy, M., A. Henry and R.L. Roseberry. 2001. Small Corpus Studies and ELT: Theory and Practice. Philadelphia, Pennsylvania: John Benjamins.

Granger, S. 1994. 'The learner corpus: a revolution in applied linguistics', English Today 10 (3), pp. 25–33.

Granger, S. and G. Leech. 2014. Learner English on Computer. New York: Routledge.

Guia do Acordo Ortográfico. 2008. São Paulo, Brazil: Editora Moderna.

Gustilo, L. and C. Magno. 2012. 'Learners' errors and their evaluation: the case of Filipino ESL writers', Philippine ESL Journal 8 (9), pp. 96–113.

Hadamitzky, W. and M. Spahn. 1997. A Handbook of the Japanese Writing System. North Clarendon, Vermont: Tuttle Publishing.

Hohenstein, S. and R. Kliegl. 2013. 'Eye movements reveal interplay between noun capitalization and word class during reading' in Proceedings of the 35th Annual Conference of the Cognitive Science Society pp. 2554–9. 31 July to 3 August 2013. Berlin, Germany.

Huitema, B. 2011. The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-case Studies. Hoboken, New Jersey: John Wiley and Sons.

Hutchinson, A.P. and J. Lloyd. 2003. Portuguese: An Essential Grammar. (Second edition.) London and New York: Routledge.

Ida, K., M. Nakayama and S.J. Lupker. 2015. 'The functional phonological unit of Japanese−English bilinguals is language dependent: evidence from masked onset and mora priming effects', Japanese Psychological Research 57 (1), pp 38–49.

Jarvis, S. 2000. 'Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon', Language Learning 50 (2), pp. 245–309.

Jarvis, S. 2015. 'The scope of transfer research' in L. Yu and T. Odlin (eds) New Perspectives on Transfer in Second Language Learning. Bristol: Multilingual Matters.

Jarvis, S. and A. Pavlenko. 2008. Crosslinguistic Influence in Language and Cognition. New York: Routledge.

Jenkins, F.M. 2004. Modern French Grammar: A Practical Guide. The French Review. (Volume 71. Second edition.) London and New York: Routledge.

Jiang, X., Y. Guo, J. Geertzen, D. Alexopoulou, L. Sun and A. Korhonen. 2014. 'Native language identification using large, longitudinal data' in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14), pp. 3309–12. 26–31 May 2014. Reykjavik, Iceland: European Language Resources Association (ELRA).

Kattán-Ibarra, J. and C.J. Pountain. 2003. Modern Spanish Grammar: A Practical Guide. (Second edition.) London and New York: Routledge.

Kaufman, A., S. Gettys and N. Wieda. 2006. Russian for Dummies. Indianapolis, Indiana: Wiley Publishing, Inc.

Kellerman, E. 1978. 'Giving learners a break: native language intuitions as a source of predictions about transferability', Working Papers on Bilingualism 15, pp. 59–92.

Kellerman, E. 1983. 'Now you see it, now you don't' in S. Gass and L. Selinker (eds) Language Transfer in Language Learning, pp. 112–34. Rowley, Massachusetts: Newbury House.

Klöter, H. 2004. 'Language policy in the KMT and DPP eras', China Perspectives 56, pp. 1–12.

Koda, K. 2005. Insights into Second Language Reading: A Cross-linguistic Approach. Cambridge: Cambridge University Press.

Kovac, M.M. 2011. 'Speech errors in English as foreign language: a case study of engineering students in Croatia', English Language and Literature Studies 1 (1), pp. 20–39.

Krummes, C. and A. Ensslin. 2014. 'What's hard in German? WHiG: a British learner corpus of German', Corpora 9 (2), pp. 191–205.

Laufer, B. and T. Waldman. 2011. 'Verb–noun collocations in second language writing: a corpus analysis of learners' English', Language Learning 61 (2), pp. 647–72.

Lee, Y.J. 2006. 'The process-oriented ESL writing assessment: promises and challenges', Journal of Second Language Writing 15 (4), pp. 307–30.

Lehtonen, T. and S. Karjalainen. 2008. 'University graduates' workplace language needs as perceived by employers', System 36 (3), pp. 492–503.

Lewis, M.P., G.F. Simons and C.D. Fennig. 2013. Ethnologue: Languages of the World. (Seventeenth edition.) Dallas, Texas: SIL International.

Lewis, M.P., G.F. Simons and C.D. Fennig. 2015. Ethnologue: Languages of the World. (Eighteenth edition.) Dallas, Texas: SIL International.

Little, D. 2011. 'The Common European Framework of Reference for Languages: perspectives on the making of supranational language education policy', The Modern Language Journal 91 (4), pp. 645–54.

Lunsford, A.A. and K.J. Lunsford. 2008. 'Mistakes are a fact of life: a national comparative study', College Composition and Communication 59 (4), pp. 781–806.

McCaskill, M.K. 1998. Grammar, Punctuation, and Capitalization: A Handbook for Technical Writers and Editors (NASA SP-70). Hampton, Virginia: Langley Research Center.

McEnery, T., R. Xiao and Y. Tono. 2006. Corpus-based Language Studies: An Advanced Resource Book. New York: Routledge.

Morris, L. 1998. 'Differences in men's and women's ESL writing at the junior college level: consequences for research on feedback', Canadian Modern Language Review 55 (2), pp. 219–38.

Murakami, A. 2016. 'Modeling systematicity and individuality in nonlinear second language development: the case of English grammatical morphemes', Language Learning 66 (4), pp. 834–71.

Murakami, A. and T. Alexopoulou. 2016. 'L1 influence on the acquisition order of English grammatical morphemes: a learner corpus study', Studies in Second Language Acquisition 38, pp. 365–401.

Myles, F. 2005. 'Interlanguage corpora and second language acquisition research', Second Language Research 21 (4), pp. 373–91.

Nagai, N. and F.O. Dwyer. 2011. 'The actual and potential impacts of the CEFR on language education in Japan', Synergies Europe 6, pp. 141–52.

Newell, A. and P.S. Rosenbloom. 1981. 'Mechanisms of skill acquisition and the law of practice' in J.R. Anderson (ed.) Cognitive Skills and their Acquisition, pp. 1–55. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Nezakatgoo, B. 2011. 'Portfolio as a viable alternative in writing assessment', Journal of Language Teaching and Research 2 (4), pp. 747–56.

Odlin, T. 1989. Language Transfer: Cross-linguistic Influence in Language Learning. Cambridge: Cambridge University Press.

Odlin, T. 2003. 'Cross-linguistic influence' in C.J. Doughty and M.H. Long (eds) The Handbook of Second Language Acquisition, pp. 436–85. Oxford: Blackwell Publishing.

Oswald, F.L. and L. Plonsky. 2010. 'Meta-analysis in second language research: choices and challenges', Annual Review of Applied Linguistics 30 (2010), pp. 85–110.

O'Sullivan, Í. and A. Chambers. 2006. 'Learners' writing skills in French: corpus consultation and learner evaluation', Journal of Second Language Writing 15 (1), pp. 49–68.

Peressotti, F., R. Cubelli and R. Job. 2003. 'On recognizing proper names: the orthographic cue hypothesis', Cognitive Psychology 47 (1), pp. 87–116.

Polio, C. 1997. 'Measures of linguistic accuracy in second language writing research', Language Learning 47 (1), pp. 101–43.

Proudfoot, A. and F. Cardo. 2005. Modern Italian Grammar: A Practical Guide. (Second edition.) London and New York: Routledge.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.r-project.org/.

Rayner, K. and E.R. Schotter. 2014. 'Semantic preview benefit in reading English: the effect of initial letter capitalization', Journal of Experimental Psychology: Human Perception and Performance 40 (4), pp. 1617–28.

Reder, S., K. Harris and K. Setzler. 2003. 'The multimedia adult ESL learner corpus', TESOL Quarterly 37 (3), pp. 546–57.

Revelle, W. 2016. psych: Procedures for Personality and Psychological Research. Evanston, Illinois: Northwestern University.

Ringbom, H. 1987. The Role of the First Language in Foreign Language Learning. Bristol: Multilingual Matters.

Ringbom, H. 1992. 'On L1 transfer in L2 comprehension and L2 production', Language Learning 42 (1), pp. 85–112.

Sawalmeh, M.H.M. 2013. 'Error analysis of written English essays: the case of students of the preparatory year program in Saudi Arabia', English for Specific Purposes World 40 (14), pp. 1–17.

Schachter, J. and M. Celce-Murcia. 1977. 'Some reservations concerning error analysis', TESOL Quarterly 11 (4), pp. 441–51.

Schafer, J.L. and J. Kang. 2008. 'Average causal effects from nonrandomized studies: a practical guide and simulated example', Psychological Methods 13 (4), pp. 279–313.

Schenke, H. and K. Seago. 2004. Basic German: A Grammar and Workbook. Die Unterrichtspraxis Teaching German. (Volume 10.) London: Routledge.

Sersen, W.J. 2011. 'Improving writing skills of Thai EFL students by recognition of and compensation for factors of L1 to L2 negative transfer', US-China Education Review A, 1 (3), pp. 339–45.

Shekhzadeh, E. and M. Gheichi. 2011. 'An account of sources of errors in language learners' interlanguage' in 2011 International Conference on Languages, Literature and Linguistics, pp. 159–62. (Volume 26.) 28–30 December 2011. Dubai, United Arab Emirates.

Sönmez, G. and C. Griffiths. 2015. 'Correcting grammatical errors in university-level foreign language students' written work', Konin Language Studies 3 (1), pp. 57–74.

Special Eurobarometer 243. 2006. Europeans and their Languages. Brussels: European Commission.

Spitkovsky, V.I. and D. Jurafsky. 2012. 'Capitalization cues improve dependency grammar induction' in Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure, pp. 16–22. 7 June 2012. Montreal, Canada: Association for Computational Linguistics.

Tolentino, L.C. and N. Tokowicz. 2014. 'Cross-language similarity modulates grammar instruction', Language Learning 64 (June), pp. 279–309.

Van Langendonck, W. 2008. Theory and Typology of Proper Names. Berlin: Mouton de Gruyter.

Wickham, H. 2007. 'Reshaping data with the reshape package', Journal of Statistical Software 21 (12), pp. 1–20.

Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H. 2011. 'The split-apply-combine strategy for data analysis', Journal of Statistical Software 40 (1), pp. 1–29.

Wilcox, K.C., R. Yagelski and F. Yu. 2014. 'The nature of error in adolescent student writing', Reading and Writing 27 (6), pp. 1073–94.

World Factbook, The. 2015. Mexico. Accessed 18 April 2015 at: https://www.cia.gov/library/publications/the-world-factbook/geos/mx.html. (Archived at: http://www.webcitation.org/6b563au0R.)

Yamada, J., N. Matsuura and Y. Yanase. 1988. 'Does knowledge of Romaji facilitate English reading?', The Journal of General Psychology 115 (3), pp. 229–39.

Yang, L. and S. Li. 2012. 'Topic prominence in typological interlanguage development of Chinese students' English', Journal of Cambridge Studies 7 (4), pp. 126–42.