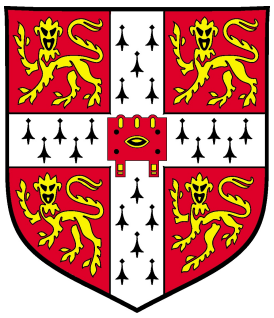
Individual Variation and the Role of L1 in the L2 Development of English Grammatical Morphemes: Insights From Learner Corpora



Akira Murakami

Department of Theoretical and Applied Linguistics

&

Hughes Hall

University of Cambridge

A dissertation submitted for the degree of Doctor of Philosophy

July 2013

Preface

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The dissertation does not exceed 80,000 words including footnotes, references, and appendices, but excluding bibliographies.

Akira Murakami
University of Cambridge
29 July, 2013

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my supervisor, Dora Alexopoulou, not only for her continuing and valuable guidance, extensive feedback, and constructive advice on the project but also for her constant encouragement, enormous patience, and unwavering support for the dissertation throughout my life as a PhD student. In particular, if any part of the data analysis is successfully communicated to the reader, I owe much of it to her. The endeavor of carrying out the project and writing up the thesis would have never been possible without her supervision.

I would also like to thank other members at the Department of Theoretical and Applied Linguistics and former Research Centre for English and Applied Linguistics. My primary appreciation goes to Henriëtte Hendriks, John Williams, and Paula Buttery for reading part of the draft of the thesis and giving me detailed and insightful comments at various stages. Their useful suggestions greatly improved the thesis in concrete ways, and the regular Research Committee Meetings with them helped my steady progress.

My thanks extend to the researchers involved in the English Profile Programme. I am especially grateful to John Hawkins and Michael McCarthy for their inspiration for and advocacy of my work at the research seminars of the programme. Their warm encouragement motivated me and increased my confidence of the work.

The dissertation also significantly benefited from the comments I received on my presentations at various conferences including EUROSLA, Corpus Linguistics, Learner Corpus Research, and Second Language Research Forum, among other smaller workshops and seminars. The practical advice from Nick Ellis at the Language Acquisition and Language Processing Research Cluster Workshop was notably helpful to polish part of Chapter 2 of the thesis.

The empirical data of the thesis came entirely from existing learner corpora. I am

thankful to Cambridge ESOL and Cambridge University Press for granting me access to the Cambridge Learner Corpus, and Education First for supplying the data for EF-Cambridge Open Language Database.

I am deeply indebted to my friends at Hughes Hall, Cambridge University Table Tennis Club, Hughes Hall Table Tennis Club, Hughes Hall Chess Club, and Cambridge University Japanese Interdisciplinary Forum (Toirokai). They brought me moments of joy in the otherwise isolated everyday life of a research student. Their diversity has enormously enhanced my life and contributed to the value of living in Cambridge. I also thank Hughes Hall Library for providing me with a pleasant environment to stay in. It is undoubtedly the place I spent the longest time in since the start of my PhD and where most part of the work for the dissertation was completed.

Financial support from Japan Student Services Organization is genuinely acknowledged. It is when I learned about their scholarship that I decided to pursue a PhD outside of Japan. I am likewise thankful to Hughes Hall Travel Grant for funding me to attend conferences.

Finally, none of these would have been possible without the long-term moral support from my family. I appreciate my parents, my sister, and my grandparents for their involvement and belief in me.

Abstract

The overarching goal of the dissertation is to illustrate the relevance of learner corpus research to the field of second language acquisition (SLA). The possibility that learner corpora can be useful in mainstream SLA research has a significant implication given that they have not been systematically explored in relation to SLA theories. The thesis contributes to building a methodological framework to utilize learner corpora beneficially to SLA and argues that learner corpus research contributes to other disciplines. This is achieved by a series of case studies that quantitatively analyze individual variation and the role of native language (L1) in second language (L2) development of English grammatical morphemes and explain the findings with existing SLA theories.

The dissertation investigates the L2 development of morphemes based on two large-scale learner corpora. It first reviews the literature and points out that the L2 acquisition order of English grammatical morphemes that has been believed universal in SLA research may, in fact, vary across the learners with different L1 backgrounds and that individual differences in morpheme studies have been relatively neglected in previous literature. The present research, thus, provides empirical evidence testing the universality of the order and the extent of individual differences.

In the first study, the thesis investigates L1 influence on the L2 acquisition order of six English grammatical morphemes across seven L1 groups and five proficiency levels. Data drawn from approximately 12,000 essays from the Cambridge Learner Corpus establish clear L1 influence on this issue. The study also reveals that learners without the equivalent morpheme in L1 tend to achieve an accuracy level of below 90% with respect to the morpheme even at the highest proficiency level, and that morphemes requiring learners to learn to pay attention to the relevant distinctions in their acquisition show a stronger effect of L1 than those which only require new form-meaning mappings. The findings are interpreted

under the framework of thinking-for-speaking proposed by Dan Slobin.

Following the first study, the dissertation exploits EF-Cambridge Open Language Database (EFCamDat) and analyzes the developmental patterns of morphemes, L1 influence on the patterns, and the extent to which individual variation is observed in the development. Based on approximately 140,000 essays written by 46,700 learners of 10 L1 groups across a wide range of proficiency levels, the study found that (i) certain developmental patterns of accuracy are observed irrespective of target morphemes, (ii) inverted U-shaped development is rare irrespective of morphemes, (iii) proficiency influences the within-learner developmental patterns of morphemes, (iv) the developmental patterns at least slightly vary depending on morphemes, and (v) significant individual variation is observed in absolute accuracy, the accuracy difference between morphemes, and the rate of development. The findings are interpreted with dynamic systems theory (DST), a theory of development that has recently been applied to SLA research. The thesis further examines whether any systematic relationship is observed between the developmental patterns of morphemes. Although DST expects that their development is interlinked, the study did not find any strong relationships between the developmental patterns. However, it revealed a weak supportive relationship in the developmental pattern between articles and plural *-s*. That is, within individual learners, when the accuracy of articles increases, the accuracy of plural *-s* tends to increase as well, and vice versa.

Table of Contents

Appendices **261**

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Introduction

Grammatical morphemes (e.g., *-ed*, *-ing*, articles) are basic building blocks of language, encoding various concepts including temporality, aspect, possession, person, and discourse features. Due to their importance, it is essential for second language (L2) learners to acquire them. However, despite their high frequency and communicative necessity, the acquisition of grammatical morphemes has been known to be notoriously difficult for L2 learners (cf., N. C. Ellis, 2008; MacWhinney, 2008). For this reason, second language acquisition (SLA) researchers have undertaken much research on their acquisition.

The thesis revisits an early series of studies on L2 morpheme acquisition, so-called "morpheme studies", which have investigated the acquisition order of grammatical morphemes and claimed that the L2 acquisition order of English grammatical morphemes is universal irrespective of learners' native language (L1) backgrounds (e.g., Dulay & Burt, 1973, 1974). The theoretical rationale behind the research was to refute Contrastive Analysis Hypothesis, a theory that is related to behaviorism and claimed that ease and difficulty in acquiring L2 is determined by similarities and differences between L1 and L2. If L1 is shown to have no influence on L2 acquisition, it constitutes evidence against the hypothesis, which in turn denies behaviorism. This is indeed what morpheme studies demonstrated, and they claimed that the invariant order of acquisition supports the view that L2 acquisition is (at least partially) driven by an innate mechanism (cf. Goldschneider & DeKeyser, 2001).

However, SLA has progressed significantly over the past 40 years, and all major current theories, generative or functional, acknowledge L1 influence (e.g., N. C. Ellis, 2006; Ionin & Montrul, 2010). As will be reviewed later, thinking-for-speaking (Slobin, 1996), for instance, makes specific predictions as to the strength of L1 influence across morphemes, and

Luk and Shirai (2009) surveyed the literature on morpheme studies and suggested that the acquisition order may vary across L1 groups. In sum, although it has been believed in SLA that the acquisition order is universal, recent theories and some empirical evidence point towards L1 influence on the issue. The thesis, therefore, investigates whether L1 affects the acquisition order of grammatical morphemes. It also looks into individual variation in the longitudinal developmental pattern of morphemes and L1 influence on the patterns.

1.2 Morpheme Studies

1.2.1 General Rationale of Morpheme Studies

Morpheme studies originate from Brown (1973), who examined the process of L1 English acquisition in three children in the U.S. As part of the study, Brown analyzed the development of 14 grammatical morphemes and found that their developmental patterns were similar across the children.

Following Brown's study, similar investigations emerged in SLA in order to test whether L1 and L2 acquisition go through similar processes. Dulay and Burt (1973) investigated the acquisition order of eight grammatical morphemes (present progressive *-ing*, plural *-s*, irregular past tense, possessive *'s*, articles, third person *-s*, copula be, and auxiliary be) by three groups of five- to eight-year-old children learning English as an L2; 95 Mexican-Americans, 26 Spanish, and 30 Puerto Ricans. The researchers predicted that the order would differ from L1 acquisition because L2 learners already possess the concepts that children have to learn at the same time with L1. They identified all the instances where the target morphemes were obligatory in English, and calculated, for each morpheme, the number of correctly supplied morphemes divided by the number of obligatory contexts. With the suppliance of a misformed morpheme (e.g., *He's eats* instead of *He's eating* for progressive *-ing*), half a point was given. This scoring method is called Suppliance in Obligatory Context (SOC). The result showed difference from the acquisition order reported in Brown

(1973) and also an overall similarity among the three groups; all three groups achieved the highest SOC score in plural *-s*, followed by present progressive *-ing*, copula be, auxiliary be, articles, third person *-s*, irregular past tense, and possessive *'s* at the lowest. This led Dulay and Burt to suggest a universal order of morpheme acquisition.

To further confirm this finding, Dulay and Burt (1974) analyzed whether the learners' L1 affects the order. They targeted 11 morphemes and tested whether their acquisition order differed in six- to eight-year-old Spanish-speaking (n = 60) from Chinese-speaking (n = 55) learners of English. They found a high correlation between the two groups of learners, from which they concluded that there is a consistent acquisition order of grammatical morphemes across L2 learners with different L1 backgrounds[1]. The same order has also been confirmed when comparing ESL (English as a Second Language) to EFL (English as a Foreign Language) learners (Pica, 1983) and instructed to non-instructed learners (Larsen-Freeman, 1975).

Criticism of morpheme studies. An early criticism of morpheme studies was that the order conceals the accuracy distance between morphemes. This distance is important regarding the development of the learner because when morphemes are ordered simply according to the order of accuracy, a 1% difference in the accuracy may make as much an impact as a 50% difference. In order to counter this claim, Krashen (1977), by reviewing the literature, grouped up the morphemes that tended to have similar accuracy ranks and proposed an acquisition order presented in Figure 1. This order has become known as the "natural order" of L2 English morpheme acquisition, and is believed to be the order that

[1]Note that all the L2 studies cited here employed accuracy order, from which they derived acquisition order on the assumption that accuracy reflects the degree of acquisition (cf. de Villiers & de Villiers, 1973). Although a few morpheme studies investigated acquisition order longitudinally (e.g., Hakuta, 1976), the majority have tackled this issue with accuracy order (e.g., Andersen, 1978, in addition to the studies cited here). An alternative criterion of acquisition is emergence of the form (e.g., Pienemann, 1998). A proper comparison of the two approaches is beyond the scope of the thesis.

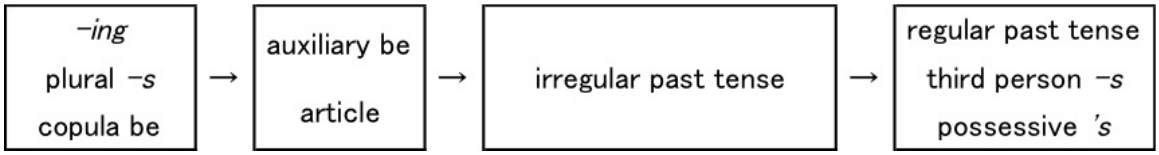


Figure 1. Natural Order of Acquisition Proposed by Krashen (1977)

L2 learners universally follow in acquiring English grammatical morphemes.

It has, thus, been fairly accepted in SLA that there is a fixed order of acquisition with regards to the grammatical morphemes, as the following citations from recent SLA textbooks show:

> These morpheme acquisition studies attracted criticism . . . . However, the basic argument that both child and adult learners of English as an L2 developed accuracy in a number of grammatical morphemes in a set order, no matter what the context of learning (classroom, naturalistic, mixed), survived the critique. . . . The existence of such an order suggested that L2 learners are guided by internal principles which are largely independent of their first language; this was a serious blow for Contrastive Analysis. (Mitchell, Myles, & Marsden, 2013, p.40)

Similarly,

> The accuracy order has been shown to be relatively similar for both young and adult L2 learners, for both naturalistic and instructed learners, and regardless of L1 background or whether the data are collected orally or via writing. (Ortega, 2009, p.124)

Meisel (2011, p.65) also supports the view.

> I believe that Larsen-Freeman and Long (1991, p.92) are right in concluding:

> In sum, despite admitted limitations in some areas, the morpheme studies provide strong evidence that ILs [interlanguages] exhibit common accuracy/acquisition orders.

Thus, as Luk and Shirai (2009) claim, "SLA textbooks mostly 'teach' introductory students of SLA that the order of acquisition is universal" (p.724).

1.2.2 Explanation of the Morpheme Acquisition Order

In the 1980's, the focus of morpheme studies shifted from description to explanation, that is, the identification of the determinants of the observed order. A number of factors have been proposed to account for the observed order, including frequency (Larsen-Freeman, 1976), syntactic properties (Zobl & Liceras, 1994), and communicative value (VanPatten, 1984), among others. Although various proposals have been made, there is to date no complete theory to explain the acquisition order of grammatical morphemes. It is in fact very likely that the observed order cannot be attributed to any single determinant but to the interaction of multiple determinants. Goldschneider and DeKeyser (2001) attempted to step forward to revealing this complexity. It meta-analyzed 12 empirical morpheme studies and claimed to have analyzed the extent to which the following five factors account for the observed acquisition order; perceptual salience, semantic complexity, morphophonological regularity, syntactic category, and frequency. The target morphemes were articles, regular past tense *-ed*, plural *-s*, possessive *'s*, progressive *-ing*, and third person *-s*. Unfortunately, though, their data analysis is flawed in a significant way. See Appendix A for the details. Nonetheless, their study remains very important in morpheme studies because of its multifactorial nature, seeking multiple sources for explanation simultaneously. Although multifactorial studies are common in recent L2 research (e.g., Blom, Paradis, & Sorenson Duncan, 2012; Gries & Deshors, in press; Gries & Wulff, 2009; Wulff, Ellis, Römer,

Bardovi-Harlig, & Leblanc, 2009), their study over a decade ago is perhaps one of the first that attempted relative weighting among predictors, with a possible exception of some variationist SLA studies (e.g., Berdan, 1996; Regan, 1996; Young, 1988, 1996).

1.2.3 L1 Influence in Morpheme Studies

The general conclusion of morpheme studies that the order of acquisition is universal is in line with literature on cross-linguistic influence. It has been accepted in cross-linguistic influence research that morphological cross-linguistic influence is weak, at least when compared to phonology or lexis (Jarvis & Pavlenko, 2007). Note that cross-linguistic influence refers to cross-linguistic influence of semantic functionality, as opposed to that of form, which can only occur between typologically similar languages. These views, however, have been seriously challenged recently especially with respect to article acquisition. A number of studies targeting a variety of L1 groups found L1 influence on the L2 acquisition of English articles by using a range of tasks (J. A. Hawkins & Buttery, 2010; Jarvis, Castañeda Jiménez, & Nielsen, 2012; Snape, 2005, 2008; Zdorenko & Paradis, 2012).

Luk and Shirai (2009) reviewed the literature on morpheme studies and investigated whether Japanese, Korean, Chinese, and Spanish learners of English acquire grammatical morphemes differently from the order predicted by the natural order. The result showed a clear tendency that Japanese, Korean, and Chinese learners of English acquire possessive *'s* earlier than, and plural *-s* later than, predicted by the natural order. Luk and Shirai (2009) concluded that "the acquisition order of grammatical morphemes is highly affected by the learner's L1 such that it is possible to predict, to some extent, what is difficult and what is easy for language learners based on their L1s" (p.742).

However, one weakness of their study is that it drew from a very diverse set of studies varying in length (longitudinal vs. cross-sectional), criteria of acquisition (accuracy order vs. the threshold of 80% or 90% correct), and mode (oral vs. written production).

Furthermore, their investigation was limited to three grammatical morphemes. Given the pervasive impact of L1 on L2 acquisition, I predict L1 influence on the acquisition of other morphemes, too. Therefore, the dissertation addresses the following question: Is the acquisition order of L2 English grammatical morphemes affected by L1? This will be empirically tested by investigating whether the morpheme accuracy order varies across learners with different L1 backgrounds but is consistent within each L1 group.

Moreover, the thesis will further explore L1 influence on morpheme accuracy by examining (i) the strength of L1 influence and (ii) whether all morphemes are equally sensitive to L1 influence. I hypothesize that the lack of the equivalent feature in L1 leads to a lower accuracy rate. Although this has been documented especially for articles (e.g., Ionin & Montrul, 2010, in addition to the studies cited earlier), no systematic analysis, to the best of my knowledge, has targeted as many as six morphemes. Additionally, as few studies have investigated the difference in the strength of L1 influence across morphemes, it is still unclear whether the effect of L1 varies across morphemes. Absence of empirical data, however, does not entail absence of hypotheses. One framework to conceptualize the nature of cross-linguistic influence is thinking-for-speaking, to which I will turn below.

1.3 Thinking-for-Speaking

Thinking-for-speaking is a framework proposed by Slobin (1996, 2003, 2008). Slobin claims that speakers of a particular language conceptualize the world in the way required by the language, and that the concepts whose verbalization the language requires draw the speakers' attention. In other words, concepts develop as people learn the dimension of the world the language they speak requires verbalization of. In this sense, concepts, or attention toward the concepts, are language-specific. Applying this idea to L1 influence on L2 acquisition, Slobin also predicts that L2 learners typically face greater difficulty acquiring the concepts that necessitate the distinctions unrequired in their L1s. The point

is well explained by Slobin (1996).

> I propose that the grammaticized categories that are most susceptible to SL [source language] influence have something important in common: they cannot be experienced directly in our perceptual, sensorimotor, and practical dealings with the world. . . . [O]nly language requires us to categorize events as ongoing or completed, objects as at rest or as at the end point of a trajectory, and so forth. . . . I would imagine, for example, that if your language lacked a plural marker, you would not have insurmountable difficulty in learning to mark the category of plurality in a second language, since this concept is evident to the non-linguistic mind and eye. . . . Plurality and manipulation are notions that are obvious to the senses. Yet there is nothing in everyday sensorimotor interactions with the world that changes . . . when you refer to the same object in successive utterances as "a car" and "the car". Distinctions of aspect, definiteness, voice, and the like are, par excellence, distinctions that can only be learned through language, and have no other use except to be expressed in language. They are not categories of thought in general, but categories of thinking for speaking. It seems that once our minds have been trained in taking particular points of view for the purposes of speaking, it is exceptionally difficult for us to be retrained. (p.91, emphasis removed)

Therefore, as far as grammatical morphemes are concerned, the framework of thinking-for-speaking predicts that articles and aspectual distinctions such as progressiveness are more sensitive to L1 influence than plural *-s*. Two other morphemes that are often targeted in morpheme studies, third person *-s* and possessive *'s*, are less likely to be affected by L1 because the concepts they encode (person, number, and possession) are common across languages. That is, even the learners whose L1s do not morphologically encode person are

still likely to be able to draw distinctions between first, second, and third person. Similarly, possession is often marked even in the languages that do not have equivalent features to English possessive *'s*, typically by the form of postnominal modification. Based on the framework, I predict that different morphemes are differentially sensitive to L1 influence. The present study will empirically test the hypothesis.

1.4 The Importance of Individual Variation

A criticism against morpheme studies is that they obscure individual differences by aggregating data over individual learners (de Bot, Lowie, & Verspoor, 2007a; Kwon, 2005). The dissertation, therefore, investigates whether and to what extent individual differences are present in the L2 acquisition of English grammatical morphemes. For this purpose, the thesis exploits longitudinal data because having multiple data points per learner greatly facilitates the examination of inter-learner differences. The following research question is thus addressed: What are the longitudinal developmental patterns of L2 English grammatical morphemes? The thesis is the first that targets a large number of learners and discloses individual differences in the developmental pattern. Additionally, the dissertation will test the effect of L1 on the pattern. Given the large impact of L1 on L2 acquisition in general (Jarvis & Pavlenko, 2007), I expect L1 to have some influence on the longitudinal developmental pattern of morphemes, too.

An appropriate framework to situate individual differences in SLA is dynamic systems theory (DST), which views inter- and intra-learner variability as an important source of information that signals development. Due to its emphasis on the role of variability in language acquisition, the framework is suited to interpret individual differences in morpheme acquisition as well. I chose DST over other paradigms because it is the only SLA theory that I am aware of that explicitly recognizes the importance of variation in L2 development. The following quotation from de Bot et al. (2007a) well summarizes the framework.

> [F]rom a DST perspective, a language learner is regarded as a dynamic subsystem within a social system with a great number of interacting internal dynamic sub-sub systems, which function within a multitude of other external dynamic systems. The learner has his/her own cognitive ecosystem consisting of intentionality, cognition, intelligence, motivation, aptitude, L1, L2 and so on. The cognitive ecosystem in turn is related to the degree of exposure to language, maturity, level of education, and so on, which in turn is related to the SOCIAL ECOSYSTEM, consisting of the environment with which the individual interacts. For any system to grow, a minimal amount of force or resources is needed. . . . Each of these internal and external subsystems is similar in that they have the properties of a dynamic system. They will always be in flux and change, taking the current state of the system as input for the next one. A small force at a particular point in time may have huge effects (butterfly effect) and a much stronger force at another point in time may not have much effect in the long run. Each system has its own attractor and repeller states; however, variation is inherent to a dynamic system, and the degree of variation is greatest when a (sub) system moves from one attractor state to the other. Flux - growth or decline - is non-linear and cannot be predicted exactly. (p.14; emphasis in the original)

The framework, thus, predicts individual variation. It also predicts L1 influence based on a feature called “attractor states,” which will be discussed later. Since the framework is relatively new in SLA, I will explain it at some length below and go on to explain why it predicts individual variation and L1 influence.

1.4.1 Characteristics of DST

DST has its roots in mathematics (de Bot, 2008; de Bot et al., 2007a) and has been applied to a range of disciplines including developmental psychology (Thelen & Smith, 1994) and language acquisition (van Geert, 1995). De Bot and Larsen-Freeman (2011, p.9) raise the following nine characteristics of dynamic systems;

1. Sensitive dependence on initial conditions
2. Complete interconnectedness
3. Nonlinearity in development
4. Change through internal reorganization and interaction with the environment
5. Dependence on internal and external resources
6. Constant change, with chaotic variation sometimes, in which the systems only temporarily settle into "attractor states"
7. Iteration, which means that the present level of development depends critically on the previous level of development
8. Change caused by interaction with the environment and internal organization
9. Emergent properties

1, 2, 3, 5, and 6 are particularly relevant to the dissertation, and I will focus on their explanation below.

Sensitive dependence on initial condition refers to the fact that even a small difference in the initial state might result in a large difference, or vice versa. In language acquisition research, this means that the minimal difference between learners may result in completely

different outcomes. We can interpret L1 influence under this idea, as the following quotation shows. “The idea that the L1 may play an important role in L2 acquisition is in line with the DST notion of sensitive dependence on initial conditions. It is much easier for learners to learn languages that are similar than languages that are different. It also implies that learners with different L1s have different problems in learning the same L2” (Verspoor & Behrens, 2011, p.30). Therefore, differences between learners, regardless of the scale of the difference, are likely to result in different outcomes under the DST.

Complete interconnectedness refers to the phenomenon where all the systems are interlinked to each other, and change in one system causes change in all the others. In L2 acquisition, the proficiency of a learner depends on various factors such as his/her proficiency at the previous state, his/her attitude, the amount of language contact, and the use of strategies or metalinguistic awareness (de Bot & Larsen-Freeman, 2011). As is claimed by de Bot, Lowie, and Verspoor (2005), the idea is also in line with multicompetence (Cook, 1995), which views multiple languages within individuals as one system. L1 and L2 should not be considered separately but the two language systems are interrelated and form one larger language system. Although the focus is somewhat different, this, again, points toward L1 and other cross-linguistic influence because L1 and L2 are intrinsically connected within learners.

Nonlinearity, as mentioned earlier, means that the change in output is not necessarily proportional to the change in input. An example is U-shaped development because the length of time and accuracy are not in a linear relationship (Larsen-Freeman, 1997).

With respect to dependence on resources, systems require resources to develop, and in this sense development is constrained by resources. It depends on both internal resources such as learners’ memory capacity, their motivation, and their attention, and external resources such as input or space to learn (de Bot, 2008; de Bot & Larsen-Freeman, 2011).

As described earlier, systems are constantly interacting with internal and external variables, and they are never the same across time. However, when a change occurs as a result of interactions, the system can settle into a new state with less variability than before. The state is called an *attractor state* and is defined as "the state the systems [*sic*] prefers to be in over other states at a particular point in time" (de Bot & Larsen-Freeman, 2011, p.14). The opposite state, where the system disfavors to settle in, is called a *repeller state*. The idea of attractor and repeller states is often illustrated with an example of a rolling ball.

> The notion of development and attractor states are somewhat analogous to a ball rolling over a surface with holes and bumps, with the ball's trajectory as development, the holes as attractor states and the bumps as repeller states. The holes can be shallow or deep, and the deeper the hole is, the more energy is needed to get the ball out of the hole and make it move on to the next hole. (de Bot et al., 2007a, p.8)

A related notion to attractor states is the *basin of attractor*, which refers to an area around attractor states where the system moves toward the attractor. Within the region, even the states with different initial conditions will converge at the attractor (Carver & Scheier, 1999; Nowak, Vallacher, & Zochowski, 2005). To continue the example of a rolling ball above, it is a slightly depressed plain that draws the ball at a fairly distant place (de Bot & Larsen-Freeman, 2011).

In psychology, goals are often viewed as attractors (Carver & Scheier, 1999; Nowak et al., 2005). In the context of SLA, this means that target languages form attractor states toward which learners' interlanguage systems are drawn. However, fossilization can also be a kind of attractor states (de Bot & Larsen-Freeman, 2011; Larsen-Freeman, 1997) because, in fossilization, the developing system (interlanguage) falls into a state where it does not seem to develop further. In a similar vein, L1 can form an attractor state (Larsen-

Freeman, 1997; Plaza-Pust, 2008) and invite similar patterns within L1 groups. If an attractor state draws the system toward the norm of the target language, the effect is called positive transfer. If it draws the system toward non-target-like forms, the effect emerges as negative transfer. Due to the influence from multiple attractor states that are difficult to foresee, the path dynamic systems take is difficult to predict, resulting in chaotic variation.

These are some of the characteristics of the DST approach to L2 acquisition.

1.4.2 Variability in DST

The typical view of variability in language acquisition research is variability as error. The so-called measurement-error-hypothesis regards variability as stemming from measurement errors, or the deviation from "true" competence (van Geert & van Dijk, 2002). DST, however, places a significant role on intra- and inter-learner variability, and DST researchers believe variability includes important information on the developmental process (van Dijk, Verspoor, & Lowie, 2011; van Geert & van Dijk, 2002).

An emerging line of research in SLA that looks into within-learner variability under the DST framework is the studies on the relationship in the developmental patterns between multiple measures (Caspi, 2010; Spoelman & Verspoor, 2010; Verspoor, Lowie, & van Dijk, 2008; Verspoor & van Dijk, 2011). The relationship can take three forms (Verspoor & van Dijk, 2011); (i) supportive, where the target linguistic features develop in a similar way, (ii) competitive, where the features compete with each other and develop in an alternating fashion, and (iii) conditional, where the development of one feature is required for the other to develop. Previous studies have identified competitive (Caspi, 2010; Spoelman & Verspoor, 2010; Verspoor et al., 2008; Verspoor & van Dijk, 2011) and supportive (Caspi, 2010; Spoelman & Verspoor, 2010; Verspoor & van Dijk, 2011) relationships between various linguistic features.

To the best of my knowledge, no study has investigated the relationships in the develop-

mental patterns of grammatical morphemes. This, however, does not mean that variability is absent in the development of morphemes (de Bot et al., 2007a). Even within the same L1 group, Koike (1983), reported in Luk and Shirai (2009), found slightly different orders of acquisition of nine grammatical morphemes between three Japanese learners of English. Regarding intra-learner variability, few studies have investigated the longitudinal development of grammatical morphemes against which the presence of intra-learner variability can be tested. The very origin of morpheme studies, Brown (1973), is one of them, however. For example, Sarah, one of the participants in Brown (1973), exhibits ups and downs in the accuracy development of present progressive *-ing*, especially when its accuracy is still low, and also in plurality marker to a lesser extent.

1.4.3 L1 Influence in DST

Although DST gives emphasis to the dynamic nature of L2 development and intra-learner variability as its result, it is capable of handling stability or systematicity as well (de Bot, Lowie, & Verspoor, 2007b; van Geert, 2008). L1, when viewed as a language system as a whole, is a dynamic system. However, as was reviewed earlier, it brings a rather systematic effect to L2 development. How can this apparent non-dynamic effect be explained under the framework of DST? A key notion is attractors. Attractor states created by learners' L1 pull the system, or interlanguage, toward certain states, and as a result, developmental patterns can be more similar within L1 groups than between them.

Dörnyei (2009) reformulated traditional individual differences research in SLA under the DST framework, and in its course claimed that individual differences variables (e.g., motivation) can act as powerful attractors. In the presence of such strong attractor states, the systems with different starting points will converge at the same state at the end. This process limits intra-learner variability and generates the inter-learner similarity in the development (Nowak et al., 2005). Even though Dörnyei's (2009) primary concern was

individual differences variables such as motivation and affect, the analysis should fully apply to L1 influence.

Another way to view L1 influence from the DST perspective is to regard L1 as an internal resource that helps learners by bringing with it the conceptual framework to the task of L2 acquisition. As Ringbom (2007) and Ringbom and Jarvis (2009) point out, cross-linguistic influence is caused by the similarities between multiple languages in that there is nothing to transfer if no similarity is assumed or perceived by the learner. Under this framework, negative L1 influence can be interpreted as the lack of perceived similarity. Based on the view, L1 and the concepts encoded in L1 provide an extremely rich resource that learners can draw on in learning and using L2. Combined with the sensitivity of the system to initial conditions, L1 is likely to exert a strong influence on L2 acquisition.

1.5 Research Questions

The thesis revisits and builds on morpheme studies by (i) focusing on L1 influence and (ii) taking advantage of large-scale learner corpora that emerged after the initiation of morpheme studies. The specific research questions addressed in the dissertation are divided into two types; empirical and modeling. The empirical research questions are as follows.

1. Is the acquisition order of L2 English grammatical morphemes affected by L1?
2. Is L1 influence equally strong across morphemes?
3. What are the longitudinal developmental patterns of L2 English grammatical morphemes?
4. Is the pattern affected by L1?

Research Question 1 directly address the issue of acquisition order, while Research Question 2 asks if different morphemes are differently sensitive to L1 influence. Research Ques-

tion 3 and 4 try to disclose the individual variation in morpheme development. The research questions on modeling are as follows.

5. How can L1 influence be modeled?

6. How can pseudo-longitudinal and longitudinal development be modeled?

Research Question 5 tries to capture transfer effect. Thinking-for-speaking suggests that learners without the target morpheme in their L1s are likely to suffer negative L1 influence, and the effect is stronger if the concept encoded by the morpheme is non-transparent. Research Question 6 attempts to model longitudinal development. The term *pseudo-longitudinal* means that data are gathered from learners at a single point in time (as in the cross-sectional design) but include the learners at successive levels of proficiency to allow virtual tracking of learners (Jarvis & Pavlenko, 2007). Pseudo-longitudinal development, therefore, refers to the difference in ability between lower proficiency learners and higher proficiency learners.

Note that the thesis analyzes the accuracy of grammatical morphemes. Comparing learner language with target language in this way has been controversial in SLA research (R. Ellis, 2008), particularly from the viewpoint of comparative fallacy (i.e., the idea that learner language should not be measured against the native speaker's norm of the target language; Bley-Vroman, 1983), and the debate fully applies to corpus-based error analysis. Whereas the analyses in the thesis are subject to the same problems and limitations, Chapter 4 employed a clustering technique by which the learners with similar developmental patterns were clustered without any reference to external variables. This can be a promising way forward to analyze learner language as it is.

Chapter 2: L1 Influence on the Acquisition Order of English Grammatical Morphemes

2.1 Introduction

The study reported in this chapter directly investigates the accuracy order of morphemes and whether the strength of L1 influence varies across L1 groups.

2.1.1 Methodological Challenges of Transfer Studies

A central issue in the thesis is L1 influence, and a crucial stage of transfer studies is the identification of transfer effects. To this end, Jarvis (2000) claims that three types of evidence are generally necessary in order to identify instances of transfer; intragroup homogeneity, intergroup heterogeneity, and crosslinguistic performance congruity. Intragroup homogeneity means that, assuming that the direction of transfer is from L1 to L2, learners from the same L1 background must exhibit a similar tendency. To give an example in morpheme studies, learners with the same L1 must show a similar order of morpheme acquisition. Intergroup heterogeneity, on the other hand, means that the learners who do not share L1s should exhibit differences. In the case of morpheme studies, learners with different L1s should show different orders. The final type of evidence, crosslinguistic performance congruity, refers to the similarity between the performance in L1 and the performance in L2. In morpheme studies, early acquisition of possessive *'s* by Japanese learners of English can be explained by the existence of a corresponding particle in Japanese (*-no*). This congruity shows that the phenomenon (early acquisition of *'s*) is attributable to (a property of) L1.

Intragroup homogeneity and intergroup heterogeneity can be relative (Jarvis & Pavlenko, 2007). In other words, there is no absolute criterion to determine whether intragroup homogeneity or intergroup heterogeneity is satisfied. Rather, the two should be compared and

the differences within groups should be smaller than those between groups.

With respect to the requirement of intergroup heterogeneity, it is desirable to compare groups in a single research design. Unfortunately, to date there have been few studies comparing morpheme acquisition orders between learners with different L1s in a single study (Kwon, 2005). The present study fills this gap.

2.1.2 Research Questions

In this chapter, I investigate the following questions.

1. Do learners' L1s affect the accuracy order of L2 English grammatical morphemes? In particular,
   (i) Is the accuracy order consistent within each L1 group?
   (ii) Is the accuracy order different between L1 groups?
   (iii) If any, is the difference in accuracy order motivated by a property of L1?
2. How strong is L1 influence in determining the accuracy of English grammatical morphemes compared to other factors such as general proficiency?
3. Are all English grammatical morphemes equally affected by L1, or are the morphemes requiring the distinctions unnecessary in learners' L1s more sensitive to L1 influence, as thinking for speaking predicts?

2.2 Method

2.2.1 Target Morphemes

The target morphemes of the present study are the ones that have been most often dealt with in the literature and were also targeted in Goldschneider and DeKeyser's (2001) meta-analysis of morpheme studies; articles, past tense *-ed*, plural *-s*, possessive *'s*, progressive

-ing, and third person *-s*. Articles included both indefinite (*a*, *an*) and definite (*the*) forms. Past tense *-ed* only included regular past tense forms that end in *-ed* and not irregular ones (e.g., *went*, *ate*) or modal verbs (e.g., *would*, *could*). The decision was based on Dulay and Burt (1974). No other use of *-ed*, such as passive voice or participial use, was targeted. Plural *-s* included both *-s* and *-es* but not irregular forms (e.g., *teeth*, *children*). Possessive *'s* included all possessive markers with an apostrophe (*'s*, *s'*). Progressive *-ing* targeted all progressive uses of *-ing* regardless of tense and aspect (e.g., present progressive, past progressive, present perfect progressive). It, however, did not target other uses of *-ing* such as gerund or participial. Third person *-s* included *-s*, *-es*, and *has*. *Don't/doesn't* as a negative marker was not counted, but *have/has* as an auxiliary verb, including the negative form (*haven't/hasn't*), or the modal use (*have/has to*) was counted. The reason for excluding *don't/doesn't* was in order to avoid doubly counting third person *-s* errors in such sentences as *He don't eats breakfast*. Here, if *don't* is counted as an omission error and *eats* as an overgeneralization error, two errors are tallied for one main verb, and it can unfairly lower the accuracy score of third person *-s*. Note that if a verb with third person *-s* was wrongly supplied instead of that with past tense *-ed* or a participial form (e.g., *plays* instead of *played*, *wakes* instead of *woken*), the instance was not counted as an error of third person *-s*. *Be* verbs (e.g., *is, am, were*) were not included in third person *-s* because it has three person contrasts unlike other verbs and the literature on morpheme studies has typically treated them separately (e.g., Dulay & Burt, 1974).

2.2.2 Target L1 Groups

L1 Japanese, Korean, Spanish, Russian, Turkish, German, and French learners of English were targeted in the study. They were selected so that learners with typologically diverse L1s could be compared.

2.2.3 Corpus

The Cambridge Learner Corpus (CLC) is a corpus consisting of learners' written responses to various examinations of Cambridge ESOL. It has been collaboratively developed by Cambridge ESOL and Cambridge University Press. The corpus currently contains 45 million words from 135,000 learner essays, and one third to one half of the corpus has been manually error-tagged by linguists at Cambridge University Press (J. A. Hawkins & Buttery, 2010; J. A. Hawkins & Filipović, 2012, see also Nicholls, 2003). Each script has several versions including an original text written by a learner, an error-tagged text, a corrected text, and texts with part-of-speech (POS) tags. POS tags were annotated with the Robust Accurate Statistical Parser (RASP; Briscoe, Carroll, & Watson, 2006), which utilizes the CLAWS tagset.

Subcorpus. The subcorpus used in the present study contains the exam scripts written by the test takers of Cambridge ESOL Main Suite Examinations. It consists of five proficiency levels, corresponding to A2 to C2 of the Common European Framework of Reference levels; KET (A2), PET (B1), FCE (B2), CAE (C1), and CPE (C2). Although all the test takers are not necessarily at the target level of the exams, it is expected that the proficiency levels of test takers are higher in exams which ask for higher proficiency.

Comparing Guiraud's index across the five levels. To verify this assumption, Guiraud's index was computed for each learner essay. This is a measure of lexical richness and is calculated by the number of types (i.e., the number of unique words) divided by the square root of the number of tokens (i.e., the number of words). Since this measure is not free from the influence of text length (Baayen, 2008; Hout & Vermeer, 2007), only the first 100 words were used for calculation. When an entire essay was shorter than 100 words, Guiraud's index was computed for the whole essay. The average values per L1 group per level are shown in Table 1. The value tends to be relatively stable within each exam level

Table 1

Mean Guiraud's Index for Each L1 and Proficiency Group

L1	CPE	CAE	FCE	PET	KET
L1 Japanese	6.95	7.20	6.92	6.43	4.47
L1 Korean	6.92	7.11	7.02	6.56	4.88
L1 Spanish	6.93	7.13	6.82	6.49	4.75
L1 Russian	6.98	7.15	7.06	6.51	5.14
L1 Turkish	6.87	6.96	6.95	6.41	4.91
L1 German	7.00	7.23	6.93	6.67	4.84
L1 French	6.90	7.22	6.93	6.65	4.67
Average	6.94	7.14	6.95	6.53	4.81

and gradually moves up from low to high levels, indicating lexical development. Although the reason for the decrease of Guiraud's index from CAE to CPE is unclear, it seems safe to assume that test takers at the same exam level are of roughly equal proficiency, and that those at different exam levels represent learners at different proficiency levels.

Note that Guiraud's index is not the only measure that captures (part of) proficiency. Lu (2010), for example, lists 14 indices that measure the syntactic complexity of L2 English, such as mean length of sentence, coordinate phrases per clause, and verb phrases per T-unit, among others. Assuming that the syntactic complexity increases as learners' proficiency rises, these and other indices (cf. Attali & Burstein, 2006) are likely to tap into aspects of proficiency as well. The study employed Guiraud's index because the measure of lexical richness (e.g., type-token ratio, Guiraud's index) has often been used to measure part of L2 proficiency in SLA literature (Michel, Kuiken, & Vedder, 2007; Robinson, 2001; Sercu, De Wachter, Peters, Kuiken, & Vedder, 2006, see also Bulté & Housen, 2012).

Table 2

Number of Scripts and Words in the Subcorpus Used in the Study

L1	Scripts/words	KET	PET	FCE	CAE	CPE	Total
L1 Japanese	# of scripts	129	190	527	251	163	1,260
	# of words	4,065	22,756	184,762	136,940	114,425	462,948
L1 Korean	# of scripts	35	147	356	147	146	831
	# of words	1,459	19,236	136,457	83,810	108,322	349,284
L1 Spanish	# of scripts	609	1,568	830	555	612	4,174
	# of words	22,732	199,874	311,251	324,357	454,919	1,313,133
L1 Russian	# of scripts	104	133	233	149	156	775
	# of words	4,562	16,104	89,376	87,288	116,902	314,232
L1 Turkish	# of scripts	157	186	234	116	88	781
	# of words	6,605	25,924	91,524	66,794	65,358	256,205
L1 German	# of scripts	99	878	323	224	282	1,806
	# of words	3,621	110,574	126,201	137,717	214,086	592,199
L1 French	# of scripts	402	664	593	280	327	2,266
	# of words	14,774	87,747	219,683	167,371	238,754	728,329
Total	# of scripts	1,535	3,766	3,096	1,722	1,774	11,893
	# of words	57,818	482,215	1,159,254	1,004,277	1,312,766	4,016,330

Note also that the development of the index is not linear. The largest jump is observed between KET and PET levels, and the increase is much smaller afterwards.

Corpus size. Table 2 shows the number of scripts and words in each L1 and proficiency level[2]. In total, the subcorpus is made up of 11,893 essays, consisting of over 4 million words.

2.2.4 Scoring Method

The present study employed accuracy as a tool to infer acquisition. As has been discussed earlier, this is a common method in morpheme studies. As a measure of accuracy,

[2]Please note that these numbers are not necessarily representative of the number of scripts and words in each L1 and proficiency level in the entire Cambridge Learner Corpus.

the study mainly employed the target-like use (TLU) score calculated by the following formulae (Pica, 1983b);

$$\text{TLU score} = \frac{\text{number of correct suppliances}}{\text{number of obligatory contexts} + \text{number of overgeneralization errors}}$$

The TLU score, unlike the SOC score, penalizes overgeneralization of a morpheme by adding the number of such cases to the denominator. This is considered to better reflect acquisition. However, since the majority of the literature on morpheme acquisition order adopts the SOC, the present study also used the SOC to facilitate the comparison with the literature. SOC is calculated by the following formulae;

$$\text{SOC score} = \frac{\text{number of correct suppliances} + 0.5 \times \text{number of incorrect suppliances}}{\text{number of obligatory contexts}}$$

where "incorrect suppliance" refers to the suppliance of misformed morphemes, as was mentioned earlier. SOC was only used in rank-order correlation analysis, the technique that has most often been employed in literature comparing acquisition orders.

2.2.5 Data Extraction

In order to obtain SOC and TLU scores, the number of obligatory contexts, the instances of overgeneralization, and those of errors were extracted using Perl scripts. Obligatory contexts were identified by looking into corrected texts. The number of instances of, say, plural *-s* in a corrected text by a native speaker was considered to be the number of obligatory contexts of the morpheme in the corresponding original essay. The number of correct suppliances was obtained by subtracting the number of errors from that of obligatory contexts.

More concretely, in the case of past tense *-ed*, for example, the following algorithm was employed. Errors in the CLC have been annotated in the form of *<NS type="(error classification)">erroneous form|correct form</NS>* (e.g., *She was one of the five daughters of a*

farmer who<NS type="TV">live|lived</NS>in a small village.). An error was counted as a non-overgeneralization error of past tense *-ed* if all of the following conditions were satisfied: (i) The error classification was either IV (incorrect verb inflection) or TV (incorrect verb tense), (ii) neither the erroneous form nor the correct form included *be* verbs, modals, or *have* and its inflected forms, (iii) the erroneous form did not include a word ending in *-ed*, and (iv) the two words immediately preceding the error tag (*farmer* and *who* in the previous example) did not include *be* verbs, *have*, *get*, *make*, *let* and their inflected forms. Condition (ii) was to exclude voice errors, the errors related to modal verbs, and aspectual errors. Condition (iii) excluded overgeneralization errors that are dealt with separately. Condition (iv) precluded passive voice and participial use linked to causative verbs. Similarly, an error was counted as an overgeneralization error of past tense *-ed* if the following conditions were met: (i) The error classification was IV, TV, or FV (wrong verb form), (ii) the correct form did not include a word ending in *-ed*, (iii) the erroneous form did not include *have* and its inflected forms, (iv) the correct form did not include *be* verbs, and (v) the two words preceding the error tag did not include *be* verbs, *have*, *get*, *make*, *let* and their inflected forms. As to (i), FV was added because preliminary manual inspection showed some cases where overgeneralization errors of past tense *ed* were classified as FV. Conditions (ii), (iii), and (iv) excluded omission, aspectual, and voice errors respectively. Condition (v) excluded participial use of *-ed* including passive voice. Finally, obligatory contexts were identified in the following way. The corrected texts are annotated with POS tags based on CLAWS tagset, and have the format of *|(word+morpheme):(word number starting at 1 at the beginning of the sentence)_(POS tag)|* (e.g., *|we:5_PPIS2| |tend+ed:6_VVD| |to:7_TO| |respect:8_VV0| |some:9_DD| |particular:10_JJ| |job+s:11_NN2|*). Obligatory contexts of past tense *-ed* were identified as those words that have the VVD tag (past tense form of lexical verbs) and have the word form of *(word)+ed*, where *(word)* is a regular verb. The Perl scripts used to extract the information are available from the author.

A potential issue with this approach and in error analysis in general is that there can be multiple ways to correct an error. For instance, the sentence *the student need help* can be corrected to either *the students need help* or *the student needs help* (Fitzpatrick & Seegmiller, 2004). The former correction conceives the sentence to lack plural *-s*, while the latter correction views it as lacking third person *-s*. Whereas I cannot reject the possibility that this causes a problem, it is also true that the learner's intention is often clear from the context, and the study assumed the annotator's correction is the solo intention of the learner.

Table 3 reports the accuracy of the scripts used to extract errors. Here, a hundred errors for each morpheme were manually identified according to the proportion of the total number of words in each L1 and proficiency level out of the whole subcorpus. Errors were identified based on the error tags in the CLC. In other words, error annotation was considered as the gold standard against which I tested the accuracy that the script correctly retrieves the intended errors. As the measure of accuracy, the precision and recall were obtained from those sampled parts of the corpus for each morpheme. Precision refers to the degree to which what the script captures accurately includes what it is intended to capture, and recall refers to the degree to which the script captures what it is intended to capture. For instance, if a script to count the frequency of past tense *-ed* errors identified 80 out of 100 instances of the errors and 70 out of the 80 correctly included the target errors, then the precision is 87.5% (70/80) and the recall is 80% (80/100). F_1 is the harmonic mean of precision and recall and represents the total accuracy. It is calculated by

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall}$$

The set of essays the accuracy of the script was tested against was different from the essays that were used to tune the script.

Table 3

Precision and Recall of the Scripts Used in the Study

Morpheme	Precision	Recall	F_1
Articles	98%	88%	0.926
Past tense *-ed*	77%	95%	0.848
Plural *-s*	99%	90%	0.942
Possessive *'s*	97%	87%	0.916
Progressive *-ing*	86%	80%	0.828
Third person *-s*	81%	95%	0.872

In order to confirm that the accuracy of the scripts is not biased for any L1 or proficiency level, I calculated the accuracy of the script for each morpheme in two proficiency levels (high vs. low) across two individual L1 groups (one L1 that lacks the equivalent morpheme and another that has the corresponding morpheme). Similarly to the above, precision and recall were calculated based on manually identified 100 errors. The result is shown in Table 4. Especially for low proficiency cells and for small-sized L1 groups, sampling had to span multiple proficiency levels. In L1 Russian third person *-s*, for instance, sampling began at KET, but since KET included only three instances of third person *-s* errors after removing the parts that were used to tune and test the script before, sampling continued to PET, FCE, and eventually into CAE. Proficiency here, therefore, should only be interpreted in a relative term. Underlined L1 groups are those that lack the corresponding morpheme in their L1s. Overall, the accuracy is fairly high and is roughly comparable across proficiency levels and L1 groups. The accuracy tends to be slightly lower at the higher proficiency level because the non-target use and the error of the target morpheme (e.g., participial *-ed*, gerund *-ing*) increase and there is a higher chance that the script mis-captures them. However, the script accuracy is generally high and it is safe to say that no strong bias is

Table 4

Accuracy of the Scripts by L1 and Proficiency

Proficiency	Morpheme	L1	Precision	Recall	F_1
High					
CPE	Articles	French	0.94	0.94	0.940
FCE-CPE	Past tense *-ed*	Korean	0.61	0.86	0.714
CPE	Plural *-s*	German	0.87	0.85	0.860
CPE	Possessive *'s*	Spanish	0.84	0.91	0.874
FCE-CPE	Progressive *-ing*	Japanese	0.96	0.79	0.867
CAE-CPE	Third person *-s*	Turkish	0.86	0.98	0.916
Low					
KET-PET	Articles	Japanese	0.98	0.92	0.949
KET-PET	Past tense *-ed*	Spanish	0.92	0.97	0.944
KET-FCE	Plural *-s*	Korean	0.94	0.89	0.914
PET-CAE	Possessive *'s*	Turkish	0.87	0.93	0.899
KET-PET	Progressive *-ing*	German	0.96	0.90	0.929
KET-CAE	Third person *-s*	Russian	0.73	0.94	0.822

introduced by the script in the comparison between L1 groups or across proficiency levels.

2.2.6 Data Analysis

Correlation analysis. Three kinds of data analyses were applied. The first analysis was Spearman's rank-order correlations. The analysis was based on SOC scores and answers the following Research Questions (RQs); 1-i, whether the accuracy order is similar within each L1 group, and 1-ii, whether it varies across L1 groups. The purpose of this analysis was to compare the result with previous studies, as this analysis has been a typical

method in the literature for comparing morpheme acquisition orders. Correlation coefficients were calculated between all the L1-proficiency pairs. If two orders were significantly correlated, the two groups from which the orders were obtained were considered to mark a similar accuracy order. In the present study, following Jarvis' (2000) intragroup homogeneity and intergroup heterogeneity, within-L1 orders (i.e., accuracy orders of the same L1, but different proficiency, group) were compared with between-L1 orders (i.e., accuracy orders of different L1 groups). If within-L1 orders are more consistent (i.e., within-L1 pairs show more similarities in the order) than between-L1 orders, then L1 is likely to affect accuracy order.

Clustering based on bootstrapping. The correlation method, although well established in the literature on morpheme studies, has a weakness in that small differences in accuracy count as heavily as large differences. Furthermore, it does not reveal which morpheme is different in the order between groups. Thus comes the clustering analysis. In the clustering analysis, the specific accuracy order per L1 per proficiency level was obtained and compared between L1 groups and proficiency levels. For this purpose, morphemes with similar TLU scores were clustered together within each L1-proficiency group. This is similar to Krashen (1977), although the method in this study is different. This clustering method was used in order to avoid, as mentioned in the last chapter, a 1% difference having as much of an impact as a 50% difference. To cluster morphemes, the accuracy of each morpheme within each L1 and proficiency level was compared with each other, and statistical significance among them was tested. However, since each morpheme in a given proficiency level in an L1 group only has one TLU score, it was not possible to conduct a formal statistical test to compare the TLU scores of two morphemes. Instead, I employed a statistical technique known as bootstrapping in order to obtain a sampling distribution of the difference in TLU scores between morphemes. Bootstrapping is a resampling technique and is used to gain insights into the population of a given sample. Intuitively, it increases

the number of data points by repeatedly drawing random learner essays and calculating a TLU score from each sample. The following describes how this technique works in more detail and what was done to obtain the confidence interval in the case of third person *-s* (or any other morpheme) in the L1 Japanese CPE data containing 163 essays.

(1) 163 essays were randomly selected from the Japanese CPE data. Single essays could be selected multiple times.

(2) A TLU score was calculated from the sample obtained in (1).

(3) (1) and (2) were repeated 10,000 times, resulting in 10,000 TLU scores.

After bootstrapping, I identified all the morpheme pairs whose TLU scores differed statistically by looking at every two-morpheme pair in the same L1 and analyzing the distribution of the difference in their TLU scores. More specifically, the 10,000 TLU scores of one morpheme were subtracted from the 10,000 TLU scores of the other morpheme, which resulted in 10,000 differences in the TLU scores between the two morphemes. I then tested whether the 95% range of the differences included zero. If it did, the accuracy difference between the two morphemes was considered non-significant and they were clustered together.

An intuition behind this process is the following. The difference in the TLU scores at hand between two morphemes might be merely due to sampling variability. Therefore, I obtained 10,000 differences from assumedly independent samples and tested if the difference is likely to be real (i.e., non-zero). If far majority of the differences have the same sign (positive or negative), then the difference in the population is unlikely to be zero.

After the clustering was completed, it was possible to compare the accuracy orders within (for intragroup homogeneity) and between (intergroup heterogeneity) L1 groups. If the orders were similar within-L1 between-proficiency groups but different between L1 groups, L1 was considered to influence accuracy order. At the same time, I identified the specific morphemes which differed in the order between L1 groups.

Regression model. Once the specific orders and their specific differences between L1 groups are recognized, we can test whether the difference is motivated by L1 (RQ1-iii), the strength of L1 influence (RQ2), and whether all the morphemes are equally sensitive to L1 influence (RQ3). These targets were achieved by the third analysis, regression modeling. The TLU score was regressed against L1 influence, L1, exam level, morpheme, and their significant two-way interactions. If the effect of L1 influence is significant in determining accuracy when the effects of other variables are partialled out, this indicates that L1 features are related to L2 performance, capturing crosslinguistic performance congruity. This analysis also made it possible to compare the strength of L1 influence in determining accuracy with the strength of other factors such as proficiency development. Furthermore, the inspection of the interaction between morpheme and L1 influence discloses whether certain morphemes are more immune to L1 influence than others.

The question was how to operationalize L1 influence in the regression model. We adopt a simple approach to explore this influence, a dichotomous variable coding L1 effect as 0 if the morpheme is absent or optional in the L1, and 1 if it is obligatory. For example, the article in L1 Japanese was coded as 0 because, although a demonstrative determiner in Japanese, *sono*, covers part of the semantics expressed by English articles, demonstratives in general are more limited in scope than articles (J. A. Hawkins, 1991). More specifically, demonstrative determiners typically cover only the anaphoric and part of the immediate situation use of the definite article but not other uses such as the larger situation use (as in [When invited to a wedding] Have you seen *the bridesmaids*?; Hawkins, 1978). It is, therefore, not obligatory to express identifiability in Japanese. Rather, the definiteness status is often determined by the hearer in article-less languages (Ekiert, 2010). It is possible and common to translate an English sentence with articles into an equivalent Japanese sentence without any explicit indication of definiteness. Therefore, the marking of definiteness, the core concept of English articles, was considered non-obligatory in Japanese and thus a 0

was given to articles. On the other hand, a 1 was assigned to past tense *-ed* in L1 Japanese because a Japanese morpheme, *-ta*, roughly corresponds to past tense *-ed* in English, and it is difficult to express past-ness without the use of the morpheme in Japanese. Hence, the past tense marker (*-ta*) was considered obligatory in Japanese and a 1 was given. Although past tense *-ed* encodes both tense and aspect, the decision was made on the basis of tense.

Table 5 shows all the values of the variable used in the study and the references that support the decision. Admittedly, this is a rather crude and oversimplified modeling of L1 influence, but as we shall see, it will prove a useful way of capturing the effect of L1.

The three analyses should disclose the similarities and differences in the accuracy order of English grammatical morphemes between the L1 groups and between proficiency levels. It should consequently show whether the concept of natural order is tenable.

2.3 Results

2.3.1 Descriptive Data

Accuracy scores of each morpheme, L1, and proficiency level. Table 6 shows the SOC and the TLU scores of the target grammatical morphemes across the seven L1 groups and five proficiency levels. For each morpheme, the table also includes the number of obligatory contexts to indicate the data size of each morpheme in each L1 and proficiency level. Two observations are worth noting here. First, the number of obligatory contexts was small in some cases (e.g., 5 in the case of the possessive *'s* in L1 Korean KET or 9 in the case of the past tense *-ed* in L1 German KET). This may mean some data points are unreliable. These data points were included in the correlation analysis because previous morpheme studies did not always have a large number of obligatory contexts and the purpose of the analysis was to conceptually replicate them. The small data size was explicitly marked in the clustering approach and those data points were excluded in the regression analysis. Second, both the SOC and the TLU scores were high overall, some achieving above 0.90

Table 5

Binary Variable Indicating Whether the Morpheme is Obligatorily Marked in Target L1s and the References Supporting the Decision

L1	Articles	Past tense *-ed*	Plural *-s*	Possessive *'s*	Progressive *-ing*	Third person *-s*
L1 Japanese	0	1	0	1	1	0
	LS	Sh-a	LS	LS	Sh-a	J
L1 Korean	0	1	0	1	1	0
	LS	Le	LS	LS	Sh-b	C
L1 Spanish	1	1	1	0	1	1
	LS	Sa	LS	LS	Sl	Hr
L1 Russian	0	1	1	1	0	1
	I6	GC	M	M	I8	AL
L1 Turkish	0	1	1	1	1	0
	Sl	SA	Gö	GK	Sl	E
L1 German	1	1	1	1	0	1
	Sl	Gr	K	Li	Sl	Gr
L1 French	1	1	1	0	0	1
	BH	BH	BH	Hw	Sl	BH

Note. One denotes that the morpheme is obligatorily marked in the language. Otherwise 0 is given. AL = Ambridge and Lieven (2011); BH = Battye and Hintze (1992); C = Choi (2005); E = Ekmekci (1982); GC = Gor and Chernigovskaya (2004); GK = Göksel and Kerslake (2005); Gö = Görgülü (2005); Gr = Graves (1990); Hr = Harvey (2006); Hw = R. Hawkins (1981); I6 = Ionin (2006); I8 = Ionin (2008); J = Jelinek (1984); K = Köpcke (1988); Le = Lee (2006); Li = Lindauer (1998); LS = Luk and Shirai (2009); M = Müller (2004); Sa = Salaberry (2002); SA = Slobin and Aksu (1982); Sh-a = Shirai (1998a); Sh-b = Shirai (1998b); Sl = Slobin (1996)

from the KET level (e.g., past tense *-ed* in L1 Japanese or plural *-s* in L1 Spanish). This implies that there might be ceiling effects that need to be taken into account when analyzing and interpreting the data. For instance, we cannot claim that a morpheme with a 0.98 TLU score is acquired earlier than that with a 0.95 TLU, as both morphemes are presumably close to the end of the acquisitional process and accuracy as a measure of the degree of acquisition might be less precise when learners reach this advanced level. This issue was dealt with in the clustering analysis by setting the TLU score of 0.90 as the criterion of acquisition.

Comparing TLU scores across L1 groups. Figure 2 contrasts the TLU scores of each morpheme across L1 groups. Only the TLU scores between 0.60 and 1.00 are displayed, as the inclusion of lower scores would push up the entire graph and cause difficulty in inspection. Although the TLU score tends to increase as the proficiency goes up in all the L1 groups, striking differences between them are also observed when attention is paid to individual morphemes. For instance, when L1 Japanese and L1 Spanish learners are compared, articles are consistently the least accurate morpheme among L1 Japanese learners of English, whereas they tend to reside in the upper half for L1 Spanish learners. Past tense *-ed*, on the other hand, is either the most or the second most accurate morpheme across proficiency in L1 Japanese, whereas it tends to be at lower half in L1 Spanish learners of English. To take another example, progressive *-ing* is constantly the second least accurate morpheme among L1 German learners of English, whereas it is consistently in the upper half with regards to many other L1 groups (e.g., L1 Russian, L1 Korean). Therefore, even from this simple graph, it can be inferred that accuracy order is likely to be different between L1 groups.

Table 6

SOC and TLU Scores of Each Morpheme, L1 Group, and Proficiency Level

Morpheme and	L1 Japanese					L1 Korean					L1 Spanish				
accuracy measure	KET	PET	FCE	CAE	CPE	KET	PET	FCE	CAE	CPE	KET	PET	FCE	CAE	CPE
Articles															
OC	218	1,258	10,696	9,283	8,035	72	1,040	8,146	5,987	7,883	1,189	12,691	18,951	23,384	31,654
SOC	0.68	0.73	0.82	0.84	0.89	0.73	0.70	0.81	0.82	0.88	0.94	0.95	0.95	0.96	0.97
TLU	0.63	0.67	0.76	0.79	0.84	0.66	0.66	0.77	0.77	0.82	0.86	0.91	0.90	0.92	0.94
Past tense *-ed*															
OC	12	297	1,831	926	1,240	5	271	1,380	638	1,222	48	2,910	3,453	2,040	5,420
SOC	0.92	0.98	0.96	0.96	0.97	1.00	0.95	0.97	0.95	0.97	0.90	0.89	0.92	0.93	0.97
TLU	0.92	0.95	0.92	0.91	0.93	1.00	0.92	0.94	0.91	0.93	0.70	0.83	0.87	0.88	0.94
Plural *-s*															
OC	114	583	7,077	7,425	5,440	25	493	5,045	4,723	5,215	602	6,328	12,393	16,806	22,048
SOC	0.85	0.85	0.91	0.93	0.94	0.76	0.86	0.91	0.92	0.93	0.95	0.95	0.96	0.97	0.98
TLU	0.84	0.82	0.88	0.90	0.91	0.73	0.85	0.89	0.90	0.90	0.91	0.92	0.93	0.96	0.96
Possessive *'s*															
OC	6	55	390	438	457	5	47	353	362	438	79	340	573	953	1,696
SOC	0.67	0.82	0.87	0.88	0.93	1.00	0.82	0.92	0.94	0.92	0.89	0.89	0.84	0.79	0.88
TLU	0.67	0.69	0.80	0.82	0.89	0.83	0.78	0.87	0.91	0.88	0.82	0.74	0.70	0.68	0.82
Progressive *-ing*															
OC	33	317	1,631	917	677	14	216	1,172	583	619	248	2,269	3,168	2,458	2,759
SOC	0.86	0.97	0.97	0.99	0.99	0.86	0.95	0.98	0.98	0.99	0.83	0.96	0.98	0.99	0.99
TLU	0.85	0.94	0.93	0.95	0.97	0.86	0.92	0.94	0.94	0.96	0.83	0.94	0.96	0.96	0.98
Third person *-s*															
OC	22	168	1,043	1,138	1,281	8	127	806	688	1,179	146	1,289	1,844	3,002	5,165
SOC	0.77	0.86	0.89	0.91	0.93	0.75	0.88	0.91	0.90	0.92	0.79	0.78	0.88	0.92	0.95
TLU	0.74	0.84	0.84	0.87	0.89	0.67	0.86	0.88	0.85	0.89	0.77	0.77	0.84	0.89	0.94

Note. OC = number of obligatory contexts

SOC and TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (continued)

Morpheme and accuracy measure	L1 Russian					L1 Turkish					L1 German				
	KET	PET	FCE	CAE	CPE	KET	PET	FCE	CAE	CPE	KET	PET	FCE	CAE	CPE
Articles															
OC	307	999	5,495	6,333	8,474	335	1,354	5,465	4,730	4,640	239	7,640	8,259	10,391	15,068
SOC	0.81	0.77	0.81	0.86	0.90	0.78	0.78	0.81	0.86	0.92	0.97	0.97	0.97	0.98	0.98
TLU	0.76	0.71	0.75	0.81	0.85	0.72	0.72	0.75	0.80	0.87	0.89	0.93	0.94	0.94	0.95
Past tense *-ed*															
OC	14	248	678	497	1,137	25	421	1,040	449	788	9	1,766	1,291	843	2,185
SOC	0.89	0.94	0.95	0.96	0.98	0.96	0.96	0.97	0.96	0.96	1.00	0.95	0.97	0.97	0.99
TLU	0.60	0.90	0.90	0.93	0.95	0.65	0.90	0.93	0.91	0.92	0.82	0.91	0.94	0.95	0.96
Plural *-s*															
OC	156	493	3,913	5,166	6,051	211	720	3,492	3,523	3,051	87	3,197	5,002	7,230	10,677
SOC	0.97	0.92	0.96	0.98	0.98	0.86	0.88	0.91	0.95	0.96	0.95	0.97	0.97	0.98	0.98
TLU	0.96	0.90	0.94	0.97	0.97	0.82	0.84	0.88	0.92	0.94	0.93	0.94	0.95	0.97	0.97
Possessive *'s*															
OC	5	22	169	285	466	14	44	204	240	250	8	75	190	317	767
SOC	1.00	0.73	0.86	0.87	0.92	0.86	0.75	0.79	0.85	0.81	0.94	0.72	0.86	0.93	0.91
TLU	0.83	0.48	0.74	0.79	0.89	0.86	0.60	0.64	0.77	0.72	0.70	0.46	0.73	0.84	0.84
Progressive *-ing*															
OC	40	182	735	538	610	38	283	813	564	468	32	848	934	665	902
SOC	0.88	0.96	0.98	0.97	0.99	0.87	0.97	0.98	1.00	0.99	0.78	0.91	0.96	0.97	0.98
TLU	0.88	0.88	0.94	0.93	0.95	0.80	0.91	0.94	0.94	0.94	0.76	0.85	0.90	0.89	0.88
Third person *-s*															
OC	15	72	573	784	1,376	38	164	526	577	630	20	763	717	1,216	2,359
SOC	0.73	0.88	0.91	0.95	0.96	0.87	0.93	0.90	0.92	0.95	0.90	0.89	0.95	0.99	0.98
TLU	0.73	0.83	0.88	0.92	0.94	0.87	0.89	0.85	0.85	0.91	0.90	0.88	0.91	0.96	0.97

SOC and TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (continued)

Morpheme and accuracy measure	L1 French				
	KET	PET	FCE	CAE	CPE
Articles					
OC	837	5,490	14,290	12,612	17,176
SOC	0.92	0.95	0.95	0.97	0.97
TLU	0.86	0.90	0.89	0.93	0.93
Past tense *-ed*					
OC	24	1,181	1,940	1,044	2,538
SOC	0.96	0.92	0.96	0.95	0.98
TLU	0.82	0.88	0.91	0.91	0.94
Plural *-s*					
OC	441	2,546	8,443	8,130	11,000
SOC	0.94	0.91	0.94	0.97	0.96
TLU	0.91	0.87	0.89	0.94	0.94
Possessive *'s*					
OC	26	103	386	395	913
SOC	0.96	0.85	0.84	0.86	0.90
TLU	0.81	0.63	0.75	0.82	0.85
Progressive *-ing*					
OC	197	741	1,683	887	1,304
SOC	0.42	0.91	0.96	0.96	0.98
TLU	0.42	0.87	0.92	0.90	0.89
Third person *-s*					
OC	79	563	1,327	1,366	2,754
SOC	0.49	0.76	0.87	0.94	0.94
TLU	0.48	0.75	0.84	0.92	0.92

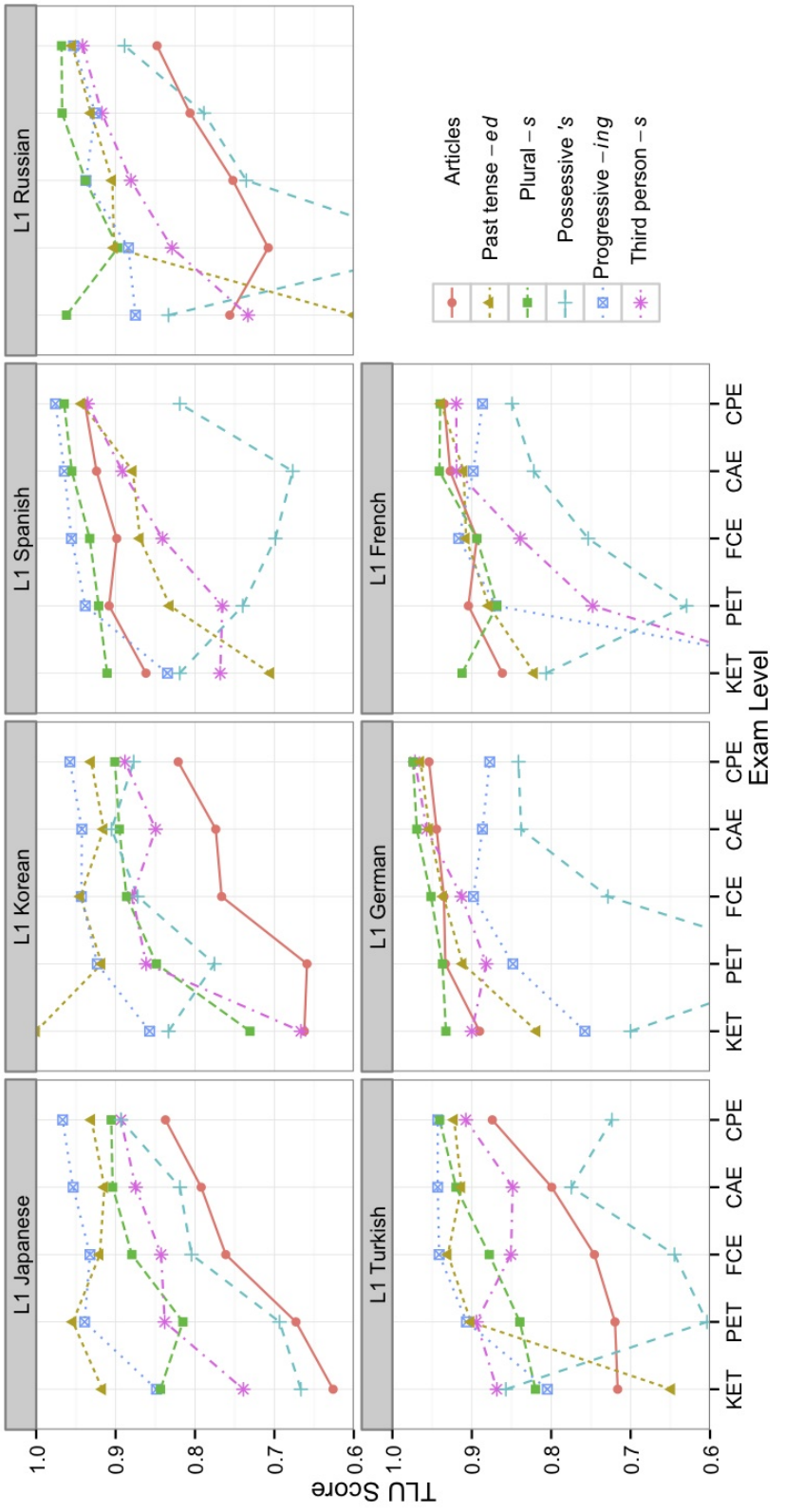


Figure 2. TLU Scores of Each Morpheme in Each L1 Group

2.3.2 SOC-Based Correlations

As previously mentioned, this analysis replicated a commonly adopted technique in comparing morpheme acquisition orders and aimed to answer the RQ1-i, which queries the consistency of accuracy order within each L1 group, and RQ1-ii regarding the differences between L1 groups. For this purpose, one accuracy order per L1 group per level was obtained from the SOC scores in Table 6. The total number of L1-proficiency pairs, or the total number of SOC orders, was 35 (7 L1s $\times$ 5 proficiency levels). The total number of possible comparisons between two orders among them was 595 ($_{35}C_2 = \frac{35!}{2! \times (35-2)!} = 595$). Out of the 595 order pairs, 70 were within-L1 (10 within-L1 between-proficiency pairs $\times$ 7 L1s), and the rest (525) were between-L1. The Spearman's rank correlation coefficient was calculated for each pair and its statistical significance was tested. The result showed that 269 out of the 525 between-L1 pairs (51.2%) were significantly correlated at $p < 0.05$, whereas 57 out of 70 within-L1 correlations (81.4%) were significant. The difference between the two (51.2% vs. 81.4%) was statistically significant ($\chi^2(1) = 22.727$, $p < .001$, $\phi = 0.195$) and indicates that within-L1 pairs show more similar accuracy orders than between-L1 pairs. This means that accuracy orders are more similar within L1 groups than between them, satisfying the first two requirements for the identification of L1 transfer by Jarvis (2000).

Moreover, although there has been a criticism that the Spearman's rank coefficient is difficult to turn out to be non-significant in morpheme studies (cf. M. H. Long & Sato, 1984), the present result does not support this idea. In fact, as much as 45% (269/595 = 45.2%) of the coefficients of the concerned pairs were non-significant. Within-L1 pairs, however, tended to have more statistically significant coefficients than between-L1 pairs. This shows the relative consistency of accuracy orders within L1 groups compared to those between these groups.

2.3.3 TLU-Based Clustering

This analysis also addressed the same research questions regarding within-L1 consistency and between-L1 differences in the accuracy order, but in a more specific way; which morphemes are different or the same in the order between L1 groups? I obtained 95% confidence intervals through bootstrapping, and identified the morpheme pairs whose TLU scores were significantly different from each other. Table 7 shows clustered TLU scores based on the confidence intervals. The morphemes in Cluster 1 marked higher TLU scores than those in Cluster 2, which in turn scored higher than those in Cluster 3 and so forth. For example, at the L1 French PET level, articles marked the highest TLU score, followed by three morphemes (past tense *-ed*, plural *-s*, progressive *-ing*) which achieved fairly similar accuracy levels. These were followed by third person *-s*, with possessive *'s* as the least accurate morpheme.

In order to alleviate the ceiling effect, the morphemes with TLU scores over 0.90 at $p < 0.05$ (marked with asterisks (*)) and those whose scores were not significantly different from them were clustered in the first clusters. The criterion of 0.90 was adopted because it has been a common benchmark of acquisition in morpheme studies, albeit in SOC (Hakuta, 1976). The morphemes which are crossed out were considered unreliable due to a small sample size (number of obligatory contexts < 100). Those underlined were the ones which could not be clustered by statistics alone, such as the case where the morpheme did not show a significant difference with any other morpheme. In this case, the morpheme was classified into the cluster including the morpheme whose accuracy was closest to the accuracy of the concerned morpheme. Note that this table denotes accuracy order but not accuracy itself, thus the morphemes in the same cluster order might have widely different accuracy rates across L1 groups. Table 8, for reference purposes, is the natural order of acquisition (Krashen, 1977) limited to the target morphemes of the present study under the

Table 7

Clustered Order of TLU Scores

CPE

Clustered order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	past tense *-ed* * progressive *-ing* *	past tense *-ed* * progressive *-ing* *	articles * past tense *-ed* * plural *-s* * progressive *-ing* * third person *-s* *	past tense *-ed* * plural *-s* * progressive *-ing* * third person *-s* *	past tense *-ed* * plural *-s* progressive *-ing* *	articles * past tense *-ed* * plural *-s* * third person *-s* *	articles * past tense *-ed* * plural *-s* * third person *-s* *
2	plural *-s* possessive *'s* third person *-s*	plural *-s* possessive *'s* third person *-s*	possessive *'s*	possessive *'s*	articles third person *-s*	possessive *'s* progressive *-ing*	progressive *-ing*
3	articles	articles		articles	possessive *'s*		possessive *'s*
4							
5							

CAE

Clustered order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	progressive *-ing* *	past tense *-ed* progressive *-ing* *	articles * plural *-s* * progressive *-ing* *	past tense *-ed* * plural *-s* * progressive *-ing* * third person *-s*	past tense *-ed* plural *-s* * progressive *-ing* *	articles * past tense *-ed* * plural *-s* * third person *-s* *	articles * past tense *-ed* plural *-s* * third person *-s* *
2	past tense *-ed* plural *-s*	plural *-s* possessive *'s*	past tense *-ed* third person *-s*	articles possessive *'s*	third person *-s*	possessive *'s* progressive *-ing*	progressive *-ing*
3	third person *-s*	third person *-s*	possessive *'s*		articles possessive *'s*		possessive *'s*
4	articles possessive *'s*	articles					
5							

FCE

Clustered order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	past tense *-ed* * progressive *-ing* *	past tense *-ed* * progressive *-ing* *	plural *-s* * progressive *-ing* *	plural *-s* * progressive *-ing* *	past tense *-ed* * progressive *-ing* *	articles * past tense *-ed* * plural *-s* *	past tense *-ed* progressive *-ing* *
2	plural *-s*	plural *-s* possessive *'s* third person *-s*	articles	past tense *-ed* third person *-s*	plural *-s* third person *-s*	progressive *-ing* third person *-s*	articles plural *-s*
3	possessive *'s* third person *-s*	articles	past tense *-ed*	articles possessive *'s*	articles	possessive *'s*	third person *-s*
4	articles		third person *-s*		possessive *'s*		possessive *'s*
5			possessive *'s*				

Clustered Order of TLU Scores (continued)

PET

Clustered order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	past tense *-ed* * progressive *-ing* *	past tense *-ed* progressive *-ing*	articles * plural *-s* * progressive *-ing* *	past tense *-ed* plural *-s* progressive *-ing* ~~third person *-s*~~	past tense *-ed* progressive *-ing* third person *-s*	articles * plural *-s* *	articles
2	plural *-s* third person *-s*	plural *-s* ~~possessive *'s*~~ third person *-s*	past tense *-ed*	articles ~~possessive *'s*~~	plural *-s*	past tense *-ed*	past tense *-ed* plural *-s* progressive *-ing*
3	articles ~~possessive *'s*~~	articles	possessive *'s* third person *-s*		~~possessive *'s*~~ articles	progressive *-ing* third person *-s*	third person *-s*
4						~~possessive *'s*~~	possessive *'s*
5							

KET

Clustered order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	~~past tense *-ed*~~ plural *-s* ~~progressive *-ing*~~ ~~third person *-s*~~	~~past tense *-ed* *~~	plural *-s*	plural *-s* * ~~possessive *'s*~~ ~~progressive *-ing*~~	plural *-s* ~~possessive *'s*~~ ~~progressive *-ing*~~ ~~third person *-s*~~	articles ~~plural *-s*~~ ~~third person *-s*~~	plural *-s*
2	articles ~~possessive *'s*~~	~~articles~~ ~~plural *-s*~~ ~~possessive *'s*~~ ~~progressive *-ing*~~ ~~third person *-s*~~	articles ~~possessive *'s*~~ progressive *-ing*	articles ~~past tense *-ed*~~ ~~third person *-s*~~	articles ~~past tense *-ed*~~	~~past tense *-ed*~~ ~~possessive *'s*~~ ~~progressive *-ing*~~	articles ~~past tense *-ed*~~ ~~possessive *'s*~~
3			~~past tense *-ed*~~ third person *-s*				progressive *-ing* ~~third person *-s*~~
4							
5							

Table 8

Clustered Natural Order of Acquisition of English Grammatical Morphemes

Clustered order	Morpheme
1	plural *-s*
	progressive *-ing*
2	articles
3	past tense *-ed*
	possessive *'s*
	third person *-s*

same format as Table 7.

Summary of the between-L1 differences. Cross-L1 differences in the accuracy order are summarized in Table 9. The L1 groups on the left marked a higher accuracy rank of the morpheme than those on the right. These differences were determined by investigating the clustered data in the following way. If it was certain that a morpheme marked a higher accuracy rank in an L1 group than in another, then the L1 group was placed on the left. For instance, in the case of the FCE level, there was considered to be no difference in the order of plural *-s* between L1 German and L1 Turkish groups because in L1 German its accuracy was the first, the second, or the third from the top, whereas in L1 Turkish it was either the third or the fourth. In other words, there is a possibility that in both L1 groups plural *-s* was the third most accurate morpheme, in which case there is no difference in the order of accuracy. In L1 Spanish or Russian, however, plural *-s* was either the most or the second most accurate morpheme, thus there could not be an overlap with its order in the L1 Turkish group. Therefore, the accuracy order of plural *-s* was considered higher in L1

Table 9

Between-L1 Differences in Clustered TLU-Score Orders

Level	Articles	Past tense *-ed*	Plural *-s*	Possessive *'s*	Progressive *-ing*	Third person *-s*
CPE	SGF > JKRT			JKR > STF	JKRT > GF	
	T > JKR					
CAE	SGF > JKRT	JKT > S		K > JSRTGF	JKSRT > GF	JRTGF > K
FCE	SGF > JKRT	JKTF > SR	SR > JKTF	JK > STGF	JKSRTF > G	RT > SF
		G > S				
PET	SGF > JKRT	JK > SG	G > JKT		JKST > G	T > SGF
		TG > S	S > T			J > SF

Note. J = L1 Japanese; K = L1 Korean; S = L1 Spanish; R = L1 Russian; T = L1 Turkish; G = L1 German; F = L1 French

L1 groups on the left mark higher accuracy ranks on the concerned morpheme than those on the right.

Spanish and Russian than in L1 Turkish. The KET level was removed from the analysis because the data were mostly too small and thus unreliable.

It is obvious from the table that the accuracy order varies across L1 groups with respect to all the concerned morphemes. We can make the following observations with respect to each morpheme in Table 7 and Table 9.

- Articles consistently rank low in Japanese, Korean, Russian, and Turkish learners of English, which partially confirms the results of Luk and Shirai's (2009) survey and runs counter to the prediction of the natural order. In other L1 groups (Spanish, German, and French), they tend to be in the upper half.

- Past tense *-ed* consistently belongs to the highest cluster in Korean and Turkish learners, and Japanese, Russian, German, and French learners show a similar trend. This again does not support the natural order, where past tense *-ed* is one of the latest morphemes to be acquired.

- Plural *-s* tends to be in higher clusters in Spanish, Russian, and German learners of English. In other L1 groups, it tends to be located somewhere in the middle.

- Possessive *'s* tends to mark a lower accuracy rank in Spanish, Turkish, German, and French learners of English than Japanese and Korean learners. Overall though, it typically ranks low and is in this sense consistent with the natural order.

- Progressive *-ing* is in the highest cluster in all L1 groups except for German and French learners, in whose groups it is one of the two morphemes which do not reach 90% of the TLU score even at the highest level, CPE.

- Third person *-s* fluctuates even within L1s over proficiency but tends to stay in the lower half where Spanish learners of English are concerned.

Summary of the within-L1 differences. Table 10 shows within-L1 differences in the accuracy order obtained and formatted in the same manner as Table 9. Few within-L1 differences could be observed in the accuracy order, indicating a consistent order across proficiency levels. There was no within-L1 between-proficiency difference in past tense *-ed*, plural *-s*, and possessive *'s*, and only one in articles. From these two tables, it is clear that the order of accuracy tends to differ between L1 groups but not within them. Therefore, both intragroup homogeneity and intergroup heterogeneity can be observed in the present data, providing empirical support for two of the three pieces of evidence necessary to claim transfer effect. The last requirement, cross-linguistic performance congruity, will be addressed in the regression analysis.

Comparison with the natural order. In order to directly analyze whether the idea of the natural order is tenable, the orders obtained from the present data were compared with the natural order, shown in Table 11. As in Table 9, when L1 groups are on the left of the natural order (NO), this indicates that the accuracy rank of the morpheme was higher in the L1 groups than the expectation of the natural order. It is clear that the orders observed in the present study often deviate from the natural order. It is also interesting to note the absence

Table 10

Within-L1 Differences of Clustered TLU-Score Orders

L1	Articles	Past tense *-ed*	Plural *-s*	Possessive *'s*	Progressive *-ing*	Third person *-s*
L1 Japanese						
L1 Korean						
L1 Spanish						
L1 Russian						
L1 Turkish						P > Ca
L1 German						
L1 French	P > F				FP > CpCa	CpCa > FP

Note. Cp = CPE; Ca = CAE; F = FCE; P = PET; Test takers mark higher accuracy order on the exam on the left than that on the right.

Table 11

Differences Between the Observed Order and the Natural Order Based on Clustered Data

Level	Articles	Past tense *-ed*	Plural *-s*	Possessive *'s*	Progressive *-ing*	Third person *-s*
CPE	NO > JKRT	JKT > NO	NO > JK		NO > GF	
CAE	NO > JKRT	JKT > NO	NO > K		NO > GF	
FCE	NO > JKRT	JKTGF > NO	NO > JKTF		NO > G	
PET	NO > JKRT GF > NO	JKTG > NO	NO > JKT		NO > G	JT > NO

Note. J = L1 Japanese; K = L1 Korean; R = L1 Russian; T = L1 Turkish; G = L1 German; F = L1 French; NO = natural order

L1 groups on the left show that they mark higher accuracy order on the morpheme concerned than predicted by the NO. L1 groups on the right show the opposite.

of difference between the natural order and the accuracy order of L1 Spanish learners. This will be further addressed in the Discussion section.

2.3.4 Regression Analyses

2.3.4.1 Graphical analysis

Now that it has been established that accuracy order clearly varies across L1 groups but not within them, the only remaining evidence necessary to claim L1 influence is cross-linguistic performance congruity (RQ1-iii). Given L1 influence on accuracy order, some

new questions arise; (1) how strong L1 influence is (RQ2), and (2) whether its strength is equal for each morpheme (RQ3). To address these questions we ran a regression analysis.

Let us begin with the question of congruency. One way to capture congruency is by looking at whether equivalent morphemes exist in L1. Figure 3 contrasts the TLU scores between the morphemes with and without equivalent forms in learners' L1s. All the unreliable data points (number of obligatory contexts < 100) were removed (34/210 = 16.2%) from the subsequent analysis including the regression model. In addition, the L1 French KET progressive *-ing* TLU score was excluded from the graphical representation (but included in mean calculation and other statistical analyses) as it was much lower than other reliable data (0.415; see Table 6). Its inclusion would have pushed up the entire graph, causing difficulty during closer examination.

In the figure, data points were plotted according to three factors; whether equivalent forms are present in L1 or not (L1 type), exam level, and morpheme. The left half represents, by exam level, the scores of the morphemes which have no equivalent form in the learners' L1 (hereafter ABSENT). The right half, on the other hand, represents the TLU scores of the morphemes which have similar forms in the learners' L1 (PRESENT). Data points were slightly shifted by morpheme within each level. Solid lines and plus signs (+) show the mean TLU scores aggregated over morphemes, and the dashed line indicates the TLU score of 0.90, a typical criterion of acquisition. Inside of the parentheses below the exam level are standard deviations of the TLU scores of each group, and the square brackets show the mean TLU scores of the ABSENT and the PRESENT groups.

A few observations are in order. First, the PRESENT L1s' morphemes clearly outperform the ABSENT morphemes. The difference in the average TLU scores is 0.084[3], a large value considering that the overall performance is approaching the ceiling in the present data. More strikingly, the mean values indicate that the TLU scores of the AB-

[3]The calculation may not seem to be correct due to rounding, but the answer is accurate. Same follows.

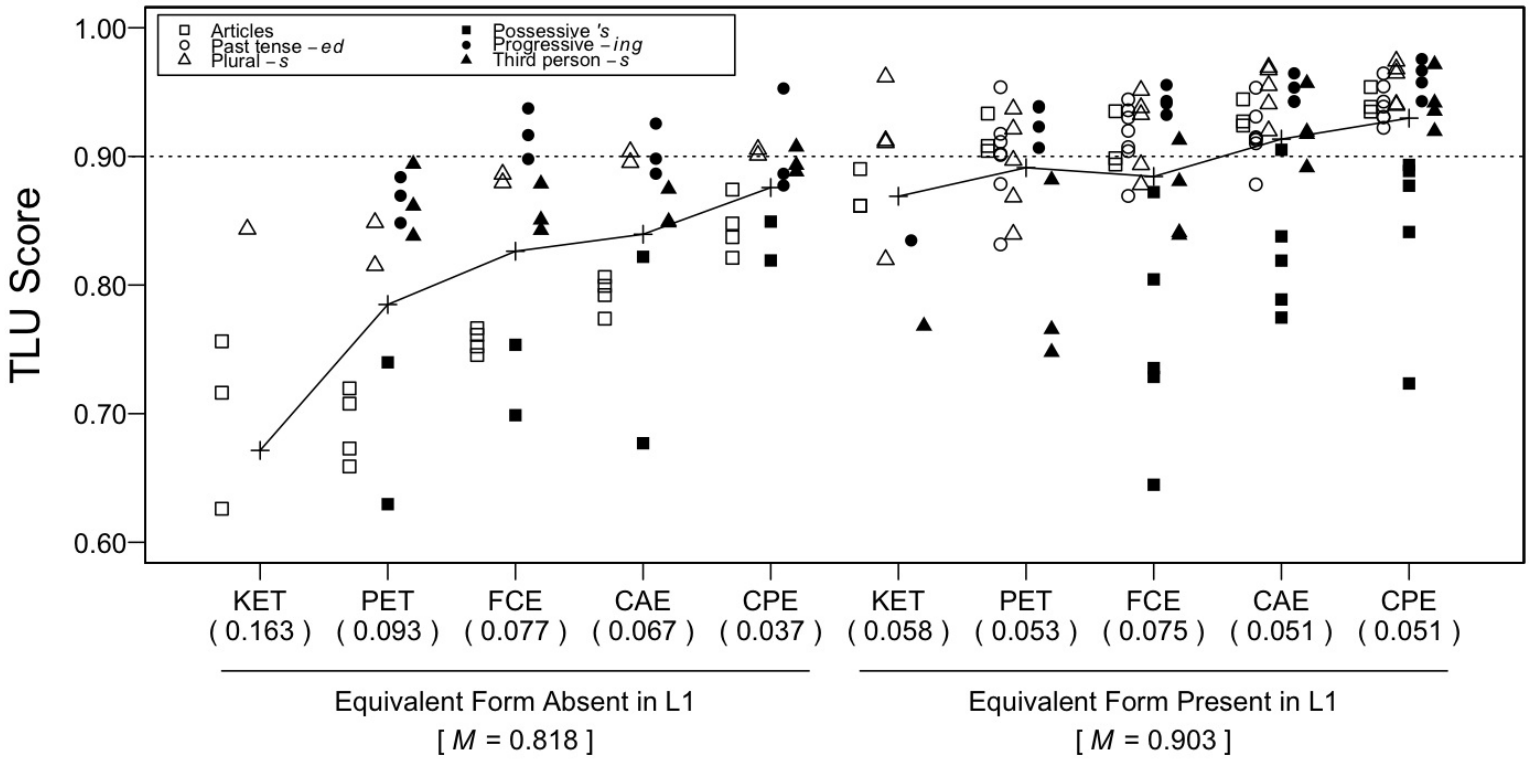


Figure 3. Distribution of TLU scores × L1 type

SENT morphemes at the CPE level are only as high as those of the PRESENT morphemes at the KET level, the difference merely being 0.007. Few ABSENT morphemes scored above 0.90 in TLU. In fact, as will be discussed later, progressive *-ing* in L1 Russian was the only morpheme over 0.90 at $p < 0.05$. Unsurprisingly, obligatory marking of the morpheme in L1 is positively correlated with the TLU score [r_{pb} = 0.474, $p < 0.001$]. The difference between the ABSENT and the PRESENT groups, however, diminishes as proficiency develops. It is understandable that the effect of L1 is stronger at the early stages of acquisition than at later stages.

Second, morphemes seem to vary with respect to the extent they enjoy benefit from the L1. The contrast between the ABSENT and the PRESENT groups for each morpheme are briefly described below. All the target L1s have past tense *-ed*, so it could not be contrasted.

- Articles are typically located below the average in the ABSENT group, whereas they are above the average in the PRESENT group, demonstrating a striking effect of L1.

- Plural *-s*, although the PRESENT group consistently scores higher, does not seem to exhibit much L1 influence until the CAE level, as the TLU score of both groups are relatively high from the beginning. However, at CAE and CPE, the difference between the ABSENT and the PRESENT groups is larger, proving L1 influence. This is contrary to the general assumption of transfer that the effect diminishes at higher proficiency levels.

- Possessive *'s* marks similar TLU scores between the PRESENT and the ABSENT groups. It scores much lower than the average in both groups, and rarely reaches 0.90 in TLU score. L1 influence is weak on this morpheme.

- Progressive *-ing* shows a clear influence from L1. Although its TLU scores are high

in the ABSENT group as well, they are even higher in the PRESENT group at all levels.

- Third person *-s* behaves somewhat differently from the other morphemes. Its accuracy is higher in the ABSENT than in the PRESENT group at the PET level and is roughly equal at FCE. The PRESENT group surpasses the ABSENT group at the CAE and the CPE levels. Therefore, on this morpheme, the effect of L1 switches the direction as the learners' proficiency rises.

These suggest that the degree of L1 influence differs across morphemes.

Third, in the ABSENT group, although the overall mean steadily increases from KET to CPE, not all morphemes act in the same way. Whereas possessive *'s* and articles demonstrate a relatively stable accuracy rise, progressive *-ing* and third person *-s*, despite their high accuracy rates from the beginning, show little improvement after FCE. As a result, a larger number of data points clustered around 0.90 of the TLU score as learners develop. Indeed, at the CPE level, it becomes the most crowded area. This is reflected in the transition of standard deviations, which consistently decrease from KET to CPE in the ABSENT group. It can be hypothesized that there is a "glass ceiling" at 0.90 of the TLU score where, without the equivalent form in L1, the learning curve reaches the asymptote with respect to grammatical morphemes. Jiang, Novokshanova, Masuda, and Wang (2011) termed it *the morphological congruency hypothesis*, which states that "when L2 learners reach an advanced or near-native level of L2 proficiency, only congruent learners (i.e., those whose L1 has a corresponding morpheme to the target L2 morpheme) are able to reach nativelike proficiency in acquiring an L2 morpheme" (p.943). The present study supports this hypothesis because a TLU score over 0.90 is hardly achievable for ABSENT (or incongruent) learners at the CPE level. This is not observed in the PRESENT group. The morphemes whose TLU scores are initially high (articles, progressive *-ing*, plural *-s*) achieve yet higher

accuracy at upper levels, even after the TLU scores pass through 0.90. In fact, at and after the PET level, the majority of data points exceed 0.90. The finding shows that even at a late stage of L2 acquisition, a clear influence of L1 is still observed for basic morphemes.

And fourth, even at the same exam level and in the same group of PRESENT/ABSENT, some morphemes, such as possessive *'s* and, to a lesser extent, third person *-s*, show larger variability than others, such as progressive *-ing*. This implies that L1 type may not fully account for L1 influence.

2.3.4.2 Logistic regression model

Model specification. In order to isolate the effect of L1 from other factors and examine its strength, a regression model was fit to TLU scores and the regression coefficient of L1 type was inspected. As TLU score is proportion data (with penalties on overgeneralization), I employed a logistic regression model, a model regressing logit-transformed TLU scores rather than TLU scores per se. The numerator of the TLU score (the number of correct suppliances) was entered as the number of successes, and the denominator minus the numerator (the number of obligatory contexts + the number of overgeneralizations - the number of correct suppliances) was entered as the number of failures. L1Type (two levels; ABSENT or PRESENT), L1 (originally seven levels; one for each L1 group, but see below), ExamLevel (linear continuous variable for exam levels), and Morpheme (six levels; one for each target morpheme) were entered into the model. The quadratic and cubic terms of ExamLevel (ExamLevel^2, ExamLevel^3) were also entered in order to capture its potentially nonlinear effect. Additionally, the model also included the two-way interactions among the above variables. Dummy variables with treatment contrasts were employed for factors. Note that L1Type and L1 are different; the former tests whether the PRESENT group scores higher than the ABSENT group, and the latter tests whether certain L1 groups score higher than others and controls for its effect when analyzing the

other variables. Although two-way interactions among the variables were generally put in the model initially, L1-Morpheme interaction was not entered as it would preclude the investigation of the concerned interaction between L1 influence and morpheme. This is because entering the L1-Morpheme interaction means allowing the effect of L1 (e.g., Korean, Russian) on the TLU score to vary across morphemes, and it fully captures the variance explained by the L1Type-Morpheme interaction, which allows the effect of L1Type (whether or not the target morpheme is obligatorily marked in L1) to vary across morphemes. Since the latter is what we are interested in, the L1-Morpheme interaction was not included in the model. Furthermore, the interaction between L1 and L1Type and that between ExamLevel variables (e.g., interaction between ExamLevel and ExamLevel^2) were also excluded since their substantial interpretation is difficult.

L1Type was, as mentioned earlier, encoded in a binary manner. ExamLevel was FCE-centered in order to facilitate interpretation; that is, -2 was allocated to KET, -1 to PET, 0 to FCE, 1 to CAE, and 2 to CPE. The interaction terms were reduced in a backward-elimination manner, dropping the one with the highest p-value calculated with Type II sum of squares and comparing the model with and without the interaction by F-tests, until the model fit became significantly worse at $p < 0.05$. With respect to L1, several levels were conflated for statistical parsimony (Crawley, 2013). I did this by looking at the estimates of each level, collapsing the two L1 groups with the closest estimates, and inspecting whether the resulting model was significantly worse than the one without the conflation. The procedure was repeated until the model became significantly worse than the previous model. The model presented below is the one with as few L1 groups as possible without losing fit. Consequently, L1 Japanese, Korean, and Spanish groups were clustered together, and so did L1 Turkish and French groups. This deletion procedure was not applied to Morpheme because conflating its levels would make it difficult to interpret the interaction terms it participates in. The reference level of L1Type was the ABSENT group, that of L1 was L1

Turkish / L1 French group, and that of Morpheme was articles.

Due to overdispersion (residual deviance = 1463.6 on 120 d.f.) quasibinomial rather than binomial distribution was assumed. Overdispersion "means that there is extra, unexplained variation, over and above the binomial variance assumed by the model specification" (Crawley, 2007, p.576). In order for the assumption of binomial distribution to be met, residual deviance should be approximately equal to the residual degrees of freedom. However, in the present model, the residual deviance was 12 times larger than the degrees of freedom (1463.6/120 = 12.197). One way to counter overdispersion is to use what is known as a scale parameter which is estimated from the observed data. Quasibinomial distribution is binomial distribution adjusted by the scale parameter (Crawley, 2013). The present model, thus, adopts quasibinomial distribution as a solution to the problem of overdispersion. When a model adopts quasibinomial distribution, an F-test, rather than a likelihood ratio test normally used in logistic regression, should be employed to formally compare the goodness of two models.

Summary of the model. Table 12 displays the estimates of each parameter. L1 German learners scored higher than the L1 Turkish / L1 French cluster when the exam level was controlled for, and so did L1 Russian learners and the L1 Japanese / L1 Korean / L1 Spanish cluster. Note that this does not invalidate the results of correlation analysis or TLU-based clustering analysis presented earlier, given the absence of within-L1 differences in the accuracy order along the learners' development (cf. Table 10). ExamLevel, together with its squared term, participate in a rather complex interaction with morpheme, which means that accuracy differences between morphemes are not equal across exam levels. Due to the complexity, its substantial interpretation will not be attempted. Certain differences were also observed among Morpheme estimates. However, as was just mentioned, it interacts with ExamLevel, which means the accuracy difference between morphemes changes as learners' proficiency goes up.

Table 12

Summary of the Logistic Regression Model Fitted to TLU Scores (n = 176)

Parameter	B	SE
Intercept	1.036 ***	0.045
L1		
German	0.479 ***	0.062
Japan/Korean/Spanish	0.092 *	0.039
Russian	0.209 ***	0.061
ExamLevel	0.179 ***	0.028
ExamLevel^2	0.032	0.018
Morpheme		
Past tense *-ed*	-0.109	0.097
Plural *-s*	0.773 ***	0.074
Possessive *'s*	-0.192	0.150
Progressive -ing	0.983 ***	0.141
Third person *-s*	0.489 ***	0.130
L1 Type		
PRESENT	1.152 ***	0.045
ExamLevel : Morpheme		
ExamLevel : Past tense *-ed*	0.090	0.091
ExamLevel : Plural *-s*	0.104 *	0.052
ExamLevel : Possessive *'s*	-0.098	0.188
ExamLevel : Progressive -ing	0.219 **	0.082
ExamLevel : Third person *-s*	0.217 *	0.090
ExamLevel^2 : Morpheme		
ExamLevel^2 : Past tense *-ed*	-0.027	0.064
ExamLevel^2 : Plural *-s*	-0.040	0.034
ExamLevel^2 : Possessive *'s*	0.072	0.104
ExamLevel^2 : Progressive -ing	-0.222 ***	0.060
ExamLevel^2 : Third person *-s*	-0.039	0.060
L1 Type : Morpheme		
PRESENT : Past tense *-ed*	NA	NA
PRESENT : Plural *-s*	-0.492 ***	0.086
PRESENT : Possessive *'s*	-0.904 ***	0.176
PRESENT : Progressive -ing	-0.193	0.167
PRESENT : Third person *-s*	-0.938 ***	0.143

Note. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$

L1 influence. L1Type is highly significant after controlling for the three variables above, which means that learners whose L1s have articles (reference-level morpheme) achieved higher accuracy in their use than those whose L1s do not. This accords with the impression from Figure 3 and reconfirms the strength of L1 effect. The significant impact of L1Type on articles is also reflected in the relatively large estimate, whose size (1.152) is even much larger than the difference between the highest and the lowest exam levels (CPE - KET = $(0.032 \times 2^2 + 0.179 \times 2)$ - $(0.032 \times (-2)^2 + 0.179 \times (-2)) = 0.714$). Thus, although the effect varies across morphemes as will be demonstrated below, L1 influence can outweigh general proficiency differences in terms of accuracy with regards to grammatical morphemes. This analysis also satisfies a piece of evidence for L1 influence, cross-linguistic congruity (Jarvis, 2000). This is because it shows an association between the presence or absence of an equivalent form in L1 (L1 performance) and the accuracy of the morpheme in L2 (L2 performance).

Varying strength of L1 across morphemes. Given that L1 influence is powerful, the next question is whether its strength differs across morphemes. The answer from the present data is affirmative, as the model with the interaction between Morpheme and L1Type had a significantly better fit than the one without the interaction term [$F(5, 150) = 18.173, p < 0.001$]. The interaction term is further examined here. The interaction parameter between past tense *-ed* and L1Type could not be estimated because all the target L1s have equivalent forms to the morpheme and it was not possible to contrast the languages with and without past tense marking. Judging from the size of the coefficients, L1Type was the strongest in articles (interaction estimate of 0), followed by progressive *-ing* (-0.193), plural *-s* (-0.492), possessive *'s* (-0.904), and third person *-s* (-0.938) as least affected by L1. With regards to third person *-s*, the accuracy difference between the ABSENT and the PRESENT groups became smaller than the difference stemming from one exam level (PRESENT-ABSENT difference = 1.152 - 0.938 = 0.214; one exam level difference =

(0.032 - 0.039) $\times$ 1^2 + (0.217 + 0.179) $\times$ 1 = 0.389). For more concrete examples, the expected difference in the TLU score of articles at the FCE level (intercept exam level in the model) produced by L1 Turkish / L1 French learners of English (reference-level L1 group) between the ABSENT and the PRESENT groups is 0.161 $\left(\frac{e^{1.036+1.152}}{1+e^{1.036+1.152}} - \frac{e^{1.036}}{1+e^{1.036}} = 0.899 - 0.738 = 0.161\right)$. Considering that the TLU scores in the present data are generally high (*M* = 0.873) and the dispersion not so large (*SD* = 0.085), this difference is prominent. However, the same expected difference of third person *-s* is merely 0.029 $\left(\frac{e^{1.036+0.489+1.152-0.938}}{1+e^{1.036+0.489+1.152-0.938}} - \frac{e^{1.036+0.489}}{1+e^{1.036+0.489}} = 0.851 - 0.821 = 0.029\right)$. Although the sources of this difference between the morphemes are not clear, a speculation is made in the Discussion section.

2.4 Discussion

Overview. The present study posed three research questions; (1) whether L1 influence alters the accuracy order of L2 English grammatical morphemes, (2) how strong L1 influence is in determining the accuracy, and (3) whether the target morphemes are equally sensitive to L1 effect. In order to argue for L1 influence (RQ 1), three pieces of evidence were necessary (Jarvis, 2000); (i) intragroup homogeneity, (ii) intergroup heterogeneity, and (iii) crosslinguistic performance congruity. The first two of these (RQ1-i and RQ1-ii) were addressed by the SOC-based correlation analysis and the TLU-based morpheme clustering approach. The last (RQ1-iii) and the remainder of the research questions (RQ2 and RQ3) were answered by regression analysis.

Summary of the correlation analysis. Regarding the correlations between SOC orders, the study illustrated the relative similarity of the accuracy order within L1 groups (corresponding to RQ1-i) compared to the order between these groups (RQ1-ii). This indicates that within-L1 SOC order pairs are more similar than between-L1 pairs, satisfying the first two requirements of Jarvis (2000).

Summary of the clustering approach. Turning to the clustering approach based on TLU scores, the present study showed the between-L1 difference in the accuracy order with respect to all of the target morphemes (RQ1-ii) and the within-L1 stability of the order across proficiency levels (RQ1-i). This again satisfies the two requirements for the identification of transfer proposed by Jarvis (2000). The similar within-L1 order shows that L2 development does not alter the accuracy order of grammatical morphemes at least in the intermediate to advanced level targeted in the present study. Together with large between-L1 differences, both intralingual similarity and interlingual heterogeneity were identified in the present data.

Differences with the natural order. This study also directly tested whether the accuracy orders obtained matched with the natural order (Table 11). It is surprising that no difference was observed between the natural order and the rank of possessive *'s* in any L1 group because Luk and Shirai (2009) found a robust effect of L1 on its acquisition. However, as shown in Table 9, the direct comparison between L1 groups still reveals noticeable differences in its accuracy order, thus it is not that L1 does not have an effect, but that the effect is not powerful enough to force an L1 group to deviate from the given order. This is still interesting because it suggests that some morphemes are more impervious to L1 than others. Another notable finding is that there is no difference between the accuracy order of Spanish learners of English and the natural order. This supports a Luk and Shirai's (2009) hypothesis that the natural order is merely the reflection of the order of acquisition by Spanish learners of English.

Strong L1 influence. The regression analysis aimed to satisfy the crosslinguistic performance similarity, which is the third requirement to identify L1 influence (RQ1-iii), evaluating the strength of L1 influence (RQ2), and examining whether all the morphemes are evenly affected by L1 influence (RQ3). This study achieved the first goal (RQ1-iii) by demonstrating that the learners whose L1s have equivalent forms to the morpheme tend to

achieve higher accuracy with regards to the morpheme. In fact, when Table 7 and Table 5 are compared, we can see that the morphemes on which learners achieved over 90% accuracy at the CPE level were almost exclusively the morphemes which are present in the learners' L1s, with the only exception being progressive *-ing* in L1 Russian. This provides support for the morphological congruency hypothesis (Jiang et al., 2011). In other words, the absence of the equivalent form in L1 nearly always leads to an accuracy below 90% in all morphemes. Among the 19 morphemes which failed to achieve 90% accuracy at the CPE level, only five had equivalent forms in learners' L1s. These were possessive *'s* in L1 Japanese, L1 Korean, L1 Turkish, and L1 German, and plural *-s* in L1 Turkish. Thus, even at the CPE level, it is difficult for learners to achieve 90% accuracy without the aid of L1. However, when their L1 assists them, they have a good chance of success with the exception of possessive *'s*.

Summary of the regression analysis. More precise answers to the three goals set out above were provided by a logistic regression analysis that controlled for the effects of L1, exam level, and morphemes. Even after controlling for these variables, L1 type was a highly significant predictor. This is interesting as the variable L1Type merely took two values; whether the morpheme is obligatorily marked in the L1 or not. The significance of this variable constitutes the evidence for cross-linguistic performance congruity (RQ1-iii) as it shows that obligatory marking of the morpheme in L1 leads to higher performance in L2 than when the morpheme is absent in L1. Together with intralingual similarity (RQ1-i) and interlingual heterogeneity (RQ1-ii), the present study provides powerful evidence for L1 influence on the accuracy order of English grammatical morphemes (RQ1 as a whole). Moreover, the effect is relatively large, exceeding the effect attributable to the difference between the lowest (KET) and the highest (CPE) level exams depending on morphemes (RQ2). Therefore, the effect of L1 influence can override the general proficiency difference, and this illustrates the powerfulness of L1 on L2 grammatical morphemes.

As to the interaction between morpheme and L1 influence (RQ3), the present study found that L1 influence is not equally strong among the target morphemes. More specifically, articles were affected the most, followed by progressive *-ing*, plural *-s*, possessive *'s*, and third person *-s*. These can be arranged into three groups according to their sensitivity to L1 influence; (1) articles and progressive *-ing*, whose TLU scores drastically change between the ABSENT and the PRESENT groups; (2) plural *-s*, which is mildly affected by L1 influence; and (3) possessive *'s* and third person *-s*, which are relatively immune to the effect of L1. Although the present analyses do not provide a reason for this order, it is possible to interpret the result under the framework of thinking for speaking.

Interpretation under thinking for speaking. Recall that thinking for speaking claims that linguistic features that only require the concepts clear to non-linguistic mind and eye (e.g., plurality) are less susceptible to L1 influence than those requiring language-specific concepts such as definiteness (e.g., articles) or aspect (e.g., progressive *-ing*). This does not mean, for instance, that German and French learners of English cannot perceive progressiveness because their languages do not regularly mark them. Rather, because their languages do not require them to look at the world from that dimension, they do not develop the "selective attention" to progressiveness (cf. Slobin, 1996), and that is why they have difficulty in acquiring the progressive aspect.

Let us now interpret the present data with this framework. As claimed above, and as the data of the present study show, definiteness and aspect are considered to be extremely difficult to acquire without the aid from L1s. According to the thinking for speaking, a reason for their difficulty lies in their language-specific nature of concept. Neither definiteness nor aspect is a clear concept from a non-linguistic point of view, and speakers give thoughts to them only when they verbalize the events. L1 Japanese, Korean, Russian, and Turkish learners of English have great difficulty in paying attention to definiteness because their languages do not force them to do so. L1 German and French learners face difficulty in

acquiring progressive *-ing* for the same reason. These can be why articles and progressive *-ing* are the two morphemes that are affected by L1 the most.

Plurality, as argued in Slobin (1996), is a clearer concept non-linguistically because the singular-plural distinction is often visible in the real world. This makes it less demanding for the learners without plural marking in their L1s to pay attention to and acquire the dimension of the L2. However, in order to use plural *-s* correctly in English, one also has to know the count-mass distinction, which Slobin does not seem to take into account in the earlier quote. Jarvis and Pavlenko (2007), citing Hiki (1991) and Yoon (1993), claim that the speakers of classifier languages such as Japanese perceive noun countability differently from those of noun class languages, and as a result, encounter difficulties in the acquisition of plurality marking. Butler (2002) also reports the difficulty for L1 Japanese speakers to correctly decide on the countability of English nouns. The count-mass distinction, therefore, seems less clear non-linguistically than plurality, and classifier language speakers, which equate to the ABSENT group in the present study, seem to have more difficulty in the distinction than noun class language speakers, or the PRESENT group. This is why plural *-s* is less prone to L1 influence but is still moderately affected.

The two least sensitive morphemes to L1 influence are possessive *'s* and third person *-s*. Possession is linguistically marked even in the L1s of the ABSENT group typically by the form of postnominal modification. This means that even the learners in the ABSENT group have been trained to pay attention to the concept of possession, and, unlike the three morphemes discussed so far, they do not need to learn to attend to it. As to third person *-s*, although the L1s of the ABSENT group do not encode person morphologically, the distinction is apparently not unique to language. Those in the ABSENT group (the speakers of Japanese, Korean, and Turkish) are likely to be able to draw distinctions between person. Thus, again, they do not have to learn to pay attention to the concept. These are supposedly the reason that these two morphemes are relatively immune to L1 influence.

To summarize, articles and progressive *-ing* are difficult for the ABSENT group learners because they have to learn both to attend to relevant distinctions and to map the form onto the concept. The task is easier in the case of plural *-s* because all the learners are trained by L1 to pay attention to the singular-plural distinction. The count-mass distinction, however, poses a unique challenge to the ABSENT group, whose L1s are classifier languages. As to possessive *'s* and third person *-s*, L1s have trained all the learners, PRESENT or ABSENT, to attend to person and possession. The remaining task for both groups to acquire the morphemes is just to map the form onto the concept, and thus, the PRESENT-ABSENT difference is relatively small, the only difference being that those in the ABSENT group have to linguistically mark a concept that is not marked in their L1s. These are fully in line with the prediction of thinking for speaking.

Limitations. One possible limitation of this study is that it investigates accuracy of morphemes but not directly their acquisition. However, the fact that there were few within-L1 differences in the accuracy order indicates that the order does not change along the developmental path. Therefore, no matter what criterion is used for acquisition, it is unlikely that the accuracy order observed in this study will be far different from the acquisition order.

Another limitation is that the data used in the study consisted of exam scripts, in writing which learners might have been form-focused and exhibited different orders than if they were meaning-focused. The present study, however, generally shows agreement with the previous studies which were conducted in a more naturalistic environment. A case in point is the uniformity between the natural order and the accuracy order of L1 Spanish learners. As Luk and Shirai (2009) show the similarity between the natural order and the acquisition order of L1 Spanish learners, this can be taken as evidence of the resemblance in the accuracy orders observed under exam situations and those in more naturalistic settings.

A further potential limitation is that exam levels may not represent learners' proficiency.

The study assumed that learners taking higher-level examinations are of higher proficiency in general, and although the assumption was partially verified by Guiraud's index, the measure only captures lexical richness, which is not equal to overall proficiency. A way to tackle this issue is to measure learners' proficiency by well-designed proficiency tests such as TOEFL and use the scores as the measure of proficiency. Although such scores are not available in the CLC, the corpus includes for each script the information on whether the learner passed or failed the exam. Limiting the analysis to the scripts of the learners who passed the exam sets a threshold that they have to have crossed. This was not attempted in the present study because it would have significantly decreased the amount of data targeted, and might have prevented fine-level analysis. Nevertheless, just to see whether the picture would change, the same analysis as the present study was carried out on the subcorpus that included only pass scripts. The result was that the overall picture does not change, and the main findings, including the claims regarding the natural order and the order of morphemes in terms of the strength of L1 influence, stay the same. This assures that the better control of learners' proficiency is unlikely to alter the findings of the present study.

2.5 Conclusion

L1 influence is pervasive in all areas of L2 acquisition (Jarvis & Pavlenko, 2007; Odlin, 1989). Despite the numerous claims for the "natural" order of L2 acquisition of English grammatical morphemes, this study has demonstrated that they are not an exception to L1 influence. A correlation analysis based on SOC scores indicated the cross-linguistic influence in the order by comparing within-L1 correlations and between-L1 correlations. Clustering of morphemes based on TLU scores directly falsified the concept of the universal accuracy order of morphemes across L1s. Regression analysis implied L1 transfer in the accuracy order by illustrating that the morphemes with equivalent forms in L1 mark higher accuracy order. These pieces of evidence should be more than sufficient to cast a

strong doubt on the universality of accuracy order, and it is very likely that "SLA textbooks underestimate the effect of L1 in their discussion of morpheme studies and need to rethink the conventional wisdom" (Luk & Shirai, 2009, p.743). The study further demonstrated that the effect of L1 varies across morphemes and that those requiring learners to draw the distinctions absent in their L1s are more prone to L1 influence.

Chapter 3: Cross-sectional Analysis of EF-Cambridge Open Language Database

3.1 Introduction

The last chapter demonstrated strong L1 influence on the L2 acquisition of some of the English grammatical morphemes. However, as the study employed the Cambridge Learner Corpus (CLC) and focused on the accuracy (order) differences between L1 groups within each proficiency level, it could not directly investigate individual variation in the developmental patterns of morpheme use by each learner. In order to complement this, the thesis also exploits EF-Cambridge Open Language Database (EFCamDat), and analyzes the development of grammatical morphemes. This chapter is primarily devoted to understanding the nature of EFCamDat by comparing it with the CLC, and the next chapter examines the longitudinal development of individual learners. The following research question is addressed: How robust are the findings based on the CLC in the last chapter? Are they replicable with EFCamDat? This question is addressed in terms of (i) the absolute accuracy of grammatical morphemes, (ii) the accuracy order among them, and (iii) the strength of L1 influence. These aspects of comparison will test the degree to which the findings reported in the last chapter are robust in that they are replicated with the data in EFCamDat.

3.2 Method

3.2.1 Target Morphemes

The study targets the same six morphemes as in the last chapter; articles, past tense *-ed*, plural *-s*, possessive *'s*, progressive *-ing*, and third person *-s*.

3.2.2 Corpus

EFCamDat is being jointly developed by Education First (EF) and the Department of Theoretical and Applied Linguistics at the University of Cambridge. The corpus contains learners' essays submitted to Englishtown, the online school of EF. A full English curriculum at Englishtown consists of 16 Lessons, corresponding to 16 proficiency levels. Each Lesson includes eight Units. A placement test can suggest an appropriate level for the learner to start the course. Each Unit covers a range of activities involving both receptive (listening and reading) and productive (speaking and writing) practice, as well as some explicit instruction on grammatical features (e.g., inflections, prepositions). At the end of each Unit is a free composition in which learners are asked to write on a certain topic. An example essay is provided for each composition task, and learners can refer to it while they write their essays. Also, because Englishtown is an online school, learners were able to consult reference materials such as dictionaries in writing their essays. Each composition task has expected length, from 20-40 words in Lesson 1 Unit 1 to 150-180 words in Lesson 16 Unit 8. Unless learners repeat the same Unit, a learner who completes the full course submits 128 (16 Lessons $\times$ 8 Units) essays in total. A variety of topics are covered including self introduction, making requests, offering an apology, and writing a story, among other things. For further information on Englishtown, visit their website at http://www.englishtown.com/. EFCamDat is publicly available at http://corpus.mml.cam.ac.uk/efcamdat/.

Learners in Englishtown receive feedback from native-speaker teachers on each essay. The feedback includes identification and correction of grammatical morphemes, among other things. I use the feedback to collect necessary information in the calculation of accuracy scores explained later. EFCamDat can, therefore, be viewed as a partially error-tagged longitudinal learner corpus. The feedback is not available for all the essays in the corpus,

however.

Apart from learners' essays, EFCamDat includes for each essay such metadata as the ID of the learner, his/her country of residence, the topic of the essay, the date and time of submission, and the Lesson and the Unit number for which the essay was written.

During the data collection phase, 76,002 learners wrote 423,117 essays in total. The total number of words in the corpus is 29,161,383, and the mean number of words for an essay is 68.911 (SD = 45.698; Median = 63). The total number of words in error-tagged texts is 11,266,085, and their mean number of words for an essay is 73.104 (SD = 44.973; Median = 66). For a reference, the number of essays in the subcorpus of the CLC used in the last chapter was 11,893, with 4,016,330 words in total. The mean number of words for an essay was 337.705 in the CLC. This shows that EFCamDat includes a larger number of learners and is larger in data size, but a typical essay is longer in the CLC than in EFCamDat.

3.2.3 Target L1 Groups and Proficiency Levels

Target L1 groups. The study targeted the same L1 groups as in the last chapter; Japanese, Korean, Spanish, Russian, Turkish, German, and French. Since no direct information on learners' L1 is available in the corpus, I used their country of residence as the closest approximation. If a learner lives in Japan, for example, his/her L1 was assumed to be Japanese. L1 Spanish includes two countries of residence; Spain and Mexico. L1 Korean, Russian, Turkish, German, and French learners correspond to those living in Korea, Russia, Turkey, Germany, and France respectively. As in the last chapter, those with the L1s that lack target morphemes are referred to as the ABSENT group, and those with the L1s that have them are referred to as the PRESENT group. The same dichotomous approach was adopted and the same coding was used as in the last chapter.

Table 13

Alignment of Englishtown Lessons and the CEFR

Englishtown Lesson number	CEFR levels
1-3	A1
4-6	A2
7-9	B1
10-12	B2
13-15	C1
16	C2

Target proficiency levels. The Lesson number the learner wrote the essay at was considered to represent the proficiency of the learner. Englishtown proficiency levels are aligned with the Common European Framework of Reference (CEFR) as shown in Table 13.

As in the last chapter, Guiraud's Index, a measure of lexical richness, was calculated for each essay. Since the value is affected by text length, an essay longer than 100 words was truncated to the first 100 words when computing the value. The data are summarized in Figure 4. The horizontal axis represents proficiency. Though the data were plotted at the level of Unit, the axis was drawn so that the numbers indicate Lesson. This is because *Lesson 3 Unit 2* is likely to be easier to understand than *Unit 26*. Each dot represents a Guiraud Index of one essay and its color corresponds to L1 groups. The seven solid and dashed lines indicate the mean Guiraud's Index of each L1 group across proficiency. Since there was no essay at some Units in some L1 groups, not all L1 groups have 128 data points. The dashed horizontal lines indicate the average Guiraud's Index adapted from the last chapter. We can make following observations on the figure. First, as expected, Guiraud Index shows an increasing tendency as proficiency goes up. However, at around Lesson

10, the value seems to have reached the ceiling and shows a flat transition afterward. The rising trend of Guiraud's Index can be interpreted to mean that the overall proficiency is generally higher at later Lessons. Second, the value does not show much variation across L1 groups. The patterns are relatively similar across L1 groups at all levels. This means that any difference found between L1 groups should not be attributed to the overall proficiency difference. It is interesting to observe the uniformity among the L1 groups, especially given the clearly bumpy transition of the measure in the figure. This indicates that what we see in the figure is partially a reflection of task effects, and Guiraud's index may be a sensitive measure to the properties of tasks. And third, in terms of Guiraud's Index, Lesson 1 through 5 roughly correspond to the KET level in the CLC and 7 through 10 or 11 to the PET. It is difficult to tell which level the FCE, CAE and CPE levels correspond to in EFCamDat due to the ceiling effect mentioned. At least for the KET and PET levels, the claimed correspondence between the Lessons at Englishtown and CEFR seems to hold. That is, Lesson 3 to 6 correspond to A2, which the KET levels is claimed to correspond to, and Lesson 7 to 9 to B1 and the PET level.

Corpus size. Table 14 shows the number of error-tagged essays, total words, and the average length of an essay in each country and proficiency level. The proficiency is represented by the Lesson at which the essay was written. Although some proficiency levels include more essays than others, it seems that the number of essays is comparable across levels until Lesson 11, with a possible exception of Lesson 1 where the volume of essays is high. After Lesson 11, the number of essays becomes increasingly smaller. The distribution of essays between L1 groups is highly skewed. Over half of the data are contributed by Russian learners of English and another 10-15% each by Spanish and German learners. This, however, does not affect our results since the data were not aggregated across L1 groups for most part of the analysis. As in the CLC, the average number of words per essay shows an upward trend as proficiency rises. Overall, the subcorpus contains over 45,000

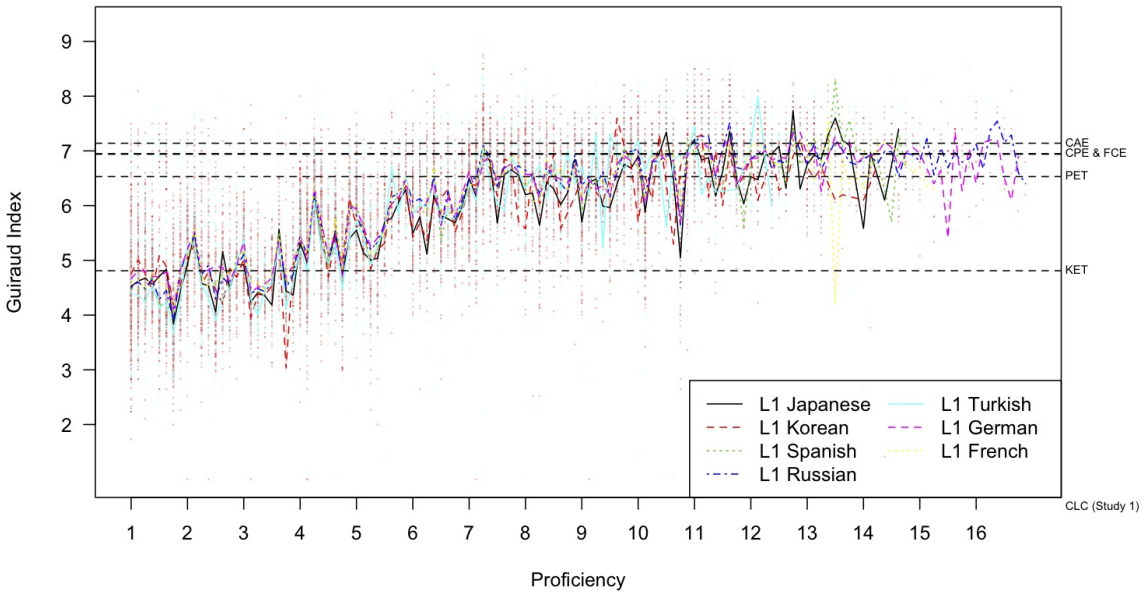


Figure 4. Development of Guiraud Index Across Proficiency

essays consisting of nearly 4 million words in total. The corpus size is comparable to the subcorpus of the CLC employed in the last chapter, which contained slightly over 4 million words. However, the number of essays is much larger in EFCamDat, compared to 11,893 essays in the CLC.

3.2.4 Scoring Method

The present study analyzed accuracy development, and as a measure of accuracy, it employed target-like use (TLU) score as in the last chapter.

3.2.5 Data Extraction

Tallying frequency necessary to compute TLU scores. In order to obtain TLU scores, the number of obligatory contexts, the instances of overgeneralization, and those of errors were extracted from error-tagged texts. Since feedback is often not provided after reaching the word limit, only the part of the essay within the word limit was targeted. For

Table 14

The Number of Error-Tagged Essays and Total Words in Each L1 and Proficiency Group

		Proficiency																Total	%
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
L1 Japanese	# Essays	452	178	134	299	193	142	208	119	133	122	81	32	23	11	5	0	2,132	4.7%
	# Words	16,397	8,782	5,455	18,289	11,934	8,616	18,770	9,540	12,130	14,474	9,387	3,690	3,525	1,604	863	0	143,456	3.7%
	# Words/Essay	36.3	49.3	40.7	61.2	61.8	60.7	90.2	80.2	91.2	118.6	115.9	115.3	153.3	145.8	172.6	NA	67.3	
L1 Korean	# Essays	190	71	53	153	89	72	108	22	40	52	28	18	15	3	5	2	921	2.0%
	# Words	7,601	3,465	2,147	9,619	5,775	4,620	10,310	1,762	4,217	5,914	3,982	2,353	2,254	454	690	318	65,481	1.7%
	# Words/Essay	40.0	48.8	40.5	62.9	64.9	64.2	95.5	80.1	105.4	113.7	142.2	130.7	150.3	151.3	138.0	159.0	71.1	
L1 Spanish	# Essays	2,012	732	563	921	446	261	447	236	204	205	116	83	49	13	2	0	6,290	13.9%
	# Words	69,884	35,646	26,381	66,248	31,876	19,734	44,265	23,222	21,411	25,605	16,068	11,850	8,455	2,276	321	0	403,242	10.4%
	# Words/Essay	34.7	48.7	46.9	71.9	71.5	75.6	99.0	98.4	105.0	124.9	138.5	142.8	172.6	175.1	160.5	NA	64.1	
L1 Russian	# Essays	2,418	1,220	1,111	1,805	1,217	1,406	1,983	1,973	2,496	2,789	1,953	1,139	640	311	146	76	22,683	50.2%
	# Words	91,517	63,953	54,199	132,755	89,481	111,239	204,749	204,559	277,977	380,614	283,836	162,398	116,756	57,338	29,606	13,019	2,273,996	58.9%
	# Words/Essay	37.8	52.4	48.8	73.5	73.5	79.1	103.3	103.7	111.4	136.5	145.3	142.6	182.4	184.4	202.8	171.3	100.3	
L1 Turkish	# Essays	438	185	106	270	96	57	134	29	22	75	9	4	11	4	0	1	1,441	3.2%
	# Words	15,094	9,065	5,293	17,858	6,651	4,416	12,859	2,832	2,491	8,670	1,198	479	1,855	657	0	189	89,607	2.3%
	# Words/Essay	34.5	49.0	49.9	66.1	69.3	77.5	96.0	97.7	113.2	115.6	133.1	119.8	168.6	164.3	NA	189.0	62.2	
L1 German	# Essays	1,184	575	442	1,156	529	327	731	262	242	473	174	123	155	52	24	14	6,463	14.3%
	# Words	49,087	31,948	21,616	79,898	37,785	23,086	69,963	24,317	25,134	60,708	23,997	16,357	27,553	8,950	3,747	2,441	506,587	13.1%
	# Words/Essay	41.5	55.6	48.9	69.1	71.4	70.6	95.7	92.8	103.9	128.3	137.9	133.0	177.8	172.1	156.1	174.4	78.4	
L1 French	# Essays	1,294	447	324	879	327	238	545	237	243	365	160	89	57	11	8	5	5,229	11.6%
	# Words	48,924	22,560	15,389	59,100	22,685	17,631	52,130	21,610	25,581	46,502	21,839	12,576	8,987	1,751	1,226	694	379,185	9.8%
	# Words/Essay	37.8	50.5	47.5	67.2	69.4	74.1	95.7	91.2	105.3	127.4	136.5	141.3	157.7	159.2	153.3	138.8	72.5	
Total	# Essays	7988	3408	2733	5483	2897	2503	4156	2878	3380	4081	2521	1488	950	405	190	98	45,159	100.0%
	Essays (%)	17.7%	7.5%	6.1%	12.1%	6.4%	5.5%	9.2%	6.4%	7.5%	9.0%	5.6%	3.3%	2.1%	0.9%	0.4%	0.2%	100.0%	
	# Words	298,504	175,419	130,480	383,767	206,187	189,342	413,046	287,842	368,941	542,487	360,307	209,703	169,385	73,030	36,453	16,661	3,861,554	100.0%
	Words (%)	7.7%	4.5%	3.4%	9.9%	5.3%	4.9%	10.7%	7.5%	9.6%	14.0%	9.3%	5.4%	4.4%	1.9%	0.9%	0.4%	100.0%	
	# Words/Essay	37.4	51.5	47.7	70.0	71.2	75.6	99.4	100.0	109.2	132.9	142.9	140.9	178.3	180.3	191.9	170.0	85.5	

the identification of obligatory contexts, error-tagged texts were first converted to corrected texts where the corrections of errors were reflected onto the texts. Obligatory contexts were then identified by looking into the corrected texts. As in the last chapter, the number of instances of target morphemes in a corrected text was assumed to be the number of obligatory contexts in the corresponding original essay. For instance, if the phrase *a farmer who live in a small village* was corrected into *a farmer who lived in a small village*, the latter phrase was used to count the obligatory contexts of the morphemes, and in this case, the *lived* made an obligatory context of past tense *-ed*.

I did not rely on error classification of the error-tags in extracting errors because feedback in Englishtown is not meant as error tags as in the CLC, and, partially due to the difference in the error tag sets employed, it was difficult to extract morpheme errors solely on the basis of error classification. For example, omission and overgeneralization errors of *-s*, identified by searching for such patterns as X $\rightarrow$ X(e)s or X(e)s $\rightarrow$ X, where X represents any number of any letters, showed that the error is variably classified as SI (singular), AG (agreement), PL (plural), WC (word choice), SP (spelling), VT (verb tense), x»y (change from x to y) and many others. Whereas there are certain patterns such as SI being more likely plurality errors and VT being likely agreement errors, some error tags such as x»y include a variety of errors and it seemed difficult to distinguish between different error types within the tags solely based on error-tagged essays. Therefore, I wrote a script that looks at both error-tagged texts and part-of-speech (POS) tagged corrected texts. For example, in the case of an *-s* omission error, the script checked the part of speech of the corrected word with *-s* in the corrected text, and if the word was tagged as a verb, the error was considered as an instance of third person *-s* error. POS tags were annotated by TreeTagger (Schmid, 1994), whose reported accuracy is over 95%. Similarly, in the identification of overgeneralization errors of third person *-s* or plural *-s*, the script looked into the original essays with POS tags assigned by TreeTagger. A potential issue with this approach was the accuracy

of POS tags on learner data. Since the tagger has been trained with native speaker data, its accuracy on learner data is not guaranteed. Given that the overgeneralization error only constitutes 30.5% and 28.4% of all the errors in third person *-s* and plural *-s* respectively, the negative effect was hoped to be small. The overall accuracy of the script reported below indeed shows that the performance of the scripts used to retrieve third person *-s* errors and plural *-s* errors was not poorer compared to the other morphemes. Correct suppliance was obtained by subtracting the number of omission errors from that of obligatory contexts.

Because the error tags in the EFCamDat were not meant to be error tags as mentioned earlier, there is a possibility that error annotation is not exhaustive. A manual given to Englishtown teachers, however, asks them to be thorough in error correction and explicitly mentions such errors as subject-verb agreement, use of articles, verb tense, and plural versus singular forms. Thus, at least four out of the six target morphemes in the present study are bought to the consciousness of the teacher, and this should make the error annotation of the target morphemes fairly comprehensive.

Precision and recall of the scripts used. As in the last chapter, I manually checked the accuracy of the scripts used to retrieve errors, viewing the error annotation as the gold standard. Table 15 reports the accuracy of the scripts used to extract errors. It is encouraging that even when error classification is not used, it is possible to achieve over 0.80 of the F_1 value in identifying errors of all the target morphemes.

3.2.6 Data Analysis

I applied three kinds of data analyses. First, in order to compare absolute accuracy between the CLC and EFCamDat, the TLU scores of each morpheme in each target L1 group were computed over each Lesson in EFCamDat. In order to directly compare the two corpora, the Lesson in EFCamDat was mapped onto the CEFR levels following the correspondence in Table 13. Similarly, the KET level in the CLC was considered to cor-

Table 15

Accuracy of the Scripts Used to Retrieve Errors

	Precision	Recall	F_1
Articles	90%	98%	0.94
Past tense *-ed*	76%	85%	0.80
Plural *-s*	75%	88%	0.81
Possessive *'s*	85%	82%	0.83
Progressive *-ing*	78%	86%	0.82
Third person *-s*	79%	81%	0.80

respond to A2, PET to B1, FCE to B2, CAE to C1, and CPE to C2. The TLU scores in each L1 group in EFCamDat were recalculated over the aggregated proficiency levels corresponding to each CEFR level and were compared to those of the CLC reported in the last chapter side by side.

Second, to compare the relative position of morphemes in terms of accuracy, the accuracy order of the target morphemes was compared across the two corpora in a similar manner to the accuracy order comparison in the last chapter. That is, morphemes with similar TLU scores were clustered within each level and L1 group based on bootstrapping samples, and the order was compared on the cluster basis.

Third, to gain insights into the similarities and differences between the two corpora in terms of L1 influence, a regression model was constructed and the coefficients were compared to those in the logistic regression model in the last chapter. Further, I built an aggregated regression model over the CLC and EFCamDat and tested whether the strength of L1 influence varies across the two corpora.

3.3 Results

3.3.1 Overall Picture of EFCamDat

Descriptive data. Appendix B shows the TLU score and the number of obligatory contexts in each L1, morpheme, and proficiency level. Although ten L1 groups are listed, the present study only targeted the seven L1 groups mentioned earlier (Japanese, Korean, Spanish, Russian, Turkish, German, and French), and the rest are for the next chapter. Proficiency here refers to the Lesson number at Englishtown. Although the table is large and not suited for detailed inspection, we note two observations. First, partly because proficiency is divided into 16 levels, as opposed to five in the last chapter, the total number of obligatory contexts at each level tends to be small. This is especially true at higher proficiency levels and in some morphemes such as possessive *'s*. Therefore, certain data points can be potentially unreliable. Second, the TLU scores are high overall. The average TLU score is over 0.80 in all the morphemes, and in three morphemes (plural *-s*, progressive *-ing*, and third person *-s*), it is above 0.90. The average score of third person *-s* is 0.99, which is surprisingly high. Due to the high accuracy of these morphemes, ceiling effects are expected.

Micro and macro averages of TLU scores. Figure 5 aggregates L1 groups over L1 type (ABSENT or PRESENT) and shows the accuracy transition of each morpheme across Lessons. The two lines correspond to the TLU scores of the ABSENT group (red solid line) and those of the PRESENT group (blue dashed line). The numbers on the lines indicate the total number of obligatory contexts where it was smaller than 100, and the score might be less reliable. The TLU scores in the figure were computed by micro-averaging, a term borrowed from computational linguistics, by which the data from all the target L1 groups were aggregated and TLU scores were calculated over the aggregated data. In other words, the calculation is essay-based. As described in the Method section, however, the data size

across L1 groups is not equal in the subcorpus used in the present study. Russian learners occupy over half of the data size both in terms of the number of learners and that of words. Therefore, Figure 5 might be a mere reflection of the development of Russian learners and their accuracy is not guaranteed to match with the accuracy of the other L1 groups. In order to test whether the pattern depicted in Figure 5 is indeed representative of all the ABSENT or PRESENT groups, macro averages were also computed and shown in Figure 6. Macro averages, also borrowed from computational linguistics, are the means over categories and are calculated by averaging the mean scores of each group. For example, in the case of articles in the ABSENT group, the TLU score of each level was calculated in each of the L1 Japanese, L1 Korean, L1 Russian, and L1 Turkish group, and the average of those four TLU scores was computed. This weighs all L1 groups equally, regardless of the group size. In other words, the calculation is L1-based.

We can make several observations with respect to the two figures.

- Micro averages and macro averages are mostly similar, which means the micro averages here are not skewed by one influential L1 group.
- Most data points are reliable in micro averages, except for those at higher proficiency levels.
- Accuracy tends to increase as proficiency rises in most of the morphemes (articles, past tense *-ed*, possessive *'s*, and progressive *-ing*).
- The rate of accuracy increase, however, varies across the morphemes mentioned above. Articles and past tense *-ed*, for example, appear to have a steeper slope than progressive *-ing*. This is possibly because articles and past tense *-ed* start at a lower TLU score than progressive *-ing* and have more room for accuracy increase.
- The pseudo-longitudinal development is rather flat in plural *-s* and third person *-s*. It

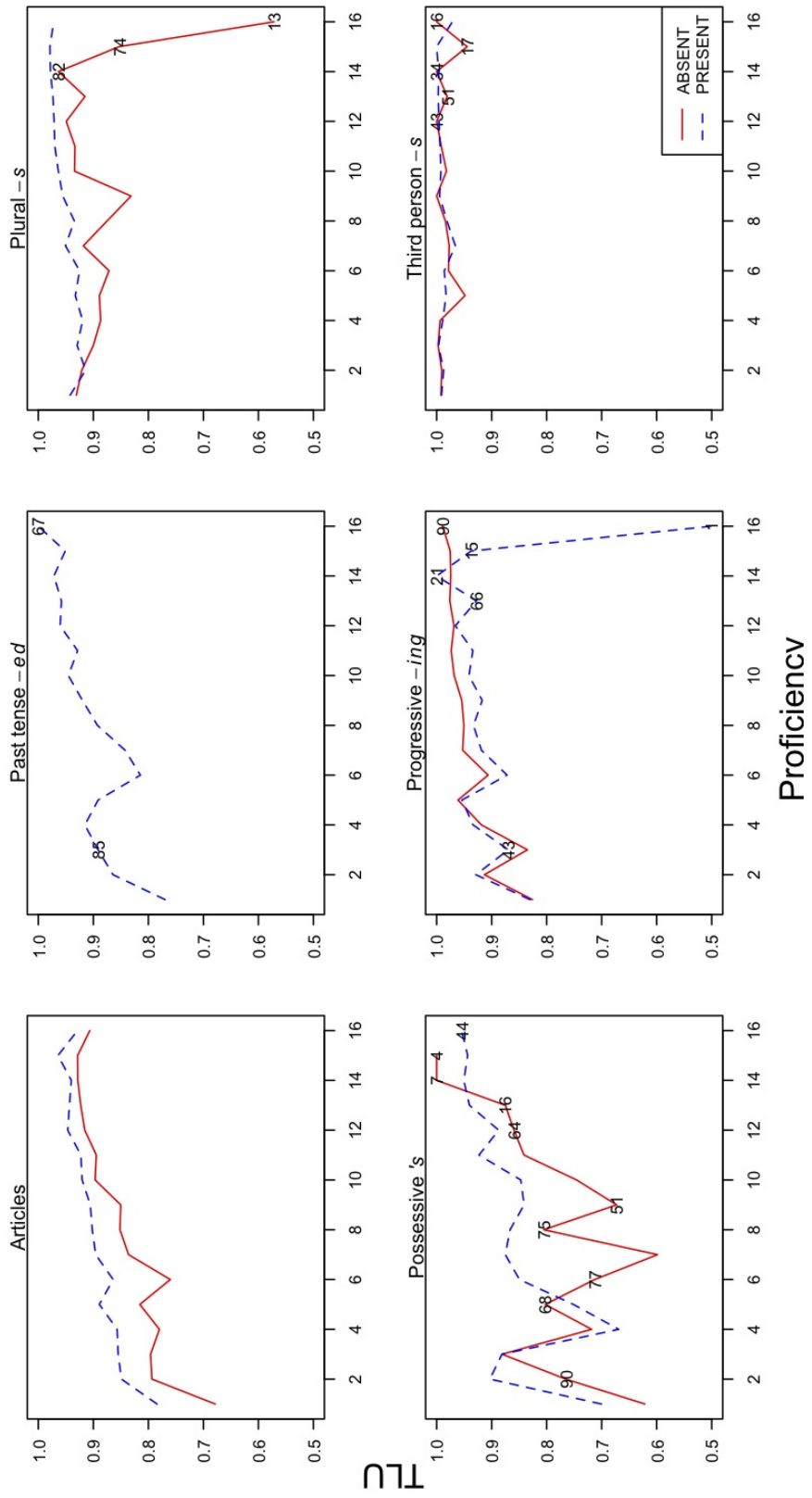


Figure 5. TLU Scores Across L1 Type (Micro Averages)

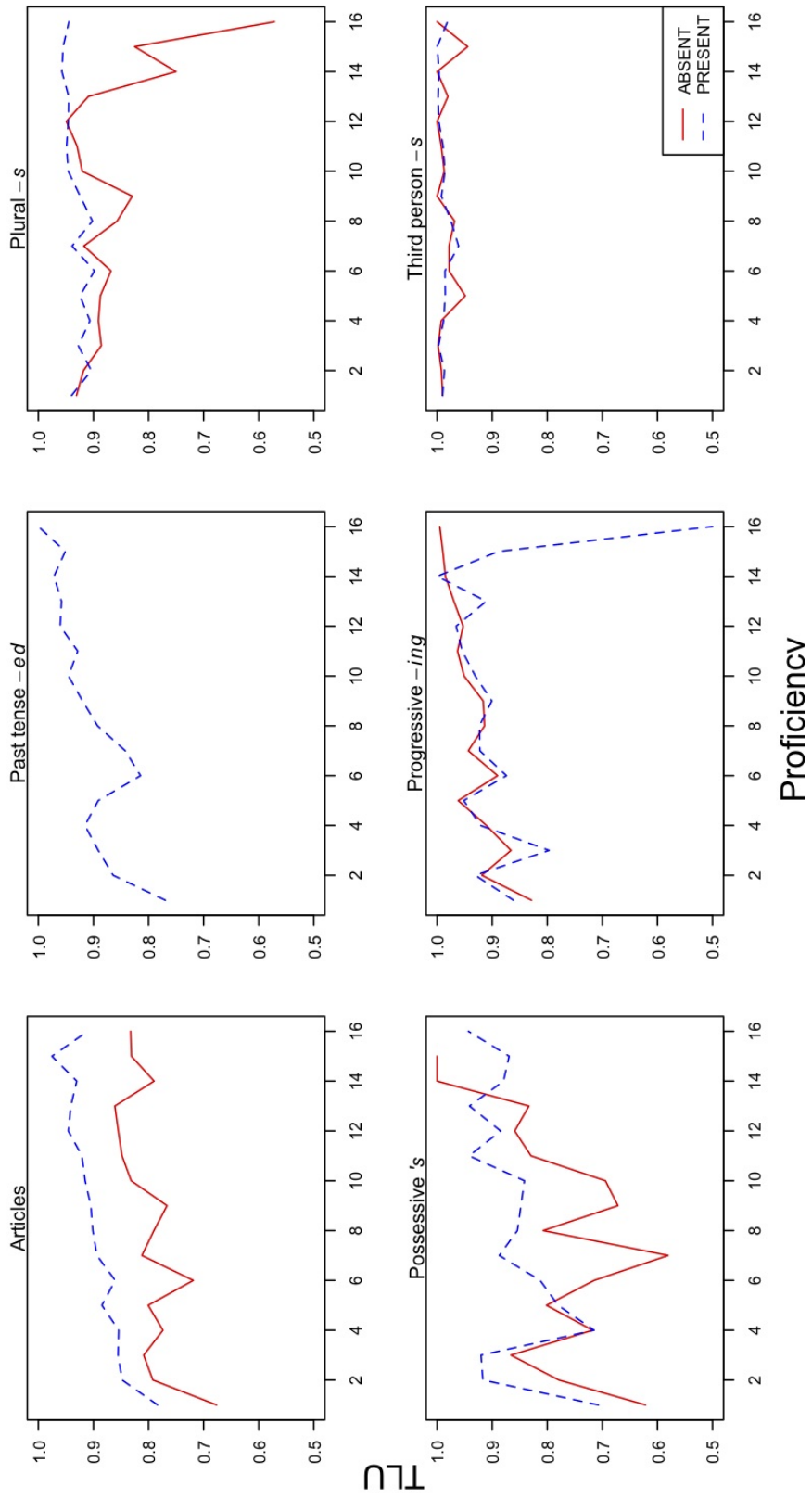


Figure 6. TLU Scores Across L1 Type (Macro Averages)

is possibly due to a ceiling effect, as both morphemes mark a high accuracy from the beginning. This is especially true for third person *-s*, whose accuracy seems to be constantly over 95%.

- L1 influence is weak, if present at all. The effect is relatively clear, though not necessarily strong, in articles and plural *-s*, as the PRESENT group almost consistently achieves a higher accuracy than the ABSENT group. However, in the other morphemes, there is no obvious effect of L1 type.

Various pseudo-longitudinal developmental patterns are observed in EFCamDat. The question now is how they are similar to or different from the CLC, to which I will turn in the following section.

3.3.2 Comparison with the CLC in the Absolute Accuracy of the Morphemes

Since the data size of individual data points tends to be smaller than in the last chapter (cf. Appendix B), the comparison between the CLC and EFCamDat was made at the level of L1 type rather than individual L1 groups. Figure 7 shows the (macro) average of TLU scores in each proficiency level, morpheme, and corpus. Since the CLC does not include the A1 level and EFCamDat has little C2 data, only A2 through C1 are targeted. Each panel shows the data of one proficiency level. In each panel, the horizontal axis represents the TLU score, and the vertical axis shows the 12 categories consisting of six morphemes times two corpora. The TLU score of the ABSENT group is represented by an "A" and that of the PRESENT group by a "P".

The relative position of the ABSENT and the PRESENT group and the difference in the absolute TLU score between the two corpora seem to vary across morphemes. More specifically, the following observations can be made regarding the comparison between the two corpora.

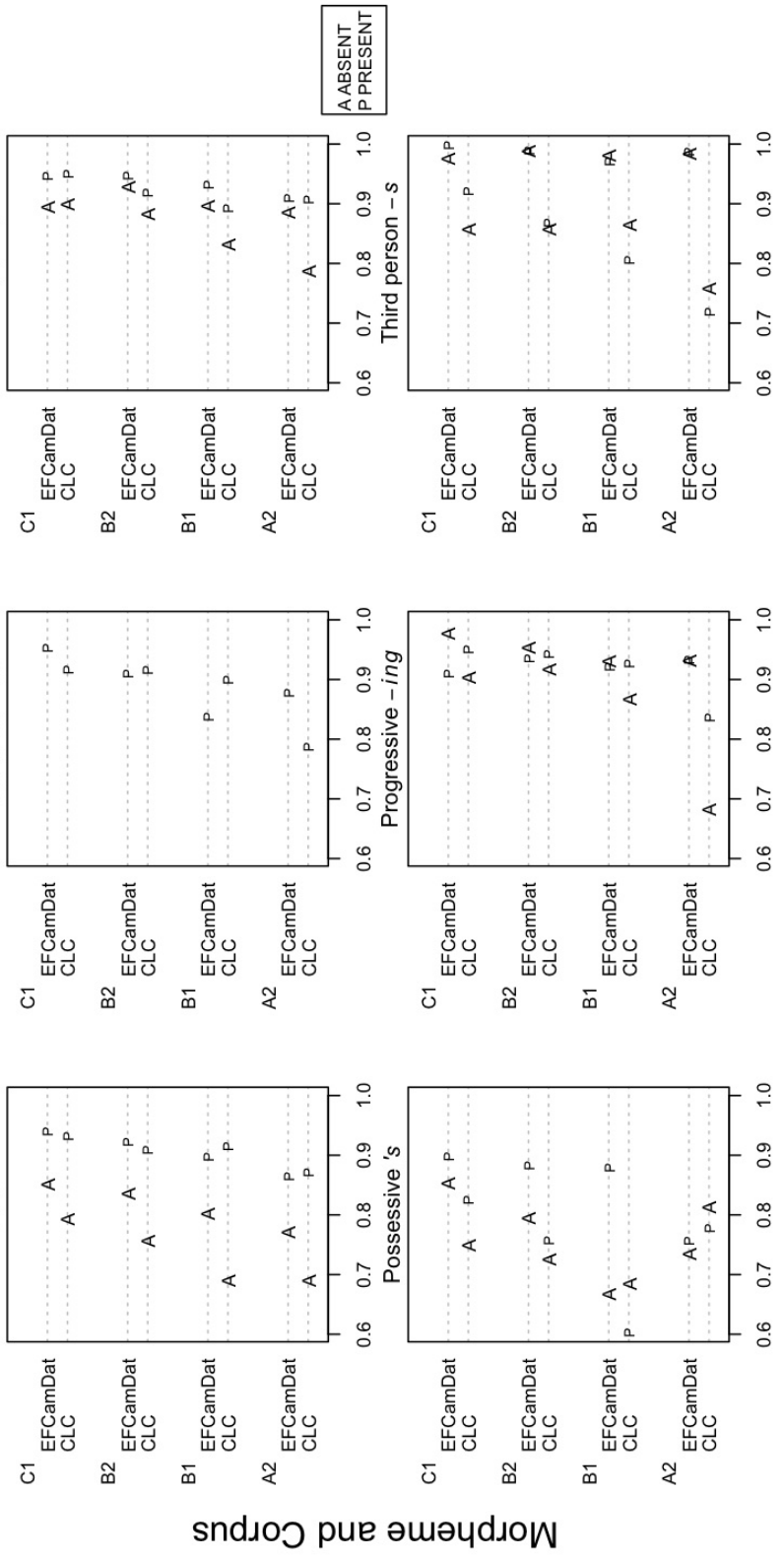


Figure 7. TLU Scores of Each Morpheme Across the Two Corpora

- With respect to articles, the accuracy order of the ABSENT and the PRESENT group is consistent across proficiency levels and across the corpora in that the PRESENT group always achieves a TLU score than the ABSENT group. The absolute accuracy of the PRESENT group is very similar across the two corpora in all levels. That of the ABSENT group, however, seems to vary to a certain extent especially in lower proficiency levels. The difference is as large as 0.1 at the B1 level.

- As to past tense *-ed*, the difference in the absolute accuracy tends to be small except for the A2 level. There is no ABSENT group because all the target L1 groups have equivalent features to English past tense *-ed*.

- With regard to plural *-s*, the accuracy order between the L1 types does not vary across proficiency levels or the corpora, and the PRESENT group marks higher accuracy than the ABSENT group. The absolute accuracy is similar in both L1 types especially at higher levels. At the A2 level, however, it shows some difference between the corpora.

- Possessive *'s* shows a wide difference between the two corpora especially at lower levels. At the A2 and B1 levels, the order of the ABSENT and the PRESENT group differs between the two corpora. In EFCamDat, the PRESENT group performs higher than the ABSENT group, as expected, but in the CLC, the ABSENT group achieves higher accuracy. At B2 and C1, the PRESENT group surpasses the ABSENT group in the CLC as well, and the order is consistent across the two groups. The absolute accuracy difference of the PRESENT group between the two corpora at B1 is huge, EFCamDat being higher than the CLC by almost 0.3. At other levels, the difference between the two corpora in both the ABSENT and the PRESENT group is within or around 0.1.

- In terms of progressive *-ing*, the accuracy order varies between the two corpora at B1 through C2 in that the ABSENT group tends to outperform the PRESENT group in EFCamDat, while the CLC shows the reverse pattern. Their absolute accuracy, however, is reasonably similar across the corpora. The absolute accuracy significantly varies across the two corpora especially in the ABSENT group at the A2 level.

- As for third person *-s*, there is a wide difference in the absolute accuracy between the two corpora in that both the ABSENT and the PRESENT group achieve very high accuracy (TLU score > 0.95) in EFCamDat throughout the (pseudo-)development whereas it is not the case in the CLC. It is not meaningful to look into the relative position of the ABSENT and the PRESENT group given the very high accuracy in both groups in EFCamDat.

Overall, the accuracy of the morphemes is relatively similar across the two corpora in articles, past tense *-ed*, and plural *-s*, but there is at least some difference in the other three morphemes (possessive *'s*, progressive *-ing*, and third person *-s*). The reason for the difference in some morphemes is unclear. It could be because an essay is typically shorter in EFCamDat than in the CLC and learners were able to sustain their attention to the accuracy of form throughout writing their essays at Englishtown, or it could be because explicit instruction learners received immediately before writing the essays affected their accuracy of writing.

3.3.3 Comparison with the CLC in the Accuracy Order of the Morphemes

In this section, the accuracy order of morphemes is compared between the CLC and EFCamDat. As in the last chapter, I performed clustering of morphemes based on bootstrapping, which produced Table 16. Since the data size was small at the KET (A2) level in the CLC, which makes the reliable comparison of orders difficult, only the B1, B2, and C1

levels were targeted in this analysis. As in Table 7 in the last chapter, morphemes whose TLU score is over 0.90 at $p < .05$ are marked with asterisks (*) and are always grouped into the highest cluster, along with the morphemes whose accuracy does not statistically differ from them. Morphemes whose number of obligatory contexts is smaller than 100 are crossed out and might be unreliable. Those underlined mean that bootstrapping-based clustering alone was not able to determine the cluster the morphemes belong to, so they were clustered into the groups that include the morpheme whose TLU score is the closest to that of the morpheme in question.

Unlike in the CLC, where the data size was small mostly at lower proficiency levels, EFCamDat has most of the data points with small data sizes at the C1 level. This is because the number of essays and that of words tend to be smaller at higher levels (cf. Table 14), which results in fewer obligatory contexts (cf. Appendix B). Let us now focus on B2 and B1 and look into the relative position of each morpheme with respect to the others. In L1 Russian learners, since all the morphemes were clustered into one at the B2 level, it is difficult to tell the relative position to other morphemes. I will therefore not interpret it.

- Articles tend to be ranked low in L1 Japanese, Korean, and Turkish learners of English. In L1 Spanish and French learners, it is located somewhere at high- to mid-rank. In L1 German learners, articles are in the highest cluster at both levels.

- Past tense *-ed* marks high- to mid-rank in L1 Japanese learners, and mid- to low-rank in L1 Korean, Turkish, and German learners. In L1 Spanish and French, it tends to be one of the least accurate morphemes. In L1 Russian, it is in the upper cluster at B1.

- Plural *-s* is consistently in the highest cluster in L1 Spanish, German, and French learners. It is in the high- to mid-rank in L1 Japanese, and in the mid- to low-rank in

Table 16

Accuracy Order in EFCamDat

C1

Clustered Order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	~~past tense *-ed*~~ plural *-s* * ~~possessive *'s* *~~ ~~progressive *-ing* *~~ ~~third person *-s* *~~	~~third person *-s* *~~	articles * ~~past tense *-ed*~~ plural *-s* * ~~possessive *'s*~~ ~~progressive *-ing*~~ third person *-s* *	articles * past tense *-ed* * plural *-s* * ~~possessive *'s* *~~ progressive *-ing* * third person *-s* *	~~past tense *-ed* *~~ ~~third person *-s*~~	articles * past tense *-ed* * plural *-s* * progressive *-ing* * third person *-s* *	articles * ~~past tense *-ed* *~~ plural *-s* * ~~possessive *'s* *~~ ~~progressive *-ing* *~~ ~~third person *-s* *~~
2	articles	articles ~~past tense *-ed*~~ plural *-s* ~~possessive *'s*~~ ~~progressive *-ing*~~			articles plural *-s* ~~possessive *'s*~~ ~~progressive *-ing*~~	~~possessive *'s*~~	
3							
4							

B2

Clustered Order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	past tense *-ed* * plural *-s* * third person *-s* * progressive *-ing* *	third person *-s* *	plural *-s* * progressive *-ing* * third person *-s* *	articles * past tense *-ed* * plural *-s* * possessive *'s* progressive *-ing* * third person *-s* *	~~possessive *'s* *~~ ~~third person *-s* *~~	articles * plural *-s* * progressive *-ing* * third person *-s* *	articles * plural *-s* * progressive *-ing* * third person *-s* *
2	articles ~~possessive *'s*~~	past tense *-ed* plural *-s* ~~progressive *-ing*~~ ~~possessive *'s*~~	articles past tense *-ed* ~~possessive *'s*~~		~~past tense *-ed*~~ plural *-s* ~~progressive *-ing*~~	past tense *-ed*	past tense *-ed* possessive *'s*
3		articles			articles	possessive *'s*	
4							

B1

Clustered Order	L1 Japanese	L1 Korean	L1 Spanish	L1 Russian	L1 Turkish	L1 German	L1 French
1	progressive *-ing* * third person *-s* *	third person *-s* *	plural *-s* * third person *-s* *	past tense *-ed* * plural *-s* * progressive *-ing* * third person *-s* *	third person *-s* *	articles * plural *-s* * progressive *-ing* third person *-s* *	plural *-s* * progressive *-ing* third person *-s* *
2	past tense *-ed* plural *-s* possessive *'s*	~~possessive *'s*~~	articles progressive *-ing*	articles possessive *'s*	plural *-s* ~~possessive *'s*~~ progressive *-ing*	past tense *-ed*	articles
3	articles	~~past tense *-ed*~~ plural *-s* ~~progressive *-ing*~~	past tense *-ed*		articles past tense *-ed*	possessive *'s*	past tense *-ed* possessive *'s*
4		articles	possessive *'s*				

the L1 Korean group. In L1 Turkish learners, plural *-s* is somewhere in the middle at both levels, and in L1 Russian it is in the higher cluster at B1.

- Possessive *'s* is in the least accurate cluster in L1 Spanish, German, and French learners. In L1 Japanese, it is ranked in the middle at B1, and in L1 Russian, it is one of the two least accurate morphemes at B1. In L1 Korean and Turkish, possessive *'s* is considered unreliable at both levels so no observation or interpretation is given.

- In L1 Japanese, German, and French learners, progressive *-ing* is consistently in the highest cluster. It is in the high- to mid-rank in L1 Spanish, and in somewhere in the middle in L1 Turkish. In L1 Russian at B1, it is in the higher cluster.

- The TLU score of third person *-s* is consistently over 0.90 at $p < .05$ and thus in the highest cluster.

There are some notable differences from the findings in the last chapter. For example, third person *-s* was not always among the most accurate morpheme in the CLC. Also, progressive *-ing* was one of the least accurate morphemes in L1 German and L1 French learners of English in the CLC, whereas in EFCamDat it is one of the most accurate ones.

Comparing accuracy orders between the corpora. Table 17 summarizes the accuracy order differences between the CLC and EFCamDat. The corpus name in the parentheses refers to the corpus in which the preceding L1 groups marked higher accuracy ranks of the morpheme. For example, ST (CLC) in articles at B2 means that L1 Spanish and L1 Turkish learners of English marked a higher accuracy rank of articles at the B2 level in the CLC than in EFCamDat.

The table shows that accuracy ranks can differ especially in past tense *-ed* and third person *-s*, whereas in plural *-s*, possessive *'s*, and progressive *-ing*, the difference was not large enough to emerge by the comparison procedure. It is natural that there are large

Table 17

Accuracy Order Comparison Between the CLC and EFCamDat

	Articles	Past tense *-ed*	Plural *-s*	Possessive *'s*	Progressive *-ing*	Third person *-s*
B2	ST (CLC)	SGF (CLC)				KSF (EFCamDat)
B1	F (CLC)	JSTGF (CLC)				JKSF (EFCamDat)

Note. J = L1 Japanese; K = L1 Korean; S = L1 Spanish; R = L1 Russian; T = L1 Turkish; G = L1 German; F = L1 French

differences in accuracy rank in third person *-s* between the two corpora since, as noted earlier, its accuracy is extremely high in EFCamDat, much higher than in the CLC. Past tense *-ed* is of a lower accuracy order in EFCamDat possibly because third person *-s* in the highest cluster pushed down the rank of the other morphemes, and since past tense *-ed* tended to be in the highest cluster in the CLC, it was affected the most in the present analysis.

It is interesting to observe the absence of differences in possessive *'s*. As seen above, the accuracy of possessive *'s* is widely different between the CLC and EFCamDat, especially in the PRESENT group (L1 Japanese, Korean, Russian, Turkish, and French). This is possibly because the accuracy of possessive *'s* is generally low to start with, and even if the difference in absolute accuracy is observed between the corpora, it might not affect the resulting accuracy order. Similar is the case in progressive *-ing*. It is in the highest cluster in L1 German and French learners in EFCamDat but tended to be in lower half in the CLC. It can be because the cluster progressive *-ing* belongs to is relatively large, including three or four morphemes, in EFCamDat, and its wide coverage makes it difficult for a difference to emerge.

Summary of the accuracy order comparison. On the whole, the two corpora are generally comparable in terms of the accuracy order of each target morpheme. Although accuracy ranks widely differ between the two corpora in past tense *-ed* and third person *-s*, it is likely due to the unusually high accuracy of third person *-s* in EFCamDat mentioned

in the last section. Apart from it, despite considerable differences in the environments in which the essays were written, the accuracy order of the target morphemes in the two corpora is relatively similar.

3.3.4 Regression Modeling

Background of the analysis. The last two sections investigated whether the accuracy and the relative position of the morphemes differ between the CLC and EFCamDat. Although informative, the analyses did not directly quantify L1 influence and test its effect. The present section builds a regression model as in the last chapter and statistically tests whether the strength of L1 influence on each morpheme varies across the two learner corpora. A potential problem, however, is the nonlinearity of proficiency effect as Figure 5 and 6 suggest. In the last chapter, I entered a linear, a quadratic, and a cubic term of ExamLevel (or proficiency) to capture nonlinearity. However, the subcorpus used in the last chapter only had six proficiency levels whereas EFCamDat has 128. This can invite the necessity for higher polynomial terms (e.g., quartic or fifth power), which will make the model overly complex and difficult to interpret. It is also unclear how to decide the appropriate degree of polynomial terms. Thus, the present study employed a generalized additive model (GAM) instead of a logistic regression model used in the last chapter.

Description of GAMs. A GAM is a regression model in which the relationship between independent and dependent variables is estimated by a nonlinear function, that is, it is not necessary to prejudge a particular form to model the relationship. In the present study, it allows nonparametric smoothers to model the nonlinear relationship between proficiency and morpheme accuracy, thereby significantly relaxing the restriction on the relationship compared to the case where proficiency is entered as polynomial terms. At the same time, other predictors such as L1Type and Morpheme can be entered as parametric terms in much the same way as in the model in the last chapter. Thus, GAMs are

semi-parametric, allowing both parametric and nonparametric terms in one model. Although new in (second) language acquisition research, the technique and its applied form have been used in other areas of applied linguistics such as psycholinguistics (Baayen, Kuperman, & Bertram, 2010; Balling & Baayen, 2012) and sociolinguistics (Wieling, Nerbonne, & Baayen, 2011). The present study, as in the studies just cited, used the `mgcv` package (Wood, 2006) in R for the analysis. The nonlinear relationship was entered by using thin plate splines with generalized cross-validation to automatically determine the optimal number of smoothing parameters. Since there appear to be interactions between proficiency and other variables, nonlinear relationships were estimated for each level of the target factors, as explained below.

Model specification. The present analysis employed a GAM with a logit link function and quasibinomial distribution. In other words, it built an additive logistic regression model. Quasibinomial distribution was assumed because a fully parametric model that included the same dependent and independent variables as the GAM below and the linear, quadratic, and cubic terms of Proficiency as well as their two-way interactions suffered overdispersion (residual deviance = 14,612 on 4,243 d.f.) after backward elimination of non-significant predictors. Following variables were included in the model. As discussed above, since finer level classification is available in EFCamDat compared to the CLC, Unit was employed as the value of Proficiency. Recall that there are 16 Lessons $\times$ 8 Units = 128 Units in total in EFCamDat. Because finer division of proficiency inevitably diminishes the size of each data point, the threshold of 100 obligatory contexts to be included in the model, which was a condition to be included in the model in the last chapter, was not applied to the present modeling. The number of correct suppliance was entered as the number of successes, and the number of omission and overgeneralization errors as the number of failures. These two formed the dependent variable. L1Type (two levels; ABSENT vs PRESENT), L1 (seven levels; one for each L1 group), Morpheme (six levels; one for each

morpheme), and the two-way interaction between L1Type and Morpheme were entered as independent variables. L1-Morpheme interaction was not included because it completely captures the variance explained by the interaction between L1Type and Morpheme, which is an interest of the analysis (cf. last chapter). The interaction between L1 and L1Type was also not included because it represents whether the effect of L1 varies across L1Type and is difficult to interpret. As to nonparametric smoothers, the median of Proficiency after excluding missing values (n = 582; 10.9%) was 58, and the value was first subtracted from Proficiency in order to make the intercept maximally meaningful. The intercept, therefore, represents Lesson 7 Unit 2. The centered Proficiency, its two-way interaction with L1, and that with morpheme were entered as independent variables with thin plate splines. This means that separate nonlinear relationship between proficiency and morpheme accuracy was estimated for each L1 and morpheme. Dummy variables with treatment contrasts were used for all the categorical variables.

Interpreting parametric terms. Table 18 shows the summary of the GAM. First, let us look at the parametric coefficients. L1 and Morpheme as such are not very meaningful because they participate in the interaction with nonlinear effects of Proficiency, which means their effect varies nonlinearly across proficiency. Their interpretation is not attempted, therefore. L1Type is significant and, as expected, the PRESENT group generally achieve higher accuracy than the ABSENT group. The Morpheme-L1Type is also significant, which means the strength of L1 influence varies across morphemes. The interaction between past tense *-ed* and L1Type is estimated but does not make sense, since all the target L1s belong to the PRESENT group. The coefficient, therefore, is not interpreted. The order of sensitivity to L1 influence is, from the most sensitive morpheme to the least, articles, plural *-s*, progressive *-ing*, possessive *'s*, and third person *-s*. The order is, in fact, very similar to that observed in the last chapter. This will be visualized later.

Table 18

Summary of the Generalized Additive Model Fitted to the TLU Score in EFCamDat (n = 4,794)

Parametric terms					
Parameter		B	SE	*t*	*p*
Intercept		1.511	0.042	35.836	0.000
L1					
	French	-0.338	0.040	-8.500	0.000
	Japanese	0.051	0.066	0.767	0.443
	Korean	-0.372	0.073	-5.123	0.000
	Russian	0.284	0.041	6.966	0.000
	Spanish	-0.317	0.048	-6.629	0.000
	Turkish	-0.326	0.069	-4.730	0.000
Morpheme					
	Past tense *-ed*	0.000	0.000	NA	NA
	Progressive *-ing*	1.308	0.072	18.207	0.000
	Possessive ’s	0.195	0.127	1.531	0.126
	Plural *-s*	0.879	0.071	12.467	0.000
	Third person *-s*	3.182	0.281	11.308	0.000
L1 Type					
	PRESENT	0.890	0.039	22.991	0.000
Morpheme : L1 Type					
	Past tense *-ed* : PRESENT	-0.109	0.068	-1.594	0.111
	Progressive *-ing* : PRESENT	-0.755	0.114	-6.645	0.000
	Possessive ’s : PRESENT	-0.904	0.139	-6.488	0.000
	Plural *-s* : PRESENT	-0.368	0.075	-4.904	0.000
	Third person *-s* : PRESENT	-1.026	0.285	-3.604	0.000

Splines				
Spline	edf	Ref.df	*F*	*p*
spline proficiency : morpheme (articles)	8.052	8.665	4.908	0.000
spline proficiency : morpheme (past tense *-ed*)	6.454	7.460	9.512	0.000
spline proficiency : morpheme (progressive *-ing*)	4.326	5.299	4.550	0.000
spline proficiency : morpheme (possessive ’s)	7.876	8.560	5.826	0.000
spline proficiency : morpheme (plural *-s*)	7.165	8.064	11.018	0.000
spline proficiency : morpheme (third person *-s*)	4.562	5.648	9.080	0.000
spline proficiency : L1 (German)	1.001	1.001	0.185	0.667
spline proficiency : L1 (French)	1.000	1.001	0.165	0.684
spline proficiency : L1 (Japanese)	2.395	3.012	2.170	0.089
spline proficiency : L1 (Korean)	3.223	3.999	1.416	0.226
spline proficiency : L1 (Russian)	3.596	4.450	2.218	0.058
spline proficiency : L1 (Spanish)	1.981	2.482	0.971	0.385
spline proficiency : L1 (Turkish)	1.001	1.001	0.334	0.564

Interpreting nonparametric terms. The lower half of Table 18 shows estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F, and p-values for the splines. When the edf is one as in L1 German, L1 French, and L1 Turkish, the relationship between Proficiency and accuracy is linear in logit scale (Baayen, 2010), and the higher its value, the more knots were necessary to model the relationship. This is illustrated in Figure 8. The horizontal axis represents proficiency, and the vertical axis represents accuracy in each panel. The figure shows the varying nonlinear effect of Proficiency across Morpheme. Note that because centering constraints are applied to the smoothers, both axes are difficult to interpret, and the figure should only demonstrate that the effect of Proficiency is nonlinear and that it further varies across morphemes. Notice that morphemes with high edf values, such as possessive *'s*, have more unpredictable effect of Proficiency. Figure 9 displays the fitted, or predicted, TLU scores based on the GAM. The lines were drawn for two morphemes (progressive *-ing* and possessive *'s*) across two L1 groups (L1 German and L1 Russian) whose edf values are relatively polar ends. We can see that, although there are slight differences in the Proficiency effect across L1 groups in that the accuracy difference between them becomes larger as proficiency goes up, a much larger difference is observed between morphemes: Progressive *-ing* and possessive *'s* have a completely different effect of proficiency on accuracy development. Note, however, that the data of possessive *'s* are sparse compared to progressive *-ing*. This might partially account for the nonlinearity of the fitted values. However, since no evidence suggests the homogeneity of the effect of proficiency on accuracy across morphemes, it is reasonable to control for the effect with the assumption that the effect might vary.

Comparing L1 influence between the CLC and EFCamDat. The question now is whether there is any difference in L1 influence between the CLC and EFCamDat. To answer this, sensitivity of each morpheme to L1 influence across the two corpora is illustrated in Figure 10. The higher a morpheme is located, the more sensitive it is to L1 influence.

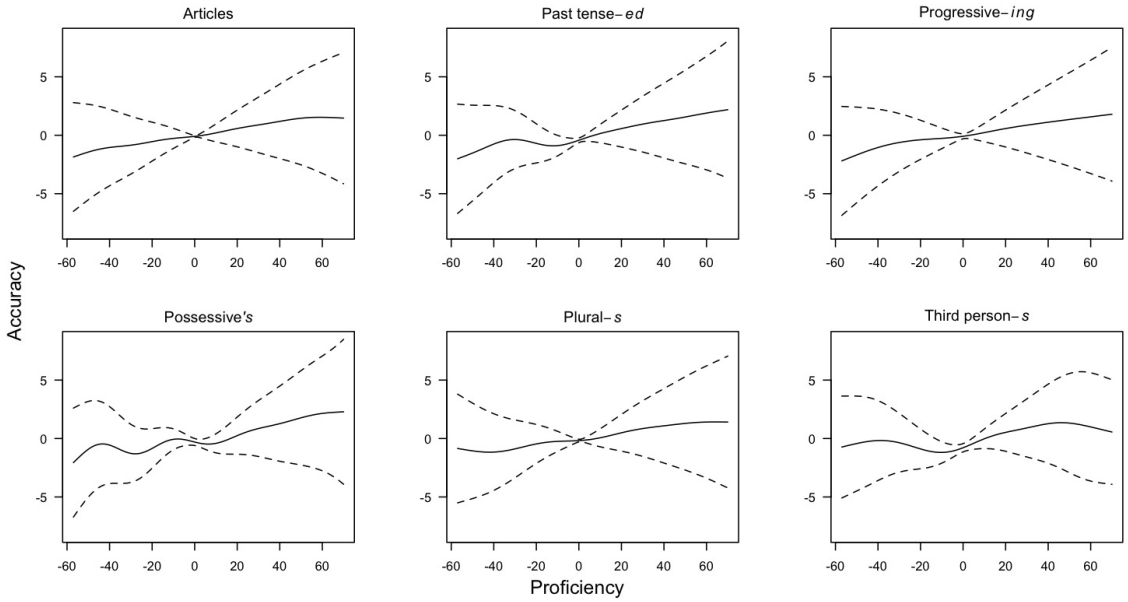


Figure 8. Nonlinear effect of proficiency across morphemes

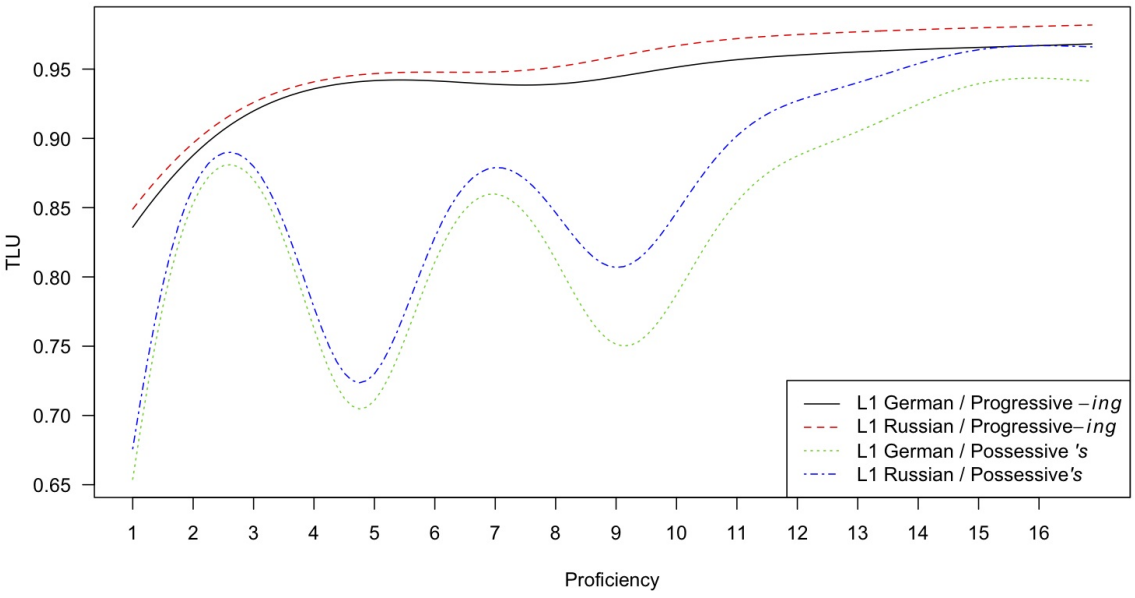


Figure 9. Fitted TLU Scores of Progressive -ing / Possessive 's by L1 German / L1 Russian Learners

Strength was computed by totaling the coefficients of L1Type and each morpheme. For instance, since the coefficient of L1Type is 0.881 and that of plural *-s* : PRESENT under the interaction between Morpheme and L1Type is -0.360 in EFCamDat, it is situated at 0.521 (0.881 - 0.360) on the figure. The value of 0.0 represents the absence of L1 influence, and the negative value as in the case of possessive *'s* and third person *-s* in EFCamDat indicates the reverse effect, that is, the ABSENT group scores higher than the PRESENT group. In the figure, L1 influence looks generally stronger in the CLC than in EFCamDat. This matches with the impression from Figure 5 and 6 that demonstrated the lack of L1 influence in some morphemes. Regarding the relative position of the morphemes, however, the similarity between the two corpora is clear, with a possible exception of progressive *-ing*. Apart from the progressive, the order of the morphemes as well as their relative distance to each other closely match between the corpora. The reason that progressive *-ing* was impervious to L1 influence in EFCamDat is unclear, but Figure 7 suggests that A2 is the only proficiency level where the distance between the ABSENT and the PRESENT group is large in the CLC. This might have caused the strong L1 influence in progressive *-ing* in the CLC. Except for the A2 level, the CLC and EFCamDat show a similar pattern with regard to the morpheme. Overall, it seems safe to claim that the regression model based on EFCamDat mostly replicated the findings in the last chapter.

Background of the aggregated model. The regression model just discussed only shows L1 influence in EFCamDat, and the comparison with the CLC was conducted manually. The question now is whether statistics confirms the similarity in the effect of L1 across the two corpora. In order to investigate it, the study constructed a logistic regression model with the data from both the CLC and EFCamDat, and checked the significance of the interaction between the corpora and L1Type. The more specific procedure is as follows. EFCamDat data used in the previous generalized additive model were aggregated over Proficiency levels corresponding to the CEFR (A1-C2) in order to match its Proficiency with

Figure 10. Strength of L1 Influence Across the Two Corpora

that in the CLC (operationalized by exam level). Since the number of levels in Proficiency is much smaller (6) than the previous analysis, GAMs were not employed. The data points whose obligatory contexts are fewer than 100 were excluded from the analysis. As in the previous model, the number of correct suppliance was entered as the number of successes and that of correct suppliance and overgeneralization as the number of failures. All the independent variables in the last model except for nonparametric smoothers were included in the present logistic regression model. In addition, Proficiency and the variable Corpus (two levels; CLC and EFCamDat) were also introduced. Proficiency was represented by CEFR levels and was B1-centered. Its linear, quadratic, and cubic terms were entered to capture nonlinearity. Two-way interactions among the independent variables, including the quadratic and cubic terms of Proficiency, were entered. However, the interactions among Proficiency itself (e.g., the interaction between Proficiency and Proficiency^2) were excluded due to the difficulty of interpretation. The interaction between L1 and Morpheme and that between L1 and L1 Type were not included for the same reason as in the previous

model. A three-way interaction between Corpus, L1Type, and Morpheme was entered into the model. The three-way interaction is the focus of the analysis because it tests whether the varying L1 strength over morphemes differs between the corpora. In other words, the term tests if the distance between the morphemes in Figure 10 varies across the corpora. Since the model suffered overdispersion (residual deviance = 2,732; df = 291), a quasibinomial distribution was assumed. Backward elimination was applied until the model became significantly worse at $p < .05$. Dummy variables with treatment contrasts were used for categorical variables.

Result of the aggregated model. Table 19 is the summary of the aggregated regression model. As expected, Proficiency and its quadratic and cubic terms are engaged in complex interactions with other variables. Due to the complexity, the substantial interpretation will not be attempted. Instead, however, the effect of Proficiency is visualized in Figure 11. As an example, the figure plots the fitted, or predicted, TLU scores of articles and third person -*s* in L1Russian and L1 Spanish learners in the two corpora across the five Proficiency levels at which EFCamDat data are rich. The L1 Russian third person -*s* data were plotted only at B1 through C1 in the CLC panel because A1 and A2 each included fewer than 100 obligatory contexts and the predicted values would have been less reliable. The morphemes and the L1 groups whose coefficients for Proficiency^3 are polar ends were selected. The figure shows that (i) Proficiency effect is nonlinear, (ii) the curvature varies across morphemes, L1 groups, and corpora, and (iii) the degree of linearity seems to be related to the ceiling effect in EFCamDat.

Let us now interpret other variables.

- As to L1, at the intercept proficiency level (B1) in the CLC (reference-level corpus), L1 German learners mark higher accuracy than L1 French, Japanese, Russian, Spanish, and Turkish learners. This partially reconfirms a finding based on the CLC that

Table 19

Summary of the Aggregated Logistic Regression Model Fitted to the TLU Score in the CLC and EFCamDat (n = 348)

Parameter		B	SE
Intercept		1.334 ***	0.094
L1			
	French	-0.545 ***	0.084
	Japanese	-0.305 **	0.109
	Korean	-0.192	0.118
	Russian	-0.457 ***	0.097
	Spanish	-0.448 ***	0.079
	Turkish	-0.442 ***	0.113
Proficiency		0.155 *	0.060
Morpheme			
	Past tense *-ed*	-0.186	0.100
	Progressive *-ing*	0.830 ***	0.140
	Possessive 's	-0.139	0.190
	Plural *-s*	0.641 ***	0.080
	Third person *-s*	0.268	0.155
L1 Type			
	PRESENT	1.208 ***	0.066
$Proficiency^2$		-0.014	0.047
$Proficiency^3$		0.009	0.012
Corpus			
	EFCamDat	0.178	0.115
L1 : Proficiency			
	French : Proficiency	0.032	0.038
	Japanese : Proficiency	-0.011	0.044
	Korean : Proficiency	-0.085	0.048
	Russian : Proficiency	0.127 ***	0.034
	Spanish : Proficiency	0.044	0.036
	Turkish : Proficiency	0.020	0.050
L1 : Corpus			
	French : EFCamDat	0.201	0.117
	Japanese : EFCamDat	0.268	0.153
	Korean : EFCamDat	-0.108	0.175
	Russian : EFCamDat	0.740 ***	0.120
	Spanish : EFCamDat	0.119	0.115
	Turkish : EFCamDat	0.219	0.172
L1Type : Morpheme			
	PRESENT : Past tense *-ed*	NA	NA
	PRESENT : Progressive *-ing*	-0.363 **	0.134
	PRESENT : Possessive 's	-1.009 ***	0.159
	PRESENT : Plural *-s*	-0.490 ***	0.068
	PRESENT : Third person *-s*	-0.960 ***	0.124
Proficiency : Morpheme			
	Proficiency : Past tense *-ed*	-0.040	0.080
	Proficiency : Progressive *-ing*	0.163	0.095
	Proficiency : Possessive 's	-0.039	0.119
	Proficiency : Plural *-s*	0.037	0.044
	Proficiency : Third person *-s*	-0.025	0.111
$Proficiency^2$: Morpheme			
	$Proficiency^2$: Past tense *-ed*	0.069	0.061
	$Proficiency^2$: Progressive *-ing*	-0.056	0.041
	$Proficiency^2$: Possessive 's	0.038	0.047
	$Proficiency^2$: Plural *-s*	0.093 ***	0.018
	$Proficiency^2$: Third person *-s*	0.258 ***	0.070
$Proficiency^3$: Morpheme			
	$Proficiency^3$: Past tense *-ed*	-0.013	0.021
	$Proficiency^3$: Progressive *-ing*	-0.005	0.018
	$Proficiency^3$: Possessive 's	-0.005	0.021
	$Proficiency^3$: Plural *-s*	-0.029 ***	0.008
	$Proficiency^3$: Third person *-s*	-0.065 **	0.023
Morpheme : Corpus			
	Past tense *-ed* : EFCamDat	-0.007	0.130
	Progressive *-ing* : EFCamDat	0.614 ***	0.170
	Possessive 's : EFCamDat	0.210	0.207
	Plural *-s* : EFCamDat	0.247 **	0.082
	Third person *-s* : EFCamDat	2.404 ***	0.211
L1 Type : Corpus			
	PRESENT : EFCamDat	-0.345 ***	0.080
Corpus : Proficiency		0.154 **	0.059
Corpus : $Proficiency^2$		0.017	0.048
Corpus : $Proficiency^3$		-0.030 *	0.014

Note. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

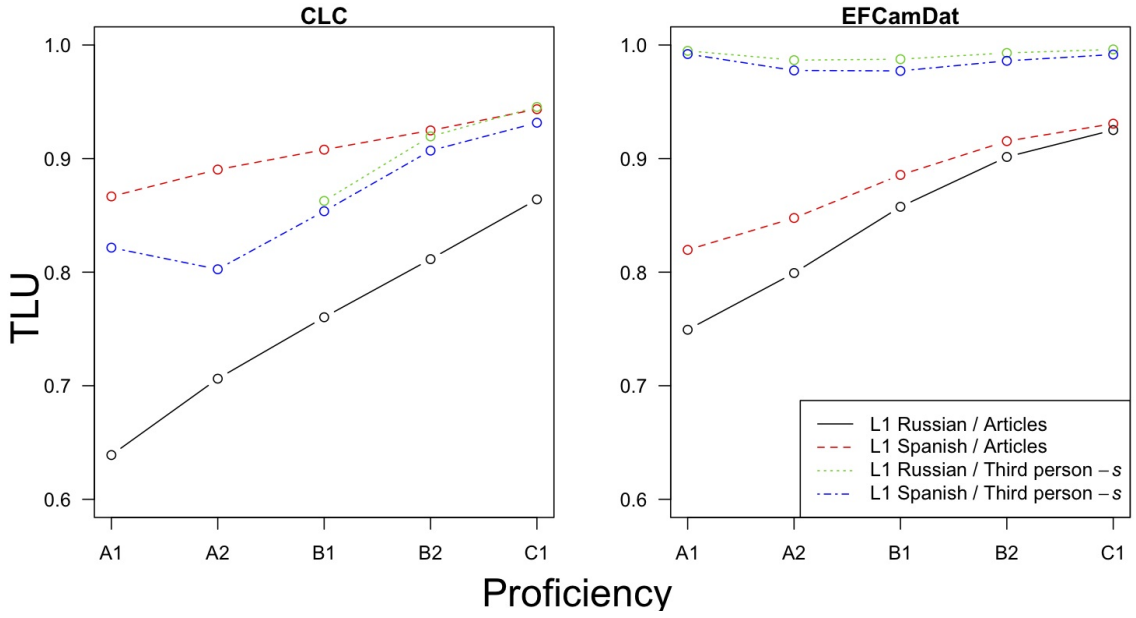


Figure 11. Fitted TLU Scores of Articles and Third Person -s by L1 Russian and L1 Spanish Learners in EFCamDat

L1 German learners are more accurate in morpheme use in general than L1 French and L1 Turkish learners. The interaction between L1 and Corpus is also significant and shows that L1 Russian learners achieve higher accuracy in EFCamDat than in the CLC.

- With respect to Morpheme,
  - Progressive *-ing* and plural *-s* are more accurate than articles in the ABSENT group (reference-level L1 type) in the CLC. This means, for example, that the use of progressive *-ing* by the ABSENT group is more accurate than the use of articles by the ABSENT group, although the specific L1 groups in the ABSENT group vary across morphemes. Morpheme interacts with L1Type, which means that the strength of L1 influence varies across morphemes.
  - When the CLC and EFCamDat are combined, the order of the strength of L1

influence is as follows; articles, progressive *-ing*, plural *-s*, third person *-s*, and possessive *'s*. The absence, or non-significance, of the three-way interaction between Corpus, Morpheme, and L1Type indicates that there is no evidence showing that the relative position of the morphemes along the strength of L1 influence varies across the corpora (cf. Figure 10).

  - The significant Morpheme-Corpus interaction shows that the relative accuracy between morphemes varies across the corpora. This is expected because, for instance, as seen earlier, the accuracy and the rank of third person *-s* is much higher in EFCamDat than in the CLC. In addition, progressive *-ing* and plural *-s* are of higher accuracy in EFCamDat than in the CLC.

- L1Type is significant, and the PRESENT group marks higher accuracy than the ABSENT group in articles (reference-level morpheme) in the CLC. In addition to the interaction with Morpheme discussed above, L1Type also interacts with Corpus, and the interaction shows that the strength of L1 influence is overall weaker in EFCamDat than in the CLC. This confirms the impression from Figure 10, which shows L1 influence is generally stronger in the CLC. It is also interesting to note the absence of the interaction with Proficiency, which means that the strength of L1 influence does not necessarily diminish as proficiency goes up.

This part of the analysis, therefore, shows that, although the overall strength of L1 influence could vary across the two learner corpora, there was no evidence that shows that the relative position in the strength of L1 influence between morphemes differs across the corpora. This supports the robustness of the finding in the last chapter on the order of sensitivity to L1 influence (e.g., articles are more sensitive to L1 influence than possessive *'s*).

3.4 Discussion

The present study asked whether the findings based on the CLC are robust in terms of (i) the absolute accuracy, (ii) the accuracy order, and (iii) the strength of L1 influence in L2 English grammatical morphemes.

Absolute accuracy in the CLC and EFCamDat. With respect to i, the difference in the absolute accuracy is similar across the two learner corpora at all the proficiency levels in articles, past tense *-ed*, and plural *-s*. In the other three morphemes, there was some apparent difference between the two at least at one of the targeted proficiency levels.

Accuracy order in the CLC and EFCamDat. In response to ii, plural *-s*, possessive *'s*, and progressive-*ing* did not show any difference in its order of accuracy between the CLC and EFCamDat, and the other three morphemes (articles, past tense *-ed*, and third person *-s*) showed some differences. I speculate, however, that the only morpheme worthy of attention is third person *-s*. The reasons follow. First, as reviewed above, articles and past tense *-ed* have similar accuracy across the two corpora. Second, since the accuracy rank of third person *-s* significantly rose due to the large increase of absolute accuracy discussed above, it perhaps pushed down the rank of other morphemes. Indeed, articles and past tense *-ed* are ranked at a higher position in the CLC in all the instances where the differences emerged, and third person *-s* is ranked higher in EFCamDat consistently. Therefore, although superficially there seem to be quite a few differences in morpheme ranks between the two corpora, the only notable point here is third person *-s*.

Strength of L1 influence in the CLC and EFCamDat. With regard to iii on the strength of L1 influence, regression models point toward a weaker L1 influence in EFCamDat than in the CLC. The reason for this is unclear, but a speculative cause is the difference in the validity of the variable L1 in the two analyses. As noted earlier, EFCamDat does not include direct information on learners' L1 and it was inferred from their country of resi-

dence. However, learners living in France, for instance, may have other L1s than French. The cases like this might have mitigated or confounded what I called L1 influence in this chapter. The relative strength of L1 influence between morphemes, however, turned out to be similar between the two learner corpora, with a possible exception of progressive *-ing*. This makes the finding in the last chapter robust and provides a further support to the thinking-for-speaking hypothesis.

Replicability of the findings in the last chapter. Major findings in the last chapter are mostly replicable in EFCamDat. More precisely, the present study demonstrated that the accuracy order of morphemes differs between L1 groups and that the strength of L1 influence varies across morphemes. Further, both studies showed that L1 influence is strong on articles, followed by plural *-s*, which in turn is followed by possessive *'s* and third person *-s*. Together with the support by the thinking-for-speaking hypothesis, this order of sensitivity to L1 influence is robust and I hypothesize that it holds in other data sources as well.

This is encouraging because the CLC and EFCamDat vary widely in terms of (i) typical essay length, (ii) condition of writing (e.g., whether there is a time limit and whether consultation to external materials such as example essays and dictionaries is permitted), (iii) the scope of target structures (Englishtown essays can aim at particular linguistic structures whereas Cambridge ESOL exams perhaps try to elicit a range of expressions), (iv) learners' involvement (they are perhaps more involved in Cambridge exams because they pay for the exam), (v) explicit instruction (learners receive explicit instruction, although not necessarily on grammatical morphemes, before writing essays in EFCamDat), and (vi) tasks. The fact that the findings in the last chapter were largely replicated despite all these differences indicates the robustness of the findings.

Given that learners in EFCamDat had access to example essays as well as the prompts, it would be interesting to analyze whether and the extent to which learners copy and paste

them in writing their essays. It is possible, for example, that lower proficiency learners use the same lexical items and syntactic structures as the example essays more often than higher proficiency learners. It is an empirical issue to test whether we can identify such a class of learners who tend to or tend not to take advantage of the given sentences, but is beyond the scope of the present study.

An interest that arises from the cross-sectional analysis is how the longitudinal development of learners maps to the cross-sectional, or pseudo-longitudinal, development shown in Figure 5 and 6. This is what I will turn to in the next chapter.

3.5 Conclusion

The present study investigated the similarities and differences between the Cambridge Learner Corpus and EF-Cambridge Open Language Database. Overall, the two learner corpora produced similar findings with regard to the accuracy of grammatical morphemes, their accuracy order, and the relative positioning of morphemes in terms of the strength of L1 influence. This is encouraging since it shows that the findings based on the CLC are robust. There were a few parts where the data from the two corpora look different. Notable are the extremely high accuracy of third person *-s* and a much weaker L1 influence on progressive *-ing* in EFCamDat.

Chapter 4: Individual Variation in the Longitudinal L2 Development of English Grammatical Morphemes

4.1 Introduction

While the last two chapters focused on L1 influence on morpheme accuracy, the present chapter focuses on individual variation in the development of morphemes. Typical cross-sectional and pseudo-longitudinal analyses do not allow to disentangle inter-learner variation from intra-learner variability. Therefore, the present study employs a longitudinal learner corpus to investigate the issue. There have been few attempts to investigate individual variation based on large-scale quantitative empirical data such as learner corpora. The present chapter and the following two chapters address this gap. More specifically, the study aims to separate systematic development and individual variation.

As to the development within individual learners, two hypotheses can be formed. First, DeKeyser's (2007) skill acquisition theory, based on Anderson's (e.g., J. R. Anderson, 1983) ACT model, predicts that the accuracy development follows the so-called power law of learning (DeKeyser, 2001; Dekeyser, 1997). A power function is a function that takes the form of $y = x^n$. In the case of accuracy development, y corresponds to accuracy, x to the amount of practice, and n to a constant. The theory, therefore, predicts that the accuracy rapidly increases initially, and the degree of accuracy increase gradually decreases as learners progress. Indeed, Dekeyser (1997) found that the development of reaction time and error rate follows a power law by targeting the grammatical morphemes of an artificial language that encode gender and instrumentality. Thus, for the present study, it is hypothesized that learners tend to follow a power law in their accuracy development of grammatical morphemes.

Second, SLA literature has documented that the development of certain features, such as *-ing* (Lightbown, 1983) or codas (Abrahamsson, 2003), follows the shape of U, where

the accuracy is high at the beginning, then decreases, and finally increases again. These studies point towards the possibility that the L2 development of (some of) the morphemes may follow U-shaped development. Note that these two hypotheses do not contradict. It is possible that some morphemes follow a power function, while others follow U-shaped development.

The following research questions are addressed:

1. Does the longitudinal transition of morpheme accuracy within individual learners show systematic patterns such as a power function or U-shaped development, or does the accuracy randomly fluctuate?

2. What is the extent of intra- and inter-learner variability observed in the development?

4.2 Method

4.2.1 Corpus

As in the last chapter, EFCamDat was employed. Table 20 demonstrates that the number of essays each learner wrote varies widely in the corpus. The left half of Table 20 shows that 69.9% of the learners wrote five or fewer essays in total, 15.3% of all the learners wrote 6-10 essays, 6.3% wrote 11-15 essays, and so forth. The median number of essays a learner wrote was three. In the corpus, the feedback is available for 154,110 essays, or 36.4% of all the essays. The right half of Table 20 displays the number of error-annotated essays per learner. There are in total 51,919 learners who have at least one of the essays error-tagged in the corpus. Out of them, 86.1% have five or fewer essays error-annotated. The median number of essays that have been error-tagged in the corpus per learner, excluding those who had none of their essays error-tagged, was two.

Table 20

Distribution of Learners According to the Number of Essays Written

# of essays	All Essays		Error-Tagged Essays	
	# of learners	%	# of learners	%
1-5	53,149	69.9%	44,684	86.1%
6-10	11,660	15.3%	5,338	10.3%
11-15	4,791	6.3%	1,377	2.7%
16-20	2,660	3.5%	366	0.7%
21-25	1,463	1.9%	102	0.2%
26 or more	2,279	3.0%	52	0.1%
Total	76,002	100.0%	51,919	100.0%

4.2.2 Target Morpheme, L1 Groups, and Proficiency Levels

Target morpheme. As in the previous chapters, the present study targets the following six grammatical morphemes; articles, past tense *-ed*, plural *-s*, possessive *'s*, progressive *-ing*, and third person *-s*.

Target L1 groups. The L1 groups investigated in previous chapters had small-sized data for the investigation of this chapter. Thus, further L1 groups were analyzed. The present study targeted Brazilian, Chinese, German, French, Italian, Japanese, Korean, Mexican, Russian, Spanish, Turkish, and Taiwanese learners of English. They are the learners in the countries of residence with the largest amount of data in EFCamDat. Their respective L1s were assumed to be Brazilian-Portuguese, Mandarin Chinese, German, French, Italian, Japanese, Korean, Spanish, Russian, Spanish, Turkish, and Mandarin Chinese. L1 Brazilian-Portuguese and L1 Mandarin-Chinese are referred to as L1 Brazilian and L1 Chinese hereafter in order to save space.

As in the previous chapters, the ABSENT group refers to those with the L1s that lack the corresponding morpheme, and the PRESENT group refers to those with the L1s that have the corresponding morpheme. Table 21 shows the dichotomous coding of L1 type employed in the present study. The Lesson and Unit number at which the learner wrote the essay was considered to represent proficiency of the learner.

Corpus size. Table 22 shows the number of error-tagged essays, that of total words, and the average length of an essay in each L1 and proficiency level. The subcorpus used in the study included nearly 140,000 essays consisting of 10 million words. There tended to be more essays, and thus a larger number of words, in lower proficiency levels, presumably because learners tended to start at lower levels and then dropped out or the data collection period for the current batch of data ended while they were progressing from lower to higher levels. The distribution between L1 groups was highly skewed as well. Over 40% of the data were contributed by L1 Chinese learners of English, and another large portion (14%-23% each) by L1 Brazilian and L1 Russian learners. This, however, should not affect our analysis because the data were not aggregated over L1 groups for most part of the analysis. As expected, the average number of words per essay shows an upward trend as proficiency rises.

4.2.3 Scoring Method and Data Extraction

The study adopted TLU scores and the same data extraction method as Chapter 3.

4.2.4 Data Analysis

I grouped up learners with similar developmental trajectories and studied the patterns. It answers Research Question 1 asking whether there are systematic patterns of accuracy development. Through the process, I addressed Research Question 2 examining the extent of individual variation.

Table 21

L1 Type over Each L1 and Morpheme and the References Supporting the Decision

L1	Articles	Past tense *-ed*	Plural *-s*	Possessive *'s*	Progressive *-ing*	Third person *-s*
L1 Brazilian	1	1	1	0	1	1
	Pr	Pr	Pr	Pp	KS	Pr
L1 Chinese	0[a]	0	0	1	0	0
	LS	DL	LS	LS	CLS	Hs
L1 German	1	1	1	1	0	1
	Sl	Gr	K	Li	Sl	Gr
L1 French	1	1	1	0	0	1
	BH	BH	BH	Hw	Sl	BH
L1 Italian	1	1	1	0	1	1
	PC	PC	PC	PC	PC	PC
L1 Japanese	0	1	0	1	1	0
	LS	Sh-a	LS	LS	Sh-a	J
L1 Korean	0	1	0	1	1	0
	LS	Le	LS	LS	Sh-b	C
L1 Russian	0	1	1	1	0	1
	I6	GC	M	M	I8	AL
L1 Spanish	1	1	1	0	1	1
	LS	Sa	LS	LS	Sl	Hr
L1 Turkish	0	1	1	1	1	0
	Sl	SA	Gö	GK	Sl	E

Note. [a]This was the initial setting. See the Results section for details.; One denotes that the morpheme is obligatorily marked in the language. Otherwise 0 is given.; AL = Ambridge and Lieven (2011); BH = Battye and Hintze (1992); C = Choi (2005); CLS = Cheung, Liu, and Shih (1994); DL = Duff and Li (2002); E = Ekmekci (1982); GC = Gor and Chernigovskaya (2004); GK = Göksel and Kerslake (2005); Gö = Görgülü (2005); Gr = Graves (1990); Hr = Harvey (2006); Hs = Hsin (2002); Hw = R. Hawkins (1981); I6 = Ionin (2006); I8 = Ionin (2008); J = Jelinek (1984); K = Köpcke (1988); KS = King and Suñer (1980); Le = Lee (2006); Li = Lindauer (1998); LS = Luk and Shirai (2009); M = Müller (2004); PC = Proudfoot and Cardo (2005); Pp = Papadopoulou (2006); Pr = Parkinson (1988); Sa = Salaberry (2002); SA = Slobin and Aksu (1982); Sh-a = Shirai (1998a); Sh-b = Shirai (1998b); Sl = Slobin (1996)

Table 22

The Number of Error-Tagged Essays and the Number of Total Words in Each L1 and Proficiency Level

	Proficiency																	
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total	%
Brazilian																		
# Essays	8,234	3,476	2,344	4,264	1,818	950	1,917	657	457	657	244	107	151	29	8	12	25,325	18.1%
# Words	276,870	154,309	103,954	288,599	124,989	64,912	184,477	60,113	47,996	83,349	32,866	13,622	25,267	4,765	1,252	2,013	1,469,353	14.3%
# Words / Essay	33.6	44.4	44.3	67.7	68.8	68.3	96.2	91.5	105.0	126.9	134.7	127.3	167.3	164.3	156.5	167.8	58.0	
Chinese																		
# Essays	10,093	4,889	8,693	15,265	7,220	2,686	6,904	2,500	1,333	2,189	486	178	124	22	6	2	62,590	44.8%
# Words	350,418	256,916	446,631	1,154,411	538,668	207,130	725,495	234,099	137,320	289,545	69,186	24,001	20,390	3,616	1,028	352	4,459,206	43.5%
# Words / Essay	34.7	52.5	51.4	75.6	74.6	77.1	105.1	93.6	103.0	132.3	142.4	134.8	164.4	164.4	171.3	176.0	71.2	
German																		
# Essays	1,184	575	442	1,156	529	327	731	262	242	473	174	123	155	52	24	14	6,463	4.6%
# Words	49,087	31,948	21,616	79,898	37,785	23,086	69,963	24,317	25,134	60,708	23,997	16,357	27,553	8,950	3,747	2,441	506,587	4.9%
# Words / Essay	41.5	55.6	48.9	69.1	71.4	70.6	95.7	92.8	103.9	128.3	137.9	133.0	177.8	172.1	156.1	174.4	78.4	
French																		
# Essays	1,294	447	324	879	327	238	545	237	243	365	160	89	57	11	8	5	5,229	3.7%
# Words	48,924	22,560	15,389	59,100	22,685	17,631	52,130	21,610	25,581	46,502	21,839	12,576	8,987	1,751	1,226	694	379,185	3.7%
# Words / Essay	37.8	50.5	47.5	67.2	69.4	74.1	95.7	91.2	105.3	127.4	136.5	141.3	157.7	159.2	153.3	138.8	72.5	
Italian																		
# Essays	984	463	597	1,379	742	614	813	268	232	294	127	60	53	16	1	18	6,661	4.8%
# Words	33,576	22,220	25,485	91,766	48,734	41,409	77,552	23,666	22,881	37,435	17,562	8,977	9,353	2,475	161	2,696	465,948	4.5%
# Words / Essay	34.1	48.0	42.7	66.5	65.7	67.4	95.4	88.3	98.6	127.3	138.3	149.6	176.5	154.7	161.0	149.8	70.0	
Japanese																		
# Essays	452	178	134	299	193	142	208	119	133	122	81	32	23	11	5	0	2,132	1.5%
# Words	16,397	8,782	5,455	18,289	11,934	8,616	18,770	9,540	12,130	14,474	9,387	3,690	3,525	1,604	863	0	143,456	1.4%
# Words / Essay	36.3	49.3	40.7	61.2	61.8	60.7	90.2	80.2	91.2	118.6	115.9	115.3	153.3	145.8	172.6	NA	67.3	
Korean																		
# Essays	190	71	53	153	89	72	108	22	40	52	28	18	15	3	5	2	921	0.7%
# Words	7,601	3,465	2,147	9,619	5,775	4,620	10,310	1,762	4,217	5,914	3,982	2,353	2,254	454	690	318	65,481	0.6%
# Words / Essay	40.0	48.8	40.5	62.9	64.9	64.2	95.5	80.1	105.4	113.7	142.2	130.7	150.3	151.3	138.0	159.0	71.1	
Russian																		
# Essays	2,418	1,220	1,111	1,805	1,217	1,406	1,983	1,973	2,496	2,789	1,953	1,139	640	311	146	76	22,683	16.2%
# Words	91,517	63,953	54,199	132,755	89,481	111,239	204,749	204,559	277,977	380,614	283,836	162,398	116,756	57,338	29,606	13,019	2,273,996	22.2%
# Words / Essay	37.8	52.4	48.8	73.5	73.5	79.1	103.3	103.7	111.4	136.5	145.3	142.6	182.4	184.4	202.8	171.3	100.3	
Spanish																		
# Essays	2,012	732	563	921	446	261	447	236	204	205	116	83	49	13	2	0	6,290	4.5%
# Words	69,884	35,646	26,381	66,248	31,876	19,734	44,265	23,222	21,411	25,605	16,068	11,850	8,455	2,276	321	0	403,242	3.9%
# Words / Essay	34.7	48.7	46.9	71.9	71.5	75.6	99.0	98.4	105.0	124.9	138.5	142.8	172.6	175.1	160.5	NA	64.1	
Turkish																		
# Essays	438	185	106	270	96	57	134	29	22	75	9	4	11	4	0	1	1,441	1.0%
# Words	15,094	9,065	5,293	17,858	6,651	4,416	12,859	2,832	2,491	8,670	1,198	479	1,855	657	0	189	89,607	0.9%
# Words / Essay	34.5	49.0	49.9	66.1	69.3	77.5	96.0	97.7	113.2	115.6	133.1	119.8	168.6	164.3	NA	189.0	62.2	
Total																		
# Essays	27,299	12,236	14,367	26,391	12,677	6,753	13,790	6,303	5,402	7,221	3,378	1,833	1,278	472	205	130	139,735	100.0%
Essays (%)	19.5%	8.8%	10.3%	18.9%	9.1%	4.8%	9.9%	4.5%	3.9%	5.2%	2.4%	1.3%	0.9%	0.3%	0.1%	0.1%	100.0%	
# Words	959,368	608,864	706,550	1,918,543	918,578	502,793	1,400,570	605,720	577,138	952,816	479,921	256,303	224,395	83,886	38,894	21,722	10,256,061	100.0%
Words (%)	9.4%	5.9%	6.9%	18.7%	9.0%	4.9%	13.7%	5.9%	5.6%	9.3%	4.7%	2.5%	2.2%	0.8%	0.4%	0.2%	100.0%	
# Words / Essay	35.1	49.8	49.2	72.7	72.5	74.5	101.6	96.1	106.8	132.0	142.1	139.8	175.6	177.7	189.7	167.1	73.4	

4.3 Results

4.3.1 Cross-Sectional View of Article Development

Appendix B presents the TLU scores and the number of obligatory contexts of each morpheme, L1, and proficiency level. Because the table is large, no interpretation is attempted. Instead, the data will be visualized as we progress.

To date, very few, if any, studies have investigated individual variation in longitudinal L2 development based on large-scale data (cf. Myles, 2008). The present study therefore first had to establish a method of analyzing the longitudinal development of a large number of learners. For this purpose, I explored the developmental patterns of articles to grasp the nature of the data and decide the techniques used to analyze the other morphemes. I chose articles because (i) they are frequent so the data size was large and (ii) the accuracy of the script to retrieve article errors is high and includes less noise in the data. Thus article data are rich and clean, and whatever technique considered appropriate in their analyses can be subsequently applied to the analyses of the other morphemes.

Micro and macro averages. Figure 12 demonstrates the micro (left panel) and macro (right panel) averages of the TLU scores of articles between L1 type over proficiency. Recall that micro averages are the averages calculated at the essay level and macro averages are those calculated at the level of each L1 group. Proficiency here refers to the Lesson number at Englishtown. All the data points of the micro averages included more than 100 obligatory contexts so are reliable. Notice that the difference between the ABSENT and the PRESENT group differs across the micro and the macro averages. L1 influence does not look very pronounced in the micro averages but is clear in the macro averages. This implies that there is one (or a few) L1 group that exerts a strong influence on the micro averages. Notice further that the ABSENT group is the one whose TLU score is different between the micro and the macro averages. Considering that L1 Chinese is the only group

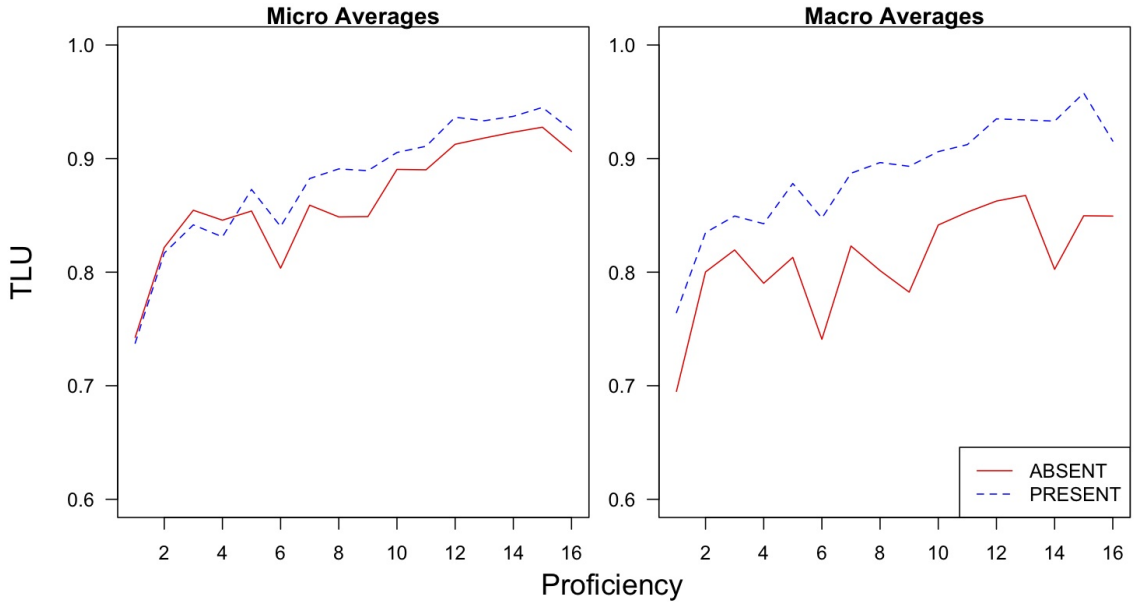


Figure 12. Pseudo-Longitudinal Development of Article Accuracy (Binary Coding of L1 Type)

that was newly added in the present chapter and belongs to the ABSENT group, and that they are the largest group and thus have a strong influence on micro averages, it is likely that L1 Chinese learners are behaving somewhat differently from the other ABSENT groups. Indeed, there are studies reporting that Mandarin-Chinese has linguistic features that play similar roles to English definite (Huang, 1999) and indefinite (Chen, 2004) articles, and it might be better not to classify them as the ABSENT group. Therefore, for the rest of the analysis, as far as articles are concerned, L1 Chinese learners constitute yet another L1 type, L1 Chinese.

Figure 13 visualizes the micro and macro averages when learners were divided into three groups, ABSENT, PRESENT, and L1 Chinese. The numbers on the graph of micro averages indicate the number of obligatory contexts. They are plotted only when the number is smaller than 100. Overall, the data are large enough to be reliable. This time, the difference between the micro and the macro averages is smaller, especially at lower

proficiency levels. Although there are still some differences in the absolute accuracy of the ABSENT group, it is perhaps because some L1 groups such as L1 Korean mark a low TLU score particularly at higher levels (see Appendix B) and they have larger effects on micro averages than on macro averages. Most of these TLU scores are based on few obligatory contexts, and they may not be reliable. We thus focus on micro averages here to abstract away from language-specific effects and focus on language type. We can make a few observations. First, although not large, L1 influence is clear and the PRESENT group consistently outperforms the ABSENT group. Second, the overall accuracy generally increases as proficiency rises. The development of L1 Chinese learners is flatter compared to the other two groups but will not be our focus here. Finally, accuracy does not increase linearly. Part of the bumps may be due to teaching materials in Englishtown, such as the timing articles are explicitly instructed. The pattern is as expected based on both the CLC study (cf. Figure 2) and the pseudo-longitudinal analysis of EFCamDat in Chapter 3 (cf. Figure 8). They both demonstrate that the accuracy development of morphemes is not necessarily linear, and Figure 13 supports the view as well. One of the central interests of the present study is to examine how this pseudo-longitudinal view of the data matches with the longitudinal view, to which I will turn now.

4.3.2 Longitudinal View of Article Development

4.3.2.1 Moving Window

Aim. Ideally, we should compute a TLU score for each essay and study the transition of the TLU scores within each learner. However, because each essay was relatively short, a single essay does not yield enough obligatory contexts to reliably calculate TLU scores. Therefore, TLU scores were computed over multiple essays written by a learner. The purpose here is to construct windows so that the shape of accuracy development is as close as the shape that would be generated if each essay included a large number of obligatory

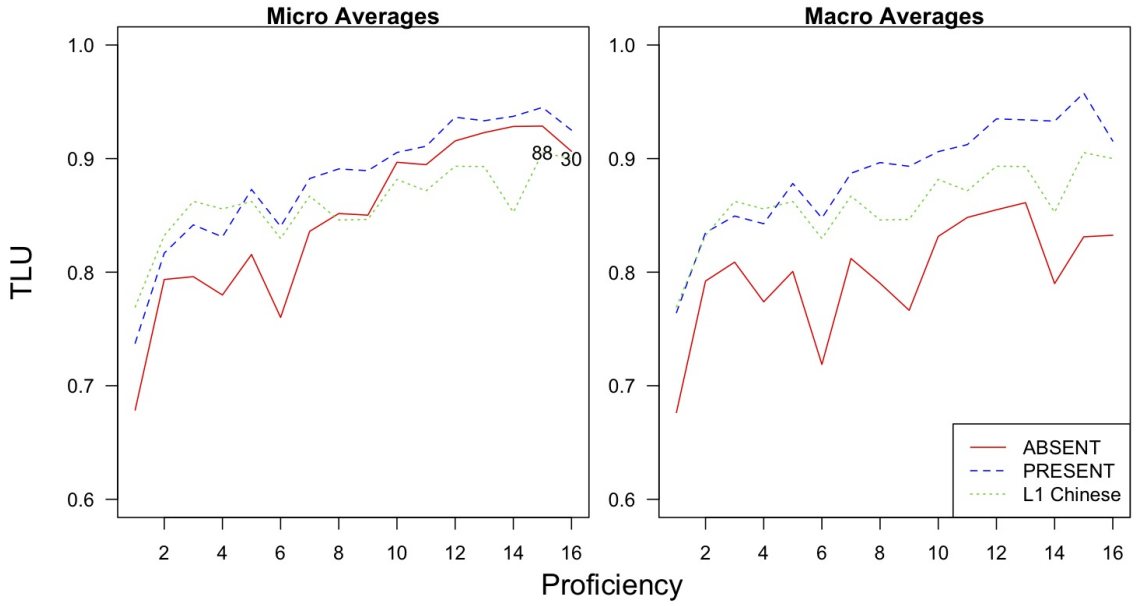


Figure 13. Pseudo-Longitudinal Development of Article Accuracy (Ternary Coding of L1 Type)

contexts.

Construction of windows. The error-tagged essays were first chronologically ordered for each learner based on their submission dates. TLU scores were then calculated in a moving-window fashion (cf. Spoelman & Verspoor, 2010; Stoll & Gries, 2009; van Geert & van Dijk, 2002; Verspoor et al., 2008). Each window included at least 15 obligatory contexts (OCs) and could cover multiple essays. For instance, let us say that a learner wrote five essays and the number of OCs each essay includes is the following;

Essay 1; 8 OCs

Essay 2; 8 OCs

Essay 3; 6 OCs

Essay 4; 9 OCs

Essay 5; 2 OCs

Here, the first TLU score would be calculated over Essay 1 and Essay 2 because Essay 1 alone does not reach 15 OCs but Essay 1 and Essay 2 combined do. The first window would thus include 16 OCs. In computing the second TLU score, the head of the first window is shifted forward by one, and the second window starts from Essay 2. The second window would cover Essay 2 through Essay 4, because Essay 2 alone or Essay 2 and Essay 3 together do not include 15 OCs, but the three essays combined do. The second window would include 23 OCs in total. Similarly, the head of the window is now shifted to Essay 3, and the third TLU score would be calculated over Essay 3 and Essay 4. This learner has these three TLU scores in total, as Essay 4 alone, Essay 4 and Essay 5 combined, or Essay 5 alone does not include 15 OCs. TLU scores were calculated over essays and not strictly over 15 obligatory contexts because learners' ability was assumed to be the same within an essay. Essays without any obligatory contexts of the target morpheme were skipped.

The number 15 was chosen in order to balance the number of TLU scores obtained from learners and the reliability of the scores. That is, on the one hand, it is better to have a narrower window because we potentially capture finer developmental patterns. On the other hand, however, narrowing the window size means fewer OCs, which then makes TLU scores less reliable.

Pros and cons of windows including overlapping essays. As is clear from the above, a window typically included essays that have also been included in the previous windows. Alternatively, it was also possible to construct windows so that no essay overlaps in any window. In the example above, Window 1 covered Essay 1 and Essay 2. Window 2 could have started from Essay 3 to avoid any overlap and the window could have covered Essay 3 and Essay 4. There are pros and cons of each approach.

An advantage of the approach adopted in this study (overlapping approach) is that we can follow more fine-grained developmental patterns. Let us assume that a learner wrote 17

essays each of which included two OCs. Let us further assume that the learner's "true" ability translated into TLU scores should result in the decrease of 0.1 per essay starting from 1.0 for the first nine essays and the increase of 0.1 per essay afterwards (1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). If we take the overlapping approach, each of the 10 windows would cover eight essays and the expected TLU scores would be {0.65, 0.55, 0.48, 0.43, 0.40, 0.40, 0.43, 0.48, 0.55, 0.65}. In other words, the developmental shape would be U-shaped. If we take the approach that does not allow overlapping essays (non-overlapping approach), however, Window 2 would cover Essay 9 through Essay 16 and the TLU score would be 0.55, while Window 1 and its TLU score would remain the same. In other words, the latter approach would fail to capture the U-shaped development that would be captured by the overlapping approach. Thus, the overlapping approach can capture the development at a higher resolution. Related to it, the overlapping approach results in a larger number of windows, and we will have more suitable data for the analysis of developmental trajectories.

On the other hand, the overlapping approach can dilute the slope of the developmental pattern. Let us now assume that a learner wrote 16 essays each of which includes two OCs, and his/her true ability in terms of TLU scores was {0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1}. If we could follow the learner's development precisely, we should observe a large jump at Essay 9. This is indeed what we see if the non-overlapping approach is taken. Window 1 would cover Essay 1 through Essay 8 and the corresponding expected TLU score is 0, while Window 2 would cover Essay 9 through Essay 18 and the corresponding TLU score is 1. However, if the overlapping approach is adopted, the expected TLU score would be a linear increase of accuracy of {0.00, 0.12, 0.25, 0.38, 0.50, 0.62, 0.75, 0.88, 1.00}, failing to capture the large jump of accuracy. There is therefore a scenario in which the non-overlapping approach can better capture the developmental pattern.

Furthermore, in the overlapping approach, the skewness of OCs may distort the devel-

opmental pattern. The direct problem is the varying number of times each essay contributes to TLU scores. Let us assume that a learner wrote nine essays including 2, 2, 2, 2, 2, 2, 2, 2, and 15 OCs. In the second TLU score computed over Essay 2 through Essay 9, the influence of Essay 9 is slightly over half (15/29) in that 15 out of 29 OCs are contributed by Essay 9. The value increases as the window progresses. The value is 15/27 in Window 3 covering Essay 3 through Essay 9, 15/25 in Window 4 covering Essay 4 through Essay 9, 15/23 in Window 5 covering Essay 5 through Essay 9, 15/21 in Window 6 covering Essay 6 through Essay 9, and so forth until it becomes 100% at Window 9 (15/15) covering Essay 9 alone. This indicates that the tail of the window does not change from Window 2 to Window 9, and that Essay 9 has an immense influence on the learner's TLU scores. The invariance of the tail is not a problem. If we assume that the nine essays above were written at Unit 1 through Unit 9 (or Unit 1 in the next Lesson), the median proficiency levels of Window 2 through Window 9 are {5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0}. In other words, the learner's proficiency level where his/her ability is estimated progresses as windows advance[4]. This therefore satisfies the purpose of tracking the accuracy development of the learner. However, a large impact of Essay 9 on the TLU scores can potentially influence the TLU scores in undesirable ways. If the same learner's true ability in TLU is a linear increase of {0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, the resulting TLU scores would be nonlinear, {0.55, 0.81, 0.84, 0.88, 0.91, 0.94, 0.97, 0.99, 1.00}. More concretely, the slope becomes flatter as windows progress. The large jump between Window 1 and Window 2 is because the essay with 15 OCs started to play a role there.

Despite the drawbacks described above, the present approach is favored for two reasons. First, finer resolution of data is indispensable in the study in order to observe change (de

[4]Using weighted mean, as will be explained later, is more appropriate to pinpoint the learner's proficiency level that a TLU score reflects the ability of. Median was used here for the sake of simplicity. The idea is the same for weighted means.

Bot et al., 2007b; Larsen-Freeman & Cameron, 2008; Siegler, 2006). Windows already contaminate the estimate of learner's ability to a certain extent by aggregating multiple essays. In order to reveal the developmental pattern while preventing further noise in the data, it is essential to have as fine-grained data as possible.

Second, the two possible drawbacks mentioned above are not major disadvantages unless OCs are extremely skewed. In terms of dilution of the shape, it is highly likely that even the non-overlapping approach does dilute the shape to a certain extent. In the example earlier where the true ability was assumed to be {0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1}, if there had been more essays before Essay 1 (or the total OCs of the first eight essays had been fewer) and a window had been constructed to cover both the essays at which the learner's true ability was 0 and those at which it was 1, the jump would have been diluted and the TLU score of the window would have fallen somewhere between 0 and 1. Dilution is thus a nearly intrinsic property of the moving-windows approach in general and not unique to the overlapping approach.

I have to partially concede a risk from the skewness of OCs. At the same time, however, its effect is not very large to the extent that it severely distorts the developmental trajectory. Even in the relatively extreme example given earlier where essays included 2, 2, 2, 2, 2, 2, 2, 2, and 15 OCs and the true ability was assumed to be linear, the development of TLU scores was not very different from linear (0.55, 0.81, 0.84, 0.88, 0.91, 0.94, 0.97, 0.99, 1.00). A large difference in absolute accuracy will not be a problem because what we will study is developmental trajectories. Also, the non-overlapping approach is not free of the effect of the skewness of OCs. If a learner's essays include OCs of {14, 1, 14, 1, 1, 14}, and his/her true ability is {0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, the learner's TLU scores are {0.51, 0.71, 0.99} in the non-overlapping approach, and {0.51, 0.69, 0.71, 0.98} in the overlapping approach. While both are nonlinear, their tendency is similar to the true ability. Therefore, in this sense, the current approach can capture the developmental pattern sufficiently well

even when OCs are skewed, and the non-overlapping approach is not necessarily better, either.

Once TLU scores were obtained for all the windows, I analyzed the development of the learners over 10 windows. The learners were then clustered according to their longitudinal developmental patterns of articles over multiple windows in order to identify typical developmental patterns.

4.3.2.2 Longitudinal Development of Articles

Descriptive data of article windows. For articles, there were in total 53,543 TLU scores (windows) by 14,144 learners. Out of the 14,144 learners, 1,044 (7.4%) had 10 or more windows. In this section, the first 10 windows of these learners will be the main target of analysis. The value 10 was chosen in order to balance the total number of learners that can be investigated and the number of data points for each learner. The average number of essays in a window was 3.426 (*SD* = 1.531). The mean number of unique essays over 10 windows was 11.730 (*SD* = 1.065). The latter is not 10 times as large as the former because there are overlaps of essays over windows. The average number of Units covered in 10 windows was 26.637 (*SD* = 12.116). It is not the same as the unique number of essays because analyzed essays were not necessarily consecutive due to the absence of error annotation in many essays in EFCamDat. Thus, a window corresponds to 3.330 Units on average for each learner.

Visualization of article development. Figure 14 depicts the transition of the first 10 TLU scores in each of the 1,012 learners. The horizontal axis shows the chronological order of the TLU score for each window (e.g., 1 for the first window, 2 for the second window), and the vertical axis represents the TLU score for that window. One learner is represented by one thin line whose color corresponds to the L1 type the learner belongs to, and thick lines are the locally weighted scatterplot smoothing (LOWESS; Larson-Hall

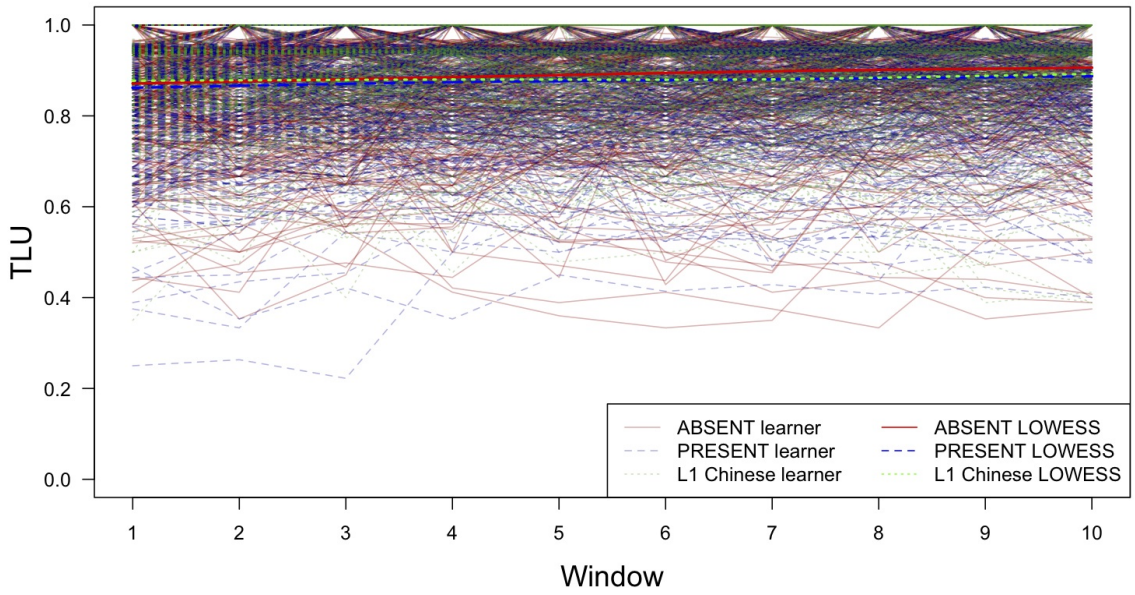


Figure 14. Accuracy Development of Individual Learners

& Herrington, 2010; Singer & Willett, 2003) lines for the ABSENT, PRESENT, and L1 Chinese learners. LOWESS fits regression models to local subsets of the data and helps to show the overall pattern of development. Simply put, it is a trend line showing the general pattern. In the present context, LOWESS shows the general accuracy transition over 10 windows by L1 type.

As can be seen in the figure, the TLU scores are overall high, centering around 0.90. The difference between the PRESENT and the ABSENT group that was relatively clear in Figure 13 is now hardly identifiable. The difference between Figure 14 and Figure 13 is that in Figure 13 scores are given for each level of proficiency. In Figure 14, proficiency is not shown, since it only shows scores of individuals over a period of time (windows) irrespective of proficiency. But as it happens, the average proficiency of the ABSENT learners included in Figure 14 (7.8 in terms of Lesson) is much higher than that of the PRESENT learners (4.5) or L1 Chinese learners (4.7). In other words, because the ABSENT learners in Figure 14 are of higher proficiency than the PRESENT learners, the accuracy difference

between the ABSENT and the PRESENT group disappeared. The figure further shows that there are significant individual differences, and that just looking at the overall pattern (three LOWESS lines) is misleading. If the average is taken, the accuracy transition of all the three groups is relatively flat. But individual lines show fluctuation.

Window proficiency. The task now is to understand the variation among learners. For this purpose, the study investigated whether learners can be clustered into groups of similar accuracy developmental pattern. It, for instance, analyzed whether learners' accuracy consistently rises, stays stable, or fluctuates, and if it fluctuates, in what shape. In other words, the main purpose of clustering is to identify whether the learners' fluctuation is random or whether we can identify distinct developmental shapes.

In order to understand whether we can group learners into different developmental patterns, we consider two variables that may determine developmental shapes; learners' L1 and proficiency. But calculation of learner proficiency is not straightforward. This is because the first essay (covered by the first window) for each learner may start at any proficiency level. In addition, each window has multiple essays and one learner may have progressed more than another over 10 windows. To overcome such issues, I define learner's proficiency as the average of the window proficiency, which in turn is defined as the weighted average Unit number (1 to 128) of the essays that the learner wrote for the window. The weighting was by the sum of the number of obligatory contexts and that of overgeneralization errors. For example, if a window included three essays whose Unit numbers were 10, 11, and 12, their unweighted mean is 11. However, if the number of obligatory contexts plus that of overgeneralization errors of the essays were 13, 1, and 1 respectively, the first essay has a much stronger influence on the TLU score of the window than the other two. It, then, might be inappropriate to take the unweighted mean as the representative Unit number of the window. Therefore, the Unit numbers were weighted by the denominator of the TLU score (i.e., the sum of the number of obligatory contexts and that

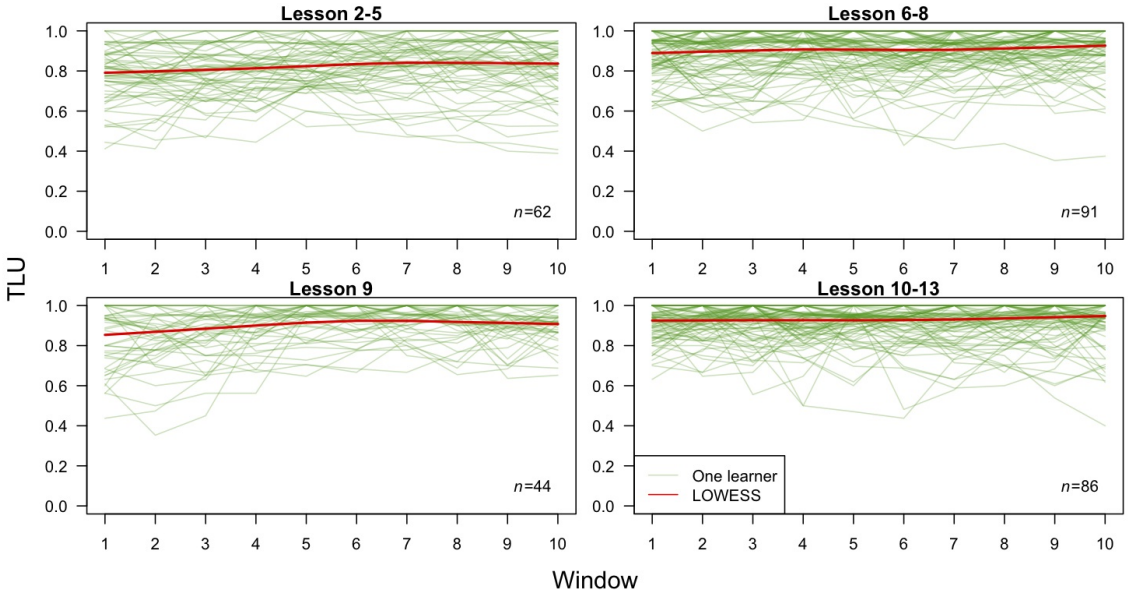


Figure 15. Accuracy Development of L1 Russian Learners by Proficiency

of overgeneralization errors) in order to make window proficiency more properly reflect the Unit that the calculation of the TLU score is based on. In the case above, the weighted mean is computed by (10 × 13 + 11 × 1 + 12 × 1) / (13 + 1 + 1) and the value is 10.2. The values for window proficiency were computed for each window and their unweighted average over 10 windows was considered the proficiency of the learner.

Top-down clustering of learners. Figure 15 and Figure 16 respectively show the accuracy development of L1 Russian and L1 Brazilian learners of English according to their proficiency level. Each green line shows the development of one learner, and thick red lines are trend lines (LOWESS). As expected, the overall accuracy, as shown by LOWESS lines, tends to increase in both L1 groups as proficiency rises. But there is no visible L1 influence, despite the fact that Russian and Brazilian-Portuguese are two typologically distinct languages. Finally, the averages seem to still hide individual differences in the development over 10 windows.

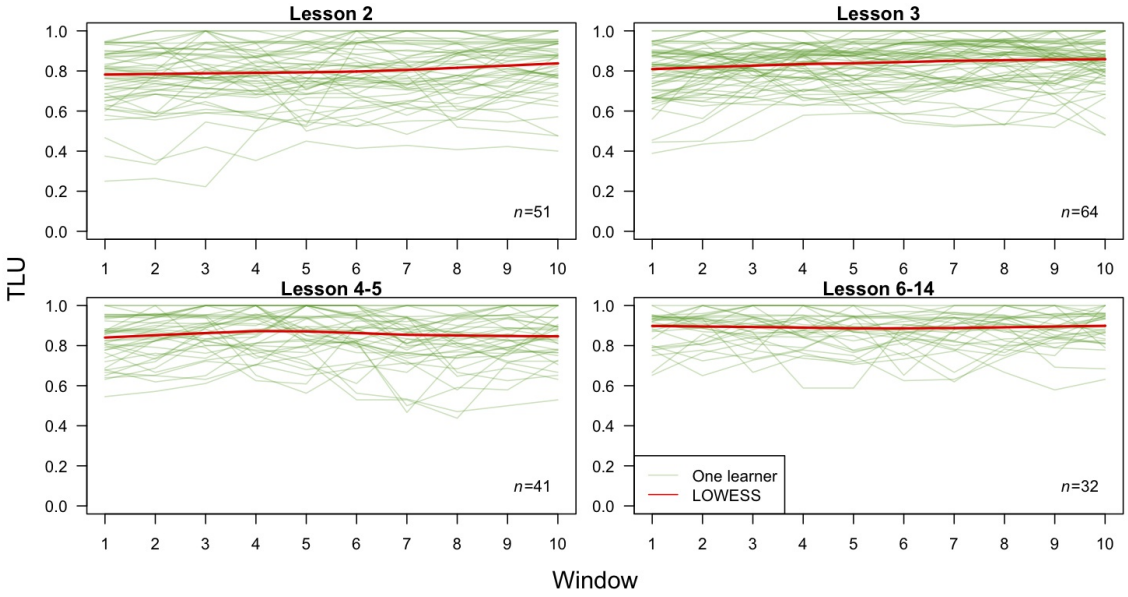


Figure 16. Accuracy Development of L1 Brazilian Learners by Proficiency

4.3.3 Clustering Learners According to Their Shapes of Article Development

What the previous section has shown is that pre-selected variables like L1 and proficiency are not straightforwardly linked to the developmental path. The task we should tackle is to figure out whether there are any systematic patterns in the development of individuals over 10 windows or whether accuracy fluctuations are completely random. We will do this without pre-specifying the variables that might affect the development but through employing data-driven clustering techniques.

Clustering has been applied to previous research in linguistics. Gries and Stoll (2009) and Hilpert and Gries (2009) employed variability-based neighbour clustering, a variant of agglomerative clustering, in order to segment the MLU development of a child and identify the child's developmental stages. Deshors and Gries (in press) used a hierarchical cluster analysis to disclose the similarities between the use of *can* and *may*, as well as the French word *pouvoir*, by native and L1 French learners of English. Ionin and Mon-

trul (2009) employed k-means clustering, the algorithm of which I will explain later, to divide learners into different proficiency groups according to cloze test scores. K-means clustering has also been used to group L2 learners according to their motivational profiles (Csizér & Dörnyei, 2005), according to their ability/aptitude profiles (Rysiewicz, 2008), according to their developmental profiles based on international posture, L2 willingness to communicate, and frequency of communication in L2 (Yashima & Zenuk-Nishide, 2008), and according to their cognitive and achievement profiles based on L1 achievement, intelligence, L2 aptitude, and L2 proficiency (Sparks, Patton, & Ganschow, 2012). Whereas some studies pre-specified the number of clusters (Ionin & Montrul, 2009; Sparks et al., 2012), others first ran a hierarchical cluster analysis and decided the number of clusters used in k-means clustering that followed (Csizér & Dörnyei, 2005; Rysiewicz, 2008; Yashima & Zenuk-Nishide, 2008).

The cluster analyses have produced some findings that are otherwise difficult to reveal. Through the clustering technique, Gries and Stoll (2009) identified developmental stages of utterance length in highly variable data. The clustering of Deshors and Gries (in press) based on a number of variables showed that the use of *can* by native and non-native speakers is relatively similar, that the same is the case for the use of *may*, and that *pouvoir* behaves differently from *can* and *may*. Csizér and Dörnyei (2005) disclosed four motivational profiles of Hungarian learners of English from *least motivated learners* to *most motivated learners*.

There are multiple approaches to clustering. The study exploits two ways to let the data cluster according to developmental shapes. The first method pre-defines the shape of the development while the second method lets learners with similar developmental patterns cluster in a data-driven way. An advantage of using the first method is that, because the shape of the development in each cluster is known beforehand, interpretation of the results is easier. An advantage of the second method, on the other hand, is that the developmental

trajectory each cluster follows is more flexible and the shape the method captures might be closer to the real developmental pattern.

4.3.3.1 Regression-Based Clustering

Shapes of clusters. In regression-based clustering, TLU scores of each learner are regressed against (or predicted by) window number. A crucial step is to decide the power of window entered into the model. If window to the first power is the only predictor, the resulting regression model ($TLU_i = \beta_0 + \beta_1 \times window_i + \varepsilon_i$, where i is the window number and ε is a normally-distributed error term. β_0 and β_1 are estimated from data.) is linear. That is, we assume the equal increase (or decrease) of TLU scores between Window 1 and Window 2 and between Window 7 and Window 8. If the second power, or quadratic term, is also included, the resulting regression model ($TLU_i = \beta_0 + \beta_1 \times window_i^2 + \beta_2 \times window_i + \varepsilon_i$) is either U-shaped ($\beta_1 > 0$) or inverted U-shaped ($\beta_1 < 0$) unless β_1 is zero. That is, the difference in TLU varies between Window 1 and Window 2 and between Window 9 and Window 10. The power indicates the complexity necessary to model TLU scores. In the present study, the first and the second powers were tested. It was statistically possible to test higher-order terms such as cubic or quartic. This was not attempted because, first, SLA research has observed U-shaped development that can be modeled by quadratic terms (Abrahamsson, 2003; Zobl, 1984), but there has been no evidence for cubic development. Second, a graph showed that the distinction between cubic development (sideways S's) and the linear developmental patterns are not visually very clear. Because the decision is potentially subjective and I have to accept the possibility that including higher-order terms produces different results, cluster analyses with varying numbers of clusters will be tested later in the second type of bottom-up clustering.

The algorithm of regression-based clustering. Now that we decided the power of window used in regression models, the next step is to perform clustering. More specifically,

the following procedure was taken.

1. For each learner, TLU scores for the first 10 windows were taken. If all the 10 scores of the learner were above 0.90, s/he was classified into the *Over 90%* cluster. This cluster acted as a reference group where the learners are considered to have acquired articles.

2. If one or more scores of the learner was below 0.90, it was checked whether the developmental shape is a straight line or (inverse) U-shaped. The TLU scores were regressed against window and its quadratic (i.e., squared) term that shows the depth of the (inverse) U shape. The regression model was then compared against the model without the quadratic term. If a model with (inverse) U-shaped curve (i.e., quadratic term) better models the development of a learner than the model without it, then the learner was classified into one of the two clusters; *Quadratic Positive*, corresponding to U-shaped development, or *Quadratic Negative*, corresponding to inverse U-shaped development. If the coefficient of the quadratic term (β_1 above) was positive, the learner was classified into the Quadratic Positive cluster, whereas if the coefficient was negative, the learner was classified into the Quadratic Negative cluster.

3. If the quadratic term is non-significant, that is, if the depth of the (inverse) U shape is not significantly different from zero, it indicates that a less complex model can predict TLU scores well. In that case, linear development was assumed. Similarly to the above, if a regression model with a linear term better models the development than the one without (i.e., intercept-only model that represents flat, horizontal development), then the learner was classified into either the *Linear Positive* cluster where the accuracy consistently increases over 10 windows or the *Linear Negative* cluster where the accuracy consistently decreases. If the coefficient of the linear term (β_2 above) was positive, the learner was classified into the Linear Positive cluster,

whereas if the coefficient was negative, the learner was classified into the Linear Negative cluster.

4. If neither the quadratic term nor the linear term was significant, it means that the developmental pattern is not significantly different from a flat line. Thus, the learner was classified into the *Horizontal* cluster.

In sum, if the shape of the development is more quadratic than linear, then the learner was classified into either the quadratic positive or the quadratic negative cluster depending on whether the shape is closer to a U or an inverted U. If the developmental shape is linear, then it was inspected which of (i) straight increase, (ii) straight decrease, or (iii) absence of change best describes the shape, and the learner was classified into one of the three groups accordingly. Note that absolute accuracy is not looked at in clustering. What we are interested in is the deviation of fluctuation.

Developmental patterns of each cluster in regression-based clustering. Once this procedure was completed for every learner, all the six clusters were realized. This means that there was at least one learner in each pattern. The result of the clustering is presented in Figure 17. Each panel contains one cluster of learners. As in Figure 14, thin lines represent individual learners, and thick lines are trend lines (LOWESS). At the bottom right corner of each panel is the number of learners in that cluster and their proportion out of those with 10 or more windows. We can make several observations. First, the largest cluster is the Horizontal cluster with 40.2% of the learners. Considering that the Over 90% cluster also marks a relatively flat development, nearly half of the learners lack the change over 10 windows. This can mean that their development has been stabilized or fossilized (Han & Odlin, 2006) or has reached a plateau. For the Over 90% cluster, this is a ceiling effect. Together the Over 90% cluster and the Horizontal cluster occupy 47.8% of the learners. This explains the little overall accuracy change in Figure 14. The next largest cluster is

the Linear Positive cluster (16%), which matches with the pseudo-longitudinal view of the data in Figure 13. The third largest cluster (15.9%) is the Quadratic Negative, with accuracy first going up and then down. The Quadratic Positive (12.5%) is the reverse of the Quadratic Negative. The smallest cluster is Linear Negative (8.1%), in which article accuracy consistently decreases. Overall, it is interesting that only 16% of all the learners exhibit a consistent accuracy increase. Also, the significant overlap of the three LOWESS lines in each panel again tells us that accuracy level is similar across L1 types when 10 windows are focused on.

4.3.3.2 KmL Clustering

Rationale for KmL clustering. The key feature of the regression-based clustering is the pre-specification of the number and shapes of clusters. However, because the technique pre-defines the developmental shape, it risks missing the shapes outside the initial hypothesis. For instance, even if a typical developmental pattern is cubic, the analysis above fails to identify the shape because it only included a linear and a quadratic terms. As a result, the learners whose developmental patterns are otherwise cubic might have been classified into, for example, the Linear Positive cluster. To obtain clusters better reflecting the data, the study explored a data-driven way of determining developmental shapes. In particular, the present study employed KmL clustering (Genolini & Falissard, 2010). The key advantage of KmL clustering over the regression-based clustering is that it determines the shape of the accuracy development in a bottom-up manner. Our goal is also to compare and contrast the two approaches to clustering and identify how similar their results are.

K-means clustering. KmL stands for k-means for longitudinal data, and is an implementation of k-means specifically for longitudinal, trajectory data. K-means is a clustering technique where an analyst specifies the number of clusters (k) to which each observation is assigned, and centroid-based clustering is performed. More specifically, in k-means

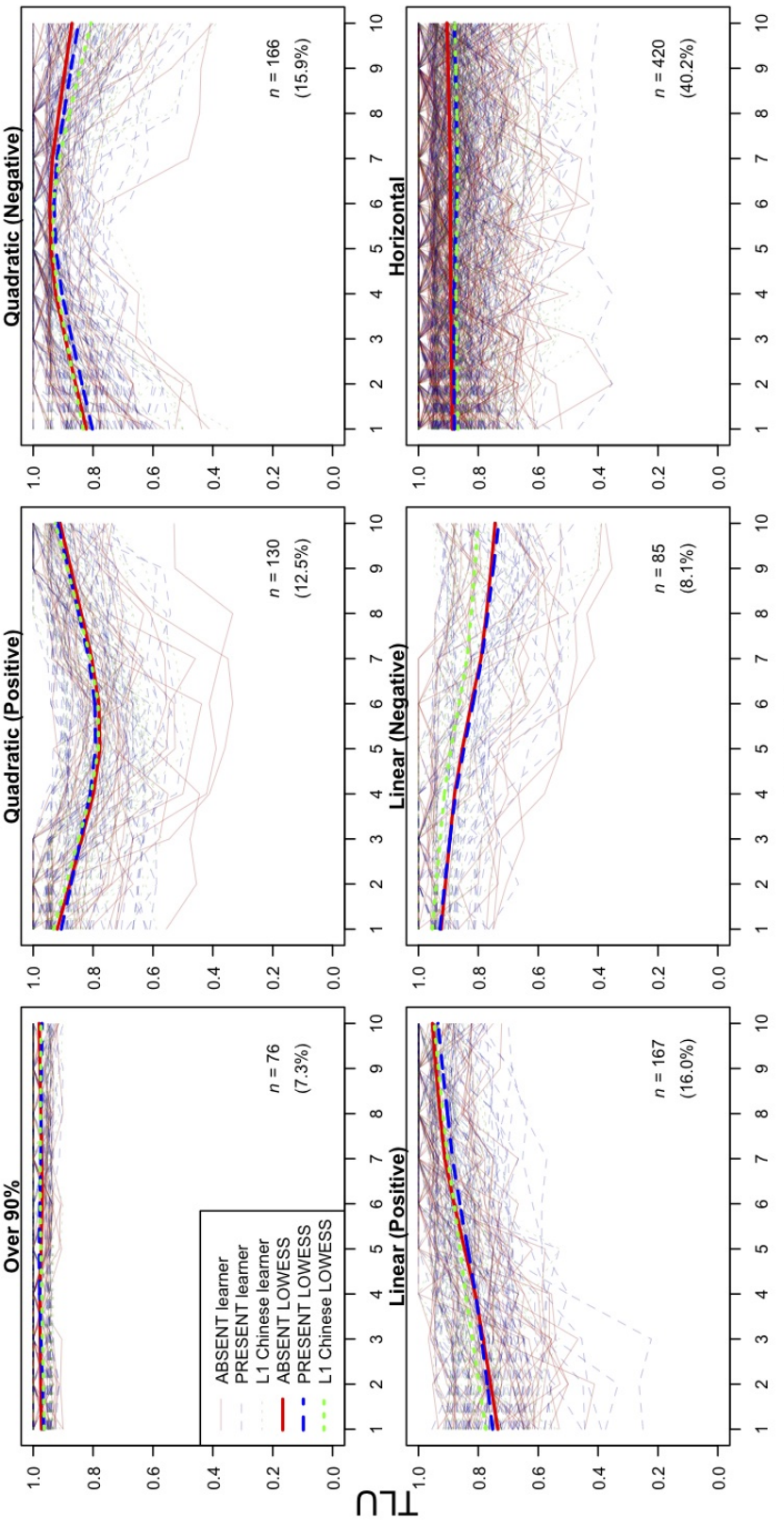


Figure 17. Regression-Based Clustering of Individual Learners According to the Shape of Accuracy Development

clustering,

> [g]roup membership is determined by calculating the centroid for each group. This is the multidimensional equivalent of the mean. Each individual is assigned to the group with the nearest centroid. The `kmeans` function [a function in R] fits a user-specified number of cluster centres, such that the within-cluster sum of squares from these centres is minimized, based on Euclidian distance. (Crawley, 2013, p.816; emphasis removed)

Euclidian distance is the most intuitive geographical distance. For instance, the Euclidian distance in a two-dimensional coordinate system between (0, 1) and (0, 2) is 1, and that between (0, 1) and (1, 0) is $\sqrt{2}$.

Figure 18 illustrates a concrete example of the algorithm taken from Shinnou (2007). The task is to cluster the five data points into two clusters. The coordinates of the data point 1 through 5 are (2, 1), (1, 2), (2, 4), (4, 4), and (4.5, 2) respectively. Step 0 (the upper left panel) shows the five data points plotted on a two-dimensional coordinate system. The first step is to randomly place k centroids. Typically, k data points are randomly selected for this purpose. Here, the first and the second data points were chosen and represented by x's (Step 1; the upper middle panel). The algorithm then calculates the distance between each data point and the centroids, which is depicted in the upper right panel (Step 2). Each data point is now assigned to the nearest cluster. In this case, data point 1, 4, and 5 are assigned to Cluster 1 represented by Centroid 1 and data point 2 and 3 are assigned to Cluster 2 represented by Centroid 2. Because data point 4 is located equidistant from both centroids, it can be assigned to either cluster. In the present case, it belongs to Cluster 1. The next step is to renew the centroids. The centroid of Cluster 1 that includes data point 1, 4, and 5 is moved to the mean coordinate of the three data points at $(\frac{2+4+4.5}{3}, \frac{1+4+2}{3})$ = (3.500, 2.333). Similarly, the centroid of Cluster 2 is moved to $(\frac{1+2}{2}, \frac{2+4}{2})$ = (1.5, 3.0). This is

illustrated in Step 3 at the lower left panel. As in Step 2, the distance from each data point to the centroids is now calculated (Step 4; the lower middle panel), and each data point is assigned to its closest centroid. In this case, assignment does not change. That is, data point 1, 4, and 5 are still assigned to Cluster 1 and data point 2 and 3 to Cluster 2. Because the assignment stays the same as the last iteration, the algorithm ends here, and Cluster 1 includes data point 1, 4, and 5, whereas Cluster 2 includes data point 2 and 3 (Step 6; the lower right panel). If the assignment changes, the algorithm repeats the procedure until no more change in the assignment of data points to clusters takes place. Although this example is a case of two-dimensional space, the same idea applies to higher dimensional space as well. In the present data, clustering is based on the distance in a 10-dimensional space because each learner has 10 TLU scores (i.e., windows).

An issue in the k-means algorithm is that it does not always give the best solution (Flach, 2012). Indeed, in the earlier example, if data point 1 and 4 had been selected as the initial centroids, the resulting two clusters would have included different data points from the clustering above. Therefore, it is recommended that we run the algorithm multiple times with different starting values (i.e., centroids) and retain the best solution that maximizes the between-cluster differences and minimizes the within-cluster differences.

Mean-centering to neutralize absolute accuracy differences. What I am interested in here is the developmental shape over 10 windows. However, if KmL is run on the present data as they are, it will take into account the absolute accuracy of each learner and may cluster learners according to their accuracy. To partial out absolute accuracy, all the data points were learner-mean-centered: The mean accuracy value of each learner was subtracted from all the data points of the learner. For example, let us suppose the TLU scores of Learner A over 10 windows were {0.40, 0.42, 0.44, 0.46, 0.48, 0.50, 0.52, 0.54, 0.56, 0.58} and those of Learner B were {0.80, 0.82, 0.84, 0.86, 0.88, 0.90, 0.92, 0.94, 0.96, 0.98}. Although the difference in accuracy scores is identical in the two learners (0.02 per

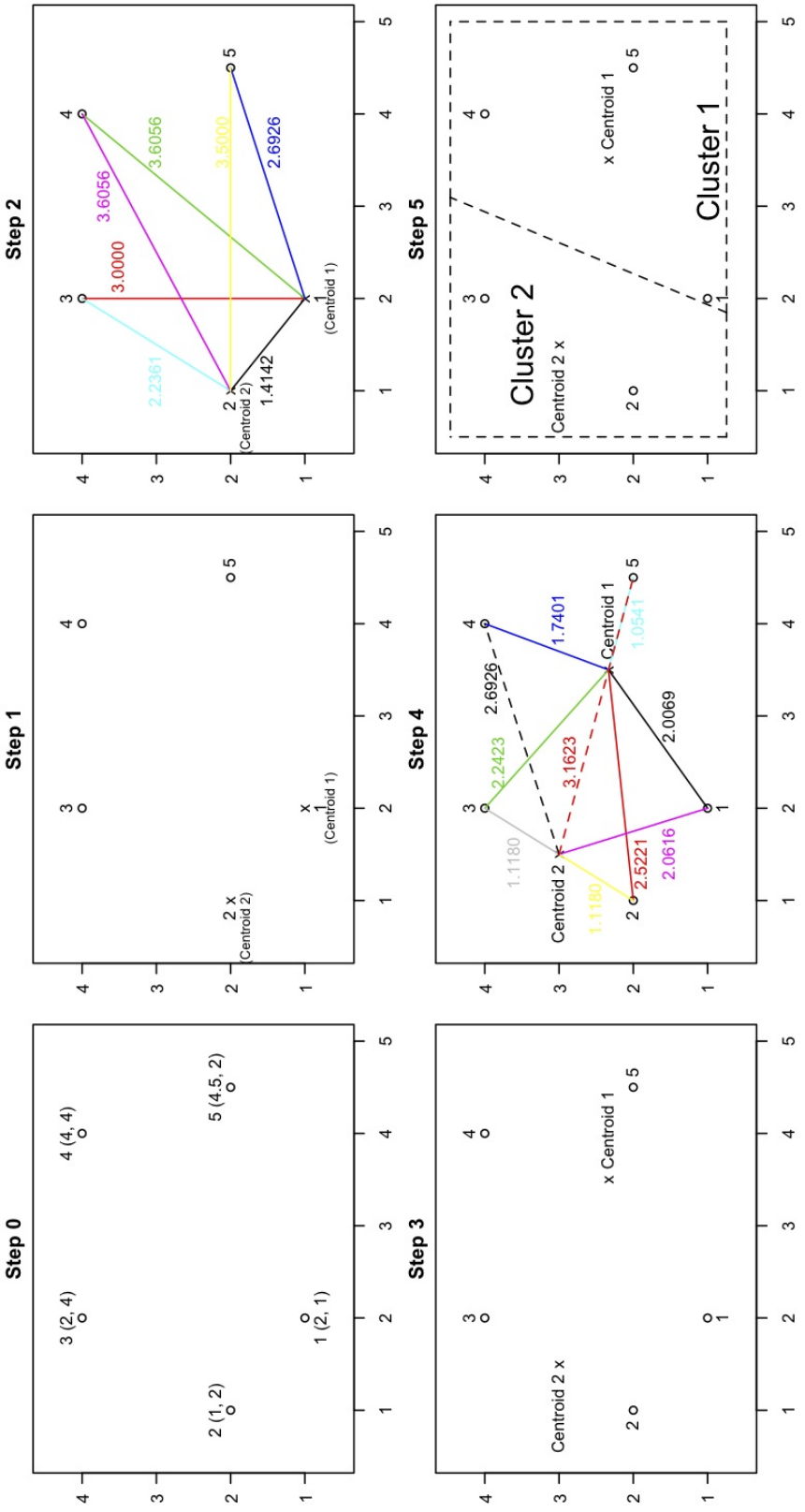


Figure 18. Description of the K-Means Algorithm

window), their absolute accuracy has a large gap of 0.40. If clustering is performed on absolute accuracy scores, the two learners are unlikely to be classified in the same cluster. Therefore, the mean value of Learner A (0.49) is subtracted from all the data points of Learner A, resulting in {-0.09, -0.07, -0.05, -0.03, -0.01, 0.01, 0.03, 0.05, 0.07, 0.09}. Similarly, the mean value of Learner B (0.89) is subtracted from all the data points of Learner B, resulting in exactly the same set of values as in Learner A. It is these values that were entered into KmL. This time, it is guaranteed that these two learners are classified into the same cluster as they have exactly the same feature values[5].

Clustering with different numbers of clusters. Clustering and determining the number of clusters are two separate issues, and there is no agreed way to formally determine the number of clusters in k-means clustering. One approach is clustering with different numbers of clusters and examining how emerging patterns vary across the clustering. For instance, we can establish which k certain patterns, such as U-shaped or inverted U-shaped development, appear at. If a pattern appears when k is small, it is a dominant pattern that applies to many learners. If, on the other hand, a pattern appears only when k is large, the pattern is likely to be a minor one. Also, if a certain pattern consistently appears with varying k's, it is likely that the pattern reflects something in learners' performance. A difficulty is that the similarity of developmental patterns of multiple learners is continuous and not dichotomous. That is, the developmental patterns of two learners are more similar or less similar, but not dichotomously similar or different. Clustering groups up the learners with similar patterns into the number of clusters set by the analyst. What differs, then, between the clustering with, say, $k = 2$ and the clustering with $k = 6$ is granularity. The former answers the question *if I am forced to choose only two article developmental patterns typical in the learners in EFCamDat, what would the two patterns be?* and, similarly, the latter

[5]Alternatively, I could cluster learners with the same starting point. But due to sparsity of data, this analysis was not possible.

responds to the case where the number of developmental patterns is six. In this respect, having multiple types of clustering does not mean one is better than the other. The choice depends on the level of granularity and the research question.

Developmental patterns when the number of clusters varies. I chose 10 for the maximum number of clusters because the total of 10 L1 groups are targeted and 10 different patterns are expected if each L1 has a distinct developmental pattern. Figure 19 shows the LOWESS of the developmental patterns of each cluster when $k = 2$ through $k = 10$ in KmL clustering. As before, the horizontal axis of the figure represents windows and the vertical axis represents TLU scores. Each panel represents the clustering when the k is the value stated above the panel. Each line is the LOWESS of the development of the learners in each cluster. Cluster A is always the largest cluster, followed by Cluster B, which in turn is followed by Cluster C, and so forth.

We can make a few observations here. First, except when $k = 3$, the largest cluster (Cluster A) shows a slight increase ($k = 4$, $k = 5$, $k = 9$) or decrease of accuracy ($k = 2$, $k = 6$, $k = 7$, $k = 8$, $k = 10$) over the 10 windows, and their accuracy does not change radically. This means that a great many learners follow a relatively flat development. Second, there is always a cluster showing an upward trend over 10 windows (e.g., Cluster F in $k = 6$) and a cluster showing decreasing accuracy over the period (e.g., Cluster C in $k = 3$). Third, U-shaped development is prevalent. It first appears in $k = 3$ as Cluster B, and can always be observed until $k = 10$ as Cluster F and H. Finally, there is no inverted U-shaped development. Possibly the closest cluster to an inverted U shape is Cluster A in $k = 10$, with the accuracy slightly going up until the fifth window and then down towards the end. The depth of the inverted U, or the accuracy difference between the first and the fifth windows, is small, and the shape is not very clearly inverted U.

In sum, the developmental patterns of (i) relatively flat, (ii) increasing or decreasing accuracy, and (iii) U-shaped development are robust and can be observed almost irrespective

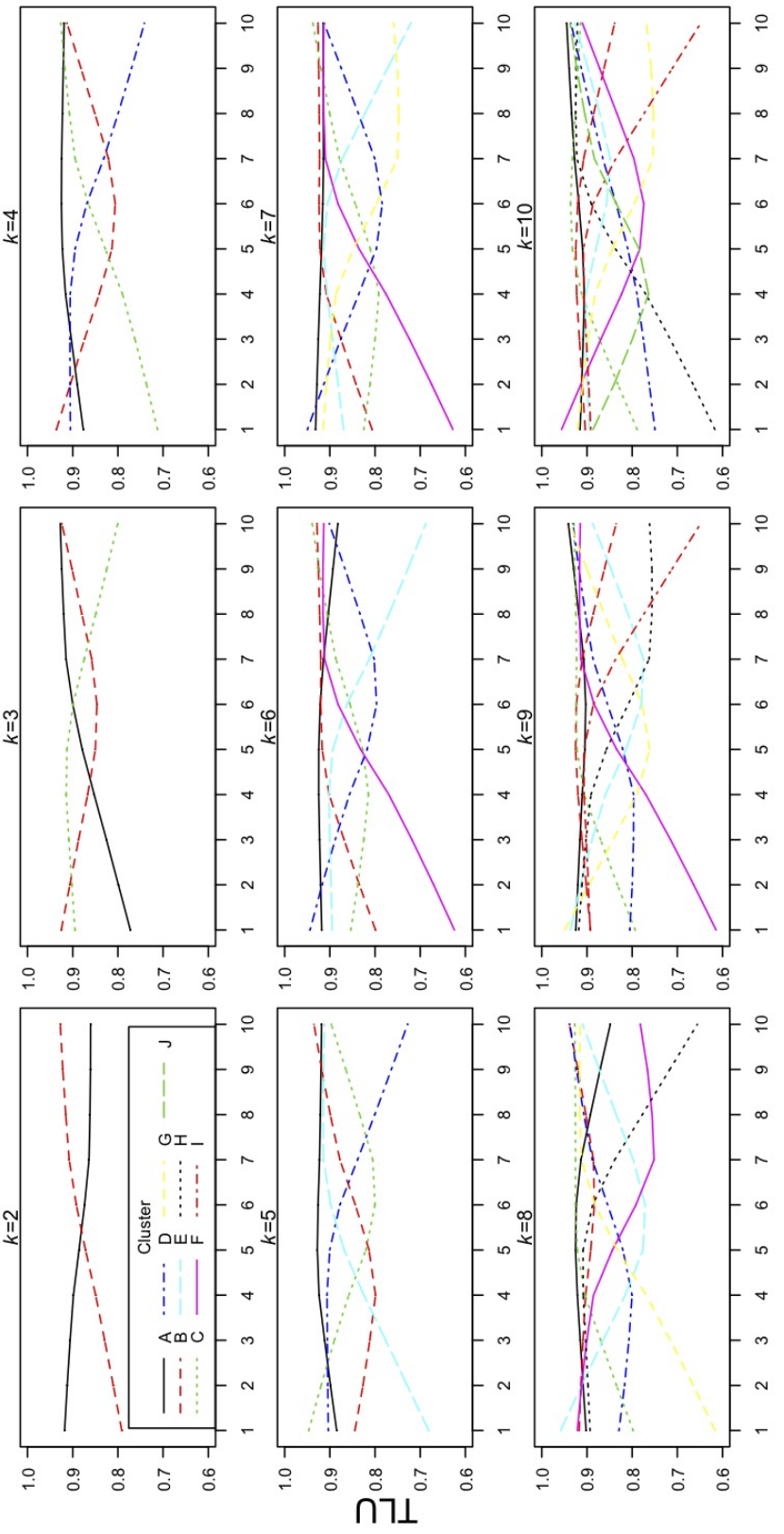


Figure 19. Developmental Patterns of Each Cluster in Varying Numbers of Clusters

of the number of clusters. On the other hand, increasing the number of clusters only results in finer splits of the same patterns and does not lead to the emergence of new patterns. For example, the U-shaped pattern is represented by two clusters as Cluster E and Cluster F in $k = 9$ or as Cluster F and Cluster H in $k = 10$.

Optimal number of clusters. As shown in Figure 19, no new pattern emerges after $k = 3$. This means that having $k = 4$ or more is unlikely to reveal important aspects of development. We thus assume $k = 3$ is optimal for article developmental patterns derived from KmL clustering. The main criterion for decision here is how informative each $k = n$ is, and whether new information or pattern can be revealed. However, in order to more directly compare the clusters with those obtained in the regression analyses, $k = 6$ was also tested and is reported.

Developmental patterns of clusters in KmL clustering ($k = 3$). Figure 20 shows the clusters of article development when $k = 3$. Note that, although clustering was performed on learner-mean-centered data, the figure shows the original non-centered data. As before, thin lines represent individual learners and thick lines represent LOWESS. The learners are approximately equally spread between the three clusters. As in the regression-based clustering, L1 type does not seem to affect clustering because the three LOWESS lines largely overlap. The accuracy of Cluster A gradually rises until around the seventh window, after which it stays relatively flat possibly due to the ceiling effect. Learners in Cluster B show a smooth U-shaped developmental curve. Their accuracy slowly decreases for the first five or six windows, after which it slowly increases. Their overall accuracy is high overall, being consistently over 0.8. Learners in Cluster C show a flat development until the fifth or the sixth window, after which their accuracy decreases. Significant individual differences can be observed in all the clusters, some learners radically going down and others rapidly going up.

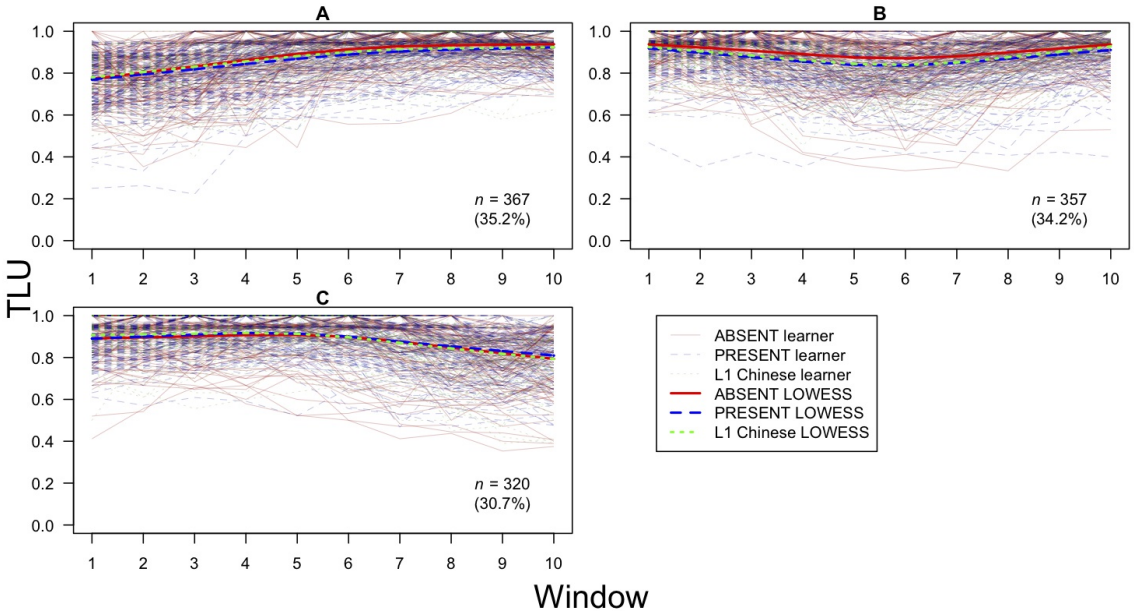


Figure 20. KmL Clustering of Articles (k = 3)

Developmental patterns of clusters in KmL clustering (k = 6). Figure 21 shows the clusters of article development when the number of clusters is six. The following observations can be made:

- Cluster A constitutes the largest cluster with 28.3% of the learners. Their development is relatively flat over 10 windows. This cluster consists of many learners in the Over 90% cluster and the Horizontal cluster in the regression-based clustering (discussed later).

- Cluster B, C, and F are similar in that they all show upward trends in accuracy. Differences pertain to the form of the development.

  – The learners in Cluster B tend to raise their accuracy only until the fifth window or around it and then seem to reach a ceiling. To substantiate the ceiling effect, the LOWESS representing the overall pattern starts at a rather high accuracy around 80% and then reaches around 90% at the fifth window, by which most

learners in the cluster have apparently acquired articles.

  - Cluster F, on the other hand, keeps increasing their accuracy until the seventh window, after which they seem to hit a plateau. Unlike those in Cluster B, they tend to start at a lower accuracy and show a steeper rise. Although this long-run rise of accuracy along the development is what one might expect, the cluster is the smallest of all and includes 7.1% of the learners.
  - Cluster C shows the opposite pattern of Cluster B. The starting point is high at about 85%, and they first exhibit a flat (or slightly downward particularly for the ABSENT learners) progression until the fourth window, after which the accuracy continues rising towards the end.

- Cluster D presents a clear pattern of U-shaped development. The accuracy first declines and then rises. The fact that both regression-based clustering and KmL clustering succeeded in extracting a number of learners whose developmental path is U-shaped means the observed U-shaped development is real and possibly reflects a certain stage of article development for certain learners.

- Cluster E demonstrates a declining trend, especially from the fifth window. This cluster seems to include many of those who were clustered in the Linear Negative cluster in the regression-based clustering.

When Figure 21 and the six clusters in regression-based clustering are compared, we can see that the Linear Positive cluster in the former (i.e., those who show steady increase in their accuracy) might have been divided into three clusters in the latter, and also the quadratic negative cluster (i.e., those whose accuracy first rises and then drops) was not observed in the KmL clustering. The three clusters emerged possibly because accuracy rise is common in article development and the learners showing the pattern were further

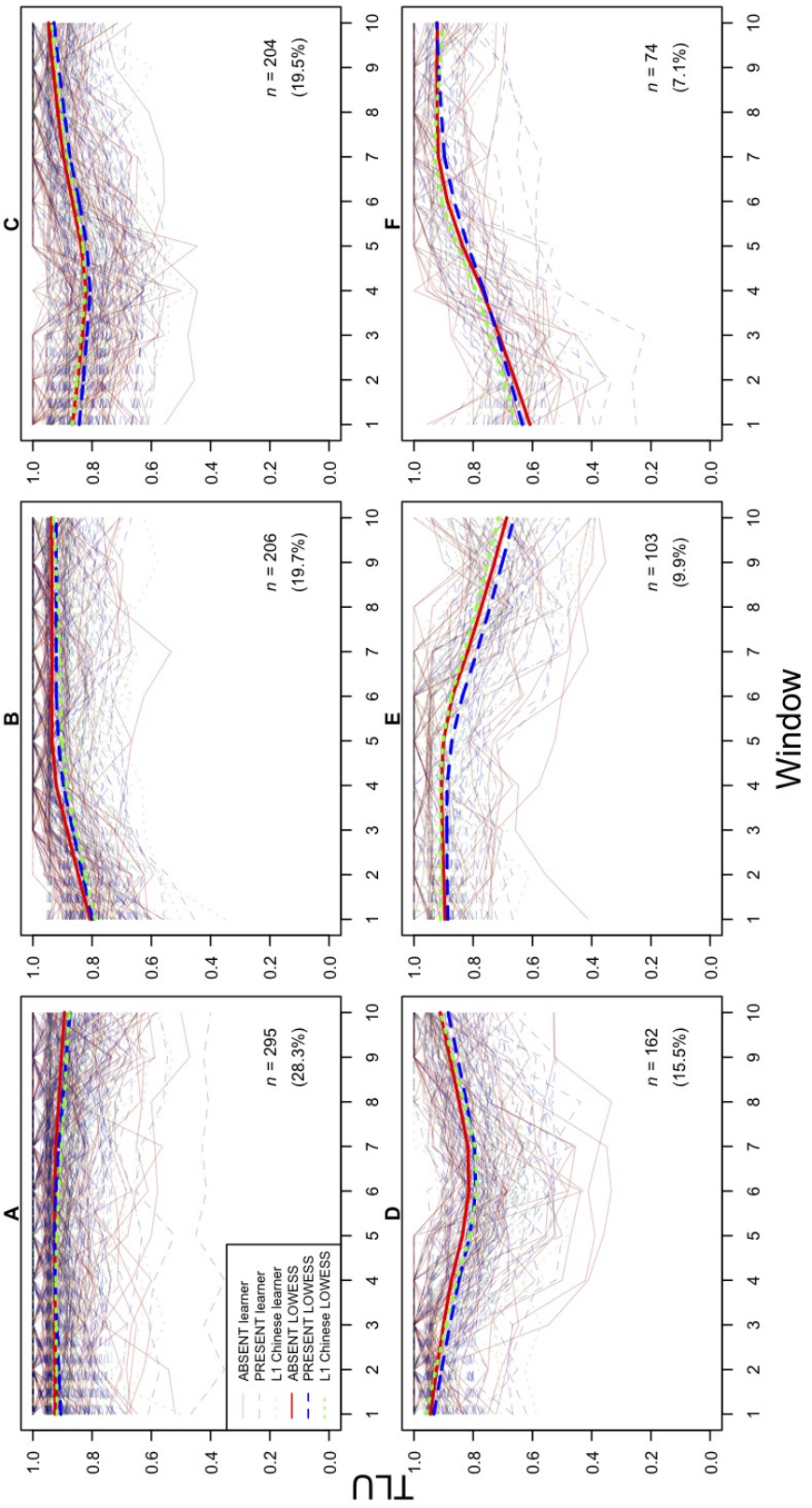


Figure 21. KmL Clustering of Articles ($k = 6$)

subdivided into three types. The Quadratic Negative cluster vanished potentially because it included learners with a variety of developmental paths and they were absorbed in Cluster B and E in the KmL clustering. This is confirmed and further discussed below.

4.3.3.3 Comparing Regression-Based Clustering and KmL Clustering

Let us compare the results of the regression-based and KmL clustering. Visual inspection indicates similarities between the two, particularly in relation to the developmental shape of relatively flat/horizontal or linear upwards or downwards pattern. In both cases there is a U-shaped developmental pattern. However, none of the KmL clustering, regardless of the value of k, yielded an inverted U-shaped development.

The analytical question is how much overlap there is between the learners that have been classified. It is expected, for example, that the learners in the Quadratic Positive cluster in regression-based clustering are mostly classified into Cluster B in KmL clustering when $k = 3$. To answer this question, the learners were cross-tabulated between regression-based clustering and the KmL clustering in Table 23 ($k = 3$) and Table 24 ($k = 6$). Below the tables are the results of χ^2 tests and Cramer's V value. Residual analyses of χ^2 tests revealed that the cells marked with a superscript H are where the number of learners is larger than expected by chance, and those with L are where it is smaller than expected, both at $p < 0.05$. For a reference, the figure showing the clusters based on the regression-based clustering (Figure 17) is redisplayed below.

With respect to Table 23, we can make following observations:

- As expected, there is a relatively strong association between two types of clustering overall. For instance, 86.2% of the learners in the Quadratic Positive cluster, 94.6% of those in the Linear Positive cluster, and 85.9% of those in the Linear Negative cluster in regression-based clustering are classified into Cluster B, Cluster A, and

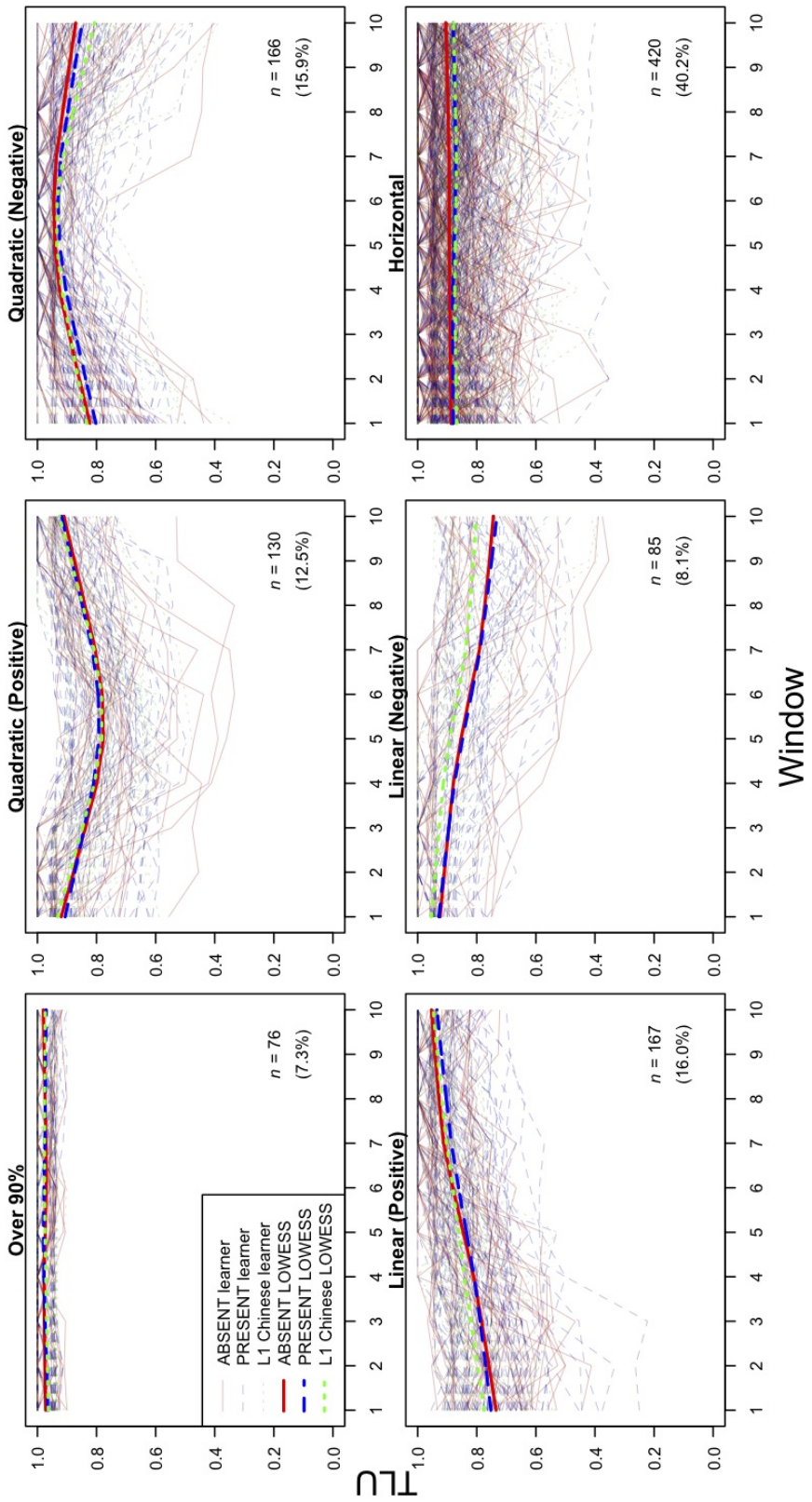


Figure 17. Regression-Based Clustering of Individual Learners According to the Shape of Accuracy Development

Table 23

Cross-Tabulation Between Regression-Based Clustering and KmL Clustering (k = 3)

KmL Clustering		Regression-Based Clustering: Over 90%	Quadratic (Positive)	Quadratic (Negative)	Linear (Positive)	Linear (Negative)	Horizontal	Total
A								
	# Learners	14^{L}	18^{L}	81^{H}	158^{H}	0^{L}	96^{L}	367
	%	18.4%	13.8%	48.8%	94.6%	0.0%	22.9%	35.2%
B								
	# Learners	34^{H}	112^{H}	0^{L}	9^{L}	12^{L}	190^{H}	357
	%	44.7%	86.2%	0.0%	5.4%	14.1%	45.2%	34.2%
C								
	# Learners	28	0^{L}	85^{H}	0^{L}	73^{H}	134	320
	%	36.8%	0.0%	51.2%	0.0%	85.9%	31.9%	30.7%
Total								
	# Learners	76	130	166	167	85	420	1,044

Note. $\chi^2(10) = 674.580$; $p < 0.001$; Cramer's $V = 0.568$

H = significantly more learners than expected at $p < 0.05$; L = significantly fewer learners than expected at $p < 0.05$

Cluster C in KmL clustering respectively. This is expected given the similarities of the relevant clusters.

- The learners in the Over 90% cluster are largely split between Cluster B and Cluster C. It is not very clear why Cluster A did not include many learners classified into the Over 90% in regression-based clustering, but possibly because those who have already reached a ceiling are less likely to show further increase of accuracy than the decrease of accuracy due to little room for improvement.

- The Quadratic Negative cluster absent in KmL clustering was almost equally split between Cluster A and Cluster C. Note that the absence of an inverted U-shape in KmL is the most substantial difference in the results of the two approaches, so it is worth asking why. The hypothesis is that in fact only a few learners in the Quadratic

Negative cluster show truly inverted U-shaped development. The majority of the learners classified into the Quadratic Negative in regression-based clustering only partially resembled the presupposed shape and really only showed the patterns in Cluster A and Cluster C of the KmL clustering. A typical developmental pattern in Cluster A and C can be classified into the Quadratic Negative cluster because cluster membership in regression-based clustering is a matter of which of the residuals of a quadratic function or a linear function is larger and, in the case of the latter, whether the decrease of residuals is worth an additional parameter to the regression model. When a linear model is fit to a typical pattern in Cluster A or C, the residuals are possibly large due to the ceiling effect and the nonlinearity of the shape as its result. No learner in the Quadratic Negative was classified into Cluster B because it shows a reverse pattern (i.e., U-shaped development).

- The Horizontal cluster was split into all the three clusters. Nearly half of the learners (45.2%) were classified into Cluster B possibly because they, in fact, show slightly U-shaped curve but were categorized into the Horizontal cluster due to small residuals stemming from the shallow depth of the U.

With respect to Table 24, the following observations can be made:

- A relatively strong association is found between two types of clustering. Far majority (76.3%) of the learners in the Over 90% cluster were classified into Cluster A in KmL clustering, which showed a flat development. All the others in the Over 90% cluster were classified into Cluster B, in which learners raise their accuracy at first but soon reach a plateau, or Cluster C, in which the rise of accuracy followed after a plateau.

- Similarly, 36.9% of those in the Horizontal cluster were classified into Cluster A, and another 21.4% and 15.2% into Cluster B and Cluster C respectively. This is

Table 24

Cross-Tabulation Between Regression-Based Clustering and KmL Clustering (k = 6)

KmL Clustering		Regression-Based Clustering						
		Over 90%	Quadratic (Positive)	Quadratic (Negative)	Linear (Positive)	Linear (Negative)	Horizontal	Total
A								
	# Learners	58^{H}	6^{L}	47	1^{L}	28	155^{H}	295
	%	76.3%	4.6%	28.3%	0.6%	32.9%	36.9%	28.3%
B								
	# Learners	10	0^{L}	67^{H}	65^{H}	0^{L}	64^{L}	206
	%	13.2%	0.0%	40.4%	38.9%	0.0%	15.2%	19.7%
C								
	# Learners	8^{L}	49^{H}	0^{L}	57^{H}	0^{L}	90	204
	%	10.5%	37.7%	0.0%	34.1%	0.0%	21.4%	19.5%
D								
	# Learners	0^{L}	74^{H}	0^{L}	0^{L}	14	74	162
	%	0.0%	56.9%	0.0%	0.0%	16.5%	17.6%	15.5%
E								
	# Learners	0^{L}	0^{L}	33^{H}	0^{L}	43^{H}	27^{L}	103
	%	0.0%	0.0%	19.9%	0.0%	50.6%	6.4%	9.9%
F								
	# Learners	0^{L}	1^{L}	19^{H}	44^{H}	0^{L}	10^{L}	74
	%	0.0%	0.8%	11.4%	26.3%	0.0%	2.4%	7.1%
Total								
	# Learners	76	130	166	167	85	420	1,044

Note. $\chi^2(25) = 886.889$; $p < 0.001$; Cramer's $V = 0.412$

H = significantly more learners than expected at $p < 0.05$; L = significantly fewer learners than expected at $p < 0.05$

because other clusters, particularly Cluster E and Cluster F show a large intra-learner variability that is far from horizontal.

- When the distribution in the Quadratic Positive cluster is analyzed, we can see that 56.9% naturally fit into Cluster D, where learners show U-shaped development. Another 37.7% of the learners were classified into Cluster C, where the accuracy development is going up from the fifth window, or slightly U-shaped.

- In the Quadratic Negative cluster, over 40% of the learners (40.4%) were classified into Cluster B, and another 28.3% and 19.9% in Cluster A and Cluster E respectively. In Cluster A and B, it is possible that some learners there show a small decline towards the end, and it makes the developmental shape close to an inverted U. It is interesting that only 19.9% of the learners were classified into Cluster E because it is the only cluster in KmL clustering that shows a clear downward trend. This, again, may mean that there are few that show truly inverted U-shaped development and the development of those who were classified into the Quadratic Negative in regression-based clustering, in fact, only partially resembled the presupposed shape. A typical developmental pattern in Cluster B can be classified into the Quadratic Negative cluster because, when a linear model is fit to a typical pattern in Cluster B, it is likely that the residuals are large due to the ceiling effect.

- All but one learners in the Linear Positive cluster were classified into Cluster B (38.6%), Cluster C (34.1%), or Cluster F (26.3%). This is natural because they are the clusters that show upward trends in accuracy. The fact that the Linear Positive cluster was divided into three clusters in KmL clustering means there might be multiple patterns within the Linear Positive cluster.

- Not surprisingly, the majority of those in the Linear Negative cluster were classified

into Cluster E, the only cluster with a declining tendency. Another 32.9% went into Cluster A, whose final two to three windows show a slight decrease of accuracy.

In sum, regression-based clustering with pre-defined hypotheses here classified some learners into some shape that is not so real. However, the differences in the two approaches should not be very significant, judging from the high association between the two sets of clustering.

4.3.3.4 Interim Summary

Let us return to our research questions. Our first question is whether the fluctuation in the accuracy of learners is random or is characterized by specific shapes. Both approaches showed clear patterns. The two approaches consistently identified three main patterns; flat development, steady increase or decrease of accuracy, and U-shaped development. This shows the robustness of these patterns. The challenging question now is whether we can identify features predicting these patterns. This issue will be investigated in the next chapter.

A further question is how much variation there is among learners. Significant inter-learner variation was observed. Even when learners were clustered according to the overall developmental patterns, individual variation was still large within each cluster. We also observed intra-learner variability. Although it was possible to cluster learners according to the broad shape of development, their patterns were rarely purely flat, purely linear, or purely U-shaped.

4.3.4 Analyses of the Other Morphemes

Let us now turn to the other morphemes. We will adopt the same methodology for the investigation of the other morphemes. However, as regression-based clustering and KmL clustering gave similar results for articles, only the latter will be employed given

its larger flexibility. The analysis had to be adapted because obligatory contexts for past tense *-ed*, possessive *'s*, progressive *-s*, and third person *-s* were sparser. Thus, if we keep the requirement of 15 obligatory contexts per window at minimum, we find 10 windows for very few learners except for plural *-s*. The study therefore set the window size to 10 obligatory contexts, instead of 15. Even with this adjustment, few learners could provide 10 windows. The analysis is thus adjusted to five windows.

4.3.4.1 Cross-Sectional View of the Other Morphemes

Figure 22 and 23 show the pseudo-longitudinal development of all the target morphemes by micro and macro averaging respectively. On Figure 22, the numbers on the graph show the number of obligatory contexts where it is smaller than 100. Although all the other morphemes are less frequent overall than articles, their data size seems large enough when looked at pseudo-longitudinally. As with articles, the difference between micro and macro averages is not very large. For the majority of the morphemes, including articles (already discussed), past tense *-ed*, possessive *'s*, and progressive *-ing*, the accuracy tends to increase over 16 Lessons. Plural *-s* and third person *-s* show relatively flat development, possibly due to the ceiling effect. The accuracy difference between the ABSENT and the PRESENT groups is not large except for a few data points (e.g., Lesson 6 to 10 in plural *-s*). There are fluctuations in accuracy, possibly due to the teaching materials used in Englishtown. We now turn to how they map onto the longitudinal development.

4.3.4.2 Clustering of the Longitudinal Development of the Other Morphemes

Descriptive data of the windows of the other morphemes. Table 25 shows basic descriptive data on windows for each morpheme. Although the window size is larger in articles and plural *-s*, they are the morphemes with by far the largest number of windows. This impact is not small, and even if five windows are taken as the period for longitudinal

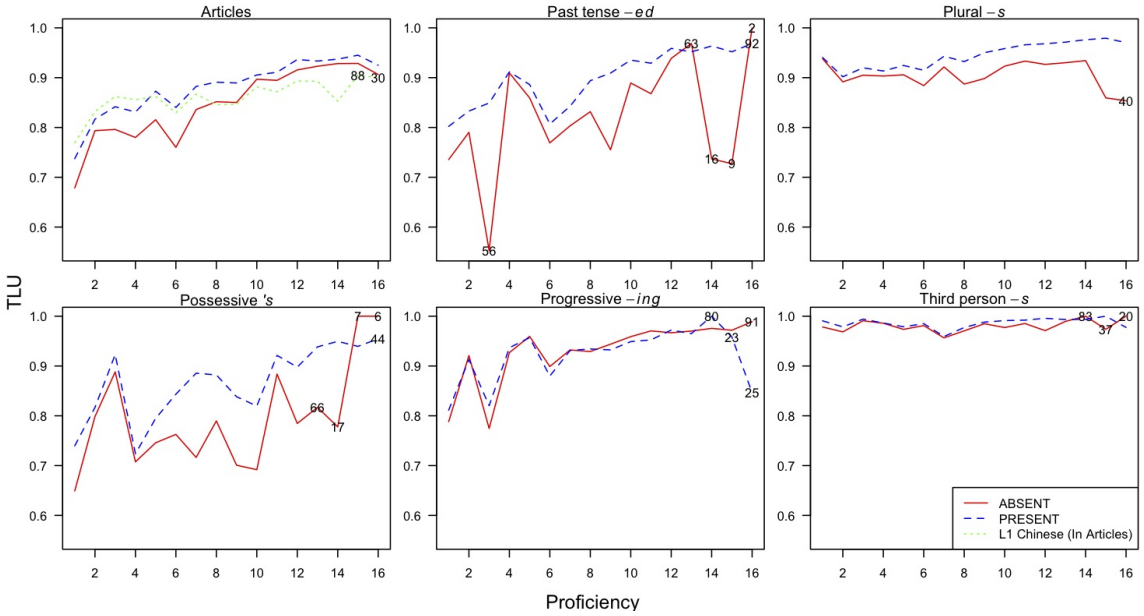


Figure 22. Pseudo-Longitudinal Development of Morpheme Accuracy (Micro Averages)

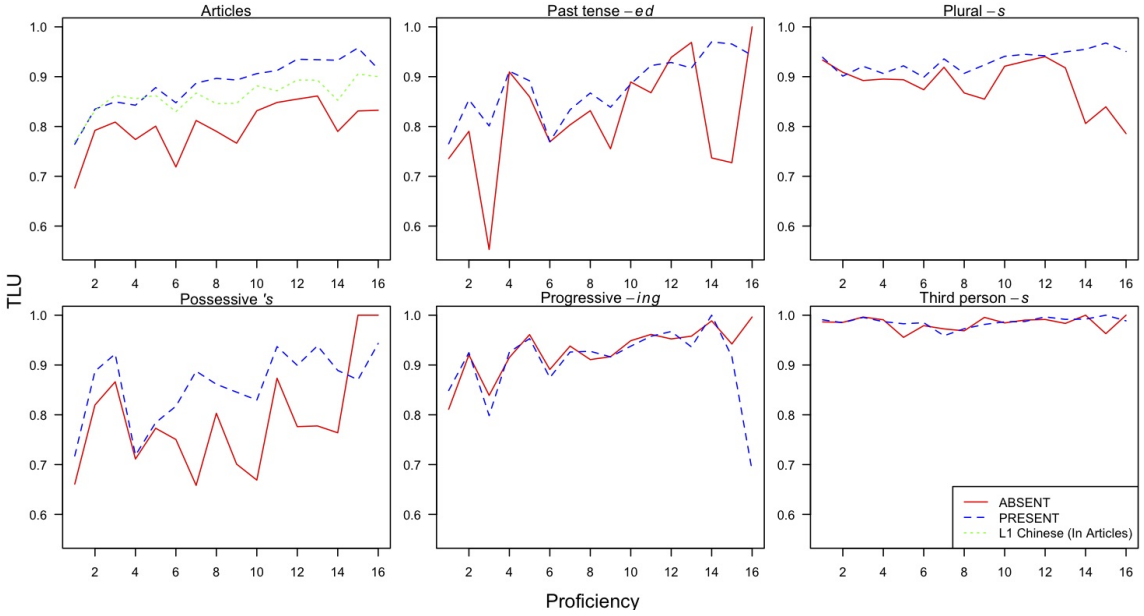


Figure 23. Pseudo-Longitudinal Development of Morpheme Accuracy (Macro Averages)

analyses, the number of learners with five or more TLU scores (windows) in past tense *-ed*, possessive *'s*, progressive *-ing*, and third person *-s* is still smaller than those who had 10 or more windows in articles and plural *-s*. Underlined numbers are those who took part in the longitudinal analysis below. Naturally, the number of essays per window is smaller in articles and plural *-s* than in the other morphemes. It is interesting to observe the relationship between the number of unique essays over 10 or 5 windows and that of Units covered by the windows. Because articles and plural *-s* require 15 obligatory contexts to form a window, and 10 windows were taken as the criterion to participate in the longitudinal analysis, the number of unique essays included over the windows is larger than the other morphemes. However, due to their high frequency, the average number of Units covered is approximately the same as the other morphemes. In other words, the number of essays learners wrote over 10 windows in articles or plural *-s*, each of which includes at least 15 obligatory contexts, is still smaller than the number of essays they wrote over 5 windows in the other morphemes with the window size of 10 obligatory contexts. Therefore, although different morphemes had different number of windows and the window size, their development is comparable at the level of Unit.

Optimal number of clusters for past tense *-ed*, plural *-s*, possessive *'s*, progressive *-ing*, and third person *-s*. As in the analysis of article development, I tested different numbers of clusters in order to visually determine the optimal number of clusters based on the emergence of different shapes as the number of cluster increases. Because possessive *'s* only included seven learners, it was not possible to apply the KmL clustering algorithm to the morpheme. As in articles, the developmental pattern of $k = 2$ through $k = 10$ was drawn for each morpheme.

Developmental patterns of the other morphemes when the number of clusters varies. Figure 24 shows the developmental patterns of each cluster of plural *-s* represented by LOWESS as k varies. We can make several observations here. As in article

Table 25

Descriptive Data on Windows

Morpheme	Total # of Windows	Total # of Learners with at Least One Window	Five Windows	Ten Windows	Window Size	Avg. # of Essays per Window (SD)	Avg. # of Unique Essays by 10/5 Windows (SD)	Avg. # of Units Covered in 10/5 Windows (SD)
Articles	53,543	14,144	4,055	1,044	15 OCs	3.426 (1.531)	11.730 (1.065)	26.637 (12.116)
Past tense *-ed*	2,467	1,305	91	9	10 OCs	4.462 (1.675)	7.813 (1.719)	31.187 (15.370)
Plural *-s*	37,067	9,860	2,738	708	15 OCs	4.240 (1.715)	12.377 (1.360)	30.384 (12.852)
Possessive *'s*	110	43	7	2	10 OCs	6.057 (1.282)	9.571 (0.535)	40.714 (25.349)
Progressive *-ing*	3,573	1,637	133	17	10 OCs	5.490 (1.627)	9.165 (1.702)	32.436 (15.729)
Third person *-s*	5,887	2,717	215	36	10 OCs	4.630 (1.630)	8.126 (1.679)	29.553 (19.067)

Note. OC = obligatory context. Underlined is the number of learners who were subsequently analyzed longitudinally.

development, the pattern of the largest cluster (Cluster A) is always smooth, demonstrating slight increase ($k = 7$), slight decrease ($k = 2$, $k = 4$, $k = 5$, $k = 6$), or flat development ($k = 3$, $k= 8$, $k = 9$, $k = 10$). No radical increase or decrease of accuracy is observed in the largest cluster. Second, accuracy increase is a common trend that is observed in all the clustering patterns, and they tend to be a major group. The second largest cluster (Cluster B) often shows the overall accuracy increase over 10 windows ($k = 2$ through $k = 8$). Third, accuracy decrease is common as well. In all the k's, there is at least one cluster whose accuracy more or less decreases over 10 windows. Fourth, as in article development, U-shaped development can be observed. Cluster C in $k = 4$ and $k = 5$, Cluster F in $k = 6$, Cluster G in $k = 7$, Cluster E and H in $k = 8$, Cluster D and I in $k = 9$, and Cluster B, I, and J in $k = 10$ all show clear U-shaped development. And fifth, similarly to article development, no cluster in any k shows inverted U-shaped development. All in all, the developmental patterns of articles and plural *-s* tend to be similar, although differences were observed as to the absolute accuracy as well as the proportion of the learners classified into each cluster. Following the same logic as in article development, $k = 4$ was selected to represent the developmental patterns of plural *-s* because no radically new pattern appeared after that.

With respect to past tense *-ed* (Figure 25), we can observe both similarities and differences from the developmental patterns in articles and plural *-s*. As in the development of articles and plural *-s*, the developmental path of the largest cluster (Cluster A) is smooth, without radical ups and downs. Also, we can observe the patterns of accuracy increase and decrease irrespective of the number of clusters. $K = 2$ is an exception to this since neither cluster shows a clear form of accuracy decrease. However, unlike in articles and plural *-s*, no cluster in any k demonstrates U-shaped development. An apparent exception is Cluster D in $k = 4$, Cluster E in $k = 5$, Cluster F in $k = 6$, and Cluster G in $k = 7$. However, these clusters just include one learner. Note that past tense *-ed* only covers five windows, each of which includes 10 obligatory contexts, and a hypothesis is that the total number of

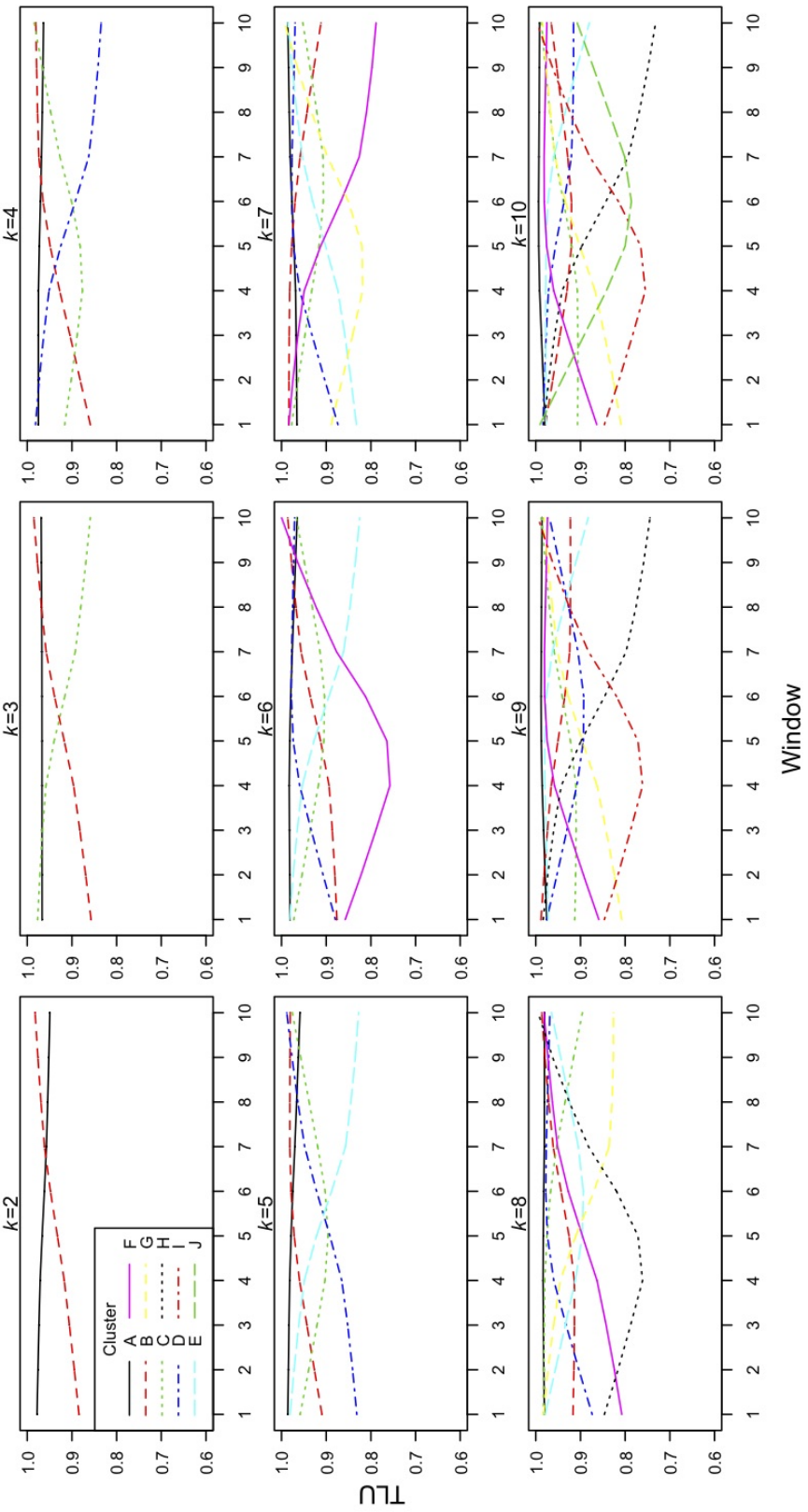


Figure 24. Varying Numbers of Clusters in Plural *-s*

obligatory contexts is not large enough to capture U-shaped development. This, however, is unlikely for two reasons. First, U-shaped development is observed in third person *-s* below, which has the same window size and the number of windows as in past tense *-ed*. Second, when articles are clustered under the same condition as past tense *-ed* (the figure not shown here), a similar U-shaped developmental pattern is observed at and after $k = 4$. Therefore, the absence of U-shaped development reflects some properties of the morphemes and not an artifact. Assuming for now that the learner showing U-shaped development is an outlier, $k = 3$ was selected to represent the developmental patterns of past tense *-ed*. Downward trend first appears at $k = 3$, but no new pattern emerges after that.

Progressive *-ing* (Figure 26) shows a similar trend with past tense *-ed*. The largest cluster (Cluster A) is always stable regardless of the number of clusters. In addition, it is flat at the TLU score of 1.00, reflecting the overall high accuracy of progressive *-ing*. Both accuracy increase and decrease are observed in all the k's at and after $k = 3$. As in past tense *-ed*, U-shaped development is absent in any k. Similarly to past tense *-ed*, $k = 3$ was selected to represent the developmental patterns of progressive *-ing*.

In terms of third person *-s* (Figure 27), the patterns look somewhat different from past tense *-ed* and progressive *-ing*. Cluster B in $k = 2$ is an outlier, and so are Cluster E, F, and G in $k = 7$. Note that these three clusters in $k = 7$ only include one learner each because only one learner is classified into Cluster E (cf. Figure 31) and Cluster F and G are of equal size to, if not smaller than, Cluster E due to the way KmL orders clusters. If those three patterns are removed, the remaining patterns look relatively similar to progressive *-ing*. The largest cluster is flat and very high in accuracy, and all the k's observe both increase and decrease of accuracy. In order to capture the potentially U-shaped development of Cluster D in $k = 5$, I chose $k = 5$ to represent the development.

Note that in all the morphemes, as in the case for articles, emerging patterns are not completely at random. Like articles, no k in any morpheme included inverted U-shaped

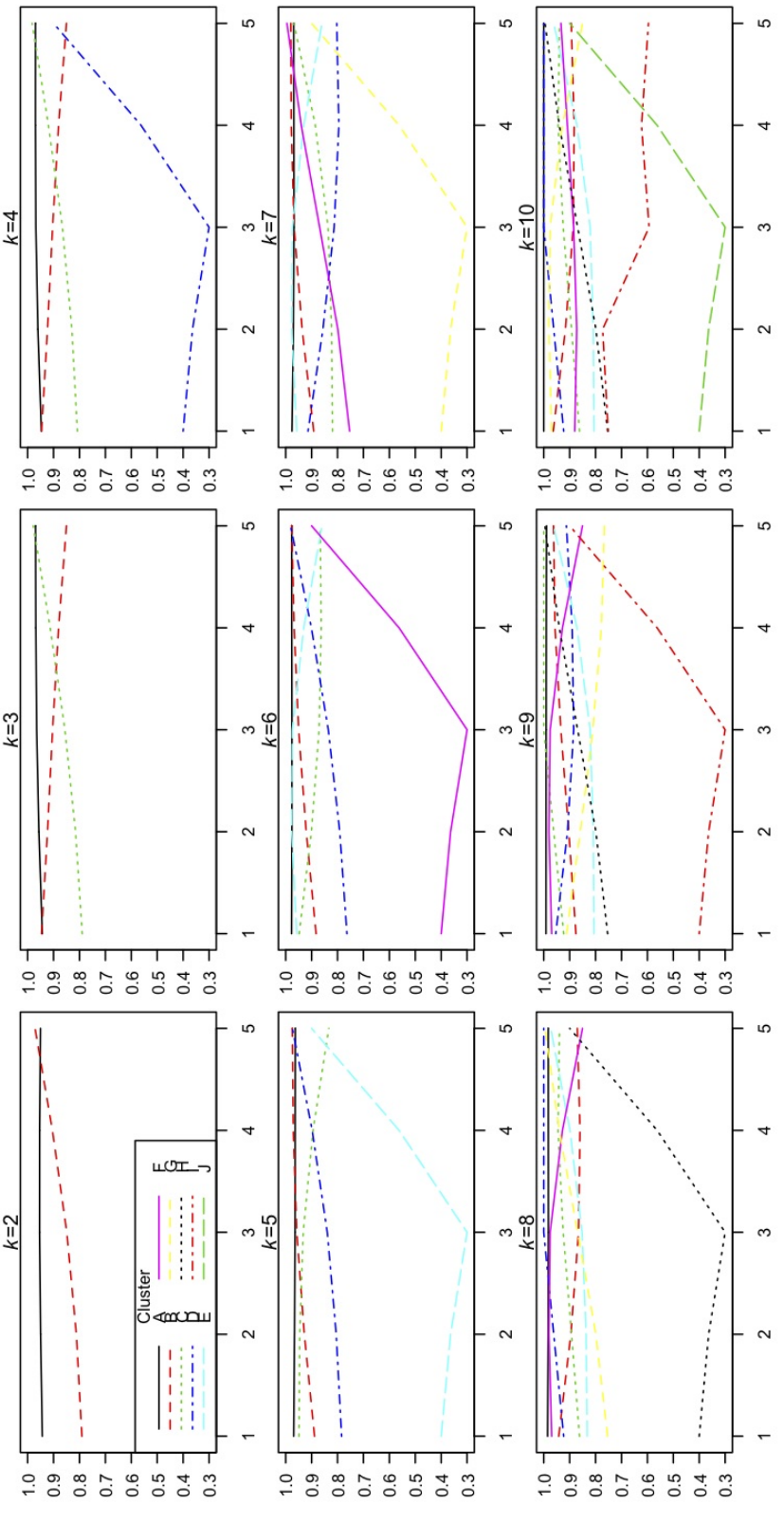


Figure 25. Varying Numbers of Clusters in Past Tense *-ed*

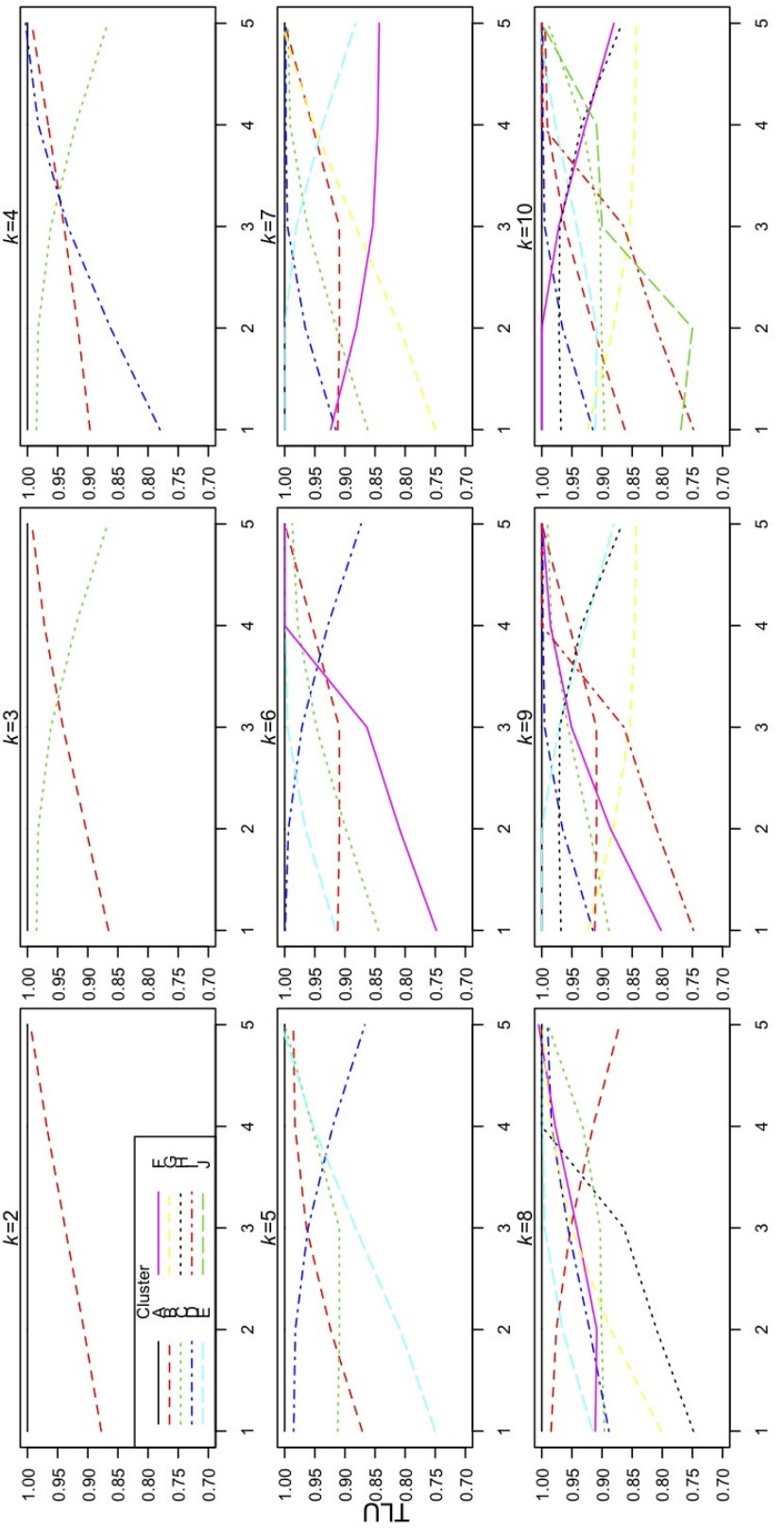


Figure 26. Varying Numbers of Clusters in Progressive -*ing*

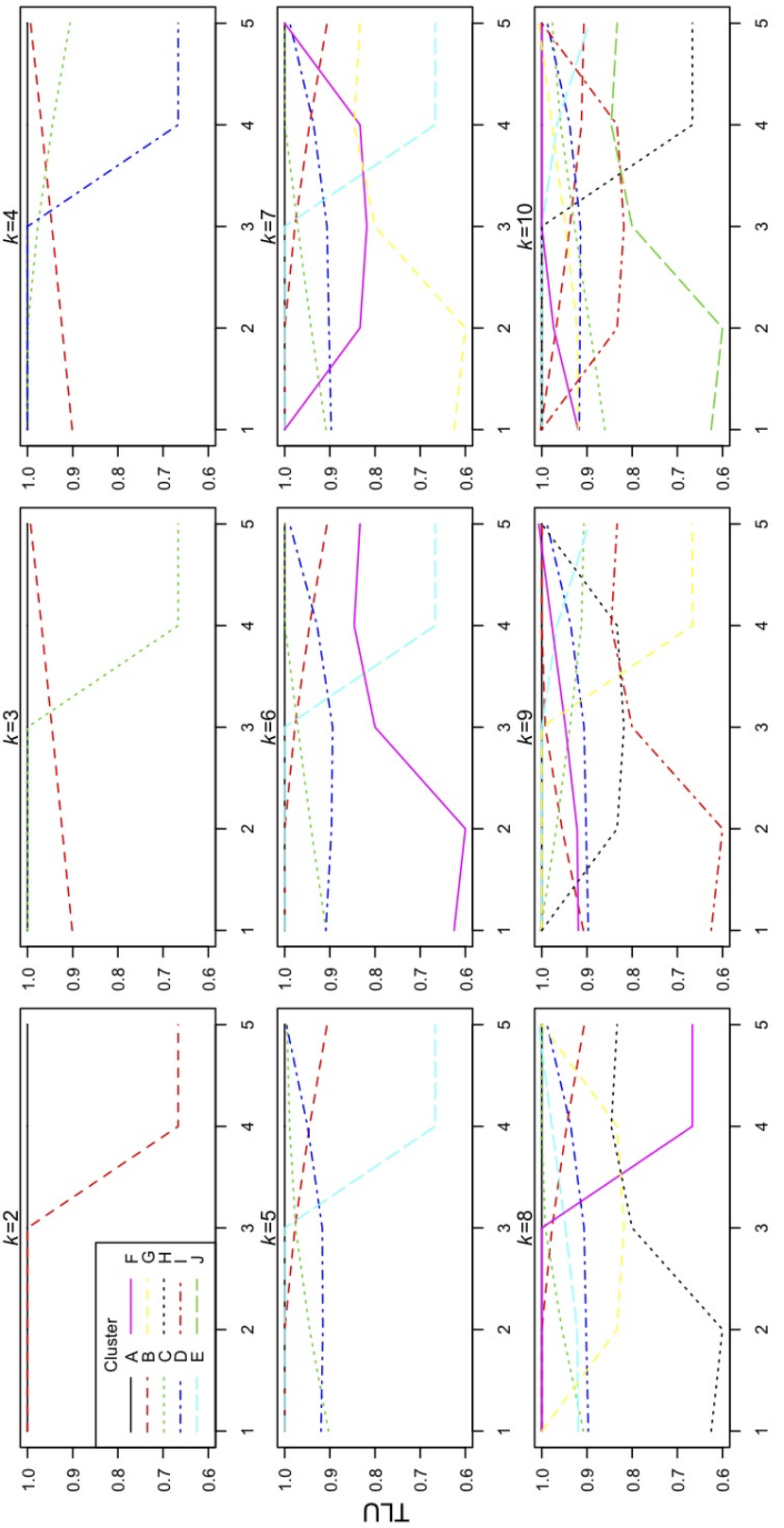


Figure 27. Varying Numbers of Clusters in Third Person -*s*

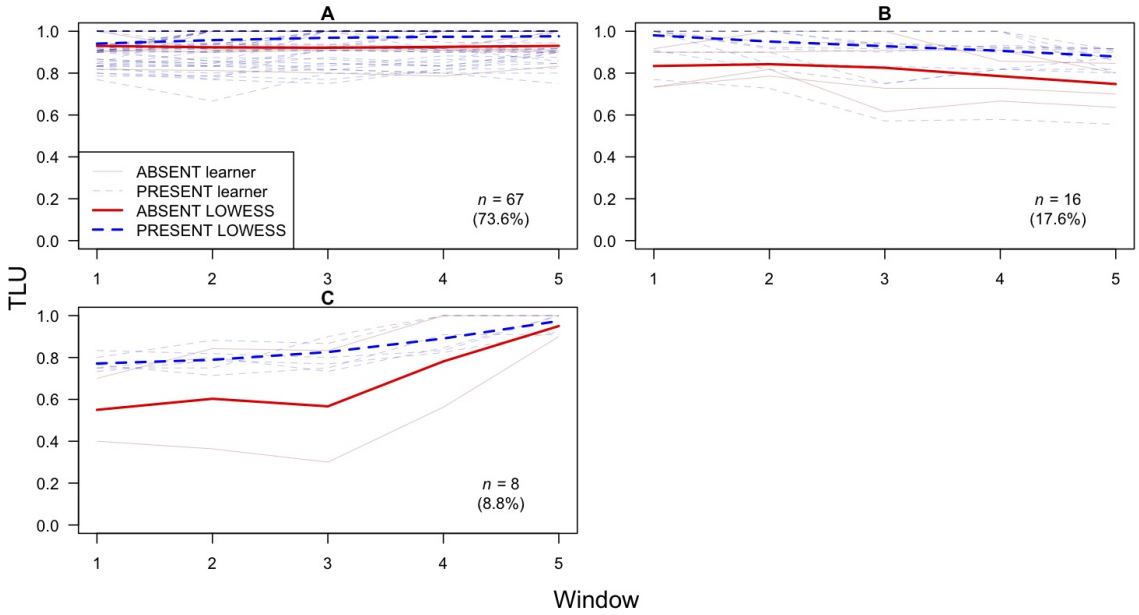


Figure 28. Clustering of Past Tense *-ed*

patterns, and increasing the number of clusters merely resulted in finer splits of the patterns already present when *k* is smaller, typically the pattern of increasing accuracy. This means that the developmental patterns of these morphemes are not random, and that the patterns in clustering are not artifacts.

KmL clustering of the other morphemes. Figure 28 to 31 show the clusters of the developmental pattern in each morpheme according to KmL clustering. Let me comment on each figure.

- The learners were clustered into three groups in past tense *-ed* (Figure 28). Those in Cluster A, in which 73.6% of the learners were classified, follow a relatively flat development. Cluster B with 17.6% show a downward trend. Cluster B and C show a clear L1 influence with the ABSENT groups showing lower accuracy. Cluster C with eight learners exhibit an increasing accuracy. Their accuracy is relatively low, starting at around 0.80 or below.

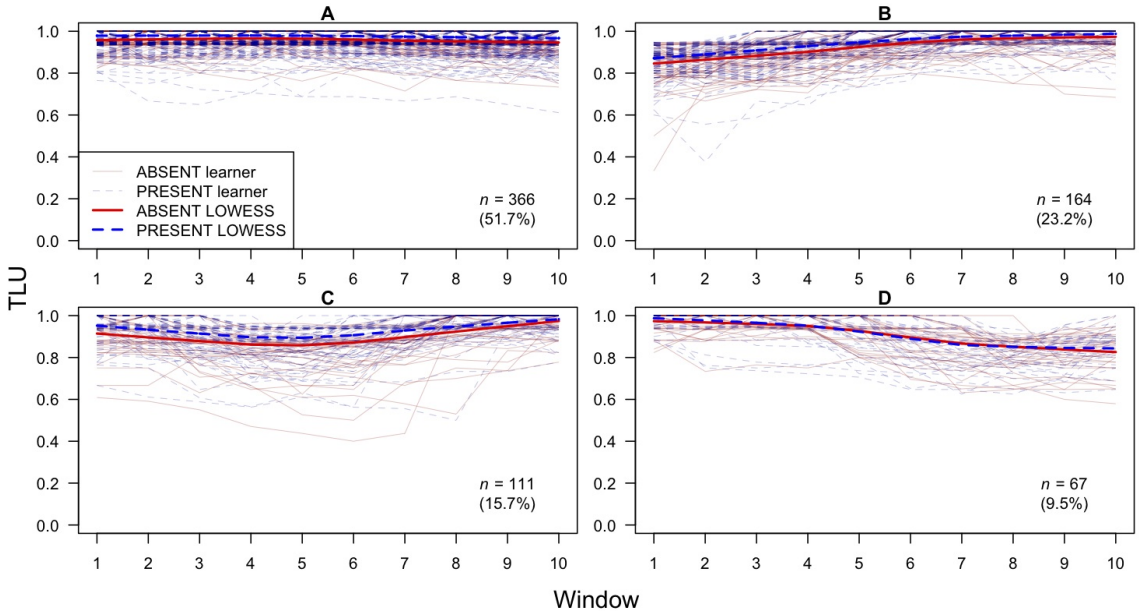


Figure 29. Clustering of Plural *-s*

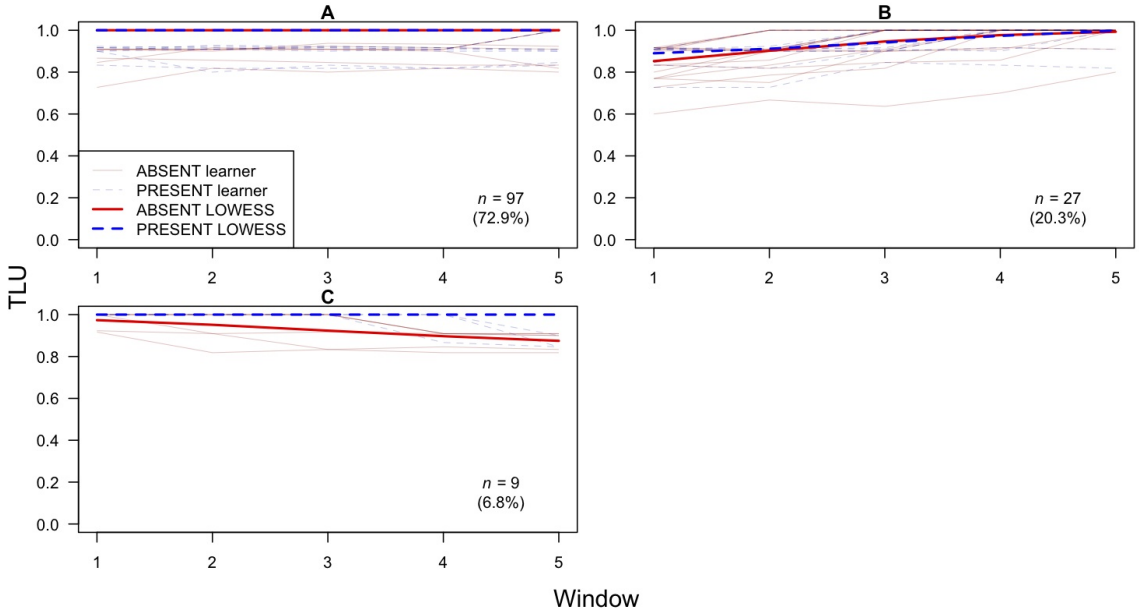


Figure 30. Clustering of Progressive *-ing*

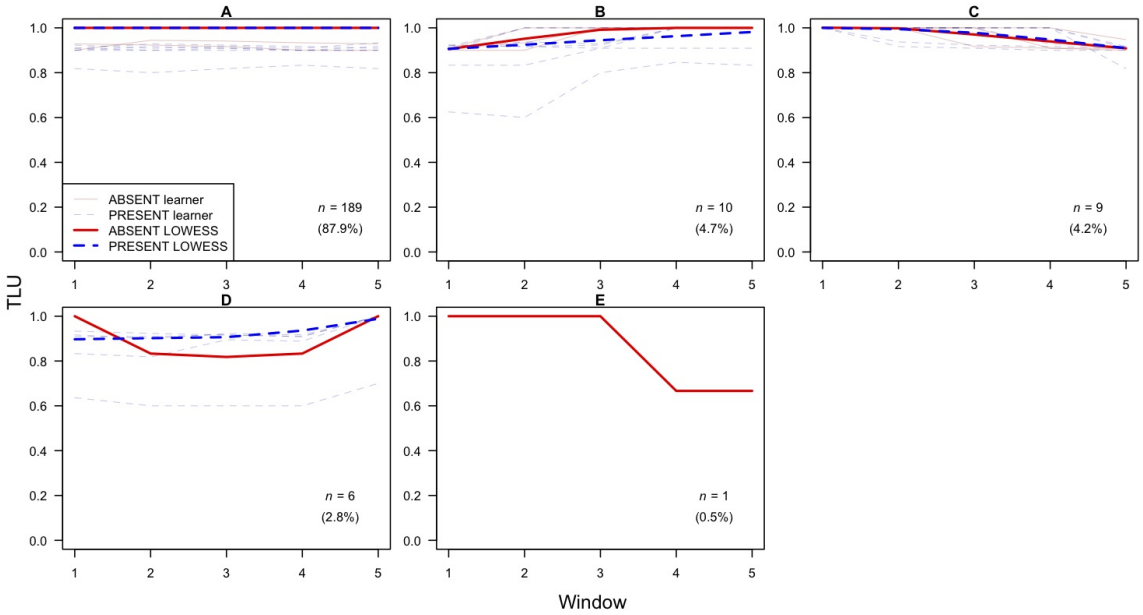


Figure 31. Clustering of Third Person *-s*

- Plural *-s* has four clusters. As in past tense *-ed*, the largest cluster (Cluster A) shows a relatively horizontal development. Although significant individual variation is recognized, their accuracy is very high overall, clearly above 0.90 throughout. The learners in Cluster B (23.2%) increase their accuracy for the first seven windows or so. Although it is lower than those in Cluster A, their accuracy is still high from the beginning (around 85-90%) and grows even higher until their LOWESS reaches almost 1.0. Cluster C (15.7%) is a smooth U-shaped developmental pattern. Their accuracy goes down for the first five windows and then rises thereafter. The accuracy of Cluster D (9.5%) keeps decreasing throughout the 10 windows, but its slope is especially steeper at the latter half.

- Three clusters emerged in progressive *-ing*. Cluster A follows an extreme pattern in that out of the 97 learners (72.9%) who were classified there, the majority seem to achieve 100% accuracy from the beginning till the end. It is reflected on the little

variation given the number of the learners and the flat LOWESS line consistently at 1.0. Those in Cluster B and Cluster C are the ones with a room for higher accuracy. Their accuracy tends to increase in Cluster B, and decrease in Cluster C in the ABSENT group. The fact that they are the minority group (27.1% together) indicates that progressive *-ing* suffers from the ceiling effect.

- Cluster A in third person *-s* is similar to Cluster A in progressive *-ing*. Their accuracy is too high for the investigation of its development. The proportion of the learners classified into Cluster A (87.9%) is even larger than that in progressive *-ing*. Ten learners in Cluster B and nine learners in Cluster C show an increasing and a decreasing trend respectively, while the PRESENT learners in Cluster D exhibit a U-shaped developmental pattern. Cluster E includes only one learner and is considered as an outlier.

One notable difference across the morphemes is the effect of L1 that is roughly absent except for the case of past tense *-ed*. The reason that L1 influence seems prominent in past tense *-ed* is perhaps due to the lack of proficiency difference between the ABSENT and the PRESENT group. Recall that a likely reason that L1 type did not influence the accuracy much in the longitudinal analysis of articles was that the ABSENT group was of generally higher proficiency than the PRESENT group. The same holds for plural *-s*, progressive *-ing*, and third person *-s*. The mean of the average window proficiency in the ABSENT group is significantly higher than that in the PRESENT group in all of the three morphemes [$t(467.624) = 7.181$, $p < 0.001$, Cohen's $d = 0.516$ for plural *-s*; $t(47.087) = 5.726$, $p < 0.001$, Cohen's $d = 1.051$ for progressive *-ing*; $t(45.747) = 4.092$, $p < 0.001$, Cohen's $d = 0.803$ for third person *-s*]. However, the difference in past tense *-ed* is non-significant [$t(15.012) = 1.279$, $p = 0.221$, Cohen's $d = 0.339$]. Thus, in past tense *-ed*, the proficiency of the ABSENT and the PRESENT groups roughly match. When the

proficiency is coincidentally controlled for in this way, the PRESENT group outperformed the ABSENT one. This is still interesting because in the pseudo-longitudinal view of the data (Figure 22 and 23), L1 influence was not obvious in past tense *-ed*. It emerged possibly because the mean of the average window proficiency in past tense *-ed* was Lesson 9 Unit 4 for the ABSENT group and Lesson 10 Unit 1 for the PRESENT group, where L1 influence is relatively clear judging from Figure 22.

4.4 Discussion

Summary. This chapter investigated the longitudinal development of morpheme use across 10 L1 groups and a wide range of proficiency levels by grouping up learners according to their developmental shapes. The results indicated that the developmental trajectories are relatively similar across morphemes and that significant individual variation is present within each cluster at the same time.

There were two research questions. The first research question asked if the longitudinal developmental of morpheme accuracy follows systematic patterns such as a power function or U-shaped development. The second addressed the extent of intra- and inter-learner variability.

General developmental patterns of grammatical morphemes. As to the first question, some commonalities in the developmental patterns were observed across the target morphemes. For example, the majority of the learners were always classified into the cluster that shows relatively flat development. This is interesting because when the data were viewed pseudo-longitudinally, at least the accuracy of articles and past tense *-ed* was on an increasing trend. However, when the same data were analyzed on an individual basis, the learners with increasing accuracy are not the largest cluster. This mismatch between the pseudo-longitudinal and longitudinal view of the data possibly stems from the length of coverage along the development by individual learners. One learner typically covered three

to four Lessons in 5 or 10 windows. This short length could have distorted the impression of the developmental slope of individual learners because the average accuracy rise over three to four lessons is small especially in relation to the large individual variation that will be discussed below. In other words, although on average learners' accuracy tends to increase from Lesson 1 to Lesson 16, the rise of accuracy is small if we focus only on three to four lessons, and given the large individual variation, the typical developmental shape can look flat. There were also learners who exhibited an increasing trend or U-shaped development. Although the developmental shapes vary, they are not completely at random because, as mentioned above, the cluster of flat or smooth development was always the largest of all. It was also observed that even when the number of clusters was manipulated, inverted U-shaped development did not emerge in any morpheme. This means that although U-shaped development is common at least in the development of articles, plural *-s*, and third person *-s*, there are few learners who follow the reverse pattern regardless of the morpheme. Thus, there are certain patterns of development that learners tend to follow or tend not to follow.

Commonalities in the developmental patterns across morphemes. As stated above, a very common developmental pattern was relatively flat development, and in this sense the developmental patterns are similar across morphemes. Also, in all the morphemes, there were many learners who showed an increasing or decreasing trend in accuracy. Thus, the clustering approach generally suggested that morphemes share at least some of the developmental patterns. However, at the same time, clustering also demonstrated that U-shaped development is only observed in articles, plural *-s*, and third person *-s*, but not in the other morphemes. There are a few possibilities to reconcile the apparently contradicting findings, but I postpone the discussion until more evidence is provided in the next chapter.

U-shaped development. Although U-shaped development has been observed in some L2 features as reviewed earlier, the present study is the first that empirically demonstrated it in the three morphemes mentioned above. An interesting question is what is behind the

U shape. I hypothesized that the accuracy decreases on the way to the bottom of U because accuracy and some other measure compete for cognitive resources. In other words, with the sacrifice of accuracy, some measure may instead show improvement during the same period as the accuracy decreases. To test this hypothesis, I looked into the difference between the regression-based article clusters in several measures including mean sentence length, Guiraud's index, essay scores assigned by the teachers, concreteness of the nouns modified by the articles (based on MRC Psycholinguistics Database; Wilson, 1988), frequency of the modified nouns (based on the same database), and error types (e.g., omission of indefinite articles, overgeneralization of definite articles, indefinite articles in place of definite articles), representing complexity, accuracy, abstractness, and other constructs that more or less require attention. The idea was that the U-shaped cluster should have a different value of the measures from, say, the linear increase cluster if those in the U-shaped cluster prioritize morpheme accuracy and their accuracy does not decrease while those in the linear positive cluster prioritize other measures with the sacrifice of morpheme accuracy. All in all, however, no clear difference was observed between clusters. It is, thus, uncertain what caused the U shape, and it should be investigated further with additional variables and/or at a finer resolution (e.g., focusing only on the essays included in the windows that are at the bottom of U instead of clusters).

Absence of the power law of learning. The observed shape largely did not lend support to the skill acquisition theory, which predicted a power function. There are two possible reasons. First, the error rate originally might not show a pure power law shape and may be difficult to identify the shape. The error rate is typically harder to fit a power function than reaction time (J. R. Anderson, 1995), and in Dekeyser (1997), too, the error rate curve tended to be bumpier than that of reaction time. Thus, the developmental pattern may not have been of a precise power function to start with. Second, the learners in the present study were not complete beginners, which may have led to the developmental shape

that is not clearly a power function. The power law of learning in the ACT model describes practice effect from the start of the learning process. Once learners reach a certain level of proficiency, their accuracy approaches the asymptote, and the amount of the increase per practice (or per window in the present study) becomes marginal. This may have prevented the identification of the shape of a power function.

Individual variation. Regarding Research Question 2 asking variability, a prominent feature of the present study is large individual variation. The clustering approach visually revealed large individual variation in the developmental shape. The large variation is as expected under the DST framework. As reviewed earlier, dependence on initial conditions as well as complete interconnectedness among the subsystems makes the development of dynamic systems sensitive to external as well as internal factors. Because learners vary with respect to various factors such as the input they have received or the working memory capacity, their overall development as well as the developmental paths naturally differ to a considerable degree. This is exactly what the present study revealed. The development did not necessarily mean increasing accuracy but could take various forms including U-shape. Additionally, even within the cluster with a similar developmental shape, there still appeared to be wide intra- and inter-learner variability.

Exploratory approach. The study also demonstrated usefulness of exploratory approaches. The clustering approach employed in the study was exploratory because the number of clusters was determined in a bottom-up manner by looking at the data. Exploratory approaches are often criticized due to their risk of overfitting (J. D. Long, 2012). However, it was possible only through clustering to reveal some of the developmental patterns like U-shaped or multiple forms of increasing accuracy (cf. van Geert & van Dijk, 2002). Whereas the above does not refute the risks associated with overfitting, it was meaningful to combine both an exploratory and a confirmatory approach to data analysis in the present study.

Limitations. As is always the case, this study is not without its limitations. First, because not all essays in the corpus have been error-tagged, the study might have failed to capture finer development of each learner which would have been possible if all the essays had been error-tagged. It is hoped that the essays currently not error-tagged will be done so in the future so that larger amount of information will be available on each learner. Second, related to it, the number of essays (or obligatory contexts) per learner might have been too small to fully capture the developmental pattern. As the larger number of essays become available per learner, we can gain the more reliable insight on the learners' development.

4.5 Conclusion

Based on a large-scale longitudinal error-tagged learner corpus, the present study investigated the longitudinal developmental pattern of morpheme accuracy. The clustering of learners according to their developmental patterns suggested that they tend to exhibit certain developmental patterns across all the morphemes. The study also quantitatively unveiled significant individual variation over the developmental patterns, as well as in the absolute accuracy.

Chapter 5: The Roles of L1 and Proficiency in the Longitudinal L2 Development of English Grammatical Morphemes

5.1 Introduction

The last chapter described the typical trajectories learners follow in the L2 development of English grammatical morphemes. The present chapter investigates the effects of L1 and proficiency on the development. Given the pervasive impact of L1 observed in Chapter 2 and 3 as well as in SLA literature in general (Jarvis & Pavlenko, 2007), and equally prevalent power of proficiency on L2 development (Thewissen, 2013), the two variables are hypothesized to play roles in the individual longitudinal development of grammatical morphemes as well. The study will also analyze whether the developmental trajectory varies across morphemes. As to L1, I hypothesize that the PRESENT learners tend to show a rapider increase of accuracy because they eventually end up at higher accuracy than the ABSENT learners. With respect to proficiency, the hypothesis is that higher proficiency learners tend to develop more flatly because they are more likely to hit the ceiling. As regards morphemes, since the last chapter revealed somewhat different patterns of development across morphemes (e.g., whether U-shaped development is observed or not), the hypothesis is that the longitudinal developmental pattern differs depending on morphemes, too. The specific research question addressed in this chapter is the following: Are the longitudinal developmental trajectories of individuals different depending on morphemes, learners' L1, and their proficiency?

5.2 Method

5.2.1 Data, Target Morpheme, L1 Groups, and Proficiency Levels

These are largely the same as the last chapter. Although the same six morphemes were the initial target, only a few of them were analyzed due to the scarcity of data. The specific target morphemes depended on the analysis, and will be clarified as we go on.

5.2.2 Data Analysis

The study conducted three kinds of analyses. The first analysis modeled the relationship between the clusters of development identified in the last chapter and the L1 and proficiency of the learners classified into the clusters. A significant association between the learners' L1 and the clusters they were classified into, for example, indicates that the developmental shape is partially predictable by L1, which I interpret to mean that L1 affects developmental trajectories.

While the first analysis predicts window-based clusters, the second and the third analyses target the essay level by regression modeling. Because windows used in the first analysis aggregate multiple essays, they may fail to capture more fine-grained developmental patterns that would have been observed if each essay had included a large number of obligatory contexts (cf. Section 4.3.2.1). For the sake of higher resolution, the proficiency level of each essay was directly specified in regression models, and the effects of L1 and proficiency were tested against the models.

5.3 Results

5.3.1 Testing the Effects of L1 and Proficiency by Predicting Cluster Membership

The first analysis tested the roles of L1 and proficiency by predicting cluster membership. I employed a multinomial logistic regression model that investigates whether learn-

ers' average proficiency level over the 10 windows and their L1 are significant predictors of the clusters learners were classified into. A multinomial logistic regression is an extension of logistic regression models and has more than two categorical variables as the response variable. In the present case, it estimates the effect of each predictor on the probability (odds) that learners are classified into each cluster compared to a baseline cluster. Average window proficiency (cf. Section 4.3.2.2) and learners' L1s were entered as predictors. The former was entered as a linear continuous variable on the assumption that the difference of, say, five Units at Lesson 1 have as much an impact as five Units at Lesson 10. Learners' L1s were entered as dummy variables. Because there were 10 L1 groups, nine dummy variables were employed, each representing one L1 group. The study used treatment contrasts, in which a one was given to a case if the learner's L1 is represented by the dummy variable and otherwise a zero was assigned. The type of contrasts remained the same for the rest of this chapter. Due to the small number of cases, it was not possible to reliably estimate the parameters of multinomial regression models for past tense *-ed*, progressive *-ing*, and third person *-s*. Therefore, the analysis only targeted articles and plural *-s*.

5.3.1.1 Testing the Effects in Articles

Given the high association between two types of clustering reported in the last chapter, we will focus on KmL clustering here. When a multinomial regression analysis was applied to three clusters of articles derived from KmL clustering, the result of Type II analyses of deviance indicated that L1 is not a significant predictor [$\chi^2(18) = 20.901$; $p = 0.284$], but Proficiency is [$\chi^2(2) = 6.210$; $p = 0.045$]. When L1 was collapsed into L1 type, the predictor remained non-significant [$\chi^2(4) = 4.462$; $p = 0.347$]. Because the residuals of a linear regression model predicting proficiency based on KmL clusters did not normally distribute, a Steel-Dwass test, a non-parametric equivalent to Tukey's HSD, was applied with an R script written by Aoki (2004) in order to identify which clusters include the

learners of higher proficiency than other clusters. The only significant difference was that learners in Cluster B tended to be more proficient than those in Cluster A [t = 2.530; p = 0.031]. The difference in the mean proficiency between the two clusters was four Units (mean proficiency level of Cluster B = Lesson 5 Unit 7; mean proficiency level of Cluster A = Lesson 5 Unit 3). This shows that the learners whose accuracy increases over the 10 windows tend to be of lower proficiency than those whose accuracy follows a U-shaped pattern. One possible reason for this is that those whose accuracy increases are still in the middle of the acquisition process so tend to be of lower proficiency, while those whose accuracy is U-shaped have already acquired articles or have stabilized and so tend to be of higher proficiency.

In sum, proficiency was a significant predictor, while L1 and L1 type were not. Since L1 is a more fine-grained category than L1, the rest of this chapter adopts L1 and not L1 type to investigate L1 influence in the hope that it will capture finer differences between L1 groups that are hidden when L1s are aggregated into L1 type.

5.3.1.2 Testing the Effects in Plural *-s*

Let us now investigate whether cluster membership is related to learners' L1 and proficiency in plural *-s*. I will focus on specific L1s rather than L1 type (PRESENT vs ABSENT). Similarly to the above, a multinomial regression model was built with cluster membership as the dependent variable and L1 and proficiency as independent variables. Independent variables were dropped in a stepwise fashion according to Type II analyses of deviance with the criterion of $p < 0.05$. The result indicated that Proficiency was non-significant and thus dropped from the model [$\chi^2(3) = 3.347$; $p = 0.341$]. L1 turned out to be significant [$\chi^2(27) = 53.793$; $p = 0.002$].

In order to identify the specific association between learners' L1 and KmL clusters in plural *-s*, the learners were cross-tabulated between the two variables. Table 26 is the con-

tingency table showing the distribution of the learners. As in an earlier table, a superscript H means that the value of the cell is larger than expected at $p < 0.05$ and a superscript L shows the value is smaller than expected based on a residual analysis. Below the table is the result of a χ^2 test examining independence of the distribution of the learners, and Cramer's V, an effect size measure. Since multiple cells have the expected value of below 5, the p-value was computed by Monte Carlo simulation (Gries, 2013). The superscripts and Cramer's V should be interpreted with caution because they assume that all cells have expected values of 5 or above. The table shows that it is mostly L1 Chinese and L1 Russian learners who behave differently from expected. L1 Chinese learners tend to avoid flat development and favor the patterns of relatively linear increase and decrease. L1 Russian learners exhibit a reverse pattern: They tend to exhibit flat development compared to the other L1 groups, and avoid linear increase or decrease patterns. Given the small effect size (Cramer's $V = 0.158$; Oswald & Plonsky, 2010), however, the results should be kept as suggestive rather than conclusive.

Overall, the effect of L1 and proficiency is not strong in determining the developmental shape. Some predictors turned out to be significant possibly due to the high frequency of articles and plural *-s* and the large number of learners who have 10 or more windows (cf. Table 25) as its result. I will now turn to an essay-level analysis.

5.3.2 Mixed-Effects Models

Background of employing regression modeling. The present and the next section aim at analyzing the effects of L1 and proficiency at a more fine-grained level. The unit of the analysis in the first analysis has been cluster based on windows. The use of windows was convenient particularly for visualization, but as it was of lower resolution than essays, it may have concealed the pattern that could have been observed at the essay level. The issue is more serious when non-error-tagged essays fall within the coverage of a window

Table 26

Distribution of Learners Between KmL Clusters in Plural -s

KmL Clustering		L1										Total
		Brazilian	Chinese	German	French	Italian	Japanese	Korean	Russian	Spanish	Turkish	
A												
	# Learners	74	66^{L}	29^{H}	12	19	3	1	142^{H}	17	3	366
	%	50.0%	35.9%	67.4%	54.5%	55.9%	50.0%	20.0%	64.3%	45.9%	37.5%	51.7%
B												
	# Learners	33	54^{H}	7	7	6	1	2	40^{L}	12	2	164
	%	22.3%	29.3%	16.3%	31.8%	17.6%	16.7%	40.0%	18.1%	32.4%	25.0%	23.2%
C												
	# Learners	28	37	3	2	5	0	1	28	5	2	111
	%	18.9%	20.1%	7.0%	9.1%	14.7%	0.0%	20.0%	12.7%	13.5%	25.0%	15.7%
D												
	# Learners	13	27^{H}	4	1	4	2^{H}	1	11^{L}	3	1	67
	%	8.8%	14.7%	9.3%	4.5%	11.8%	33.3%	20.0%	5.0%	8.1%	12.5%	9.5%
Total												
	# Learners	148	184	43	22	34	6	5	221	37	8	708

Note. $\chi^2 = 53.135$; $pMC = 0.001$; Cramer's $V = 0.158$

H = significantly more learners than expected at $p < 0.05$; L = significantly fewer learners than expected at $p < 0.05$

and the aggregated essays are not adjacent to each other in terms of the level. It aggravates differences between windows and makes tracking the development somewhat tricky. Analyses should therefore directly specify the proficiency level at which each essay was written so that we can follow the developmental pattern at that level.

Also, the above tested the effects indirectly through examining the relationship between clusters and the characteristics of those who belong to the clusters. This, however, is identical to assuming a uniform developmental pattern of learners within each cluster, although in reality we saw large within-cluster individual variation. It is desirable to perform a more direct test independently of clusters.

Two types of regression modeling employed in the study. To fill these gaps, the present study employed mixed-effects regression models and generalized additive models (GAMs). An advantage of mixed-effects models over GAMs is that they take into account the dependency of data within learners (e.g., Some learners tend to achieve higher accuracy of articles than others.) and tests of predictor effects are more accurate in general. This will be discussed in detail below. On the other hand, an advantage of GAMs over mixed-effects models is that they can better handle nonlinear effects of predictors. It is important because we observed clear nonlinearity in the pseudo-longitudinal as well as longitudinal development of morphemes in the previous chapters.

5.3.2.1 Description of Mixed-Effects Models

Introduction. A mixed-effects model is an extension of regression models and is suited for analyzing hierarchical data (Baayen, 2008; Baayen, Davidson, & Bates, 2008; Cunnings, 2012). The following example and explanation are largely based on Hox (2002). Let us suppose that we want to investigate the effect of students' social economic status (SES) on their scores of an exam. We sampled 50 classes from different schools, each with 20 pupils. The data have a multilevel structure in that pupils are nested in classes. The

total sample size of the data is 1,000. Now, simply regressing pupils' test scores against their SES (operationalized as, say, parents' income) can be problematic as it likely violates the assumption of independent observation. Pupils at the same school tend to be similar to each other because some schools attract pupils at a particular SES level and because pupils at the same school tend to share same experiences by attending the same school. Consequently, pupils at the same school have similar values of the variables (e.g., SES, test score) compared to those at other schools. This means an observation is not independent of others since the value of the response variable (i.e., test score) can be more or less predicted based on which school the pupil attends. Ignoring the assumption of independence leads to unjustifiably small standard errors, which in turn invites spurious "significant" results.

A mixed-effects model takes into account the school difference by allowing the intercept and the slope of the regression model to vary across schools. Allowing the intercept to vary between schools is called *random-intercepts*, and allowing the slope to vary between schools is called *random-slope*. By adding these random-effects, it is possible to model the relationship between pupils' SES and their test scores when the variance at school level is accounted for. Intuitively, this is similar to having one regression model per school that predicts pupils' test scores based on their SES. More precisely, mixed-effects models partition the total variance into between-school variance and within-school variance. The former refers to the variance explained by school differences (e.g., School A achieves higher marks overall than School B), whereas the latter refers to the variance within school (e.g., some pupils score higher than others in the same school). Predictors can be entered at both levels. At the level of pupil, SES might be a good predictor explaining within-school variance. At the level of school, average class size or the number of classes one teacher has to teach might be able to explain intercept differences (i.e., absolute differences in exam scores) between schools. Notice that the values of the school-level predictors remain the same across pupils in the same school. This is why they explain between-school variance

and not within-school variance, unlike pupils' SES which aims to explain within-school variance. The slope of SES can vary across schools as well. This means that the effect of SES on pupils' performance can depend on schools. This slope variance can be explained by entering into the model an interaction between pupil-level predictors and school-level predictors. In this example, school is called a *random-effect* variable, and SES a *fixed-effect* variable. The term *mixed-effects* stems from the feature of the model that the two effects are put into a model at the same time.

Mixed-effects models in longitudinal data analysis. Mixed-effects models are useful in longitudinal data analysis as well (Collins, 2006; Hox, 2002; J. D. Long, 2012; Meunier & Littré, 2013; Singer & Willett, 2003). Here, instead of schools, we have individuals, and instead of pupils, we have data points. In the present study, essays are nested within learners (as in pupils were nested in schools in the previous example), and the total accuracy variance between essays is divided into between-learner, learner-level variance (e.g., some learners tend to be more accurate in morpheme use than others in general) and within-learner, essay-level variance (e.g., a learner's accuracy in each essay changes as s/he develops). For examples of predictors, learners' L1 might explain between-learner variance, and the number of essays the learner has written might explain within-learner variance. Note that the value of L1 does not change within individuals and thus explains between-learner variance, whereas the number of essays a learner has written changes as s/he progresses and thus explains within-learner variance.

Illustration of random-effects. Figure 32 visualizes the point of random-effects in the present context. The vertical axis represents the TLU score, and the horizontal axis represents the number of essays a learner has written (EssayNum). Note that "time" in longitudinal development is operationalized by the essay number in the present study. Within-learner longitudinal development is represented as the effect of EssayNum on TLU scores because it is a phenomenon in which a learner's TLU scores change as the value of Es-

sayNum changes (i.e., as the learner writes essays). Here, let us suppose that a number of hypothetical learners wrote five essays and the accuracy development of each learner is depicted by one line in the upper left panel of Figure 32. In this case, all the learners consistently increase their accuracy at exactly the same rate over the five essays. In other words, the slope of EssayNum is constant across learners. What differs is the absolute accuracy of each learner. Some learners start at 0.4 and others at 0.7. These differences in the absolute accuracy between learners should be taken into account in modeling and that is what random-intercepts do. Random-intercepts allow learners to be of different overall accuracy. The accuracy of the learners at the intercept (where EssayNum = 0) is {0.72, 0.70, 0.68 . . . 0.36, 0.34}, and the variance is 0.014. This is the variance between learners at the intercept, and one question we can ask is how much of it can be explained by the predictors. Let us say that the learners, in fact, belonged to two L1 groups, L1 German and L1 Spanish, and in the upper right panel L1 German learners are represented by blue dashed lines and L1 Spanish learners by red solid lines. Here, L1 explained some portion of the variance in the random-intercept. Now the between-learner variance at the intercept should be computed within each L1 group, and the value (0.004 for both groups) is much smaller than the original variance (0.014). The reduction is achieved by taking L1 into account. The lower two panels illustrate an example of random-slope. Let us suppose here that all the learners start at the same accuracy, but their rate of development varies. This is what by-EssayNum random-slope means because the effect of EssayNum varies across learners. Introducing random-slope takes the difference into account. Again, this difference can be explained by L1. However, this time, it is not L1 that explains the difference in the rate of development but the interaction between L1 and EssayNum. L1 as a predictor only allows the overall accuracy to vary across L1 groups while other variables are held constant. This was fine to explain random-intercepts because random-intercepts only take care of the overall accuracy and are not related to EssayNum. However, the varying rate of development

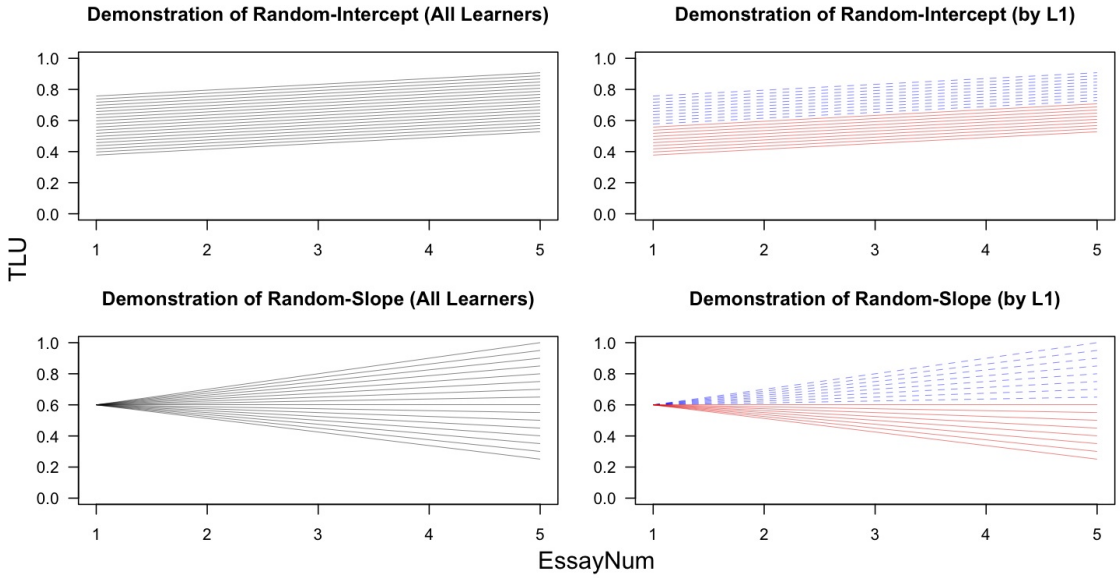


Figure 32. Illustration of Random-Effects

means varying effect of EssayNum depending on learners' L1. This type of effect can only be captured by cross-level interactions of predictors, which are the interactions between learner-level and essay-level predictors (Hox, 2002). Introduction of the EssayNum-L1 interaction allows the effect of EssayNum to vary across L1, and this is exactly what we want in order to capture the random-slope variance. As will be explained later, an interest of the present study is how much increase of variance in the random-slope of EssayNum can be observed when the EssayNum-L1 interaction is removed from the model, and similarly, how much increase can be observed when the interaction between EssayNum and learners' proficiency is removed. Figure 32 is, of course, a highly idealized scenario and the real data are much messier, especially because both random-intercept and random-slope are entered simultaneously. But hopefully the point is clear as to what random-effects mean and how they can be explained.

The study employed mixed-effects logistic regression models in order to model the relationship between accuracy, proficiency, L1, and morpheme (cf. Barr, 2008; Dixon,

2008; Jaeger, 2008).

5.3.2.2 Model Specification

Target data. The analysis targets the error-tagged essays written by the learners who wrote at least 10 error-tagged essays (but not necessarily windows) that include one or more obligatory contexts or the instances of overgeneralization errors of one of the target morphemes. Not all the essays were targeted in order to reliably account for individual variation (i.e., difference in the intercept across learners). Because progressive *-ing* and third person *-s* exhibited an extremely high accuracy development in the clustering approach in the last chapter, they were excluded from the analysis. So was possessive *'s* due to small data size. Therefore, the target morphemes were articles, past tense *-ed*, and plural *-s* in the present analysis. In total, the dataset consists of 89,448 observations over 2,234 learners. Because 30,086 cases had zeros for obligatory contexts and both types of errors, only 59,362 observations contributed to the model.

The purpose of the analysis here is to investigate whether the developmental patterns are affected by L1 and/or proficiency. This is achieved by comparing the models with and without the variables of interest (J. D. Long, 2012). The maximal model (i.e., a model with all possible predictors and the largest random-effects structure; Barr, Levy, Scheepers, & Tily, 2013) includes four fixed-effects variables, one random intercept, and two random slopes. This model is referred to as Model 1. The number of correct suppliance was considered as the number of successes and the number of omission and overgeneralization errors was the number of failures. Four fixed-effects were entered; proficiency (Proficiency), the number of essays the learner has written up to the point (EssayNum), L1 (10 levels, one for each L1 group with L1 Brazilian as the reference group), and morpheme (three levels, one for each morpheme with articles as the reference level). A binomial distribution was assumed.

Independent variables. Proficiency was represented by the average level of the essays written by the learner in terms of Unit. The variable is meant to capture the variance of accuracy between learners. In order to account for intra-learner variance, the present model employed essay number. All the essays within individuals were chronologically ordered regardless of whether they are error-tagged, and one was assigned to the first essay of the learner, two to the second essay, and so forth. This variable represents the progress of learners and is meant to explain the accuracy development over the essays within each learner. Although the developmental pattern of each morpheme is unlikely to be linear, polynomial terms of Proficiency and EssayNum were not included since their introduction invited errors in the `lmer` function in the lme4 package (Bates, 2010) in R. Also, in order to avoid a convergence problem, both Proficiency and EssayNum were standardized to the values with the mean of zero and the standard deviation of one (i.e., z-score). The mean and the standard deviation of Proficiency were 37.523 and 23.456 respectively, and those of EssayNum were 17.801 and 12.402.

Interaction terms. In addition to the three predictors above, their two-way interactions were also included. The Proficiency-EssayNum interaction tests whether the longitudinal developmental pattern (EssayNum) differs across the overall proficiency. The hypothesis is that the ceiling effects leads to a smaller accuracy rise in learners of higher proficiency. This interaction is also hoped to capture the by-EssayNum random-slope for learners, as was explained earlier. The Proficiency-Morpheme interaction tests whether the rate of accuracy increase over proficiency (i.e., pseudo-longitudinal development) differs across morphemes. It can be the case, for example, that the rate is slower in plural *-s* because its accuracy is high overall and has little room for accuracy increase (cf. Figure 22). The Proficiency-L1 interaction tests whether the rate of pseudo-longitudinal development varies across L1 groups. The interaction was complex in the aggregated regression model in the last chapter, and I expect the interaction turns out to be significant in the present

model as well. The EssayNum-Morpheme interaction tests if the rate of the longitudinal developmental differs across morphemes. As mentioned at the beginning of this chapter, this is also likely given that the clustering observed some differences in the longitudinal developmental patterns across morphemes. The EssayNum-L1 interaction similarly tests whether the longitudinal developmental pattern varies across L1 groups. This interaction attempts to capture the by-EssayNum random-slope for learners. By the random-effect structure explained later, the model allows the effect of EssayNum (i.e., longitudinal developmental pattern) to vary across individuals. The interest here is the extent to which allowing the effect of EssayNum to vary across L1 groups decreases the variance. Finally, the Morpheme-L1 interaction tests whether the effect of L1 differs across morphemes. Since L1 Type was significant in the regression models in the pseudo-longitudinal analysis, I expect this interaction to be significant as well.

Random-effects. For the random-effects structure, Learner was the only factor entered as a random-effect. Adding the by-learner random-intercept allowed the accuracy to vary across individual learners. EssayNum and Morpheme were entered as random-slopes. The by-EssayNum random-slope allows the rate of development to vary across learners. It is very likely that some learners show a rapid increase of accuracy, while others stay flat, as was shown in Figure 20, 28, and 29. The by-Morpheme random-slope allows the accuracy difference between morphemes to vary across individuals. Certain learners, for example, can be more accurate at using articles than using past tense *-ed*, whereas others might show the reverse pattern.

Multimodel inference. The effects of L1, Proficiency, and Morpheme on the developmental patterns were tested through multimodel inference (D. R. Anderson, 2004). In this approach, we build and compare multiple models with and without the parameters of interest. Through the process, we can assess the importance of the predictors. Let us call the maximal model that was just described Model 1. Four other models were constructed.

Model 2 is the same as Model 1 except that it did not include the EssayNum-L1 interaction. The comparison between Model 1 and Model 2 tells us whether and the extent to which L1 explains the variance in the slope of EssayNum. Model 3 is the same as Model 1 except that it did not include the EssayNum-Proficiency interaction. Similarly, it tests whether and the extent to which Proficiency explains the variance in the slope of EssayNum. Model 4 is the same as Model 1 but did not include the EssayNum-Morpheme interaction. It examines whether and the extent to which the slope of EssayNum varies across Morpheme. These together test the effects of the three predictors on the longitudinal developmental patterns. In addition, the Null Model with only the random-effects structure and without any fixed-effects predictors was constructed as a reference to see how much variance is reduced by the fixed-effects in Model 1 through 4.

5.3.2.3 Pros and Cons of Mixed-Effects Models

Impact of the lack of polynomial terms. Since failure to include polynomial terms of Proficiency and EssayNum has a nontrivial impact, I will elaborate the potential effects here. Including polynomial terms (e.g., $Proficiency^2$, $EssayNum^3$) allows the relationship between the predictors and TLU scores (in the logit scale) to be nonlinear. Given that we observed nonlinear relationships between Proficiency and TLU scores in logit in Chapter 3, and non-straightforward shapes of development in Chapter 4, it is expected that entering polynomial terms of Proficiency and EssayNum would make a better model. A consequence of this is explained with Figure 33. In all the panels, the horizontal axis represents essay number and the vertical axis represents TLU scores. The examples given below are simplified mainly in two ways. First, I assume that each essay included a large number of obligatory contexts to allow reliable calculation of TLU scores. The illustrative task is to estimate the effect of L1 and proficiency on the assumption that all data points are equally reliable. Second, nonlinearity discussed below is the nonlinearity between EssayNum and

TLU scores in the probability scale. However, as mentioned in Chapter 2, the logistic regression analysis models the relationship between variables in a logit scale so that the model does not predict the value below zero or above one. That is, since TLU scores cannot be less than zero or above one, we need to apply a function to the expected value so that the predicted value fits between zero and one. What is (non)linear is the relationship between EssayNum and TLU scores in the logit scale (i.e., the original scale before the function is applied) but not in the probability scale that corresponds to TLU scores. The point is illustrated in the upper left panel in the figure. In this panel, all the green solid lines demonstrate linear effects of EssayNum on logit-scaled TLU scores drawn in the probability scale, and all the purple dashed lines represent nonlinear effects. Note that the solid lines are not necessarily linear in the probability scale, and their slope tends to become smoother as they get closer to the TLU score of one. This is in order to avoid the predicted line to go above one. Notice also that although they are not necessarily linear in TLU scores, they are always monotonic, which means that their value either consistently increases or consistently decreases. There is no solid line that is U-shaped, for example. Thus, linear effects in the logit scale have monotonous impacts in the probability scale. This is not the case for nonlinear effects. The dashed lines show that when the effects are nonlinear in the logit scale, the accuracy in the probability scale goes up and down. This indicates that U-shaped developmental patterns observed in the last chapter are nonlinear in the logit scale as well, which suggests that the true relationship between EssayNum and TLU scores can be nonlinear in the logit scale used in logistic regression models. Below, I set aside this complexity and illustrate the outcome of the absence of polynomial terms using the probability scale.

The upper right panel shows the data used in the example. J1 and J2 represent the accuracy scores of two L1 Japanese learners, while S1 and S2 represent those of two L1 Spanish learners. They are hypothetical learners and are used for the purpose of illustration only.

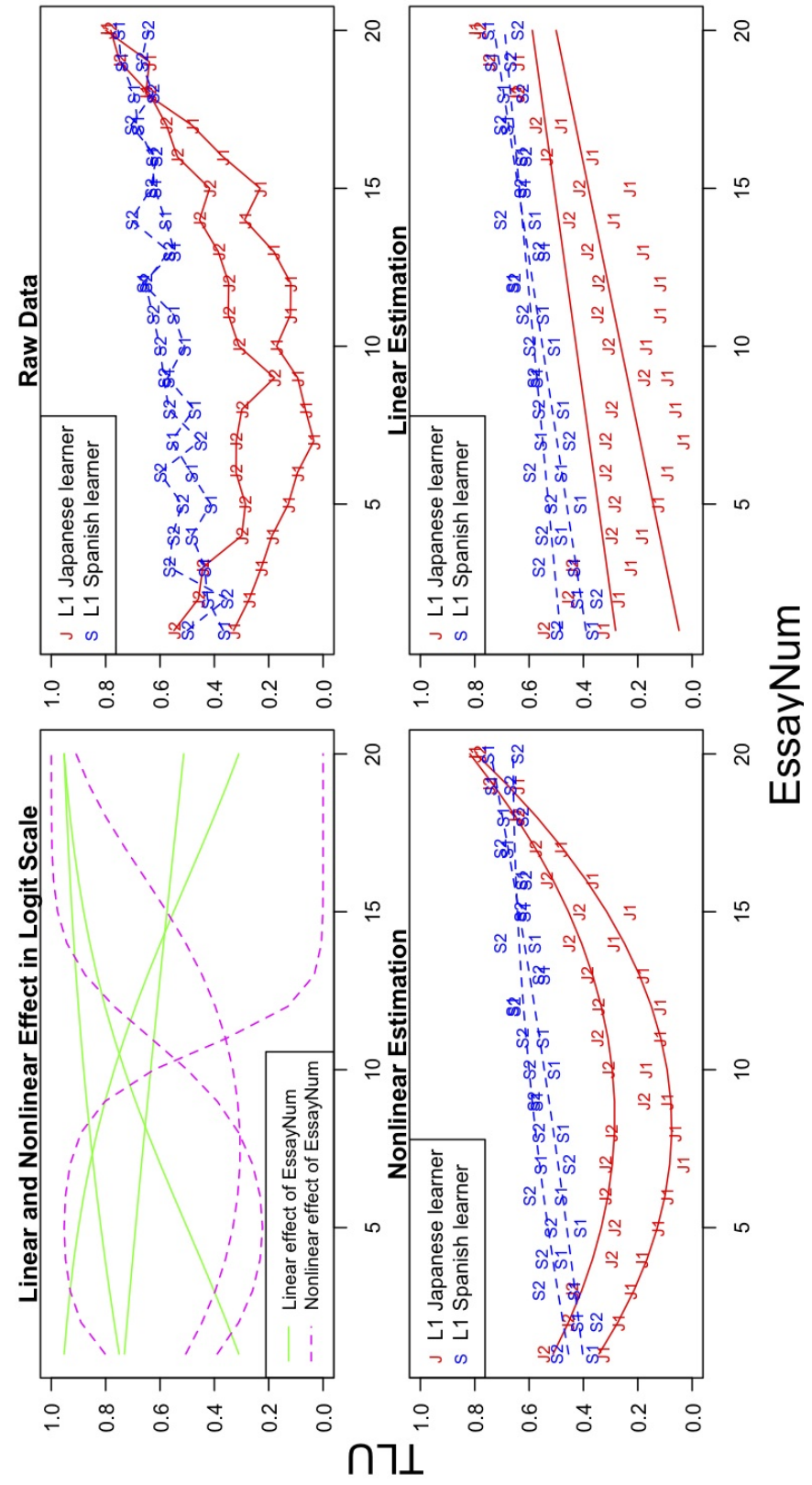


Figure 33. Illustration of Random-Effects

The red solid lines indicate the developmental patterns of the Japanese learners, and the blue dashed lines indicate those of the Spanish learners. Notice that the two developmental patterns of Japanese learners are better characterized by U-shaped development than linear development, while those of Spanish learners show the reverse tendency. The point of the illustration is how their differences are identified or fail to be identified with or without polynomial terms in the mixed-effects model. The lower left panel demonstrates a case where both the quadratic and the linear terms of EssayNum were entered into the model. That is, for a given learner, $TLUscore_i = \beta_0 + \beta_1 \times EssayNum_i^2 + \beta_2 \times EssayNum_i + \varepsilon_i$, where i is the essay number and ε is a normally-distributed error term. β_0 through β_2 are estimated from data. Here, the developmental pattern of each learner is modeled by a quadratic function (U-shaped or inverted U-shaped). We can see that Japanese learners are successfully modeled by the function and the difference between raw data points and predicted data points is not very large. Spanish learners are modeled by quadratic functions as well. Since their observed developmental pattern is almost linear, the estimate (or coefficient) of the quadratic term (β_1 above) is closer to zero, which makes the quadratic function close to linear. Their data, too, are well modeled in this case. This shows that quadratic terms combined with linear terms can model both U-shaped and linear development. Now, when testing the effect of L1, what the model does is to see if the β_1 value varies across learners' L1. In this example, the coefficients of J1, J2, S1, and S2 are 0.005, 0.004, 0.000, and -0.001 respectively. It can be seen that the Spanish learners show more linear development than the Japanese learners because their values are closer to zero. Although statistical tests are not possible due to the small data size, Japanese learners appear to show more U-shaped development (i.e., have larger values of β_1) than Spanish learners, which indicates that L1 affects developmental patterns.

The mixed-effects models used in the present study, however, were not able to incorporate quadratic terms and had to assume linear effects of Proficiency and EssayNum. The

case is illustrated in the lower right panel. Here, linear relationships between EssayNum and TLU scores are estimated so as to minimize the squared difference between the observed TLU scores and predicted TLU scores[6]. In other words, the following equation is fit and the β_0 and the β_1 are estimated; $TLUscore_i = \beta_0 + \beta_1 \times EssayNum_i + \varepsilon_i$. We can see that the difference between the drawn lines and the raw data is small in Spanish learners because their accuracy development is approximately linear. However, the same difference in Japanese learners is large because linear shapes are imposed on what appears to be U-shaped. When testing the effects of L1 under this assumption, the slope (β_1 above) is compared. The coefficients of J1, J2, S1, and S2 are 0.024, 0.016, 0.018, and 0.011 respectively. This time, the difference in the values between Japanese and Spanish learners is less clear compared to the earlier example where quadratic functions were fitted. Therefore, absence of polynomial terms might obscure and interfere with the tests of predictors. This is the reason that I later employed a GAM that allows the concerned relationship to be more flexible.

Defense of the mixed-effects approach. Nevertheless, mixed-effects models should not be abandoned in the present study for the sake of higher resolution (cf. Section 4.3.2.1). Also, compared to the GAM discussed later, taking into account data dependency within learners in the mixed-effects approach can prevent spurious significant results or rightly find significant results when GAMs fail to. GAMs in the present case violate the assumption of independence (i.e., all the observations have to be independent of each other). The assumption is not uncommon in statistical models (Crawley, 2013; Field, 2012). The mixed-effects model, as explained earlier, can handle the correlation among subjects (or among schools in the example earlier).

[6]The algorithm is more complicated in logistic regression. See, for instance, Myung (2003) for the details. The idea is the same.

5.3.2.4 Results of Mixed-Effects Models

Model selection. The lme4 package (Bates, 2010) in R was used to run the analysis, and maximum likelihood estimation was employed to allow model comparison based on likelihood ratio tests (J. D. Long, 2012). Table 27 compares the five models. AIC stands for Akaike Information Criterion and is a measure of predictive accuracy. The smaller the value, the more plausible the model is. AIC shows that Model 2, the model without the EssayNum-L1 interaction, is the best model, closely followed by the maximal model (Model 1). Likelihood ratio tests agree with AIC. They indicate that there is no significant difference in the goodness of fit between Model 1 and Model 2, but Model 1 has a better fit than Model 3 and Model 4. They also suggest that Model 1 through Model 4 are all better than the Null Model, which means that Morpheme, L1, Proficiency, EssayNum, and their interactions significantly account for the variance in the data. All things considered, Model 1 and Model 2 are of roughly equal goodness, and have better fits than Model 3 and Model 4. This means that the EssayNum-L1 interaction does not help to predict accuracy, but the EssayNum-Proficiency interaction and the Essay-Num-Morpheme interaction do. It in turn means that the developmental patterns within individuals cannot be claimed different across L1 groups but differ across learner' proficiency and across morphemes. We will focus on Model 2 hereafter.

Interpreting random-effects. Table 28 presents the random-effects structure of the mixed-effects models. For a reference purpose, it also shows the random-effects of the Null Model. Notice that the values are generally large compared to the fixed-effects terms discussed later (Table 29), which indicates large individual variation in the overall accuracy of articles (by-learner random-intercepts), accuracy difference between articles and the other morphemes (by-Morpheme random-slopes), and the rate of development (by-EssayNum random-slope). Note also that a feature of mixed-effects models called *shrinkage* (Baayen,

Table 27

Comparison of Mixed-Effects Models

			Likelihood ratio test against			
Model	Model description	AIC	Model 1		Null Model	
Model 1	Maximal model	72,607			$\chi^2(54) = 2462.675$	$p < 0.001$
Model 2	EssayNum-L1 interaction excluded	72,603	$\chi^2(9) = 14.210$	$p = 0.115$	$\chi^2(45) = 2448.465$	$p < 0.001$
Model 3	EssayNum-Proficiency interaction excluded	72,614	$\chi^2(1) = 9.225$	$p = 0.002$	$\chi^2(53) = 2453.450$	$p < 0.001$
Model 4	EssayNum-Morpheme interaction excluded	72,614	$\chi^2(2) = 11.256$	$p = 0.004$	$\chi^2(52) = 2451.419$	$p < 0.001$
Null Model	No fixed-effects	74,961	$\chi^2(54) = 2462.675$	$p < 0.001$		

2008) complicates the interpretation of the absolute value of random-effects. It is thus better to interpret the values in comparison to the other values. The values of the Null Model should serve as a reference point against which the effects of predictors in Model 2 are tested. The Null Model tells us that the standard deviation of the random-intercept is 0.758, which is how disperse the learners are in terms of absolute accuracy in logit scale. The random-slope for past tense *-ed* is 1.140 and that of plural *-s* is 1.071. They represent individual differences in the accuracy difference between articles and the morphemes. The by-EssayNum random-slope is 0.327, which represents individual differences in the rate of development (cf. the lower left panel of Figure 32). When the values in Model 2 are examined, they are generally lower than those in the Null Model. This means that the predictors (Morpheme, L1, Proficiency, EssayNum, and their two-way interactions apart from the EssayNum-L1 interaction) explain the variance to a certain degree. However, the extent that they account for the variance differs across random-effects terms. The random-slope for plural *-s*, for example, shows a relatively large drop from 1.071 in the null model to 0.510 in Model 2. This means roughly half of the variance in the accuracy difference between articles (reference-level morpheme) and plural *-s* is explained by L1 and proficiency (i.e., L1-Morpheme and Proficiency-Morpheme interactions). In other words, L1 and proficiency affect the accuracy difference between articles and plural *-s*, which is also shown by the fixed-effects structure explained later. On the other hand, L1 and proficiency explain little variance in the random-slopes of past tense *-ed* and EssayNum. This means that the two predictors hardly affect the accuracy difference between articles and past tense *-ed*. The by-learner random-intercept representing individual differences in the overall accuracy of articles seems to fall somewhere in between.

Effect of morpheme. Let us now turn to the fixed-effects part (Table 29). The main effect of Morpheme suggests that the accuracy of both past tense *-ed* and plural *-s* is generally higher than articles. Their extent, however, depends on other factors because Mor-

Table 28

Random-Effects Structure of the Mixed-Effects Models

		Model 2	Null Model
Factor	Random Effects	SD	SD
Learner	Intercept	0.501	0.758
	Morpheme		
	Past tense *-ed*	0.919	1.140
	Plural *-s*	0.510	1.071
	EssayNum.Standardized	0.297	0.327

pheme participates in multiple interactions. This therefore only holds for L1 Brazilian learners at the mean proficiency level after having written the mean number of essays. The Morpheme-L1 interaction shows that the accuracy difference between articles and past tense *-ed* is smaller in L1 Chinese, German, and Spanish learners than in L1 Brazilian learners, and is larger in L1 Japanese and Russian learners, possibly because Japanese and Russian lack articles. In fact, in L1 Spanish learners, the accuracy order is flipped and articles mark a higher accuracy than past tense *-ed* on average. The difference between articles and plural *-s* is smaller in L1 Chinese, German, French and Italian groups and larger in L1 Russian learners than L1 Brazilian learners (baseline level). The fact that the random-slope for past tense *-ed* is 0.919 and its estimate in the fixed-effects structure is 0.670 means that, at least for L1 Brazilian learners at the mean proficiency level and the mean essay number, the standard deviation of the inter-learner accuracy difference between articles and past tense *-ed* overweighs the mean accuracy difference between the two morphemes, which in turn indicates that, although past tense *-ed* tends to be more accurate than articles, the accuracy order between the two morphemes heavily depends on learners. This is not the case for the difference between articles and plural *-s*, however. Because its random-slope

(0.510) is much smaller than the fixed-effects coefficient (0.946), plural *-s* can be claimed to be usually (though not necessarily always) more accurate than articles, irrespective of learners.

Effects of L1, proficiency, and within-learner development. When L1 is looked at, many L1 groups (e.g., L1 German and L1 Italian) mark higher accuracy in articles than L1 Brazilian, and others (e.g., L1 Korean and L1 Japanese) mark a lower accuracy. It interacts with Morpheme as explained above, and the difference between L1 groups can vary across morphemes. Both Proficiency and EssayNum are positive and significant, which means that accuracy increases as proficiency goes up and as learners write more essays. However, they interact with other variables. The Proficiency-Morpheme interaction indicates that the effect of proficiency is weaker in past tense *-ed* and plural *-s* than articles. That is, the slope of proficiency tends to be flatter in the two morphemes than in articles. Similar is the case for the EssayNum-Morpheme interaction. The rate of accuracy increase within individuals is slower in the two morphemes than in articles. Proficiency is engaged in an interaction with L1 as well. The interaction says that, in pseudo-longitudinal development, L1 Chinese learners increase their accuracy more slowly than L1 Brazilian learners, and L1 Russian learners raise their accuracy faster than L1 Brazilian learners. The Proficiency-EssayNum is significant and negative, indicating that the accuracy rise by EssayNum becomes smaller as proficiency goes up. This means that the accuracy rise within individuals over essays shrinks as their overall proficiency goes up, assuming that the number of essays written are constant. In other words, the developmental patterns within individuals are affected by proficiency. The comparison between the by-EssayNum random-slope and EssayNum in the fixed-effects part is again interesting. The fact that the random-slope (0.297) is larger than the fixed-effect (0.140) means that, at least for L1 Brazilian learners on articles at the mean proficiency, the standard deviation of the inter-learner accuracy change over essays is larger than the mean accuracy change over essays. This means that the development can

Table 29

Fixed-Effects Structure of Model 2

Parameter		B	SE
Intercept		1.692 ***	0.032
Morpheme			
	Past tense *-ed*	0.670 ***	0.089
	Plural *-s*	0.946 ***	0.041
L1			
	Chinese	0.211 ***	0.040
	German	0.579 ***	0.068
	French	0.271 **	0.083
	Italy	0.503 ***	0.075
	Japanese	-0.355 **	0.122
	Korean	-0.431 **	0.160
	Russian	-0.065	0.051
	Spanish	0.221 **	0.078
	Turkish	-0.217	0.160
Proficiency.Standardized		0.247 ***	0.036
EssayNum.Standardized		0.140 ***	0.012
Morpheme : L1			
	Past tense *-ed* : Chinese	-0.416 ***	0.108
	Plural *-s* : Chinese	-0.326 ***	0.051
	Past tense *-ed* : German	-0.595 **	0.182
	Plural *-s* : German	-0.221 *	0.095
	Past tense *-ed* : French	-0.268	0.238
	Plural *-s* : French	-0.407 ***	0.109
	Past tense *-ed* : Italian	-0.310	0.210
	Plural *-s* : Italian	-0.438 ***	0.101
	Past tense *-ed* : Japanese	1.001 *	0.435
	Plural *-s* : Japanese	0.315 .	0.172
	Past tense *-ed* : Korean	0.507	0.389
	Plural *-s* : Korean	0.301	0.214
	Past tense *-ed* : Russian	0.664 ***	0.140
	Plural *-s* : Russian	0.553 ***	0.067
	Past tense *-ed* : Spanish	-0.732 ***	0.191
	Plural *-s* : Spanish	0.006	0.097
	Past tense *-ed* : Turkish	0.270	0.453
	Plural *-s* : Turkish	0.109	0.192
Morpheme : Proficiency.Standardized			
	Past tense *-ed* : Proficiency.Standardized	-0.125 **	0.049
	Plural *-s* : Proficiency.Standardized	-0.131 ***	0.022
Morpheme : EssayNum.Standardized			
	Past tense *-ed* : EssayNum.Standardized	-0.091 **	0.034
	Plural *-s* : EssayNum.Standardized	-0.044 **	0.016
L1 : Proficiency.Standardized			
	Chinese : Proficiency.Standardized	-0.175 ***	0.045
	German : Proficiency.Standardized	0.013	0.063
	French : Proficiency.Standardized	-0.026	0.078
	Italian : Proficiency.Standardized	-0.029	0.074
	Japanese : Proficiency.Standardized	-0.060	0.105
	Korean : Proficiency.Standardized	-0.216 .	0.125
	Russian : Proficiency.Standardized	0.170 ***	0.043
	Spanish : Proficiency.Standardized	0.088	0.091
	Turkish : Proficiency.Standardized	-0.200	0.191
Proficiency.Standardized : EssayNum.Standardized		-0.025 *	0.011

Note. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.10$

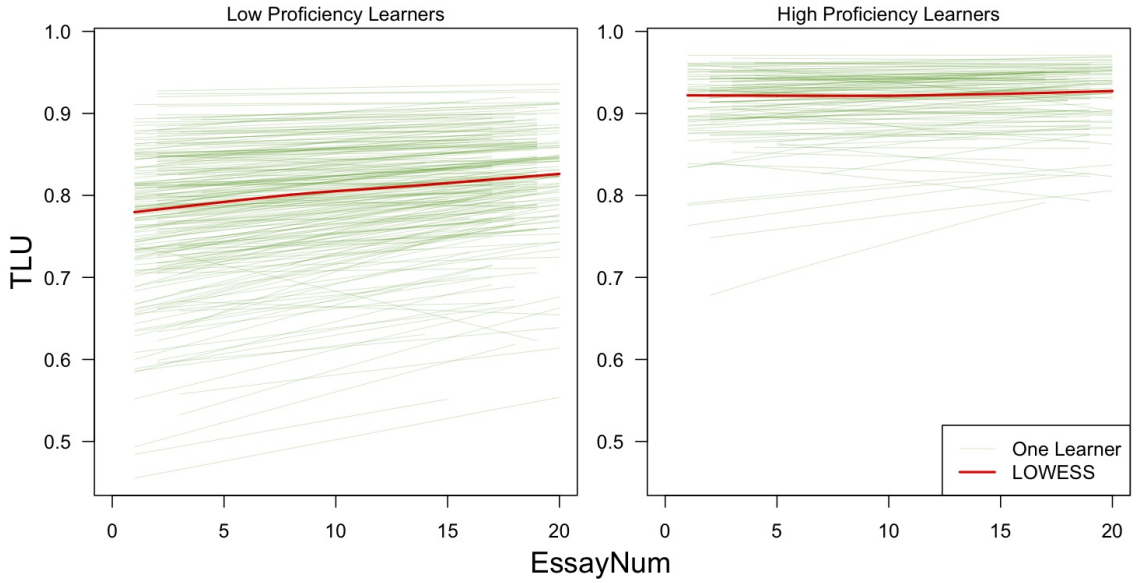


Figure 34. Fitted Values of Model 2 for High- vs Low-Proficiency Learners

be characterized by the decrease of accuracy depending on learners. This large individual variation is discussed in the Discussion section.

Looking into fitted values. Figure 34 shows the fitted values for some learners and demonstrate the difference in the developmental shape over up to 20 essays between low- and high-proficiency learners. Low-proficiency learners are those whose average proficiency (Proficiency in the models above) is between Lesson 1 Unit 5 and Lesson 2 Unit 6, and high-proficiency learners are those whose average proficiency is between Lesson 10 Unit 4 and Lesson 13 Unit 6. You can see from the figure that within-learner developmental patterns differ across proficiency levels in that lower proficiency learners exhibit a clear accuracy rise overall whereas the development of higher proficiency learners is relatively flat. This is because higher proficiency learners start at a high accuracy rate and have less room for accuracy rise compared to lower proficiency learners.

Figure 35 contrasts the fitted developmental patterns between articles and past tense *-ed*. The Proficiency of the learners in the figure is between Lesson 3 Unit 1 and Lesson 6

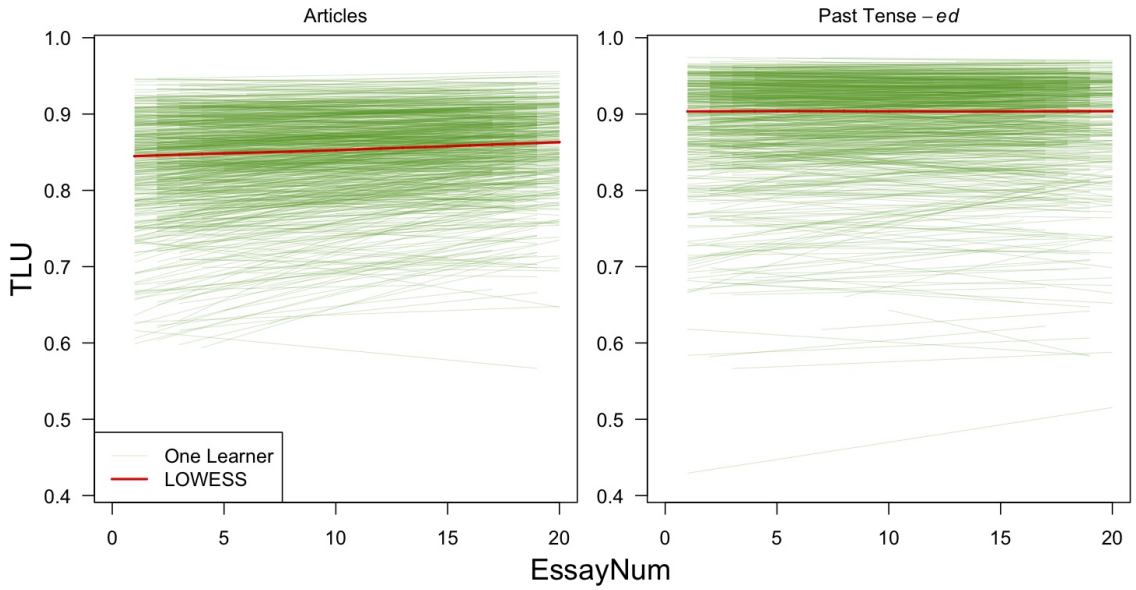


Figure 35. Fitted Values of Model 2 for Articles and Past Tense *-ed*

Unit 6 in the case of articles and between Lesson 3 Unit 1 and Lesson 6 Unit 1 in the case of past tense *-ed*. They are both around the mean proficiency of the learners in the data. The coverage of articles is slightly larger in order to match the proficiency of the learners. The difference in the mean proficiency is not significantly different [$t(4151.538) = 0.443$; $p = 0.658$; Cohen's $d = 0.010$], and therefore the difference in the developmental pattern cannot be attributed to the proficiency difference of the learners. Apart from the difference in the absolute accuracy level, the trajectories are slightly different as well. The development of past tense *-ed* tends to be flatter than article development. This, too, is because past tense *-ed* is more accurate in general than articles and it has less room for accuracy rise compared to articles.

Summary of the mixed-effects approach. On the whole, the mixed-effects approach above indicated that proficiency and morpheme affect the longitudinal developmental pattern, whereas L1 does not. It also showed large individual variation in the absolute accuracy of morphemes, the accuracy difference between morphemes, and the rate of within-learner

accuracy development.

5.3.3 Generalized Additive Models

Background for generalized additive models. A potential shortcoming in the mixed-effects models above is that they had to assume linear effects of overall proficiency and the longitudinal, within-learner development due to computational issues mentioned earlier. However, the relationship between the overall proficiency and morpheme accuracy is not necessarily linear. It is also likely that the longitudinal development is not strictly linear given the U-shaped developmental patterns that emerged in the clustering approach. Therefore, as a complement to the mixed-effects models, the generalized additive model (GAM) was employed to control for nonlinear effects in the investigation of whether the accuracy developmental pattern varies across morphemes, learners' L1, and their proficiency.

The GAM employed in the study, however, did not take into account the dependency of essays within individual learners, as the mixed-effects models did. While this assumption as well as the linearity assumption in the mixed-effects models were not strictly met, the hope was that the two modeling approaches complement each other and provide useful insights into the roles of L1 and proficiency in the longitudinal development of morphemes.

Specification of the GAM employed. A GAM with the logistic link function and binomial distribution was used to model the effect of overall proficiency, within-learner development, L1, and morpheme on morpheme accuracy. As in the previous regression models, the number of correct suppliance was entered as successes, and the number of errors plus the instances of overgeneralization were entered as failures. L1, morpheme, and their interaction were entered as predictors. A tensor product spline for the interaction between the overall proficiency (Proficiency) and the within-learner development (EssayNum) was obtained for each L1-Morpheme pair. This bivariate spline allows nonlinearity of the effects of Proficiency and EssayNum and their nonlinear interaction. The operationalization

of the variables was the same as the operationalization in the mixed-effects models above. The same standardization procedure was applied to Proficiency and EssayNum as well.

Because the computation suffered from the lack of memory, the analysis used only a subset of the data. For L1 French, Italian, Japanese, Korean, Spanish, and Turkish learners, the full data set was used as their data size was not large and further trimming of their data may invite unreliability of the result. For L1 Brazilian, Chinese, German, and Russian learners, the target data were those of the 100 learners who wrote the largest number of essays within the L1 group. In total, the data set consisted of 24,582 non-zero cases by 744 learners.

Model comparison. Similarly to the mixed-effects models, I adopted the multimodel inference approach, by which multiple GAM models were compared against each other to identify significant variables. In order to formally test whether the nonlinear developmental pattern varies across L1, proficiency, and morphemes, four GAMs were constructed. Model 1 is the maximal model described above. Model 2 tested the effect of L1 on the developmental pattern. The GAM included the same bivariate tensor product splines for each morpheme, but not for each L1-Morpheme pair as in the above, and an additional tensor product spline of Proficiency for each L1. This model assumed nonlinear relationships in the logit scale between Proficiency and TLU scores and between EssayNum and TLU scores for each morpheme, and it also assumed nonlinearity between Proficiency and TLU scores for each L1. However, nonlinearity between EssayNum and TLU scores was not assumed for each L1. By comparing this model with Model 1, we can tell whether assuming different nonlinear effects of EssayNum (i.e., nonlinear within-learner longitudinal development) across L1 groups improves the model. Model 3 tested the effect of proficiency on the developmental trajectory. The GAM included a tensor product spline of Proficiency and that of EssayNum for each L1-Morpheme pair, but did not allow them to interact. In other words, for each L1-Morpheme pair (e.g., articles for L1 Chinese), the effect of the

Table 30

Comparison of GAMs

Model	Model description	AIC	Likelihood ratio test against Model 1	
Model 1	Maximal model	43,774		
Model 2	Test the effect of L1	43,970	$\chi^2(179.16) = 554.61$	$p < 0.001$
Model 3	Test the effect of Proficiency	43,875	$\chi^2(125.34) = 351.99$	$p < 0.001$
Model 4	Test the effect of Morpheme	43,887	$\chi^2(129.14) = 371.34$	$p < 0.001$

overall proficiency as well as the longitudinal development were assumed to be nonlinear, but the nonlinear shape of the longitudinal development was assumed to be independent of the overall proficiency. That is, the overall proficiency affects the overall accuracy, but does not affect the accuracy difference between, say, the second and the eighth essay. By comparing this model with the original model, we can tell whether the learner's proficiency affects the developmental pattern. Model 4 is similar to Model 2, except that the model was constructed to test the effect of morpheme on the developmental shape. The GAM included the same bivariate tensor product splines as in the original model for each L1 and an additional tensor product spline of Proficiency for each morpheme, and compared it against the original model. This makes it possible to tell whether assuming separate nonlinear effects of EssayNum across morphemes improves the model.

Table 30 compares the four GAMs. Both likelihood ratio tests and AIC agree that Model 1 the best model to consider[7]. The GAMs therefore suggest that the developmental patterns vary across morphemes, learners' L1, and their proficiency. Thus, the maximal model (Model 1) is studied in more detail below.

Effect of L1 according to the GAM. Table 31 shows the summary of the GAM in concern. The results are mostly in line with those of the mixed-effects models. In articles, L1 Chinese, German, French, Italian, and Spanish learners achieve higher accuracy than

[7]Note, however, that the p-values here are only approximate (Wood, n.d.).

L1 Brazilian learners of English. L1 Brazilian learners, in turn, mark higher accuracy than L1 Korean learners. Both past tense *-ed* and plural *-s* are more accurate than articles in L1 Brazilian learners. L1 and morpheme interact, however. For example, the accuracy of articles and past tense *-ed* are roughly similar in L1 German learners, and the difference between articles and plural *-s* is even larger in L1 Russian learners. Since the data, although not identical, were analyzed from the viewpoint of L1-Morpheme interaction in Chapter 3, the interaction will not be further analyzed here. Overall, the effect of L1 is clear when the nonlinear effect of proficiency and within-learner development are taken into consideration.

Nonlinear proficiency effect and longitudinal development. What is further notable is the lower table showing splines. Although the values of the estimated degrees of freedom (edf) vary across L1s and morphemes, they are generally much larger than 1, which indicates nonlinear effects. Figure 36, 37, and 38 visualize the nonlinear pseudo-longitudinal and longitudinal development of the accuracy of articles, past tense *-ed*, and plural *-s* respectively. For each panel, the horizontal axis represents the overall proficiency of learners (i.e., pseudo-longitudinal development), and the vertical axis represents essay number (i.e., longitudinal development). Color indicates accuracy. Green or yellow corresponds to lower accuracy and pink or white represents higher accuracy. Because 95% of the data fall within ± 1.96 standard deviation from the mean (0.0) on the assumption of normal distribution, only the area that falls within the value both in Proficiency and in EssayNum are drawn. Two standard deviations cover from Lesson 1 Unit 1 to Lesson 11 Unit 4 in Proficiency, and Essay 1 to Essay 43 in EssayNum. A dot (.) on the figures represents an essay, and areas with denser dots are, thus, more reliable than those with less dense dots. When you focus on the L1 Japanese panel of past tense *-ed* (Figure 37), you can see that, from left to right, the color tends to become lighter, which indicates that as learners' overall proficiency goes up, so does the accuracy of past tense *-ed*. What is important in the present longitudinal analysis is the vertical view. If you see from the lower to the upper

Table 31

Summary of the Generalized Additive Model Fitted to Morpheme Accuracy

Parametric terms				
Parameter	B	SE	*z*	*p*
Intercept	1.477	0.035	42.050	0.000
L1				
Chinese	0.470	0.058	8.059	0.000
German	0.749	0.053	14.239	0.000
French	0.447	0.055	8.121	0.000
Italian	0.601	0.103	5.832	0.000
Japanese	0.000	0.162	-0.001	0.999
Korean	-0.485	0.098	-4.956	0.000
Russian	0.064	0.054	1.182	0.237
Spanish	0.256	0.097	2.641	0.008
Turkish	-0.044	0.097	-0.451	0.652
Morpheme				
Past tense *-ed*	0.568	0.158	3.605	0.000
Plural *-s*	1.068	0.071	15.136	0.000
L1 : Morpheme				
Chinese : Past tense *-ed*	-0.196	0.279	-0.702	0.483
German : Past tense *-ed*	-0.590	0.219	-2.699	0.007
French : Past tense *-ed*	-0.469	0.281	-1.669	0.095
Italian : Past tense *-ed*	0.414	0.403	1.028	0.304
Japanese : Past tense *-ed*	0.778	0.465	1.673	0.094
Korean : Past tense *-ed*	0.678	0.397	1.705	0.088
Russian : Past tense *-ed*	0.095	0.250	0.381	0.703
Spanish : Past tense *-ed*	-0.248	0.343	-0.721	0.471
Turkish : Past tense *-ed*	1.391	0.788	1.765	0.078
Chinese : Plural *-s*	-0.600	0.097	-6.208	0.000
German : Plural *-s*	-0.325	0.106	-3.053	0.002
French : Plural *-s*	-0.574	0.106	-5.420	0.000
Italian : Plural *-s*	-0.633	0.186	-3.400	0.001
Japanese : Plural *-s*	0.089	0.234	0.380	0.704
Korean : Plural *-s*	0.253	0.202	1.247	0.212
Russian : Plural *-s*	0.515	0.116	4.418	0.000
Spanish : Plural *-s*	0.103	0.199	0.517	0.605
Turkish : Plural *-s*	-0.162	0.177	-0.915	0.360

Splines				
Spline	edf	Ref.df	χ^2	*p*
L1 : Morpheme				
spline (Proficiency : EssayNum) Brazilian : Articles	12.288	14.107	101.468	0.000
spline (Proficiency : EssayNum) Chinese : Articles	16.120	17.710	61.116	0.000
spline (Proficiency : EssayNum) German : Articles	16.967	18.371	99.313	0.000
spline (Proficiency : EssayNum) French : Articles	5.217	5.957	83.224	0.000
spline (Proficiency : EssayNum) Italian : Articles	11.304	12.952	55.495	0.000
spline (Proficiency : EssayNum) Japanese : Articles	11.509	12.924	31.725	0.003
spline (Proficiency : EssayNum) Korean : Articles	9.387	7.000	15.655	0.028
spline (Proficiency : EssayNum) Russian : Articles	6.488	8.028	344.889	0.000
spline (Proficiency : EssayNum) Spanish : Articles	10.450	12.124	78.577	0.000
spline (Proficiency : EssayNum) Turkish : Articles	8.895	8.997	58.111	0.000
spline (Proficiency : EssayNum) Brazilian : Past tense *-ed*	6.162	6.942	14.735	0.038
spline (Proficiency : EssayNum) Chinese : Past tense *-ed*	7.977	9.574	35.023	0.000
spline (Proficiency : EssayNum) German : Past tense *-ed*	3.409	3.703	9.434	0.042
spline (Proficiency : EssayNum) French : Past tense *-ed*	7.824	9.119	15.359	0.086
spline (Proficiency : EssayNum) Italian : Past tense *-ed*	8.166	9.499	16.956	0.062
spline (Proficiency : EssayNum) Japanese : Past tense *-ed*	3.001	3.002	1.628	0.653
spline (Proficiency : EssayNum) Korean : Past tense *-ed*	3.001	3.002	10.836	0.013
spline (Proficiency : EssayNum) Russian : Past tense *-ed*	4.613	5.349	19.424	0.002
spline (Proficiency : EssayNum) Spanish : Past tense *-ed*	9.867	11.614	21.557	0.037
spline (Proficiency : EssayNum) Turkish : Past tense *-ed*	6.610	7.265	12.868	0.085
spline (Proficiency : EssayNum) Brazilian : Plural -	12.177	14.083	36.953	0.001
spline (Proficiency : EssayNum) Chinese : Plural *-s*	4.343	5.018	11.948	0.036
spline (Proficiency : EssayNum) German : Plural *-s*	11.606	13.603	17.173	0.224
spline (Proficiency : EssayNum) French : Plural *-s*	9.551	11.378	10.955	0.479
spline (Proficiency : EssayNum) Italian : Plural *-s*	10.669	12.323	17.124	0.160
spline (Proficiency : EssayNum) Japanese : Plural *-s*	6.921	7.573	15.868	0.036
spline (Proficiency : EssayNum) Korean : Plural *-s*	3.853	4.300	8.537	0.088
spline (Proficiency : EssayNum) Russian : Plural *-s*	3.138	3.262	34.026	0.000
spline (Proficiency : EssayNum) Spanish : Plural *-s*	10.185	11.993	40.207	0.000
spline (Proficiency : EssayNum) Turkish : Plural *-s*	8.360	9.000	15.178	0.086

side, the shade again changes from greenish to pinkish. It indicates that as learners write more essays, the accuracy of past tense *-ed* rises.

We can make four observations on the figures. First, as in the pseudo-longitudinal development in the last chapter, the longitudinal development is nonlinear in the probability scale. For example, in the L1 Italian panel of articles, at 0.5 of the standardized Proficiency (Lesson 7 Unit 1), the accuracy decreases first and then increases as learners write more essays. In other words, the longitudinal development tends to be U-shaped. Second, the nonlinear developmental pattern interacts with the overall proficiency. In the L1 Russian panel of articles, the accuracy of the learners at the standardized Proficiency of -1.0 (Lesson 2 Unit 6) tends to increase as they write more essays. However, at the standardized Proficiency of 1.0 (Lesson 8 Unit 5), their accuracy shows relatively flat development. This indicates that the developmental pattern differs across the overall proficiency of learners. Third, the two nonlinear effects further interact with L1. When, again in articles, the L1 German and the L1 Russian panels are compared, you can see that both within-learner and pseudo-longitudinal accuracy development is much more stable in L1 German than those of L1 Russian. It is possibly because L1 German learners of English already mark high accuracy of articles at a lower proficiency level and there is little room for accuracy increase initially. Fourth, the pattern also varies across morphemes. Even within the same L1 group, the longitudinal as well as pseudo-longitudinal developmental patterns vary across morphemes. For example, in L1 Brazilian panels, the development of articles seems more stable than the development of past tense *-ed* and plural *-s*. Therefore, both proficiency effects and longitudinal development are nonlinear and interact with various other factors.

Summary of the GAM approach. The GAM above revealed striking nonlinearity in morpheme accuracy development. The nonlinear effects further interact with L1, proficiency, and morpheme, that is, the developmental patterns vary across learners' L1s, their proficiency, and morphemes. As mentioned earlier, a shortcoming of the approach is that it

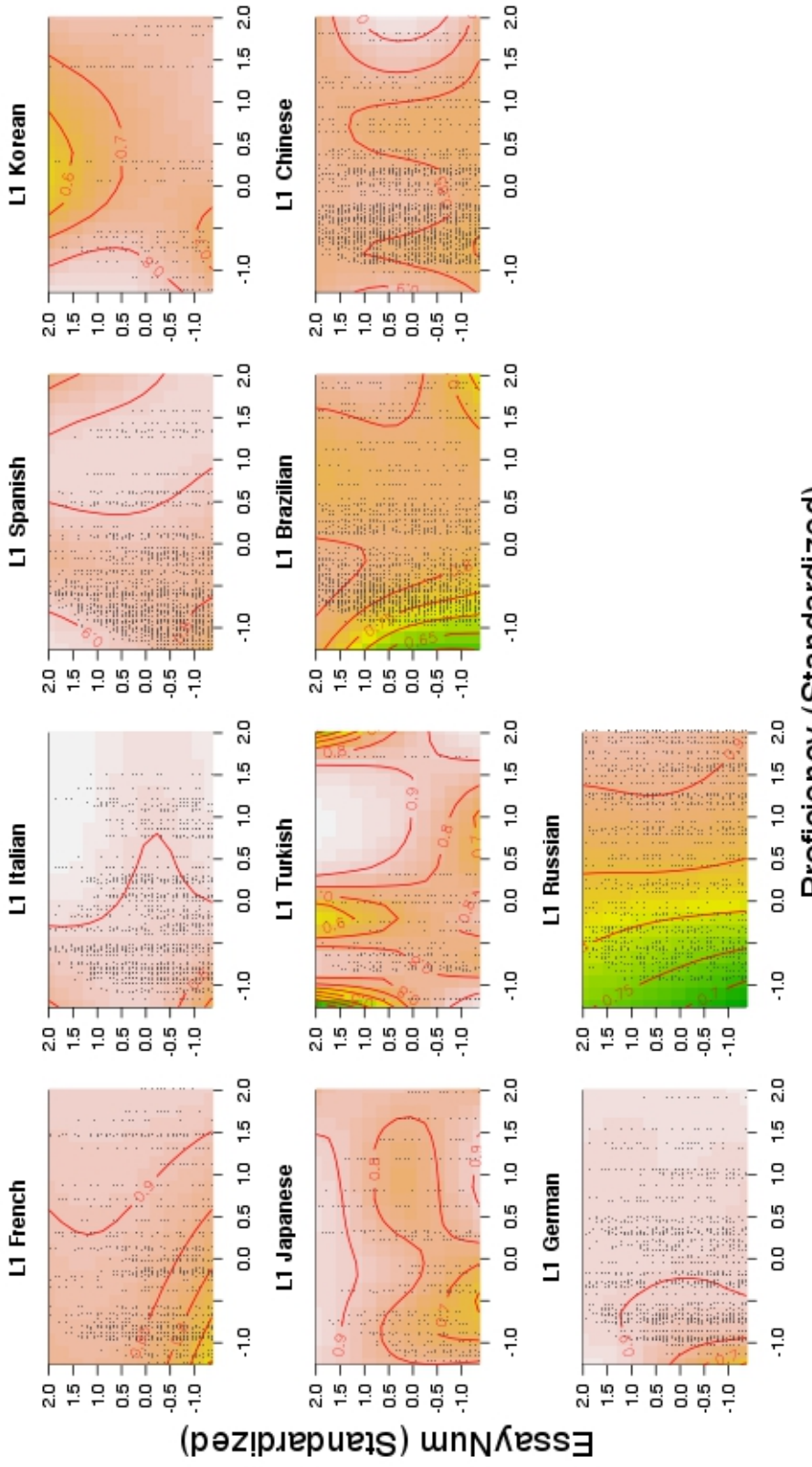


Figure 36. Nonlinear Effect of Proficiency and Longitudinal Development of Articles

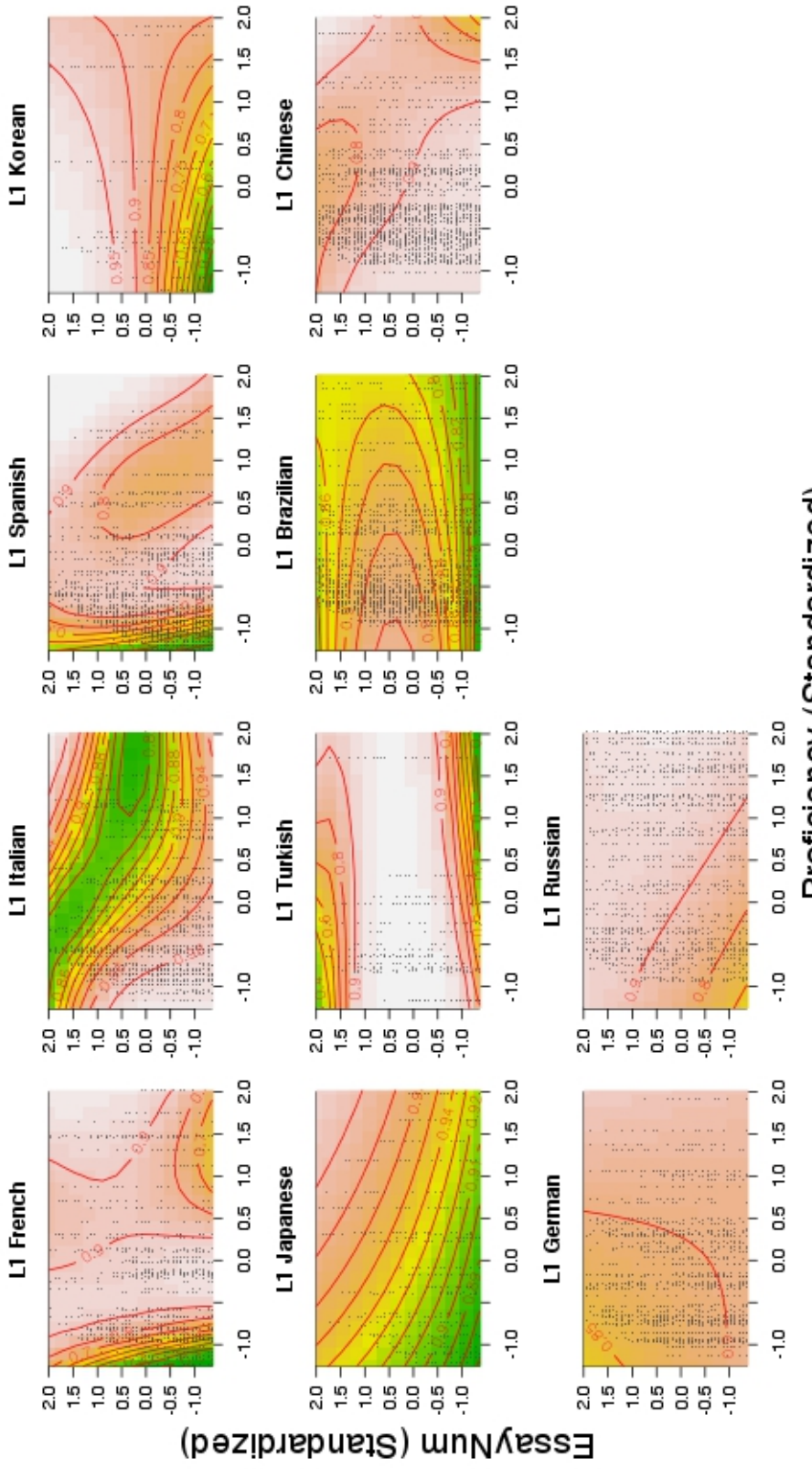


Figure 37. Nonlinear Effect of Proficiency and Longitudinal Development of Past Tense *-ed*

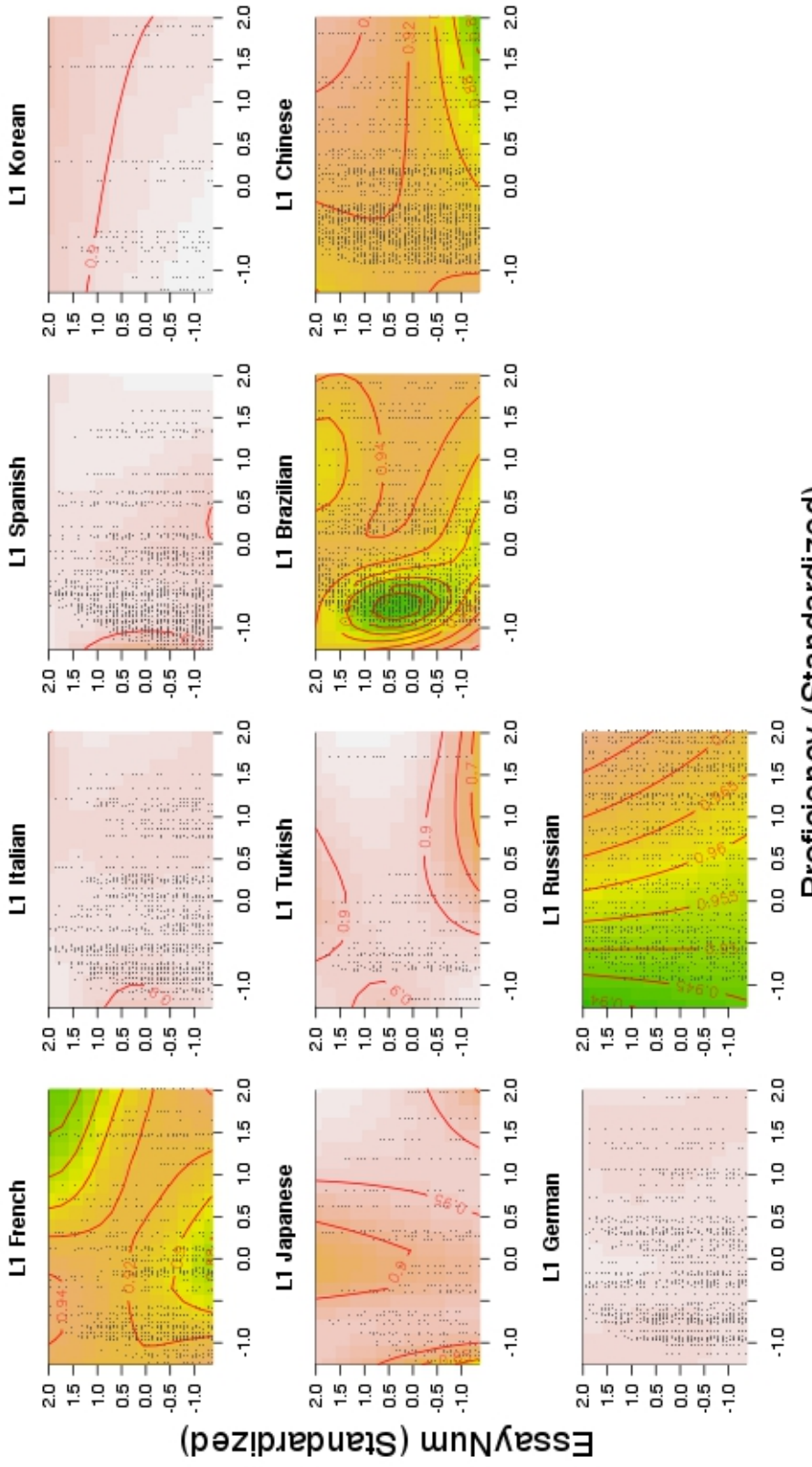


Figure 38. Nonlinear Effect of Proficiency and Longitudinal Development of Plural -*s*

did not take into account the dependency of essays within individual learners. In the mixed-effects models employed in the last section, it was, in fact, only when the dependency of data within learners was taken into account through the introduction of random-effects that an interaction of interest, EssayNum-L1, turned out to be non-significant. There is thus a risk that the lack of the control in the GAM invited spurious significant results. Despite this drawback, the present study benefited from the flexibility the GAM offers in the relationship between independent and dependent variables.

5.4 Discussion

Summary. The present study investigated whether the longitudinal developmental pattern varies across morphemes, learners' L1s, and their proficiency. The question was addressed by the prediction of cluster membership and by two types of regression modeling. The results are not straightforward and each factor will be discussed in turn.

Effect of morpheme. With respect to morphemes, there are certain commonalities but also some differences. Chapter 4 mainly revealed the commonalities in the developmental patterns across morphemes. However, in the present chapter, the mixed-effects models showed a significant interaction between morpheme and within-learner development, which means that the developmental rate varies across morphemes. The GAM also demonstrated that different morphemes exhibit different patterns of accuracy transition across pseudo-longitudinal development and across essays. A possible reason for the contrast between the clustering approach and regression modeling is the difference in the proportion of learners who share similar developmental patterns in the clustering approach. For example, in the clustering last chapter, 51.7% of the learners are classified into the cluster with flat development in plural *-s* (Figure 29), whereas 73.6% of the learners are in the cluster representing a similar developmental pattern in past tense *-ed* (Figure 28). Therefore, when we look at the overall developmental patterns for the morphemes, the pat-

tern for past tense *-ed* might be flatter than that for articles. It is, thus, safe to say that the overall pattern varies across morphemes.

Effect of L1. As to L1, the results are mixed. The clustering approach suggested that L1 affects the developmental pattern of plural *-s*. When articles, past tense *-ed*, and plural *-s* were combined, the mixed-effects models indicated that L1 is not a significant variable influencing the developmental shape. The GAM, however, supports the view that L1 affects developmental trajectories.

What caused this contradiction? Apart from the possibility that the employed techniques resulted in erroneous outcomes for the reasons explained earlier (e.g., unaccounted data dependency), I can make a following speculation. Aggregation of the morphemes in the mixed-effects models might have obscured between-L1 differences in the developmental patterns by between-morpheme differences within each L1 group. For example, if a larger proportion of a particular L1 group was classified into a flat cluster in Morpheme X than in Morpheme Y, then the large intra-L1, inter-morpheme variance might make it difficult to identify inter-L1 differences, which represent L1 influence. In other words, although there might be systematic patterns of development for each L1 group, it may have been confounded with the difference in the developmental patterns of morphemes. Therefore, it is possible that L1 affects the developmental patterns of at least some of the morphemes. This can and should be tested by including the L1-EssayNum-Morpheme three-way interaction into the mixed-effects model, which was not attempted in the present study because it would have increased the number of parameters to a great extent and the model would have been both more unreliable and more difficult to interpret. The GAM, however, supports the view because it shows that in each morpheme, even at the same overall proficiency level, the developmental pattern seemed different depending on learners' L1. The absence of the effect in the KmL clustering of articles could be because within-cluster differences were ignored, as was mentioned earlier. The above is a speculation, and given that the multiple

pieces of evidence did not converge, I will not strongly claim for the effect of L1 on the developmental path. What it does show, however, is the importance of triangulation at the level of data analysis. Applying only one of the three analyses would have certainly led to inaccurate conclusions.

Apart from the effect of L1, the clustering of article development failed to provide any evidence for the effect of L1 type (a dichotomous variable indicating whether the learner's L1 has an equivalent form to the target morpheme), which suggests that the effect is not attributable to the presence or the absence of the morpheme in learners' L1. This is interesting because L1 type was a powerful predictor especially in articles in the pseudo-longitudinal analyses reported in Chapter 2 and Chapter 3. The result suggests that we observed the effect in the earlier chapters because learners were aggregated. This further supports the idea that in the longitudinal development, individual variation outruns (part of) typological differences between L1s.

Effect of proficiency. Finally, with respect to proficiency, the answer is again mixed but is more conclusive; proficiency affects intra-learner developmental patterns. The clustering approach showed that the developmental trajectories vary across proficiency levels in articles but not in plural *-s*. The mixed-effects models suggested that proficiency is a significant predictor of the shape and higher proficiency learners tend to show flatter development. The GAM also indicated that the developmental pattern varies across overall proficiency when morpheme and L1 are controlled for. How can these partially contradictory results be reconciled? There are at least two possibilities. The first possible reason concerns the difference in the number of essays included in the analyses between the two approaches. The clustering approach exploited the first 5 or 10 windows, whereas the mixed-effects models and the GAM targeted all the essays of the learners who wrote 10 or more essays. If inclusion of all the essays made it possible to track longer developmental patterns than was possible in the clustering approach, it might have led to more

reliable modeling of individual learners. In other words, by including all the essays, the learners who appeared to show a steep increase of accuracy in the first 5 or 10 windows may show a rather flat development, or vice versa. At the same time, it is also possible that including the data of the learners who wrote 10 or more essays but did not reach 5 or 10 windows might have strengthened the relationship between proficiency and the developmental patterns. The second possible reason is the difference in the sensitivity to the rate of development. As mentioned earlier, while the clustering approach only divides the rate of development across essays into a few types, the mixed-effects models fully exploit the information available on the rate of development. This difference in the sensitivity to the developmental shape might have caused the different results. Thus, I conclude that proficiency influences the developmental shape of morphemes and that higher proficiency learners tend to show flatter developmental patterns.

Looking at individual variation through mixed-effects modeling. As in the clustering in the last chapter, the mixed-effects modeling in the present study also demonstrates large individual variation. High variance in random-intercepts and random-slopes in the mixed-effects models indicate individual variability in absolute accuracy, the accuracy difference between morphemes, and the rate of development over essays. However, the inter-learner variation is not completely at random. Some systematicity brought by such factors as proficiency is present at the same time.

Limitations. As usual, there are a few limitations of the study. First, the present study suffered from the ceiling effect. Because their starting point is already high, there is ultimately not much development in the sense of changing accuracy. This can make it difficult to detect the variation uniquely attributable to factors like L1 and proficiency because learners at the ceiling are likely to be more immune from the influence of various factors that might otherwise affect their performance.

Second, large individual variation even after controlling for the effects of L1 and profi-

ciency means that we failed to take into account a number of other factors that play roles in determining the accuracy of morphemes. The noise is potentially a reason that led to the non-straightforward results discussed earlier. In order to facilitate the test of L1 and proficiency on the developmental path, one way forward is to include other factors such as contextual ones (Gries & Wulff, in press) in the analysis, and another is to collect data in a more controlled way, such as having learners at different proficiency write essays under the same task.

5.5 Conclusion

The present study investigated the effects of L1 and proficiency on the developmental patterns of morphemes and whether the pattern varies across morphemes. The results indicated that the learners' proficiency affects the developmental patterns, and that higher proficiency learners tend to show flatter development due to the ceiling effect. The study also showed that the typical developmental shape differs across morphemes, although this is likely to be a matter of quantity (the number of learners following a particular pattern), rather than quality (shapes of the development learners can follow). No strong evidence suggested that L1 affects the developmental pattern. The study further demonstrated large individual variation in terms of the absolute accuracy, the accuracy difference between morphemes, and the rate of development. All the findings combined, the development of morpheme accuracy is a complex process influenced by a variety of factors.

Chapter 6: The Relationships in the Developmental Patterns Between English Grammatical Morphemes

6.1 Introduction

The last chapter investigated whether L1 and proficiency affect the developmental patterns of morphemes. In this chapter, we ask the question of whether there is an interrelation in the development of different morphemes or whether they develop individually.

Under the DST framework, the study investigates whether the development of the target morphemes is in competitive or supportive relationships. Recall that in a competitive relationship, the developmental patterns of two features show a complementary pattern such that when a value (e.g., accuracy) of one feature increases, that of the other decreases, and vice versa. In a supportive relationship, the patterns are similar in that when a value of one feature increases, so does the value of the other feature. The theoretical motivation behind the analysis is the competition of attentional resources. When a learner pays attention to plurality, for example, the learner might not be able to allocate sufficient attentional resources to tense in order to use past tense *-ed* correctly. In that case, plural *-s* and past tense *-ed* are likely to be in a competitive relationship.

In the present study, there are potentially three scenarios. First, all the morphemes can be in supportive relationships. All the target morphemes share a common underlying trait, and they are all connected through it. The accuracy increase in one morpheme is thus likely to be positively correlated with the accuracy increase of another morpheme. This, however, is unlikely in the present study. It is, for example, difficult to come up with a common construct between articles and past tense *-ed*, and even if there is one, the relation is unlikely to be strong enough to be empirically detectable.

Second, all the morphemes can be in competitive relationships. When learners plan to write a sentence, they may consciously or unconsciously pay more attention to tense

than to plurality. Similarly, they may focus on having plural *-s* correct and unintentionally sacrifice the accuracy of articles. In this scenario, accuracy increase in any morpheme invites accuracy decrease of the other morphemes. This, however, is also unlikely. While attentional resources can be in competition, there is no a priori reason to assume that all the morphemes compete.

Third, the relationship can depend on morpheme pairs. Supportive relationships are found between the morphemes that strongly draw on the same concept, and competitive relationships are found between the morphemes for which the use of one interferes with the allocation of attentional resources to the other. This is what I predict in the present study. In particular, I hypothesize a supportive relationship between articles and plural *-s* because the correct use of both morphemes partially depends on the correct judgments of the count-mass distinction. Thus, acquiring the distinction should contribute to the accuracy of both morphemes. On the other hand, I hypothesize a weaker relationship between nominal morphemes (e.g., articles, plural *-s*) and verbal morphemes (e.g., past tense *-ed*) because they correspond to very different aspects of grammar. Such relationships, however, have not been empirically investigated. The research question addressed is the following: Is there a systematic relationship between the developmental patterns of multiple morphemes?

6.2 Method

6.2.1 Data, Target Morpheme, L1 Groups, and Proficiency Levels

These are largely the same as the last chapter. Given the extremely high accuracy of progressive *-ing* and third person *-s*, only articles, past tense *-ed*, and plural *-s* were targeted.

6.2.2 Data Analysis

Two kinds of data analyses were performed; (i) correlation of random-effects of learners between morphemes in mixed-effects models and (ii) correlation between the development

of morphemes based on detrended data. (i) tests whether a learner who shows a more rapid rate of accuracy increase than the average in a morpheme tends to show a rapider rate of accuracy increase in another morpheme as well. (ii) investigates whether an increase of accuracy of a morpheme in a window is correlated with an increase or decrease of accuracy in another morpheme.

6.3 Results

6.3.1 Mixed-Effects Approach to Identifying the Correlation Among the Developmental Patterns of Morphemes

The present section looks into the correlation between the random-slopes of multiple morphemes to investigate whether a morpheme that shows a higher rate of accuracy increase than the average in one morpheme in a learner tends to also show a higher rate of accuracy increase in another morpheme in the same learner.

Model specification. For this purpose, a mixed-effects logistic regression model was constructed for each morpheme with the same data set as in the mixed-effects models in the last chapter. Recall that the models targeted the error-tagged essays written by the learners who wrote 10 or more error-tagged essays with one or more obligatory contexts or overgeneralization errors of one of the target morphemes. The data set consisted of 2,234 learners. Proficiency, essay number, L1, and their two-way interactions were entered as fixed-effects. Learner was included as a random-effect, and essay number was entered as a random-slope. Dummy variables with treatment contrasts were employed for categorical variables. The random-slope represents the extent to which the rate of development of each learner is higher or lower compared to the mean rate specific to his/her L1 and proficiency level. One random-slope value was obtained per learner per morpheme, and the values were correlated between morphemes. It indicates whether a learner who shows a rapider rate of accuracy increase in one morpheme than the expected rate based on his/her L1

Table 32

Correlation in the Developmental Patterns Based on Random-Effects (n = 2,234)

Morpheme pair	Pearson's *r*	*p*
Articles - Past tense *-ed*	0.038	0.076
Articles - Plural *-s*	0.144	< 0.001
Past tense *-ed* - Plural *-s*	0.037	0.079

and proficiency tends to exhibit a rapider or slower rate of accuracy increase in another morpheme.

Results of the model. The results are presented in Table 32. The correlation between articles and past tense *-ed* and that between past tense *-ed* and plural *-s* are non-significant. The correlation between articles and plural *-s* turned out to be highly significant. The positive correlation means that a learner whose accuracy increase is steeper than expected tends to have a steeper slope of accuracy increase of plural *-s* than expected as well. However, it is a rather weak correlation ($r = 0.144$).

6.3.2 Correlation Between the Development of Multiple Morphemes Based on Detrended Data

The mixed-effects models formally tested the relationship between the developmental patterns of morphemes within learners, but only assumed linear development. In order to complement it, the present study employed a correlation analysis. Through a correlation analysis, we can examine whether the accuracy increase in one morpheme is accompanied by the accuracy increase or decrease in another. The unit of the analysis reverts to windows obtained in Chapter 4, as reliable TLU scores are necessary for the analysis.

A potential problem with this approach is that in the long run accuracy may show a particular trend (e.g., accuracy increase) within individuals. If so, taking a correlation

of absolute accuracy between morphemes is interfered with the overall tendency in accuracy transition. A way to avoid this is *detrending*, a means to take away an overall trend from the data. As Verspoor et al. (2008) put it, "[d]etrending the data is an important step when focusing on intra-individual variability because otherwise the actual local variability is overestimated since a general trend by definition also consists of small local increases (or decreases, when there is a general downward trend)" (pp. 223-224). In the present study, the developmental pattern of individual learners was detrended based on LOWESS trend lines.

6.3.2.1 Description of Correlation with Detrended Data

Residuals as detrended TLU scores. Let us suppose that an interest lies in the relationship in the developmental pattern between articles and past tense *-ed* in a learner (Member ID = 18444468). The upper left panel of Figure 39 plots the TLU score of each window for articles across proficiency. The horizontal axis represents proficiency and the vertical axis represents TLU scores. The upper middle panel adds a LOWESS line onto the first panel. The LOWESS line shows that the learner's article accuracy tends to increase as his/her proficiency goes up. The dashed lines in the upper right panel show the differences between each TLU score and the LOWESS line (i.e., residuals). The residuals represent the uniqueness of each window when compared to the overall trend. Positive residuals (i.e., the residuals drawn upward from the LOWESS line) indicate that the TLU score of the window is higher than the expected value based on his/her individual developmental pattern. The residuals are the detrended TLU scores. The lower left panel plots the detrended TLU scores against proficiency. The windows above zero (solid horizontal line) are of higher accuracy than expected, and those below are of lower accuracy. The lower middle panel adds the detrended TLU scores of past tense *-ed* obtained in the same way. Note that because articles have much more windows than past tense *-ed*, the number of data points is smaller

for past tense *-ed*.

Predicting proficiency. The question is whether there is any correlation between the detrended TLU scores of articles and those of past tense *-ed*. An issue here is that the average proficiency of two morphemes does not necessarily match. If the detrended TLU score of articles at Proficiency 10.5 (in terms of Unit) is 0.02 and that at Proficiency 11.3 is 0.01, and the detrended TLU scores of past tense *-ed* at Proficiency 8.7 is -0.01 and that at Proficiency 13.8 is -0.03, it is not straightforward to compute the correlation between the developmental patterns of the two morphemes. To tackle the issue, I used the morpheme with fewer windows in total (in this case past tense *-ed*) as the basis. For each window of the morpheme, I connected the TLU scores of the two adjacent windows of the other morpheme (articles) that include the concerned window. Simple linear regression models were then constructed to predict the detrended TLU score of articles when the proficiency level equals to the average proficiency of past tense *-ed*, and the predicted scores were used as the detrended TLU scores of the article. This is illustrated in the lower right panel of Figure 39. Two TLU scores of articles are connected by a red line so that it horizontally includes one window of past tense *-ed*. From the target window of past tense *-ed*, a vertical line in blue was drawn until it crosses the connected line. The point the two lines cross is the predicted detrended TLU score of articles when proficiency is the same as the window of past tense *-ed*. Because there are eight windows for past tense *-ed*, the correlation is based on eight data points. The correlation in this case was -0.406. Although non-significant due to a small sample size, the negative correlation means that when article use is more accurate than expected, past tense *-ed* tends to be less accurate, and vice versa. This, in turn, might indicate that the two morphemes are in a competitive relationship. When a data point of the morpheme with fewer windows is beyond the proficiency end of the morpheme with more windows, the data point was not included in the analysis.

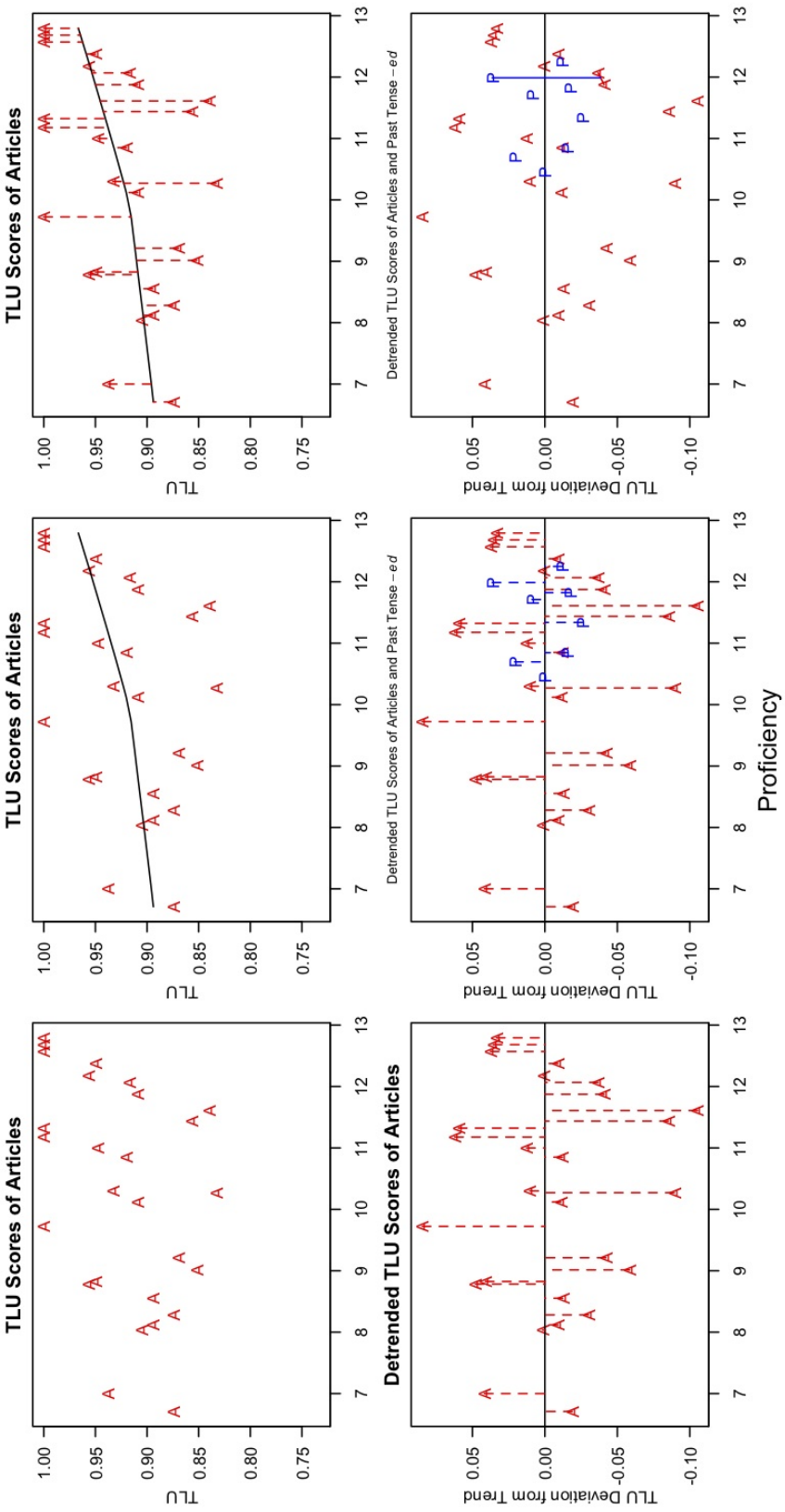


Figure 39. Example of Detrending

Matching proficiency. The analysis above, however, sacrifices matching proficiency in favor of the reliability of each data point. For example, detrended TLU scores of past tense *-ed* in a learner did not usually have corresponding TLU scores of articles computed exactly from the same set of essays. This was inevitable because a window covered a different set of essays depending on morphemes and learners. However, it resulted in unmatched comparison between morphemes. To complement this issue, the study performed another correlation analysis based on detrended data. In this analysis, TLU scores of the morpheme with a larger number of windows in total were recalculated based on the essays that were included in the windows in the other morpheme, and the two sets of TLU scores were correlated after being detrended. In other words, with the example of Figure 39, TLU scores of articles were recomputed based on the essays included in each window of past tense *-ed*. This guarantees the identity in the essays covered by both morphemes in each window, and proficiency-matched comparison is achieved. On the other hand, each data point of the morpheme with a larger number of windows could be less reliable because it may not include as many obligatory contexts as required in the previous analysis.

6.3.2.2 Results of Correlation Analyses Based on Detrended TLU Scores

The target windows were those of the learners who had at least 10 (in the case of articles and plural *-s*) or 5 (in the case of past tense *-ed*) windows. The study targeted all the windows of the learners. In the calculation of LOWESS lines, the smoothing span of 2/3 was used for articles and plural *-s*, and it was set to 1 for past tense *-ed*. The smoothing span determines the size of the window in which local regression lines are fit. A wider span produces a smoother LOWESS line. A different value was set for past tense *-ed* from articles and plural *-s* because its number of windows was smaller and the smoothing span of 2/3 could go through all the data points (i.e., overfit), yielding the residuals of zero at all the data points.

Table 33

Correlation in the Developmental Patterns Based on Detrended TLU Scores (Reliability Favored)

Morpheme pair	# Learners	Mean *r*	SD	*t*-test	
				p	Cohen's *d*
Articles - Past tense *-ed*	85	-0.082	0.440	0.088	0.188
Articles - Plural *-s*	644	0.071	0.360	< 0.001	0.197
Plural *-s* - Past tense *-ed*	80	-0.041	0.461	0.435	0.089

Reliability favored detrended correlation analysis. Table 33 shows the number of learners included in the analysis, their mean Pearson's *r*, its standard deviation, the results of a *t*-test that examines whether the mean correlation is significantly different from zero, and its effect size. For example, in the pair of article and past tense *-ed*, there were 85 learners who had both 10 or more windows in articles and 5 or more windows in past tense *-ed*, and their mean correlation in the TLU scores was -0.082 when the overall developmental pattens were detrended in each morpheme in each learner. Their standard deviation is high, however (0.440), and the mean value (-0.082) is not statistically significantly different from zero ($p = 0.088$). The effect size is small (0.188; Oswald & Plonsky, 2010) as well. Notice that in all the three morpheme pairs, the average correlation is less than 0.1 in absolute value. The weak correlations suggest no apparent supportive or competitive relationship in the developmental patterns of any morpheme pair. Although a *t*-test indicated that the relationship between articles and plural *-s* is weakly supportive, the significant result can be partially due to a large sample size, and indeed, the effect size is small (0.197).

Proficiency-matched detrended correlation analysis. Table 34 shows the result of the correlation on detrended data when the essays were exactly matched. The results are similar to Table 33, in that the mean correlations are weak and the article-plural *-s* is the

Table 34

Correlation in the Developmental Patterns Based on Detrended TLU Scores (Proficiency-Matched)

				t-test	
Morpheme pair	# Learners	Mean *r*	SD	*p*	Cohen's *d*
Articles - Past tense *-ed*	85	0.080	0.559	0.189	0.144
Articles - Plural *-s*	644	0.067	0.404	< 0.001	0.158
Plural *-s* - Past tense *-ed*	80	0.109	0.596	0.105	0.183

only significant pair. Thus, in neither of the correlation analyses, we found strong systematic relationships in the developmental patterns of multiple morphemes.

6.4 Discussion

No strong relationship identified. All in all, no strong relationship was identified in the developmental patterns between articles, past tense *-ed*, and plural *-s*. No evidence suggested any systematic relationship between verbal (past tense *-ed*) and nominal (articles and plural *-s*) morphemes. This is expected because they draw on very different aspects of language. The relationship between articles and plural *-s* was only weakly supportive, and the statistical significance was perhaps due to their large sample size. The lack of strong relationships between the two is interesting because they both require the count-mass distinction for correct use and, if so, the accuracy of both morphemes is expected to partially reflect learners' understanding of the concept. There are a few possible reasons for the absence of the strong relationships between the two morphemes.

One possible reason is that, despite the presence of systematic relationships, it may not be empirically identifiable. It might be that, although their development is interlinked somehow, the relationship is not strong enough to be reflected on the accuracy of the use of

those morphemes. For example, the correct use of articles requires not only the count-mass distinction but also the distinction of definiteness, which can make articles more complex than plural *-s*. As a result, it is possible that until learners acquire the concept of definiteness and its mapping to articles to a certain degree the acquisition of the count-mass distinction might not lead to higher accuracy of articles.

Another possible reason for the absence is potential individual differences in the relationships in the developmental patterns of morphemes. Since all the three analyses attempted to identify a certain, uniform relationship between morphemes, individual differences might have prevented the identification of meaningful patterns. For example, if the relationship between two morphemes is competitive in some learners and is supportive in others, the two patterns cancel out. More likely, even if a meaningful relationship is present under some conditions (e.g., learners with particular L1 or proficiency level), the method might fail to detect it should the relationship be absent in other conditions, especially if the former condition forms a rather minor group.

That being said, in the previous research that investigated the relationships in the developmental patterns of multiple linguistic features, the identified relationships were not strong, either. For instance, in Verspoor et al. (2008), the correlation between sentence length and type-token ratio was -0.33 and non-significant (p = 0.087), possibly due to a small sample size (n = 18). The detrended correlations reported in Spoelman and Verspoor (2010) are -0.357 (p = 0.005 according to a Monte Carlo simulation), -0.357 (p = 0.003), -0.451 ($p < 0.001$), 0.228 (p = 0.054), 0.226 (p = 0.054), and -0.022 (the p-value not reported). The absolute size of the correlation was 0.451 at the largest and was generally between 0.2 and 0.4, which is not very strong in general. Although the correlations between articles and plural *-s* in the present study are even smaller (0.144, 0.071, and 0.067), it is better not to discard the result of the present analysis outright merely because of the small effect sizes.

Limitations. As always, there are limitations to the present study. In particular, although it has been a common practice in the previous DST literature (e.g., Spoelman & Verspoor, 2010; Verspoor et al., 2008; Verspoor & van Dijk, 2011), looking at learners' written production might not be the best approach to investigate a phenomenon that has attention as its cause. It could be better if the study directly observes the deployment or the allocation of attentional resources through online measures.

6.5 Conclusion

The present study investigated whether the developmental patterns of morphemes are interrelated within individual learners. The random-effects from mixed-effects models and the correlation analysis based on detrended TLU scores indicated that the development of nominal and verbal morphemes is independent. The study identified a weak supportive relationship between articles and plural *-s*. The weak effect, however, is possibly due to the methodology employed. Although the relationship is weak, the results indicate that more targeted data and analysis could reveal some relation.

Chapter 7: General Discussion

7.1 Summary of the Findings

I will first summarize the main empirical findings of the thesis. This thesis explored the L2 development of English grammatical morphemes and factors that affect the development. The study based on the CLC in Chapter 2 showed that the morpheme accuracy order remains relatively stable within each L1 group (intragroup homogeneity) but differs across them (intergroup heterogeneity). The analysis also revealed that the difference is motivated by the difference in learners' L1s, thus satisfying crosslinguistic performance congruity stating that we have to show the similarity between the performance in L1 and the performance in L2 in claiming L1 influence (Jarvis, 2000). When learners were dichotomously divided into those whose L1s have the target morpheme and those whose L1s lack them, the former, PRESENT group outperformed the latter, ABSENT group in accuracy. The effect of L1 is robust in that the ABSENT group did not reliably achieve 90% accuracy in any of the target morphemes even at the highest proficiency level. By contrast, the PRESENT group achieved 90% accuracy in most of the target morphemes besides possessive *'s* at the highest level. Another notable finding was the complete absence of the difference between the accuracy order of L1 Spanish learners and the natural order that has been claimed universal across L1 groups. It suggests that the natural order is possibly the mere reflection of the acquisition order of L1 Spanish learners. Furthermore, the study revealed that L1 influence is so strong that the effect can be stronger than the effect of general proficiency depending on morphemes. However, L1 influence was not equally strong across all the target morphemes. Specifically, the study indicated that the morphemes that encode perceptually non-transparent concepts are more prone to L1 influence than those that encode non-linguistically clear concepts.

Chapter 3 tested the robustness of the findings in Chapter 2 with EFCamDat. The

comparison between the two corpora in absolute accuracy of the morphemes showed that articles, past tense *-ed*, and plural *-s* are of similar accuracy between the two corpora, while the accuracy differs in possessive *'s*, progressive *-ing*, and third person *-s*. Particularly in third person *-s*, the accuracy in EFCamDat was much higher than the accuracy in the CLC. In terms of the accuracy order of morphemes, the two corpora were relatively similar. No difference in the order was observed in plural *-s*, possessive *'s*, and progressive *-ing*. Although some difference was observed in the other three morphemes, it was likely to be caused by the unusually high accuracy of third person *-s*. It seems, therefore, that the accuracy order is similar between the two corpora apart from third person *-s*. Lastly, L1 influence was found somewhat stronger in the CLC than in EFCamDat. However, the order of the strength of L1 influence as well as the relative strength difference between morphemes was similar across the two corpora with a possible exception of progressive *-ing*. All in all, the two corpora produced similar findings and the results based on the CLC reported in Chapter 2 were found fairly robust.

Chapter 4 utilized EFCamDat and investigated longitudinal developmental patterns of morphemes. The clustering approach disclosed certain universal tendencies of the developmental patterns of the morphemes regardless of morphemes. For instance, the relatively flat development was always the commonest pattern among learners irrespective of morphemes. This is interesting particularly because the accuracy of some morphemes tended to increase when the data were viewed pseudo-longitudinally, but the tendency was weaker in individual development. Although flat development was common, the development was also characterized by other shapes such as decreasing accuracy in many learners. The study further observed large individual variation within and across clusters.

Chapter 5 investigated whether the developmental patterns vary depending on morphemes, learners' L1, and their proficiency. The study demonstrated that developmental patterns of morphemes are affected by proficiency and that the patterns differ across mor-

phemes. Higher proficiency learners, for example, tended to exhibit flatter developmental patterns. What was striking in the analysis, however, was significant individual variation observed at every level. It was large to the extent that the accuracy order that seemed robust in the CLC and the pseudo-longitudinal EFCamDat study, in fact, could vary across learners. The findings based on the CLC and the pseudo-longitudinal analysis of EFCamDat, as well as those in most of the previous morpheme studies, are, therefore, the reflection of the 'average' learner in the sample. It does not mean that group-level morpheme studies are of less value because it still reveals the pervasive impact of L1 on L2 acquisition in general.

Chapter 6 analyzed a factor that potentially affects within-learner variability, the within-learner developmental patterns of other morphemes. The random-effects approach and the correlation analysis based on detrended TLU scores did not find strong relationships between the developmental patterns of morphemes. It is interesting especially in the case of the article-plural *-s* relationship because the use of both morphemes partially draws on the count-mass distinction. It is, however, not that there is no relationship in the developmental patterns of morphemes at all but that the relationship is weak and requires further research to clarify the issue.

7.2 Answers to the Research Questions

The dissertation proposed six overarching research questions (RQs), restated below.

1. Is the acquisition order of L2 English grammatical morphemes affected by L1?

2. Is L1 influence equally strong across morphemes?

3. What are the longitudinal developmental patterns of L2 English grammatical morphemes?

4. Is the pattern affected by L1?

5. How can L1 influence be modeled?

6. How can pseudo-longitudinal and longitudinal development be modeled?

RQ1: L1 influence on the acquisition order. Research Question 1 asked whether the L1 affects the acquisition order. The CLC-based study provided three pieces of evidence to argue that it does; (i) the order varies across L1 groups, (ii) it is similar within L1 groups, and (iii) the difference in the order between L1 groups is motivated by the properties of the L1. The findings based on EFCamDat further reinforced the varying order between L1 groups. Both the pseudo-longitudinal and longitudinal analyses revealed different accuracy orders between L1 groups. For example, overall, L1 Spanish learners tended to be more accurate in the use or articles than in past tense *-ed*, whereas L1 Japanese or Russian learners had the opposite pattern. Therefore, although L1 influence was harder to identify for individual learners, it could still be detected when we looked at the average. The thesis, thus, argues against the universal, natural order of acquisition prevalent in recent standard SLA textbooks. The CLC-based study also demonstrated that the order that has been believed universal only applies to L1 Spanish learners of English. This possibly occurred because the main target participants of morpheme studies in the 70's were L1 Spanish learners (Luk & Shirai, 2009).

RQ2: Varying sensitivity of morphemes to L1 influence. Research Question 2 asked whether all the morphemes are equally sensitive to L1 influence. It was answered in negative. Both the CLC study and the pseudo-longitudinal analysis of EFCamDat revealed that articles are most sensitive to L1 influence, followed by plural *-s*, and possessive *'s* and third person *-s* as the morphemes relatively immune to L1 influence. Progressive *-ing* turned out to be strongly influenced by L1 in the CLC study, but not in the EFCamDat study. Apart from this discrepancy, the tendency was rather clear: L1 more strongly affects the morphemes that encode the concepts marked only in the use of language than

those whose concepts are non-linguistically clear. For instance, articles encoding definiteness are prone to L1 influence because you cannot perceive definiteness in that the same object can be referred to as *a book* and *the book* without any change to its physical being, and it is only language use that requires the distinction. On the other hand, third person *-s* encoding person is relatively robust against L1 influence because the concept of person is non-linguistically clear in that the distinction between the first and the second person is not specific to language. Non-linguistically clear concepts can, thus, mute negative L1 influence. The finding provides empirical support for the thinking-for-speaking (Slobin, 1996) and suggests that L1 affects the aspects of the world people pay attention to.

RQ3: Longitudinal developmental patterns. Research Question 3 asked the typical L2 developmental trajectories of grammatical morphemes. An important point to make here is that, while pseudo-longitudinal analyses covered a wide range of proficiency levels, the clustering analysis only covered three to four Lessons in Englishtown, corresponding to approximately one CEFR level. Therefore, although at the group level the accuracy generally increases as learners' proficiency rises, the increase over 5 or 10 windows is not large, possibly to the extent that they can look flat. This is part of the reason that flat development was prominent in all the morphemes. It was often the shape with the largest number of learners, which indicates that, although from the pseudo-longitudinal viewpoint most of the morphemes progress in terms of accuracy, longitudinally it is not always the most common developmental shape among individual learners. The mapping between the pseudo-longitudinal transition of accuracy and the longitudinal development is, therefore, not straightforward, especially when it is shaded by large individual variation. It has been argued and documented that cross-sectional data do not necessarily equate with longitudinal data (Musher-Eizenman, Nesselroade, & Schmitz, 2002; van Geert, 2008), but the present research is the first that empirically demonstrated it in the field of SLA. Besides the lack of clear correspondence between the cross-sectional and the longitudinal view, the

thesis also illuminated large intra- and inter-learner variation in morpheme development. It, again, indicates that the pseudo-longitudinal transition of accuracy is not the whole story. Moreover, the thesis found that the developmental pattern is typically nonlinear. Even though researchers have claimed nonlinearity of L2 development (e.g., Larsen-Freeman, 1997), little empirical evidence has demonstrated it. Nonlinearity of accuracy development suggests that accuracy of a single feature alone is not a good index of development, as we cannot infer one's developmental stage solely from the accuracy of the feature. Instead, data triangulation is necessary. Researchers should collect and compare multiple sources of evidence when analyzing L2 development.

RQ4: L1 influence on the developmental pattern. Research Question 4 asked if the developmental trajectory varies across morphemes, learners' L1, and their proficiency. The dissertation produced mixed results. There is some evidence pointing toward a significant effect of L1 on the developmental paths, whereas some analyses failed to identify the significance. The effect was absent when L1 was operationalized dichotomously. The lack of clear L1 influence on the developmental path is interesting because pseudo-longitudinally, although not as strong as in the CLC, we observed the effect of dichotomously operationalized L1 in EFCamDat, and the PRESENT learners achieved higher accuracy.

Why is it, then, that L1 influence that was apparent in the pseudo-longitudinal analysis was not as pronounced in the longitudinal analysis? What it indicates is that whereas there is a clear effect of L1 on the absolute accuracy of grammatical morphemes, the effect is not obvious in their developmental paths. A possible reason for this is that the developmental path is similar across L1 groups. In this scenario, although the starting point of the PRESENT group is higher than that of the ABSENT group, their accuracy develops in parallel from the beginning till the end. Learners' L1 only contributes to their starting point of accuracy but not how morphemes develop afterwards. A corollary is that the difference between the ABSENT and the PRESENT groups does not diminish even when learners'

proficiency rises. This is, in fact, what we found in Chapter 2 and is explained by the morphological congruency hypothesis (Jiang et al., 2011), a hypothesis that claims that only the PRESENT learners can attain the nativelike proficiency. Under this account, the ABSENT-PRESENT difference remains until the highest proficiency because the ABSENT group reaches the ceiling that is lower than the PRESENT group.

Another possible reason for the relative absence of L1 influence in the longitudinal development is that individual variation was so large that it shadowed between-L1 differences. Under this interpretation, although there might be individual variation in absolute accuracy as well, the effect is not strong enough to overweigh the PRESENT-ABSENT differences. However, individual variation in developmental patterns is large enough to make it difficult to detect L1 influence there. Note the two possibilities are not mutually exclusive, and it can well be the case that both simultaneously entailed weaker L1 influence in longitudinal developmental patterns than in absolute accuracy of grammatical morphemes.

RQ5: Modeling L1 influence. Research Question 5 asked how L1 influence can be modeled. In the CLC study, L1 influence was operationalized dichotomously. Although it is a rather crude measure, the variable accounted for a significant part of accuracy variance, which indicates that whether the target morpheme is superficially marked in L1 makes a difference in using it in L2. Now that the binary variable is effective, it is meaningful to investigate exactly which properties of L1 affect L2 use. In the EFCamDat study, a similar dichotomous L1 influence did not seem to explain much variance in the developmental patterns of articles. When each L1 was directly specified in the model, however, it was more effective than the dichotomous L1 influence. This empirically demonstrates that the effect of L1 is richer than binary.

RQ6: Modeling development. Research Question 6 asked how we can model pseudo-longitudinal and longitudinal development. Chapter 2 and Chapter 3 mainly focused on pseudo-longitudinal data, while Chapter 4 through Chapter 6 modeled longitudinal devel-

opment. Two important issues involved in modeling development are what variable to use in order to represent development (e.g., essay number) and the granularity of analysis (e.g., window vs. essay). As to the former issue, essay number was used to model development because, unlike Lesson and Unit number, it only shows a linear increase as learners write more essays. However, its disadvantage is that it does not include the information of the learner's overall proficiency exemplified by the Lesson and Unit number. Therefore, when the dissertation used essay number in modeling, the model simultaneously included the learner's average proficiency level over his/her essays.

The second issue is the granularity of analysis. A challenge in tracking the development of individual learners was that each essay did not include a large number of obligatory contexts of target morphemes to allow the reliable calculation of accuracy scores. The strategy taken in part of the thesis was to concatenate multiple essays and calculate accuracy scores over the set of essays (i.e., windows). Chapter 4 through Chapter 6 employed both windows and essays as the unit of analysis. Chapter 4 and Chapter 5 used clustering based on windows and regression modeling based on essays. Chapter 6 exploited windows for the correlation of detrended data, and essays for regression modeling. The primary issues involved in deciding which to use were reliability of data points and resolution of analysis. The window is more reliable because it includes a larger number of obligatory contexts, while it is of lower resolution because it may include multiple essays written at very different time points. On the other hand, the essay is of higher resolution while its reliability is a potential issue. The dissertation used both to address the same questions. In the choice of granularity, it is often worth trying to approach the same question by multiple methods, given that there are pros and cons for each.

7.3 Morpheme Development under DST

The findings above are fully compatible with the DST account of L2 acquisition. As reviewed in Chapter 2, if one's linguistic system is viewed as a dynamic system, at least four characteristics of DST predict L1 influence; sensitive dependence on initial condition, complete interconnectedness, attractor state, and resource exploitation. By sensitive dependence on initial condition, learners with different L1s have different starting points, and their interlanguages develop in different ways. For instance, being used to paying attention to the concept of definiteness, L1 Spanish learners are likely to be in a different starting position from L1 Japanese learners, which invites different outcomes throughout their development, and the difference is called L1 influence. Because systems and their components are fully connected with other systems and components, L1, which is a subsystem of one's linguistic system, inevitably affects the development of L2, which is also a subsystem of one's linguistic system. Regular attention to a concept in L1 facilitates drawing on the same concept in L2 because the two subsystems are connected. Also, because L1 is connected to various parts of the system, the impact of L1 on L2 development is pervasive (Jarvis, 2007; Odlin, 1989). L1 as an attractor state or a basin of attractor influences L2 development. Regular attention to a certain concept in L1 forms an attractor state that pulls learners' interlanguages to more target-like or non-target-like states. L1 is also a vital resource that learners can exploit in the course of L2 development. The concept of definiteness that L1 Spanish learners gained in L1 acquisition can be utilized and relied on when they use L2 English articles. Together, L1 exerts influence strong enough to alter the accuracy order among grammatical morphemes.

The varying strength of L1 influence across morphemes can be explained in the following way under the DST framework. It also shows one way to incorporate thinking-for-speaking into the DST framework. In general, input, along with other factors, leads

the system (interlanguage) to a certain direction, likely so that the system becomes more target-like. In other words, interlanguage develops with input as a vital resource. At the same time, input, or the information provided by input, interacts with various factors inside and outside of learners. One such factor is the dimension of the world that learners regularly pay attention to, which they learned through the acquisition and use of their L1. The entrenched tendency of attention allocation generates a possibly huge basin of attractor, which can facilitate or hinder L2 acquisition. The strength of L1 influence, or depth of the attractor state, is not determined by whether learners' L1s have an equivalent feature to the target L2 feature, but by whether learners regularly pay attention to the concept encoded by the feature in their daily lives. Dichotomously operationalized L1 influence appears to affect L2 development because it partially captures the amount of attention learners pay to the concept. If Ringbom's (2007; 2009) view is taken and embedded into DST, this means that regular attention to a concept present in L2 constitutes an attractor state that pulls the system to a target-like state. For the learners without the equivalent feature in their L1, lack of attention to the concept encoded by the feature means the absence of the force to draw the system. This shows that attention influences the emergence of attractor states. The strength of L1 influence varies across morphemes because some morphemes encode the concepts that are accessed only in using the language (e.g., definiteness), whereas others encode those that people generally pay attention to regardless of the language they speak (e.g., person). More generally, regular attention to the concepts present in L1 predicted by thinking-for-speaking entrenches the direction of attention, which in turn brings about an attractor state people's view is biased toward. Thus, thinking-for-speaking and DST are linked through attention and attractor states, both of which partially determines the tendency of human cognition.

The above illustrates systematicity in L2 development. However, the dissertation also demonstrated large individual variation in absolute accuracy, the accuracy difference be-

tween morphemes, and their rate of development, after controlling for the L1 and the overall proficiency of learners. The individual variation was large to the extent that it outweighed the strength of L1 influence in certain cases. It is not only inter-learner variation that was observed. Significant intra-learner variability was also present. The clustering approach in the longitudinal analysis of EFCamDat visually demonstrated the variability within and between individual learners. DST dictates that a dynamic system like one's linguistic system undergoes constant change and naturally results in large inter- and intra-individual variability because a system is sensitive to various internal and external factors due to the DST's features of dependence on initial conditions and complete interconnectedness. Learners vary in nearly all the possible aspects of L2 acquisition including their developmental paths because there are significant cognitive (e.g, working memory capacity), environmental (e.g., the amount of input one has received), and social (e.g., ESL vs EFL) differences between learners, which collectively influence L2 acquisition in complex ways. Large individual variation does not mean that group factors (e.g., L1) are unimportant because they can still be powerful and certainly essential to model L2 development, as mentioned above. However, at the same time, large individual variation should not be disregarded, either, in modeling and theorizing L2 development (cf. R. Ellis, 2008).

7.4 Future Research

There are a few possible directions for future research in relation to the dissertation. First, now that the presence of L1 influence is clear in the pseudo-longitudinal development, one can research on the properties of L1 that affect acquisition. For instance, does the number of meanings represented by corresponding L1 expressions influence the degree of transfer? Does the frequency of corresponding expressions in L1 affect transfer? When the morpheme is absent in L1, does it make a difference if the concept can be linguistically marked in a different way, for example by discourse features? Answering these questions

should lead to a more comprehensive understanding of L1 influence on the acquisition of grammatical morphemes.

Second, we can seek further explanation of morpheme accuracy and its development. In particular, contextual influence (Gries & Wulff, in press) on errors is a promising area of research. For example, does the number of words between the subject and the verb affect the accuracy of third person *-s* of the verb? Does the frequency ratio of the plural versus the singular form of a noun affect the accuracy of plural *-s* on the noun? Does article accuracy vary depending on the nouns the article modifies? Does the effect of these factors vary across learners' L1 and proficiency? These questions should lead to a better explanation of learners' use of morphemes and may partially account for the large individual variation observed in the dissertation. At the same time, it will disclose the complexity involved in the morpheme use by L2 learners.

Third, we can investigate the developmental trajectories of other linguistic features with the tools used and/or developed in the thesis. As noted earlier, we can apply the methodology in the dissertation to other linguistic features and other languages. The dissertation observed U-shaped development in articles and plural *-s*. Can we observe the same pattern in other linguistic features? What are the commonalities among the features that share developmental patterns of accuracy? The advent of large-scale learner corpora and the relevant data analytic techniques have made it possible to address the questions of this kind, through which we can tease apart idiosyncrasies from systematicity.

7.5 Concluding Remarks

Overall, the thesis empirically established that, contrary to a widely held belief among SLA researchers, the L2 acquisition order of English grammatical morphemes is not consistent across the learners with different L1 backgrounds and that the strength of L1 influence varies across morphemes. It also demonstrated that large individual variation is present

in the accuracy development of morphemes but that various factors such as learners' proficiency bring it some systematicity. Together, the findings presented in the dissertation point to complex, dynamic, and nonlinear development of morphemes.

References

Abrahamsson, N. (2003). Development and recoverability of L2 codas: A longitudinal study of Chinese/Swedish interphonology. *Studies in Second Language Acquisition*, *25*(3), 313–349. doi: 10.1017/S0272263103000147

Ambridge, B., & Lieven, E. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press.

Andersen, R. W. (1978). An implicational model for second language research. *Language Learning*, *28*(2), 221–282.

Anderson, D. R. (2004). *Model based inference in the life sciences: A primer on evidence*. Fort Collins, CO: Springer.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA.: Harvard University Press.

Anderson, J. R. (1995). *Learning and memory. An integrated approach*. New York, NY: Wiley.

Aoki, S. (2004). *Steel-Dwass no hoho ni yoru taju hikaku [Multiple comparision by the Steel-Dwass method].* Retrieved 25 April, 2013, from `http://aoki2.si.gunma-u.ac.jp/R/Steel-Dwass.html`

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3). Retrieved from `http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/1492`

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Cross-disciplinary issues in compounding* (pp. 257–270).

Amsterdam: John Benjamins.

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106. doi: 10.1016/j.cognition.2012.06.003

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. doi: 10.1016/j.jml.2007.09.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D. M. (2010). *Lme4: Mixed-effects modeling with R*. Springer. Retrieved from `http://lme4.r-forge.r-project.org/book/`

Battye, A., & Hintze, A. (1992). *The French language today*. London: Routledge.

Berdan, R. (1996). Disentagling language acquisition from linguistic variation. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 203–244). Amsterdam: John Benjamins.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*(1), 1–17. doi: 10.1111/j.1467-1770.1983.tb00983.x

Blom, E., Paradis, J., & Sorenson Duncan, T. (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular *-s* in child L2 English. *Language Learning*, *62*(3), 965–994. doi: 10.1111/j.1467-9922.2012.00715.x

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the coling/acl 2006 interactive presentation sessions* (pp. 77–80). Sydney, Australia.

Brown, R. (1973). *A first language: The early stages*. London: George Allen & Unwin.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Compmlexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.

Butler, Y. G. (2002). Second language learners' theories on the use of English articles: An

analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition*, *24*(3), 451–480. doi: 10.1017/S0272263102003042

Carver, C. S., & Scheier, M. F. (1999). Themes and issues in the self-regulation of behavior. In R. S. J. Wyer (Ed.), *Perspectives on behavioral self-regulation* (pp. 1–105). Mahwah, NJ: Lawrence Erlbaum.

Caspi, T. (2010). *A dynamic perspective on second language development*. Unpublished doctoral dissertation, University of Groningen.

Chen, P. (2004). Identifiability and definiteness in Chinese. *Linguistics*, *42*(6), 1129–1184. doi: 10.1515/ling.2004.42.6.1129

Cheung, H. S., Liu, S., & Shih, L. (1994). *A practical Chinese grammar*. Hong Kong: The Chinese University Press.

Choi, M.-h. (2005). Testing Eubank's optional verb-raising in L2 grammars of Korean speakers. *Proceedings of the 7th Generative Approaches to Second Language Acquisition Conference*, 58–67.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528. doi: 10.1146/annurev.psych.57.102904.190146

Cook, V. (1995). Multi-competence and the learning of many languages. *Language, Culture and Curriculum*, *8*(2), 93–98. doi: 10.1080/07908319509525193

Crawley, M. J. (2007). *The R book*. West Sussex: John Wiley & Sons.

Crawley, M. J. (2013). *The R book (second edition)*. West Sussex: John Wiley & Sons.

Csizér, K., & Dörnyei, Z. (2005). Language learners' motivational profiles and their motivated learning behavior. *Language Learning*, *55*(4), 613–659. doi: 10.1111/j.0023-8333.2005.00319.x

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369–382. doi: 10.1177/0267658312443651

de Bot, K. (2008). Introduction: Second language development as a dynamic process. *Modern Language Journal*, *92*(2), 166–178. doi: 10.1111/j.1540-4781.2008.00712.x

de Bot, K., & Larsen-Freeman, D. E. (2011). Researching second language development from a dynamic systems theory perspective. In M. H. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Method and techniques* (pp. 5–23). Amsterdam: John Benjamins.

de Bot, K., Lowie, W., & Verspoor, M. (2005). *Second language acquisition: An advanced resource book*. Oxon: Routledge.

de Bot, K., Lowie, W., & Verspoor, M. (2007a). A Dyamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7–21. doi: 10.1017/S1366728906002732

de Bot, K., Lowie, W., & Verspoor, M. (2007b). A dynamic view as a complementary perspective. *Bilingualism: Language and Cognition*, *10*(1), 51–55. doi: 10.1017/S1366728906002811

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge: Cambridge University Press.

DeKeyser, R. (2007). Skill Acquisition Theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Deshors, S. C., & Gries, S. Th. (in press). A case for the multifactorial assessment of learner language: The uses of *may* and *can* in French-English interlanguage. In D. Glynn & J. Robinson (Eds.), *Polysemy and synonymy: Corpus methods and applications in lognitive linguistics.* Amsterdam: John Benjamins. Retrieved from `http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html`

de Villiers, J. G., & de Villiers, P. A. (1973). A cross-sectional study of the acquisition of gram-

matical morphemes in child speech. *Journal of Psycholinguistic Research*, *2*(3), 267–278.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. doi: 10.1016/j.jml.2007.11.004

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford: Oxford University Press.

Duff, P. A., & Li, D. (2002). The acquisition and use of perfective aspect in Mandarin. In R. Salaberry & Y. Shirai (Eds.), *The L2 acquisition of tense-aspect morphology* (pp. 415–453). Amsterdam: John Benjamins.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning*, *23*(2), 245–258.

Dulay, H. C., & Burt, M. K. (1974). Natural sequences in child second language acquisition. *Language Learning*, *24*(1), 37–53.

Ekiert, M. (2010). Linguistic effects on thinking for writing: The case of articles in L2 English. In Z. H. Han & T. Cadierno (Eds.), *Linguistic relativity in SLA: Thinking for speaking* (pp. 125–153). Bristol: Multilingual Matters.

Ekmekci, F. O. (1982). Acquisition of verbal inflections in Turkish. *METU Journal of Human Sciences*, *1*(2), 227–241.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194. doi: 10.1093/applin/aml015

Ellis, N. C. (2008). The Dynamics of second language emergence : Cycles of language use , language change , and language acquisition. *Modern Language Journal*, *92*(2), 232–249. doi: 10.1111/j.1540-4781.2008.00716.x

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

Field, A. (2012). *Discovering statistics using R*. London: Sage Publications.

Fitzpatrick, E., & Seegmiller, M. S. (2004). The Montclair electronic language database project.

In U. Connor & T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 223–237). Amsterdam: Rodopi.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

Genolini, C., & Falissard, B. (2010). KmL: K-means for longitudinal data. *Computational Statistics*, *25*(2), 317–328. doi: 10.1007/s00180-009-0178-4

Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar*. Oxon: Routledge.

Goldschneider, J., & DeKeyser, D. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*(1), 1–50. doi: 10.1111/1467-9922.00147

Gor, K., & Chernigovskaya, T. (2004). Generation of complex verbal morphology in first and second language acquisition : Evidence from Russian. *The proceedings of the 19th Scandinavian Conference of Linguistics*, 819–833.

Görgülü, E. (2005). Plural marking in Turkish: Additive or associative ? *Working papers of the Linguistics Circle: Proceedings of the 27th Northwest Linguistics Conference*, *21*(1), 70–80.

Graves, P. G. (1990). *German grammar*. Hauppauge, NY: Barron's Educational Series.

Gries, S. Th. (2013). *Statistics for linguistics with R: A practical introduction (second edition)*. Berlin: De Gruyter Mouton.

Gries, S. Th., & Deshors, S. C. (in press). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*. Retrieved from `http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html`

Gries, S. Th., & Stoll, S. (2009). Finding developmental groups in acquisition data: Variability-based neighbour clustering. *Journal of Quantitative Linguistics*, *16*(3), 217–242. doi: 10.1080/0929617090297569

Gries, S. Th., & Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*, 163–186. doi: 10.1075/arcl.7.07gri

Gries, S. Th., & Wulff, S. (in press). The genitive alternation in Chinese and German ESL learners: Towards multifactorial notion of *context* in learner corpus research. *International Journal of Corpus Linguistics*. Retrieved from http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html

Hakuta, K. (1976). A case study of a Japanese child learning English as a second language. *Language Learning*, *26*(2), 321–351.

Han, Z. H., & Odlin, T. (2006). *Studies of fossilization in second language acquisition*. Clevedon: Multilingual Matters.

Harvey, W. C. (2006). *Complete Spanish: Grammar review*. Hauppauge, NY: Barron's Educational Series.

Hawkins, J. A. (1991). On (in)definite articles: Implicatures and (un)grammaticality prediction. *Journal of Linguistics*, *27*(2), 405–442. doi: 10.1017/S0022226700012731

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*(1), e5. doi: 10.1017/S2041536210000103

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.

Hawkins, R. (1981). Towards an account of the possessive constructions: NP's N and the N of NP. *Journal of Linguistics*, *17*(2), 247. doi: 10.1017/S002222670000699X

Hiki, M. (1991). *A study of learners' judgments of noun countability*. Unpublished doctoral dissertation, Indiana University.

Hilpert, M., & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literacy and Linguistic Computing*, *24*(4), 385–401. doi: 10.1093/llc/fqn012

Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Dallar, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. New York: Lawrence Erlbaum Associates.

Hox, J. (2010). *Multilevel analysis: Techniques and applications (second edition)*. New York: Lawrence Erlbaum Associates.

Hsin, A.-l. (2002). On indefinite subject NPs in Chinese. *Chinese Studies*, *20*(2), 353–376. Retrieved from http://ccs.ncl.edu.tw/chinese_studies_20_2/353_376.pdf

Huang, S. (1999). The emergence of a grammatical category *definite article* in spoken Chinese. *Journal of Pragmatics*, *31*(1), 77–94. doi: 10.1016/S0378-2166(98)00052-6

Ionin, T. (2006). *This* is definitely specific: Specificity and definiteness in article systems. *Natural Language Semantics*, *14*(2), 175–234. doi: 10.1007/s11050-005-5255-9

Ionin, T. (2008). Progressive aspect in child L2 English. In B. Haznedar & GavrusevaE. (Eds.), *Current trends in child second language acquisition: A generative perspective* (pp. 17–53). Amsterdam: John Benjamins.

Ionin, T., & Montrul, S. (2009). Article use and generic reference: Parallels between L1- and L2-acquisition. In M. García-Mayo & R. Hawkins (Eds.), *Second language acquisition of articles: Empirical findings and theoretical implications* (pp. 147–173). Amsterdam: John Benjamins.

Ionin, T., & Montrul, S. (2010). The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language Learning*, *60*(4), 877–925. doi: 10.1111/j.1467-9922.2010.00577.x

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi: 10.1016/j.jml.2007.11.007

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

Jarvis, S. (2007). Theoretical and methodological issues in the investigation of conceptual transfer. *Vigo International Journal of Applied Linguistics*, *4*, 43–71.

Jarvis, S., Castañeda Jiménez, G., & Nielsen, R. (2012). Detecting L2 writers' L1s on the basis of their lexical styles. In S. Jarvis & S. A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 34–70). Bristol: Multilingual Matters.

Jarvis, S., & Pavlenko, A. (2007). *Crosslinguistic influence in language and cognition*. New York: Routledge.

Jelinek, E. (1984). Empty categories, case, and configurationality. *Natural Language & Linguistic Theory*, *2*(1), 39–76. doi: 10.1007/BF00233713

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological congruency and the acquisition of L2 morpheme. *Language Learning*, *61*(3), 940–967. doi: 10.1111/j.1467-9922.2010.00627.x

King, L. D., & Suñer, M. (1980). The meaning of the progressive in Spanish and Portuguese. *Bilingual Review*, *7*(3), 222–238.

Koike, I. (1983). *Acquisition of grammatical structures and relevant verbal strategies in a second language*. Tokyo: Taishukan.

Köpcke, K. (1988). Schemas in German plural formation. *Lingua*, *74*(4), 303–335. doi: 10.1016/0024-3841(88)90064-2,

Krashen, S. D. (1977). Some issues relating to the Monitor Model. In H. D. Brown, C. A. Yorio, & R. H. Crymes (Eds.), *On TESOL '77: Teaching and learning English as a second language: Trends in research and practice* (pp. 144–158). Washington D. C.: TESOL.

Kwon, E. (2005). The "natural order" of morpheme acquisition: A historical survey and discussion of three putative determinants. *Columbia University Working Papers in TESOL & Applied Linguistics*, *5*(1). Retrieved from http://journals.tc-library.org/index.php/tesol/article/view/112/110

Larsen-Freeman, D. E. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly*, *9*(4), 409–430.

Larsen-Freeman, D. E. (1976). An explanation for the morpheme acquisition of second language

learners. *Language Learning*, *26*(1), 125–134.

Larsen-Freeman, D. E. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics*, *18*(2), 141–165. doi: 10.1093/applin/18.2.141

Larsen-Freeman, D. E., & Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *Modern Language Journal*, *92*(2), 200–213. doi: 10.1111/j.1540-4781.2008.00714.x

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368–390. doi: 10.1093/applin/amp038

Lee, E.-H. (2006). Dynamic and stative information in temporal reasoning: Interpretation of Korean past markers in narrative discourse. *Journal of East Asian Linguistics*, *16*(1), 1–25. doi: 10.1007/s10831-006-9003-z

Lightbown, P. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. Seliger & M. H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 217–243). Rowley, MA: Newbury House.

Lindauer, T. (1998). Attributive genitive constructions in German. In A. Alexiadou & C. Wilder (Eds.), *Possessors, predicates, and movement in the determiner phrase* (pp. 109–140). Amsterdam: John Benjamins.

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Thousand Oaks, CA: Sage Publications.

Long, M. H., & Sato, C. (1984). Methodological issues in interlanguage studies: An interactionist perspective. In A. Davies, C. Criper, & A. Howatt (Eds.), *Interlanguage* (pp. 253–279). Edinburgh: Edinburgh University Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. doi: 10.1075/ijcl.15.4.02lu

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural *-s*, articles, and possessive *'s*. *Language*

Learning, *59*(4), 721–754. doi: 10.1111/j.1467-9922.2009.00524.x

MacWhinney, B. (2008). A unified model. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). New York: Routledge.

Meisel, J. M. (2011). *First and second language acquisition*. Cambridge: Cambridge University Press.

Meunier, F., & Littré, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *Modern Language Journal*, *97*(S1), 61–76. doi: 10.1111/j.1540-4781.2012.01424.x

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, *45*(3), 241–259. doi: 10.1515/IRAL.2007.011

Mitchell, R., Myles, F., & Marsden, E. (2013). *Second language learning theories (second edition)*. Oxon: Routledge.

Müller, G. (2004). A distributed morphology approach to syncretism in Russian noun inflection. In O. Arnaudova, W. Browne, M. L. Rivero, & D. Stojanovic (Eds.), *Formal approaches to slavic linguistics #12: The Ottawa meeting 2003* (pp. 353–374). Ann Arbor, MI: Michigan Slavic Publications.

Musher-Eizenman, D. R., Nesselroade, J. R., & Schmitz, B. (2002). Perceived control and academic performance: A comparison of high- and low-performing children on within-person change patterns. *International Journal of Behavioral Development*, *26*(6), 540–547. doi: 10.1080/01650250143000517

Myles, F. (2008). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capabilities* (pp. 58–72). London: Routledge.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. doi: 10.1016/S0022-2496(02)00028-7

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography

and ELT. In *Proceedings of the corpus linguistics 2003* (pp. 572–581). Lancaster.

Nowak, A., Vallacher, R. R., & Zochowski, M. (2005). The emergence of personality: Dynamic foundations of individual variation. *Developmental Review*, *25*(3-4), 351–385. doi: 10.1016/j.dr.2005.10.004

Odlin, T. (1989). *Language transfer*. Cambridge: Cambridge University Press.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. doi: 10.1017/S0267190510000115

Papadopoulou, D. (2006). *Cross-linguistic variation in sentence processing: Evidence from RC attachment preferences in Greek*. Dordrecht: Springer.

Parkinson, S. (1988). Portuguese. In M. Harris & N. Vincent (Eds.), *The Romance languages* (pp. 131–169). Oxon: Routledge.

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning*, *33*(4), 465–497.

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: John Benjamins.

Plaza-Pust, C. (2008). Dynamic systems theory and Universal Grammar: Holding up a turbulent mirror to development in grammars. *Modern Language Journal*, *92*(2), 250–269. doi: 10.1111/j.1540-4781.2008.00717.x

Proudfoot, A., & Cardo, F. (2005). *Modern Italian grammar: A practical guide (second edition)*. Oxon: Routledge.

Regan, V. (1996). Variation in French interlanguage. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 177–201). Amsterdam: John Benjamins.

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Clevedon: Multilingual Matters.

Ringbom, H., & Jarvis, S. (2009). The importance of cross-linguistic similarity in foreign language

learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 106–118). West Sussex: Blackwell Publishing.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. doi: 1093/applin/22.1.27

Rysiewicz, J. (2008). Cognitive profiles of (un)successful FL learners: A cluster analytical study. *Modern Language Journal*, *92*(1), 87–99. doi: 10.1111/j.1540-4781.2008.00688.x

Salaberry, R. (2002). Tense and aspect in the selection of Spanish past tense verbal morphology. In R. Salaberry & Y. Shirai (Eds.), *The L2 acquisition of tense-aspect morphology* (pp. 397–415). Amsterdam: John Benjamins.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing* (pp. 44–49). Manchester, UK.

Sercu, L., De Wachter, L., Peters, E., Kuiken, F., & Vedder, I. (2006). The effect of task complexity and task conditions on foreign language development and performance: Three empirical studies. *ITL - International Journal of Applied Linguistics*, *152*, 55–84. doi: 10.2143/ITL.152.0.2017863

Shinnou, H. (2007). *R de manabu kurasuta kaiseki [Learning cluster analysis with R]*. Tokyo: Ohmsha.

Shirai, Y. (1998a). The emergence of tense-aspect morphology in Japanese: universal predisposition? *First Language*, *18*(3), 281–309. doi: 10.1177/014272379801805403

Shirai, Y. (1998b). Where the progressive and the resultative meet: Imperfective aspect in Japanese, Chinese, Korean and English. *Studies in Language*, *22*(3), 661–692. doi: 10.1075/sl.22.3.06shi

Siegler, R. S. (2006). Microgenetic analyses of learning. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Volume 2: Cognition, perception, and language (sixth edition)* (pp. 464–510). Hoboken, NJ: Wiley.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and*

event occurrence. New York: Oxford University Press.

Slobin, D. I. (1996). From "thought to language" to "thinking for speaking". In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge: Cambridge University Press.

Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 157–192). Cambridge, MA: MIT Press.

Slobin, D. I. (2008). The child learns to think for speaking: Puzzles of crosslinguistic diversity in form-meaning mappings. In T. Ogura, H. Kobayashi, S. Inagaki, M. Hirakawa, S. Arita, & Y. Terao (Eds.), *Studies in language sciences 7* (pp. 1–13). Tokyo: Kurosio.

Slobin, D. I., & Aksu, A. A. (1982). Tense, aspect and modality in the use of the Turkish evidential. In P. J. Hopper (Ed.), *Tense-aspect: Between semantics and pragmatics* (pp. 185–200). Amsterdam: John Benjamins.

Snape, N. (2005). The certain uses of articles in L2-English by Japanese and Spanish speakers. *Durham and Newcastle Working Papers in Linguistics*, *11*, 155–168.

Snape, N. (2008). Resetting the nominal mapping parameter in L2 English: Definite article use and the count-mass distinction. *Bilingualism: Language and Cognition*, *11*(1), 63–79. doi: 10.1017/S1366728907003215

Sparks, R. L., Patton, J., & Ganschow, L. (2012). Profiles of more and less successful L2 learners: A cluster analysis study. *Learning and Individual Differences*, *22*(4), 463–472. doi: 10.1016/j.lindif.2012.03.009

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics:*, *31*(4), 532–553. doi: 10.1093/applin/amq001

Stoll, S., & Gries, S. Th. (2009). How to measure development in corpora? An association strength approach. *Journal of Child Language*, *36*(5), 1075–1090. doi: 10.1017/S0305000909009337

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Massachusetts, MA: MIT Press.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *Modern Language Journal*, *97*(S1), 77–101.

van Dijk, M., Verspoor, M., & Lowie, W. (2011). Variability and DST. In M. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 55–84). Amsterdam: John Benjamins.

van Geert, P. (1995). Growth dynamics in development. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: The dynamics of cognition* (pp. 313–337). Cambridge, MA: MIT Press.

van Geert, P. (2008). The dynamic systems approach in the study of L1 and L2 acquisition : An introduction. *Modern Language Journal*, *92*(2), 179–199. doi: 10.1111/j.1540-4781.2008.00713.x

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior & Development*, *25*(4), 340–374. doi: 10.1016/S0163-6383(02)00140-6

VanPatten, B. (1984). Morphemes and processing strategies. In F. Eckman, L. Bell, & D. Nelson (Eds.), *Universals and second language acquisition* (pp. 88–98). Cambridge, MA: Newbury House.

Verspoor, M., & Behrens, H. (2011). Dynamic Systems Theory and a usage-based approach to second language development. In M. H. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 25–38). Amsterdam: John Benjamins.

Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal*, *92*(2), 214–231. doi: 10.1111/j.1540-4781.2008.00715.x

Verspoor, M., & van Dijk, M. (2011). Visualizing interactions between variables. In M. H. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development:*

Methods and techniques (pp. 85–98). Amsterdam: John Benjamins.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), e23613. doi: 10.1371/journal.pone.0023613

Wilson, M. (1988). The MRC Psycholinguistic Database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments and Computers*, *20*(1), 6–11. doi: 10.3758/BF03202594

Wood, S. (n.d.). *Approximate hypothesis tests related to GAM fits.* Retrieved 22 April, 2013, from `http://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/anova.gam.html`

Wood, S. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wulff, S., Ellis, N. C., Römer, U., Bardovi-Harlig, K., & Leblanc, C. (2009). The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. *Modern Language Journal*, *93*(3), 354–369. doi: 10.1111/j.1540-4781.2009.00895.x

Yashima, T., & Zenuk-Nishide, L. (2008). The impact of learning contexts on proficiency, attitudes, and L2 communication: Creating an imagined international community. *System*, *36*(4), 566–585. doi: 10.1016/j.system.2008.03.006

Yoon, K. K. (1993). Challenging prototype descriptions: Perception of noun countability and indefinite vs. zero article use. *International Review of Applied Linguistics*, *31*(4), 269–289.

Young, R. (1988). Variation and the Interlanguage Hypothesis. *Studies in Second Language Acquisition*, *10*(3), 281–302. doi: 10.1017/S0272263100007464

Young, R. (1996). Form-function relations in articles in English interlanguage. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 135–175). Amsterdam: John Benjamins.

Zdorenko, T., & Paradis, J. (2012). Article in child L2 English: When L1 and L2 acquisition meet at the interface. *First Language*, *32*(1-2), 38–62. doi: 10.1177/0142723710396797

Zobl, H. (1984). Cross-language generalizations and the contrastive dimension of the interlan-

guage hypothesis. In A. Davies, C. Criper, & A. Howatt (Eds.), *Interlanguage* (pp. 79–97). Edinburgh: Edinburgh University Press.

Zobl, H., & Liceras, J. (1994). Functional categories and acquisition orders. *Language Learning*, *44*(1), 159–180.

Appendix A

A Critical Appraisal of Goldschneider and DeKeyser's (2001) Statistical Analysis

Summary of the Relevant Part of Goldschneider and DeKeyser (2001)

The purpose of Goldschneider and DeKeyser (2001) was to analyze "whether a combination of five determinants (perceptual salience, semantic complexity, morphophonological regularity, syntactic category, and frequency) accounts for a large part of the total variance found in [the] acquisition order" (p.1) of L2 English grammatical morphemes. To this end, they quantified the above-mentioned five independent variables for six target morphemes (present progressive *-ing*, plural *-s*, possessive *'s*, articles, third person *-s*, and past tense *-ed*), and ran a multiple regression analysis with the SOC (suppliance in obligatory context) scores as the dependent variable. They reported that the five determinants together explain 71% and 64% of the variance in weighted and unweighted accuracy scores respectively[8]. A key feature of their study in the present context is that it was a meta-analysis, that is, for each morpheme, they had multiple (10 or more) SOC scores from different primary studies, resulting in 72 (rather than six) cases with four missing values in the dependent variable.

Theoretical Issues

The statistical analysis Goldschneider and DeKeyser (2001) conducted is unfortunately flawed in a significant way. The fundamental problem in their analysis is the small number of morphemes compared to the number of independent variables. To understand this, however, let us first have a look at the structure of their data. Their data have a multilevel, hierarchical structure of SOC scores nested within morphemes (Hox, 2010). In other

[8]In their study, "the accuracy scores for each functor from each study were weighted according to the number of subjects in each study. This was done to balance the extremes in the primary studies, where the number of subjects ranged from 6 (Rosansky, 1976) to 422 (Mace-Matluck, 1979)" (p.34).

words, there are multiple SOC values for one morpheme. The SOC scores are also nested within primary studies, but to keep things simple, I will focus only on the nesting within morphemes here.

A corollary of this multilevel structure is that the variance can be divided into multiple levels. In the present case, it can be divided into two parts:

- between-morpheme variance, that is, the variance attributable to the difference of the morphemes (e.g., progressive -ing generally has higher SOC scores than past tense -ed). This is also known as group-level variance.
- and within-morpheme variance, that is, the variance within morphemes (e.g., articles were more accurate in Study X than in Study Y).

Similarly, independent variables can also be divided into two types; ones that explain the between-morpheme variance and the ones that explain within-morpheme variance. Goldschneider and DeKeyser (2001) only have group-level independent variables that show the characteristics of each morpheme (perceptual salience etc.). This means that the values of independent variables only change between morphemes but not within them. Group-level independent variables only explain between-morpheme variance because the values of the set of independent variables are the same within morphemes and there is no information to explain the within-morpheme variance.

The authors were aware of the multilevel structure of their data, stating in a footnote that

> [s]cores on the predictor (independent) variables are "clustered"; that is, the same scores for a given predictor are used repeatedly in the multiple regression equation. This is not a problem when scores on the independent variables are clustered. There would be a problem for the analysis if the criterion (dependent) variable did not represent independent observations, but this is not

> the case. . . . A multi-level analysis would be preferable from certain points of view, but would require a larger *n* (in this case, number of studies) (p.32).

However, the multilevel structure in their data that required attention is not that of primary studies and SOC scores, as is indicated in the last sentence above, but that of morphemes and SOC scores[9], as explained above.

The crucial problem in the multiple regression model of Goldschneider and DeKeyser (2001) is that it has only six groups (or morphemes) while there are five group-level independent variables. This is virtually the same as having six cases with five independent variables in an ordinary, single-level multiple regression analysis. It is statistically impossible to estimate six parameters (estimates for five independent variables and the intercept) based on six cases.

Illustration

Table A1 through A3 are cases where R^2 is 1 in multiple regression models. In the tables, IV stands for independent variable and DV dependent variable. The values were randomly drawn from one to nine (integers only). Note that the number of observations equals to the number of independent variables plus one. In these cases, the R^2 is always 1, regardless of the values of independent and dependent variables.

Similar is the case in Goldschneider and DeKeyser (2001), although the story is a little more complicated due to the multilevel structure of their data. The following shows that replacing the five independent variables of Goldschneider and DeKeyser (2001) with some arbitrary numbers still produces the same R^2 value as their study. The illustrations below employ unweighted multiple regression for the sake of simplicity. The point is the same in the case of weighted regression as well.

[9] This may be what was implied in the first sentence of the quote, but I cannot be certain.

Table A1

Example 1

	IV-A	IV-B	DV
Observation 1	4	7	4
Observation 2	5	2	3
Observation 3	7	3	7

Table A2

Example 2

	IV-A	IV-B	IV-C	DV
Observation 1	1	8	6	4
Observation 2	8	1	9	9
Observation 3	4	1	3	2
Observation 4	1	4	2	8

Table A3

Example 3

	IV-A	IV-B	IV-C	IV-D	DV
Observation 1	8	7	2	4	9
Observation 2	3	9	4	6	4
Observation 3	7	4	8	1	3
Observation 4	2	6	3	8	8
Observation 5	1	4	2	3	5

Table A4

Original Independent Variables of Goldschneider and DeKeyser's (2001) Regression Model

	Perceptual salience	Semantic complexity	Morphophonological regularity	Syntactic category	Frequency
Articles	1.008	1	0.334	2	552
Past tense *-ed*	-0.761	2	-0.456	1	44
Plural *-s*	-0.578	1	0.456	3	147
Possessive *'s*	-0.578	1	0.456	1	71
Progressive *-ing*	1.486	2	-1.247	3	160
Third person *-s*	-0.578	3	0.456	1	89

Table A4 is the summary of the values of independent variables employed in their study. It is noticeable that the number of true observations, or the number of unique set of independent variables, equals to the number of independent variables plus one. When running a multiple regression, each row was repeated multiple times with different values of the dependent variable (SOC scores), and there were 72 cases with six patterns of independent variables.

Table A5 shows the result of the Goldschneider and DeKeyser's (2001) unweighted regression model. Because there are multiple data points per morpheme (or per certain set of independent variables), unlike Table 1 through 3, the predicted value does not exactly match with the observed values, which means that the total residual is not zero and, accordingly, the R^2 is not 1. In other words, because their model only includes group-level independent variables, it only explains the between-morpheme variance, but not the within-morpheme variance. What this means is that the R^2 value of 63.7% is the amount of variance attributable to the difference of morphemes.

Table A6 shows the independent variables whose values are randomized from the

Table A5

Summary of the Unweighted Linear Multiple Regression Model of Goldschneider and DeKeyser (2001)

Parameter	B	SE	*t*	*p*	β
Intercept	50.044	11.115	4.502	0.000	
Perceptual salience	-2.007	5.759	-0.349	0.729	-0.080
Semantic complexity	-6.988	3.002	-2.328	0.023	-0.231
Morphophonological regularity	-14.690	5.786	-2.539	0.014	-0.421
Syntactic category	10.011	2.708	3.696	0.000	0.394
Frequency	0.045	0.024	1.871	0.066	0.345

Note. $R^2 = 0.637$; $adj.R^2 = 0.608$; $F(5, 62) = 21.750$; $p < 0.001$

original values. The randomization is column-wise so the means and standard deviations of each independent variable is the same as the original ones. The values of independent variables in Table A7 were randomly chosen from 1 to 9.

Table A8 and A9 are the summaries of the regression models developed with the values of independent variables in Table A6 and A7 respectively. Although the coefficients (B) are widely different from the original model, the R^2 values are identical to the original one, 0.637. This demonstrates that, for the R^2 value, the quantification of the independent variables was of no importance in Goldschneider and DeKeyser (2001), and that what mattered was merely the number of independent variables. It should be sufficient to disprove the main claim of Goldschneider and DeKeyser (2001) that the five determinants they employed account for a large portion of the variance of SOC scores because the five determinants did not have to be the five variables they used in order to obtain the same result.

Finally, Table A10 is a set of independent variables in which morphemes are directly

Table A6

Randomized Independent Variables

	Perceptual salience	Semantic complexity	Morphophonological regularity	Syntactic category	Frequency
Articles	-0.578	1	-1.247	1	160
Past tense *-ed*	-0.578	1	0.456	2	71
Plural *-s*	-0.761	1	0.456	3	147
Possessive *'s*	-0.578	2	0.334	1	552
Progressive *-ing*	1.486	3	-0.456	1	89
Third person *-s*	1.008	2	0.456	3	44

Table A7

Random Values of Independent Variables

	Perceptual salience	Semantic complexity	Morphophonological regularity	Syntactic category	Frequency
Articles	8	6	4	5	1
Past tense *-ed*	2	6	9	4	8
Plural *-s*	9	7	8	5	3
Possessive *'s*	1	9	3	8	2
Progressive *-ing*	5	2	1	4	9
Third person *-s*	8	3	4	6	7

Table A8

Summary of the Regression Model with Randomised Independent Variables

Parameter	B	SE	*t*	*p*	β
Intercept	-105.790	28.112	-3.763	0.000	
Perceptual salience	-92.624	12.475	-7.425	0.000	-3.687
Semantic complexity	105.945	14.538	7.288	0.000	3.506
Morphophonological regularity	-35.028	5.761	-6.080	0.000	-1.007
Syntactic category	16.620	5.153	3.225	0.002	0.663
Frequency	-0.206	0.029	-7.104	0.000	-1.482

Note. $R^2 = 0.637$; $adj.R^2 = 0.608$; $F(5, 62) = 21.750$; $p < 0.001$

Table A9

Summary of the Regression Model with Random Values as Independent Variables

Parameter	B	SE	*t*	*p*	β
Intercept	68.381	36.251	1.886	0.064	
Perceptual salience	5.516	1.782	3.095	0.003	0.730
Semantic complexity	18.801	4.592	4.094	0.000	1.935
Morphophonological regularity	-10.658	1.828	-5.829	0.000	-1.290
Syntactic category	-21.488	2.207	-9.736	0.000	-1.240
Frequency	5.743	2.664	2.155	0.035	0.783

Note. $R^2 = 0.637$; $adj.R^2 = 0.608$; $F(5, 62) = 21.750$; $p < 0.001$

Table A10

Dummy Variables of Morphemes

	Progressive *-ing*	Plural *-s*	Possessive *'s*	Past tense *-ed*	Third person *-s*
Articles	0	0	0	0	0
Past tense *-ed*	0	0	0	1	0
Plural *-s*	0	1	0	0	0
Possessive *'s*	0	0	1	0	0
Progressive *-ing*	1	0	0	0	0
Third person *-s*	0	0	0	0	1

encoded as dummy variables. This produces a model that fully explains the between-morpheme variance because the independent variables directly show which morpheme the SOC score is nested within. Table A11 is the summary of the model. As expected, the R^2 value is the same as the other models, which indicates that all the models discussed so far fully explain the between-morpheme variance.

Table A11

Summary of the Regression Model with Dummy Variables

Parameter	B	SE	*t*	*p*
Intercept	80.986	4.162	19.457	0.000
Past tense -*ed*	-24.696	6.174	-4.000	0.000
Plural -*s*	-6.828	5.886	-1.160	0.250
Possessive *'s*	-30.270	6.174	-4.903	0.000
Progressive -*ing*	7.649	5.886	1.299	0.199
Third person -*s*	-43.437	5.886	-7.379	0.000

Note. $R^2 = 0.637$; $adj.R^2 = 0.608$; $F(5, 62) = 21.750$; $p < 0.001$

Appendix B

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level

Table B1

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level

	Articles																
	Proficiency																Mean TLU
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	& Total OC
Brazilian-Portuguese																	
TLU	0.70	0.79	0.83	0.81	0.86	0.82	0.87	0.87	0.86	0.88	0.89	0.91	0.92	0.92	0.86	0.91	0.86
OC	7,916	9,093	6,070	18,441	10,989	4,287	14,972	4,268	3,477	6,117	2,546	1,014	1,809	329	103	188	91,619
Mandarine-Chinese																	
TLU	0.77	0.83	0.86	0.86	0.86	0.83	0.87	0.85	0.85	0.88	0.87	0.89	0.89	0.85	0.91	0.90	0.86
OC	11,033	14,199	28,650	74,848	44,315	12,820	51,596	16,086	9,431	20,839	4,988	1,732	1,502	244	88	30	292,401
German																	
TLU	0.86	0.88	0.88	0.89	0.92	0.90	0.91	0.93	0.93	0.93	0.93	0.96	0.95	0.95	0.96	0.94	0.92
OC	2,118	2,203	1,315	5,459	3,621	1,664	5,774	1,783	2,033	4,671	2,011	1,197	1,969	666	315	261	37,060
French																	
TLU	0.77	0.84	0.85	0.84	0.87	0.85	0.89	0.89	0.89	0.91	0.92	0.94	0.92	0.96	0.96	0.88	0.89
OC	2,014	1,514	885	3,908	2,016	1,233	4,404	1,498	2,003	3,453	1,605	1,039	610	110	109	54	26,455
Italian																	
TLU	0.78	0.84	0.85	0.84	0.88	0.84	0.89	0.91	0.89	0.90	0.91	0.92	0.93	0.96	1.00	0.93	0.89
OC	1,174	1,364	1,580	6,288	4,414	2,807	6,487	1,776	1,840	2,790	1,411	637	772	208	7	237	33,792
Japanese																	
TLU	0.73	0.82	0.81	0.80	0.79	0.70	0.84	0.78	0.74	0.86	0.83	0.88	0.89	0.95	0.88	NA	0.82
OC	598	543	333	1,168	992	536	1,601	626	788	1,008	620	268	247	117	58	0	9,503
Korean																	
TLU	0.64	0.76	0.80	0.75	0.81	0.65	0.79	0.70	0.78	0.78	0.79	0.79	0.83	0.47	0.68	0.59	0.73
OC	321	221	121	589	473	299	860	95	283	347	250	176	158	17	35	17	4,262
Russian																	
TLU	0.68	0.79	0.79	0.78	0.82	0.77	0.84	0.86	0.86	0.90	0.90	0.92	0.93	0.93	0.93	0.91	0.85
OC	3,170	3,912	3,016	8,228	7,579	6,483	14,601	13,514	17,679	27,498	19,647	10,905	7,423	3,365	2,093	1,054	150,167
Spanish																	
TLU	0.72	0.82	0.83	0.83	0.86	0.84	0.88	0.89	0.89	0.90	0.91	0.94	0.95	0.88	1.00	NA	0.88
OC	1,941	2,113	1,458	4,422	2,719	1,256	3,550	1,699	1,649	1,879	1,276	867	634	188	19	0	25,670
Turkish																	
TLU	0.66	0.80	0.83	0.76	0.78	0.76	0.78	0.82	0.68	0.79	0.88	0.83	0.80	0.81	NA	1.00	0.80
OC	481	540	322	1,121	560	256	995	166	134	562	87	45	109	33	0	25	5,436
Mean TLU & Total OC (Seven L1 groups targeted in Chapter 3)																	
Mean TLU	0.72	0.82	0.83	0.81	0.84	0.78	0.85	0.84	0.82	0.87	0.88	0.89	0.90	0.85	0.90	0.86	0.84
Total OC	10,643	11,046	7,450	24,895	17,960	11,727	31,785	19,381	24,569	39,418	25,496	14,497	11,150	4,496	2,629	1,411	258,553
Mean TLU & Total OC (All L1 groups)																	
Mean TLU	0.73	0.82	0.83	0.82	0.85	0.80	0.86	0.85	0.84	0.87	0.88	0.90	0.90	0.87	0.91	0.88	0.85
Total OC	30,766	35,702	43,750	124,472	77,678	31,641	104,840	41,511	39,317	69,164	34,441	17,880	15,233	5,277	2,827	1,866	676,365

Note. OC = Number of obligatory contexts

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (Continued)

Past tense *-ed*																	
	Proficiency																Mean TLU
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	& Total OC
Brazilian-Portuguese																	
TLU	0.84	0.75	0.78	0.92	0.87	0.80	0.85	0.89	0.85	0.88	0.93	0.95	0.93	0.84	1.00	0.83	0.87
OC	236	91	38	3,164	825	533	991	366	383	494	223	126	123	28	6	10	7,637
Mandarine-Chinese																	
TLU	0.74	0.79	0.55	0.91	0.86	0.77	0.80	0.83	0.76	0.89	0.87	0.94	0.97	0.74	0.73	1.00	0.82
OC	184	130	56	12,141	3,060	1,235	3,029	1,475	1,076	1,874	444	207	63	16	9	2	25,001
German																	
TLU	0.73	0.92	0.92	0.91	0.88	0.78	0.81	0.94	0.88	0.89	0.91	0.90	0.96	0.99	0.95	1.00	0.90
OC	45	73	22	772	233	130	347	114	208	412	119	152	71	68	36	17	2,819
French																	
TLU	0.78	0.80	0.93	0.89	0.88	0.74	0.81	0.76	0.84	0.87	0.89	0.91	0.95	1.00	1.00	NA	0.87
OC	43	49	26	523	109	88	258	121	208	247	120	125	21	13	33	0	1,984
Italian																	
TLU	0.78	0.81	0.82	0.88	0.93	0.79	0.84	0.91	0.86	0.88	0.92	0.94	0.93	0.94	NA	0.94	0.88
OC	21	24	18	1,057	259	266	429	119	229	165	89	105	27	15	0	15	2,838
Japanese																	
TLU	0.72	1.00	0.75	0.90	0.87	0.80	0.91	0.88	0.84	0.98	0.92	0.97	0.83	1.00	1.00	NA	0.89
OC	23	26	4	155	64	82	89	31	114	91	67	36	5	8	13	0	808
Korean																	
TLU	0.62	1.00	0.50	0.92	0.93	0.66	0.88	0.85	0.86	0.81	0.94	0.93	0.75	1.00	0.86	NA	0.83
OC	9	5	1	93	26	38	52	11	41	22	30	51	7	1	21	0	408
Russian																	
TLU	0.85	0.84	0.92	0.93	0.92	0.86	0.88	0.91	0.95	0.97	0.94	0.97	0.96	0.97	0.95	1.00	0.93
OC	63	78	24	1,424	519	750	974	1,090	2,535	2,757	1,593	1,462	392	416	266	50	14,393
Spanish																	
TLU	0.73	0.81	0.80	0.90	0.84	0.76	0.72	0.81	0.83	0.78	0.85	0.92	0.94	1.00	1.00	NA	0.85
OC	41	35	8	682	214	146	206	81	151	91	114	103	29	7	9	0	1,917
Turkish																	
TLU	0.83	0.75	NaN	0.94	0.92	0.73	0.80	0.86	0.66	0.91	1.00	0.86	1.00	1.00	NA	NA	0.87
OC	10	10	0	130	35	48	81	13	37	52	9	6	10	5	0	0	446
Mean TLU & Total OC (Seven L1 groups targeted in Chapter 3)																	
Mean TLU	0.75	0.87	0.80	0.91	0.89	0.76	0.83	0.86	0.84	0.89	0.92	0.92	0.91	0.99	0.96	1.00	0.88
Total OC	234	276	85	3,779	1,200	1,282	2,007	1,461	3,294	3,672	2,052	1,935	535	518	378	67	22,775
Mean TLU & Total OC (All L1 groups)																	
Mean TLU	0.76	0.85	0.77	0.91	0.89	0.77	0.83	0.86	0.83	0.89	0.92	0.93	0.92	0.95	0.94	0.95	0.87
Total OC	675	521	197	20,141	5,344	3,316	6,456	3,421	4,982	6,205	2,808	2,373	748	577	393	94	58,251

Note. OC = Number of obligatory contexts

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (Continued)

	Plural -*s*																
	Proficiency																Mean TLU
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	& Total OC
Brazilian-Portuguese																	
TLU	0.94	0.89	0.92	0.91	0.92	0.89	0.94	0.93	0.93	0.94	0.94	0.94	0.96	0.97	0.99	0.95	0.94
OC	19,371	5,958	3,973	13,573	6,154	2,600	12,099	2,431	1,912	3,509	1,617	532	1,495	252	66	124	75,666
Mandarine-Chinese																	
TLU	0.94	0.89	0.91	0.90	0.91	0.88	0.92	0.89	0.91	0.92	0.93	0.92	0.93	0.92	0.87	1.00	0.92
OC	23,413	9,042	15,064	47,958	23,934	7,519	43,134	7,613	4,680	11,040	3,588	943	1,062	143	53	27	199,213
German																	
TLU	0.94	0.94	0.94	0.92	0.93	0.92	0.96	0.94	0.93	0.96	0.96	0.95	0.96	0.98	0.96	0.94	0.95
OC	3,076	1,320	903	3,666	1,749	920	4,674	886	868	2,402	1,203	629	1,423	414	197	126	24,456
French																	
TLU	0.92	0.91	0.93	0.89	0.93	0.88	0.92	0.88	0.90	0.94	0.93	0.93	0.95	0.97	0.88	0.86	0.91
OC	3,191	912	568	2,728	1,094	707	3,505	763	963	1,727	1,043	458	463	34	46	42	18,244
Italian																	
TLU	0.93	0.90	0.89	0.90	0.92	0.91	0.92	0.90	0.92	0.91	0.93	0.92	0.96	0.92	1.00	0.97	0.93
OC	2,234	785	975	4,000	2,152	1,424	4,826	828	819	1,365	882	335	534	89	4	147	21,399
Japanese																	
TLU	0.93	0.93	0.91	0.88	0.89	0.88	0.92	0.88	0.83	0.95	0.94	0.95	0.93	1.00	0.94	NA	0.92
OC	1013	345	190	849	558	322	1,311	335	377	585	578	116	203	76	46	0	6,904
Korean																	
TLU	0.93	0.91	0.86	0.90	0.88	0.86	0.91	0.83	0.83	0.89	0.92	0.95	0.89	0.50	0.71	0.57	0.83
OC	438	156	57	390	266	135	685	48	136	218	186	77	118	6	28	13	2,957
Russian																	
TLU	0.96	0.93	0.92	0.94	0.94	0.94	0.96	0.94	0.96	0.97	0.97	0.98	0.98	0.98	0.99	0.98	0.96
OC	6,018	2,555	1,868	6,332	4,641	4,600	13,109	7,471	10,753	16,581	16,121	7,400	6,536	2,470	1,395	812	108,662
Spanish																	
TLU	0.94	0.88	0.93	0.92	0.93	0.91	0.95	0.93	0.95	0.95	0.96	0.96	0.96	0.95	1.00	NA	0.94
OC	4,834	1,256	1,101	3,169	1,528	771	2,761	839	869	1,086	811	441	505	101	10	0	20,082
Turkish																	
TLU	0.94	0.86	0.91	0.86	0.89	0.84	0.90	0.82	0.87	0.92	0.92	0.90	0.86	0.90	NA	1.00	0.89
OC	1058	320	204	728	295	139	821	85	105	354	49	19	86	41	0	16	4,320
Mean TLU & Total OC (Seven L1 groups targeted in Chapter 3)																	
Mean TLU	0.94	0.91	0.91	0.90	0.91	0.89	0.93	0.89	0.90	0.94	0.94	0.95	0.93	0.90	0.91	0.87	0.91
Total OC	19,628	6,864	4,891	17,862	10,131	7,594	26,866	10,427	14,071	22,953	19,991	9,140	9,334	3,142	1,722	1,009	185,625
Mean TLU & Total OC (All L1 groups)																	
Mean TLU	0.94	0.90	0.91	0.90	0.91	0.89	0.93	0.89	0.90	0.94	0.94	0.94	0.94	0.91	0.93	0.91	0.92
Total OC	64,646	22,649	24,903	83,393	42,371	19,137	86,925	21,299	21,482	38,867	26,078	10,950	12,425	3,626	1,845	1,307	481,903

Note. OC = Number of obligatory contexts

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (Continued)

Possessive *'s*																	
	Proficiency																Mean TLU
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	& Total OC
Brazilian-Portuguese																	
TLU	0.64	0.79	0.91	0.70	0.72	0.76	0.77	0.74	0.69	0.69	0.88	0.76	0.86	0.62	1.00	1.00	0.78
OC	328	228	349	426	183	120	287	89	48	136	101	40	40	8	3	1	2,387
Mandarine-Chinese																	
TLU	0.77	0.74	0.93	0.74	0.80	0.84	0.89	0.90	0.83	0.77	0.91	0.97	0.93	0.94	0.88	NA	0.86
OC	634	225	1,536	1,501	761	457	1,583	513	227	512	225	72	38	16	7	0	8,307
German																	
TLU	0.50	0.85	0.83	0.56	0.71	0.66	0.79	0.81	0.56	0.63	0.90	0.78	0.85	0.94	0.71	0.80	0.74
OC	118	39	49	86	31	42	94	40	16	77	59	58	33	18	7	5	772
French																	
TLU	0.64	0.85	0.82	0.67	0.76	0.70	0.69	0.86	0.69	0.78	0.87	0.86	0.92	1.00	1.00	NA	0.81
OC	100	35	52	83	31	42	68	37	26	83	66	36	12	7	4	0	682
Italian																	
TLU	0.76	0.93	0.82	0.71	0.77	0.81	0.70	0.86	0.77	0.59	0.96	0.62	0.58	0.67	NA	1.00	0.77
OC	93	44	89	149	65	89	81	54	25	53	72	24	10	2	0	5	855
Japanese																	
TLU	0.80	1.00	1.00	0.93	0.78	0.89	0.96	0.93	0.83	0.85	0.88	0.93	1.00	1.00	1.00	NA	0.92
OC	77	25	18	42	23	19	73	26	21	33	48	15	4	5	5	0	434
Korean																	
TLU	0.83	1.00	1.00	0.81	0.88	0.73	0.96	0.67	1.00	0.87	1.00	0.81	0.90	1.00	0.80	1.00	0.89
OC	29	15	10	16	17	14	54	3	6	15	16	19	9	10	5	1	239
Russian																	
TLU	0.78	0.91	0.88	0.67	0.73	0.88	0.87	0.87	0.85	0.87	0.92	0.90	0.95	0.95	0.96	0.97	0.87
OC	218	97	197	180	80	252	352	394	310	791	940	399	235	124	106	37	4,712
Spanish																	
TLU	0.60	0.71	0.91	0.77	0.85	0.73	0.47	0.75	0.65	0.61	0.79	0.86	0.75	NA	NA	NA	0.73
OC	96	55	108	96	37	35	53	38	25	21	36	28	4	0	0	0	632
Turkish																	
TLU	0.62	0.82	0.89	0.60	0.80	0.89	0.85	1.00	1.00	1.00	1.00	1.00	1.00	0.50	NA	1.00	0.86
OC	38	28	18	33	10	19	32	2	5	17	5	1	2	2	0	1	213
Mean TLU & Total OC (Seven L1 groups targeted in Chapter 3)																	
Mean TLU	0.68	0.88	0.90	0.72	0.79	0.78	0.80	0.84	0.80	0.80	0.91	0.88	0.91	0.90	0.89	0.94	0.83
Total OC	676	294	452	536	229	423	726	540	409	1,037	1,170	556	299	166	127	44	7,684
Mean TLU & Total OC (All L1 groups)																	
Mean TLU	0.69	0.86	0.90	0.72	0.78	0.79	0.80	0.84	0.79	0.77	0.91	0.85	0.87	0.85	0.91	0.96	0.83
Total OC	1,731	791	2,426	2,612	1,238	1,089	2,677	1,196	709	1,738	1,568	692	387	192	137	50	19,233

Note. OC = Number of obligatory contexts

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (Continued)

Progressive *-ing*																	
	Proficiency																Mean TLU
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	& Total OC
Brazilian-Portuguese																	
TLU	0.80	0.91	0.81	0.94	0.96	0.89	0.94	0.93	0.93	0.95	0.96	0.99	0.98	1.00	1.00	0.92	0.93
OC	756	1,087	131	3,415	2,578	356	1,080	441	267	695	207	116	120	40	8	12	11,309
Mandarine-Chinese																	
TLU	0.76	0.92	0.76	0.93	0.96	0.89	0.92	0.90	0.92	0.94	0.96	0.95	0.92	1.00	0.80	1.00	0.91
OC	712	1,573	373	14,197	12,226	879	3,270	1,262	662	1,939	412	183	75	20	4	1	37,788
German																	
TLU	0.88	0.91	0.92	0.90	0.96	0.88	0.92	0.89	0.89	0.93	0.95	0.96	0.97	0.98	1.00	1.00	0.93
OC	182	194	34	797	743	114	323	129	96	424	135	138	117	56	34	16	3,532
French																	
TLU	0.82	0.94	0.93	0.89	0.96	0.88	0.95	0.89	0.89	0.94	0.96	0.92	0.96	1.00	1.00	1.00	0.93
OC	157	145	14	676	565	67	233	90	96	344	107	113	27	10	16	5	2,665
Italian																	
TLU	0.85	0.92	0.79	0.93	0.95	0.86	0.93	0.94	0.96	0.95	0.97	0.95	1.00	1.00	NaN	0.83	0.92
OC	86	149	27	1,040	1,039	150	348	145	76	272	89	59	39	19	0	12	3,550
Japanese																	
TLU	0.92	0.98	0.78	0.92	0.96	0.79	0.94	0.94	0.96	0.97	0.94	0.97	1.00	1.00	1.00	NA	0.94
OC	104	53	7	243	325	42	95	61	54	126	52	32	15	9	7	0	1,225
Korean																	
TLU	0.85	0.88	0.50	0.89	0.93	0.91	0.92	0.87	0.88	0.91	0.96	0.92	1.00	NaN	0.67	0.00	0.81
OC	38	23	1	94	139	34	58	13	16	54	23	13	5	0	2	0	513
Russian																	
TLU	0.79	0.90	0.75	0.94	0.96	0.92	0.96	0.96	0.96	0.98	0.98	0.97	0.98	0.97	0.97	0.99	0.94
OC	228	404	52	1,680	1,991	446	1,114	1,404	1,468	2,852	1,616	1,449	477	319	192	69	15,761
Spanish																	
TLU	0.72	0.92	0.91	0.94	0.95	0.89	0.91	0.94	0.91	0.95	0.92	0.97	0.93	1.00	1.00	NA	0.92
OC	164	255	33	834	631	95	208	125	95	240	93	97	40	8	6	0	2,924
Turkish																	
TLU	0.96	0.95	1.00	0.93	0.96	0.90	0.93	0.95	0.86	0.89	1.00	1.00	0.71	1.00	NA	1.00	0.94
OC	66	78	2	299	204	20	80	21	13	82	9	4	6	4	0	1	889
Mean TLU & Total OC (Seven L1 groups targeted in Chapter 3)																	
Mean TLU	0.85	0.93	0.83	0.92	0.95	0.88	0.93	0.92	0.91	0.94	0.96	0.96	0.94	0.99	0.94	0.80	0.92
Total OC	939	1,152	143	4,623	4,598	818	2,111	1,843	1,838	4,122	2,035	1,846	687	406	257	91	27,509
Mean TLU & Total OC (All L1 groups)																	
Mean TLU	0.84	0.92	0.82	0.92	0.96	0.88	0.93	0.92	0.92	0.94	0.96	0.96	0.95	0.99	0.93	0.84	0.92
Total OC	2,493	3,961	674	23,275	20,441	2,203	6,809	3,691	2,843	7,028	2,743	2,204	921	485	269	116	80,156

Note. OC = Number of obligatory contexts

TLU Scores of Each Morpheme, L1 Group, and Proficiency Level (Continued)

Third person -*s*																	
	Proficiency																Mean TLU
L1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	& Total OC
Brazilian-Portuguese																	
TLU	0.99	0.97	0.99	0.98	0.97	0.98	0.95	0.96	0.95	0.98	0.97	0.99	0.98	0.97	1.00	1.00	0.98
OC	1,340	1,328	2,950	4,145	608	792	1,719	523	316	767	501	126	214	62	8	21	15,420
Mandarine-Chinese																	
TLU	0.98	0.97	0.99	0.99	0.98	0.98	0.95	0.97	0.98	0.98	0.98	0.97	0.99	1.00	1.00	1.00	0.98
OC	1,812	1,870	10,716	18,169	1,834	2,202	5,823	1,921	966	2,429	917	230	138	49	20	4	49,100
German																	
TLU	0.98	0.99	1.00	0.99	0.99	0.98	0.97	0.97	0.99	1.00	0.99	0.99	1.00	0.99	1.00	0.97	0.99
OC	361	283	583	1,223	212	274	669	214	174	601	469	126	218	79	52	37	5,575
French																	
TLU	0.99	0.99	1.00	0.99	0.99	0.99	0.97	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00	1.00	0.99
OC	283	200	350	793	100	256	468	152	131	381	299	94	58	8	30	6	3,609
Italian																	
TLU	0.99	0.99	0.99	0.99	0.98	0.98	0.96	0.98	0.97	0.99	1.00	1.00	0.98	1.00	NA	1.00	0.99
OC	187	166	678	1,403	263	480	725	214	149	335	247	75	82	17	0	36	5,057
Japanese																	
TLU	1.00	0.99	1.00	1.00	0.95	0.99	0.98	1.00	1.00	0.98	0.99	1.00	1.00	1.00	0.89	NA	0.98
OC	110	106	162	412	57	96	212	93	109	121	149	32	19	24	8	0	1,710
Korean																	
TLU	0.98	1.00	1.00	0.99	0.94	0.98	0.99	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.99
OC	51	36	39	142	32	50	113	11	38	31	62	9	16	8	9	10	657
Russian																	
TLU	1.00	0.99	1.00	0.99	0.98	0.99	0.97	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.97	0.99
OC	572	519	1,580	2,037	358	1,111	1,643	1,989	1,781	3,658	4,514	1,109	710	429	333	164	22,507
Spanish																	
TLU	1.00	0.98	1.00	0.98	0.99	0.99	0.93	0.94	0.98	0.97	0.99	1.00	1.00	1.00	1.00	NA	0.98
OC	355	240	663	1,009	136	218	389	232	139	228	286	124	73	31	2	0	4,125
Turkish																	
TLU	0.99	0.99	0.99	0.99	0.96	0.97	0.97	0.90	1.00	0.99	1.00	1.00	0.94	1.00	NA	1.00	0.98
OC	93	72	182	229	24	86	152	19	11	69	23	2	16	2	0	6	986
Mean TLU & Total OC (Seven L1 groups targeted in Chapter 3)																	
Mean TLU	0.99	0.99	1.00	0.99	0.97	0.98	0.97	0.97	1.00	0.99	0.99	1.00	0.99	1.00	0.98	0.99	0.99
Total OC	1,825	1,456	3,559	5,845	919	2,091	3,646	2,710	2,383	5,089	5,802	1,496	1,110	581	434	223	39,169
Mean TLU & Total OC (All L1 groups)																	
Mean TLU	0.99	0.99	1.00	0.99	0.97	0.98	0.96	0.97	0.99	0.99	0.99	1.00	0.99	1.00	0.99	0.99	0.99
Total OC	5,164	4,820	17,903	29,562	3,624	5,565	11,913	5,368	3,814	8,620	7,467	1,927	1,544	709	462	284	108,746

Note. OC = Number of obligatory contexts