Longitudinal L2 Development of the English Article in Individual Learners

Akira Murakami (am933@cam.ac.uk)
Theodora Alexopoulou (ta259@cam.ac.uk)
Department of Theoretical and Applied Linguistics, 9 West Road,
University of Cambridge, Cambridge, CB3 9DP, United Kingdom

Abstract

We investigate the accuracy development of the English article by learners of English as a second language. The study focuses on individual learners, tracking their learning trajectories through their writings in the EF-Cambridge Open Language Database (EFCAMDAT), an open access learner corpus. We draw from 17,859 writings by 1,280 learners and ask whether article accuracy in individual learners fluctuates randomly or whether learners can be clustered according to their developmental trajectories. In particular, we apply k-means clustering to automatically cluster in a bottom up fashion learners with similar learning curves. We follow learners for a period covering one CEFR level. Given the relatively short learning window, the majority of learners follow a horizontal line. Nevertheless, we also identify groups of learners showing a power-function and U-shaped curve. Crucially, these groups are 'hidden' when the aggregate of learners is considered, a finding highlighting the importance of individual level analysis.

Keywords: learning curve; clustering; individual variation; second language

Introduction

Since the early days of second language acquisition (SLA) research, learner production data have been an important empirical base for developmental research (Selinker, 1972). Early studies revealed the non-linearity of the learning process and, in particular, the existence of U-shaped learning curves (Lightbown, 1983). However, despite important insights regarding learning curves, a number of important issues remain unaddressed. Most studies based on production data tend to be small scale (Lightbown, 1983) and tend to lack longitudinal information for large numbers of individuals. As a result, developmental research has predominately employed cross-sectional designs while it has generally not been possible to investigate the relation between individual learning curves and cross-sectional or aggregate patterns. This has been an important limitation in SLA research in view of evidence from psychology that the averaged pattern can conceal individual developmental trajectories (e.g., Heathcote, Brown, & Mewhort, 2000). These limitations are acknowledged by SLA researchers calling for a stronger focus on individual learners (e.g., Larsen-Freeman, 1997). However, the absence of sufficient amounts of longitudinal data for individuals is an important practical obstacle (Larsen-Freeman & Cameron, 2008).

Recently developed large-scale learner corpora (e.g., Granger, Dagneaux, Meunier, & Paquot, 2009) provide production data rich enough to allow analysis of individuals at various proficiency levels. In the present study, we exploit one such resource, EFCAMDAT, as it enables tracking the productions of sufficient numbers of individual learners. We focus on the accuracy in the use of the English article by learners of English as a second language (L2 learners). Accuracy is an important (though not sole) indicator of acquisition and the article has been central to accuracy investigations from early days (e.g., Dualy & Burt, 1973). In recent work, we show that, despite systematic effects of the native language on the accuracy of the L2 English article, there exists significant individual variation (Murakami, 2016). In the present study, we investigate whether we can identify learning curves in individuals that are absent from the aggregate analysis.

Method

Corpus

The empirical data for the study are drawn from the pre-release version of EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2014) The corpus contains learners' writings submitted to *Englishtown*, the online school of EF Education First. The 16 teaching levels of the EF curriculum range from beginners to advanced proficiency (aligned to A1-C2 in the Common European Framework Reference (CEFR) levels). Writings are on a variety of topics and range from 20-40 words at beginner levels to 150-180 words each at the highest levels. The scripts in the corpus are accompanied by error corrections by teachers. EFCAMDAT is publicly available at http://corpus.mml.cam.ac.uk/efcamdat/.

Each writing includes an anonymous learner ID, national language, the topic of the writing, the date and time of submission, the level and the unit/lesson number the writing was submitted to and error-corrections by teachers (see Geertzen et al, 2014 for further details on EFCAMDAT).

Target Linguistic Feature and L1 Groups

We targeted English articles, indefinite (*a*, *an*) and definite (*the*), because they are highly frequent and therefore provide us with dense longitudinal data while we could calculate their accuracy automatically with high precision.

The effect of mother tongue/First Language/L1 is pervasive in L2 use (Jarvis & Pavlenko, 2007) and specifically shown for the L2 English article (e.g., Murakami & Alexopoulou, 2015; Snape, 2008). Our study targeted 10 L1 groups; Brazilian-Portuguese, Mandarin-Chinese, German, French, Italian, Japanese, Korean,

Russian, Spanish, and Turkish. We used their country of residence as the closest approximation to L1 (Murakami, 2016). L1 Spanish includes two countries of residence; Spain and Mexico. L1 Mandarin-Chinese includes both those living in China and those in Taiwan. L1 Brazilian-Portuguese, German, French, Italian, Korean, Russian, and Turkish learners correspond to those living in Brazil, Germany, France, Italy, Korea, Russia, and Turkey, respectively. This method has yielded reliable L1 effects (Murakami, 2016).

We grouped L1s in two L1-types, PRESENT and ABSENT, depending on whether or not they have an article. The PRESENT group included L1 Brazilian, German, French, Italian, and Spanish. The ABSENT group included L1 Chinese, Japanese, Korean, Russian, and Turkish.

The subcorpus used in the study included nearly 140,000 writings consisting of 10 million words. There were more writings, and thus a larger number of words, at lower levels. The distribution between L1 groups was 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.

Scoring Method

To investigate accuracy in use we need to identify contexts of obligatory use and measure correct suppliance while also capturing erroneous use of the article in contexts where it is not needed (overgeneralization errors). We thus employed the target-like use (TLU) score calculated by the following formula (Pica, 1983);

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

This formula assesses the proportion of correct use and at the same time penalizes the unnecessary use of the article.

Data Extraction

To measure accuracy, we first need to obtain obligatory contexts of article use. We retrieved error-tagged texts and converted them to corrected texts (incorporating the corrections). The number of article occurrences in a corrected text was taken as the number of obligatory contexts in the learner's original writing. For instance, if the phrase *farmer who lived in a small village* was corrected into *a farmer who lived in a small village*, the two instances of *a* in the latter made two obligatory contexts. R scripts were written to automatically retrieve errors using the teacher error tags.

Tracking Development Through Moving Windows

To track individuals over time, we consider consecutive writings and compare TLU scores from one writing to next. However, writings are too short to provide enough obligatory contexts to reliably calculate the TLU score. We thus computed TLU scores over multiple writings constructing windows as close to a single writing with enough obligatory contexts as possible. Our goal was to ensure that the shape of accuracy development is as close as possible to the shape that would be generated if each essay included a large number of obligatory contexts.

We then calculated TLU scores in a moving-window fashion (e.g., van Geert & van Dijk, 2002). Each window included at least 10 obligatory contexts (OCs) and could cover multiple writings. Consider a learner who wrote five writings with OCs as in the following example:

Writing 1; 6 OCs
Writing 2; 6 OCs
Writing 3; 3 OCs
Writing 4; 7 OCs
Writing 5; 2 OCs

Here, the first TLU score would be calculated over Writing 1 and Writing 2 because Writing 1 alone does not include 10 OCs but Writing 1 and Writing 2 combined do. The first window would thus include 12 OCs. In computing the second TLU score, the head of the first window shifts forward by one, and the second window starts from Writing 2. The second window would cover Writing 2 through Writing 4, because Writing 2 alone or Writing 2 and Writing 3 together do not include 10 OCs, but the three writings combined do. The second window would include 16 OCs in total. Similarly, the head of the window now shifts to Writing 3, and the third TLU score would be calculated over Writing 3 and Writing 4. This learner has these three TLU scores in total, as Writing 4 alone, Writing 4 and Writing 5 combined, or Writing 5 alone does not include 10 OCs. Once TLU scores were obtained for all the windows, we analyzed the learners who had 10 or more windows.

Alternatively, it was also possible to construct windows so that no writing overlaps in any window. We chose the overlapping approach because it generally tracks more fine-grained developmental patterns and finer resolution of data is indispensable in the study in order to observe change (Larsen-Freeman & Cameron, 2008). See Murakami (2014) for further discussion.

There were in total 70,879 TLU scores (windows) by 20,394 learners. Out of the 20,394 learners, 1,280 (6.3%) had 10 or more windows. Among those with the minimum of 10 windows, the average number of writings in a window was 2.4 (*SD* = 1.2). The median number of windows was 12. The total number of writings by them was 17,859. The mean number of unique writings over 10 windows was 11.0 (*SD* = 0.9). This is not 10 times as large as the average number of writings in a window because there are overlaps of writings over windows. The average number of Units covered in 10 windows was 25.6 (*SD* = 10.1). This corresponds to just over three EF teaching levels — if the writings always start at levels 1, 3, 7, 10, etc., then it corresponds to one CEFR level. Note that three levels correspond to 24 writings (3 Levels × 8 Units), but the average number of unique writings over 10 windows is less than half of this (11.0). Recall that we selected only error-tagged writings for analysis. As not all writings are error-tagged, consecutive writings in our windows do not

necessarily correspond to adjacent Englishtown tasks and, indeed, span just over three teaching levels on average.

Longitudinal View of Development

There are multiple ways to visualize longitudinal accuracy in use. Figure 1 lists two of them for the 1,280 learners. In both panels, each thin line represents the accuracy development of one learner, and thick lines are locally weighted scatterplot smoothing (LOESS; Singer & Willett, 2003) lines showing the overall trend. Solid, dashed, and dotted lines correspond to the development of the ABSENT, the PRESENT, and the L1 Chinese groups respectively. L1 Chinese was separated because visual inspection (not shown) suggested that they behave differently from the other ABSENT groups, and there are indeed studies claiming that Mandarin-Chinese has linguistic features that play similar roles to English definite (Huang, 1999) and indefinite (Chen, 2004) articles.

The left panel demonstrates the accuracy development of individual learners across Englishtown teaching levels. Englishtown levels 1-16 are shown on the horizontal axis and accuracy is on the vertical axis. Each data point is plotted according to its window proficiency, defined as the weighted average Unit number (1 to 128) of the writings included in the window. Each learner horizontally covers only a short span because a learner typically covers three to four levels. The trend line indicates that accuracy increases as learners' proficiency rises. Individually, however, we can observe that some learners radically go up and down in their accuracy. Since each learner covers a short span in this panel, it is difficult to analyze the development of individual learners here. The trend lines are the cross-sectional trend lines and do not represent longitudinal development.

In order to better understand longitudinal development, the right panel visualizes the development of individual learners so that each learner covers the entire span from left to right. For the sake of comparability, it only targets the first 10 windows of each learner. The horizontal axis here represents the window number. The panel shows that the overall pattern is relatively horizontal, which may appear to contrast with the left panel that shows the overall accuracy increase across Englishtown levels but is in fact consistent, given the short span covered in the longitudinal development in this study. The same panel also demonstrates large individual variation in the development. The accuracy of some learners appears to increase over time, while the accuracy of others seems to fluctuate widely. This suggests that there can further be a difference between the average longitudinal pattern and the learning curves of individual learners.

Clustering Learners According to Their Developmental Shapes

K-Means Clustering and Number of Clusters

Now that we visually grasped the longitudinal development of the morphemes, we will explore the learning curves that characterize the learners we target. To obtain the groupings that reflect the data, we employed a data-driven way of determining developmental shapes, namely k-means clustering.

As the input data, we used the first 10 windows of each learner. Our interest here is the developmental trajectory. However, if k-means clustering 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 neutralize the effect of 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 within the first 10 windows.

Because the bottom-up approach is open not only to the shape but also number of clusters, determining the shape and number of clusters is two separate issues. The approach we take 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 development, appear at. If a certain pattern consistently appears with varying k's, it is likely that the pattern reflects some kind of mechanism in learners' performance development.

We chose 10 for the maximum number of clusters because there are 10 L1 groups in total and 10 different patterns are expected if each L1 has a distinct developmental pattern. Figure 2 shows the LOESS of the developmental patterns of each cluster when k varies from 2 to 10. Note that although k-means clustering was performed on learner-mean-centered data, we show the result in the original scale (i.e., TLU scores) for the ease of interpretation. The horizontal axis of the figure represents window number 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 LOESS of the development of the learners in each cluster. Cluster A is always the largest cluster in terms of the number of included learners, followed by Cluster B, which in turn is followed by Cluster C, and so forth.

We can make a few observations here. First, from $k = 4$ onwards, the accuracy of the largest cluster (Cluster A) is relatively unchanged. Second, there is always a cluster showing an upward trend over 10 windows (e.g., Cluster B in $k = 2$) and a cluster showing decreasing accuracy over the period (e.g., Cluster C in $k = 3$). Third, the U-shaped pattern is prevalent. It first appears in $k = 3$ as Cluster A, and can always be observed until $k = 10$ as Clusters C, H, and I. Finally, some clusters are reminiscent of the power and the exponential function. For instance, Cluster B in $k = 3$, Cluster C in $k = 4$, and Cluster D in $k = 5$ are all more or less similar to the concave function in that their accuracy increase slows down as learners progress.

In sum, the patterns of (i) relatively horizontal, (ii) increasing or decreasing accuracy including the power-law and the exponential shape, and (iii) U-shaped curve are robust and can be observed almost irrespective of the number of clusters. On the other hand, increasing the

number of clusters only results in finer splits of the same patterns and hardly leads to the emergence of new patterns. As shown in Figure 2, few new patterns emerge after $k = 3$. This means that having $k = 4$ or more is unlikely to reveal important longitudinal patterns. We thus assume $k = 3$ is optimal for article patterns derived from k-means clustering. The main criterion for decision here is how informative each k is, and whether new information or pattern can be revealed. Note that although we chose $k = 3$ as the number of clusters, we repeated the procedure for $k = 2$ and $k = 4$ and confirmed that our results, including cluster validation discussed later, remain the same. The point here is not to argue that $k = 3$ is the best number of clusters, or subpopulations, but to illustrate that there are indeed subgroups and that they can be hidden from the aggregated data.

Individual Developmental Patterns of Clusters

Figure 3 shows the clusters of article development when $k = 3$. The upper three panels present the clusters situated in the entire Englishtown course, and the lower three panels show the same data over 10 windows. In other words, the upper panel corresponds to the left panel in Figure 1 divided into the three clusters, while the lower panel corresponds to the right panel in Figure 1, again divided into the three clusters. As before, thin lines represent individual learners and thick lines represent LOESS. The learners are approximately equally spread between the three clusters. L1 type does not affect clustering because the three LOESS lines largely overlap. Learners in Cluster A 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, being consistently over 0.8. The accuracy in Cluster B gradually rises until around the seventh window, after which it levels off possibly due to the ceiling effect. The pattern is reminiscent of the power/exponential developmental pattern. Learners in Cluster C show a horizontal development until the fifth window, after which their accuracy decreases. Significant individual differences can be observed in each cluster, some learners radically going down and others rapidly going up.

Cluster Validation

An important question at this point is whether the identified clusters are 'real'. In other words, is it the difference in the true learning curve that the k-means clustering above reflects or is it just the random noise? We empirically tested the cluster validity in the following manner. The overall idea is similar to the usual statistical testing procedure: We compute the null distribution of the metric that measures goodness of clustering based on random data, and if the value of the metric in the observed data falls outside of its 95% range, we consider it as the evidence that the observed clusters are too good to be derived from random data and conclude that our clusters indeed reflect the difference in the learning trajectory. The key decisions we need to make are (i) what metric to use to measure goodness of clustering and (ii) how to conceptualize the null hypothesis and derive the null distribution, both of which are detailed below.

As the metric for goodness of clustering, we used a measure called the *silhouette* (Rousseeuw, 1987). Intuitively, the silhouette value is large if within-cluster dissimilarity is small (i.e., learners within each cluster have similar developmental trajectories) and between-cluster dissimilarity is large (i.e., learners in different clusters have different learning curves). The silhouette is given to each data point, and all the silhouette values are averaged to measure the cluster distinctiveness of a cluster analysis. Any distance can be used to calculate dissimilarity, and squared Euclidian distance was used in the present case. The mean silhouette value in our clustering was 0.151. If this value is higher than the 95% range of the null distribution of the mean silhouette value, we conclude that there are multiple learning curves.

The null hypothesis is that there is no systematic pattern in intra-learner variability. We obtained the distribution of the mean silhouette value under this null hypothesis and tested the significance of our observed value. The idea here is that we practically randomize the order of the writings within individual learners and follow the same procedure as our main analysis. Since the order of writings is random, there should not be any systematic pattern of development observed. The clusters obtained in this manner thus captures noise alone. We calculate the mean silhouette value on the noise-only, random clusters, and obtain its distribution by repeating the whole procedure a large number of times.

More specifically, the following procedure was employed.

1. For each learner,
    (a) We randomly sampled a large number (e.g., 100) of his/her writings. Here, the same writing could be selected multiple times.
    (b) We then calculated the TLU scores of the first 10 windows.
2. With the data obtained in 1, we ran a k-means cluster analysis with k = 3 and calculated its mean silhouette value.
3. 1 and 2 were repeated 1,000 times, resulting in 1,000 mean silhouette values that we consider as the null distribution.
4. We examined whether the 95% range of 3 included 0.151, the observed mean silhouette value in the present study.

The resulting null distribution showed that the upper bound of its 95% range was 0.144. Therefore, our clusters with the mean silhouette value of 0.151 is considered non-random, and what our clusters capture is not only noise.

Discussion

The present study investigated the longitudinal accuracy use of the L2 English article. We observed large individual variation in the developmental pattern. We, therefore, investigated whether we can identify systematic learning curves in a bottom-up manner. A cluster analysis identified

three learning curves that are followed by approximately an equal number of learners.

We demonstrated that there can be differences between average longitudinal development of a group of learners and the individual learning curves that constitute the group. When we aggregated the data and looked at the developmental trajectory of the average accuracy, it was fairly horizontal, (falsely) implying constant accuracy over the course of development. However, clustering demonstrated that there are indeed subgroups hidden in the aggregate pattern. This suggests that the two types of learning curves can differ, and that we cannot necessarily infer the development of individual learners based on the longitudinal data aggregated over multiple learners. Whereas prior research in psychology has shown the mismatch between aggregated and individual learning curves a number of times (Heathcote, et al., 2000), we believe the present study is the first that empirically demonstrated it in the field of SLA.

The obvious question now is why should learners show these different developmental patterns, and, in particular what is the cause of the U-shape pattern. Admittedly, we can offer no insight at this point to this question but we can note at least two potential sources of variation: (i) changes in the internal knowledge of learner related to reanalysis or increased complexity which is known to impact negatively on accuracy accuracy (Skehan & Foster, 1997); (ii) learner individual variation (e.g., working memory, aptitude, etc.). The investigation of these potential sources requires investigation of specific hypotheses (e.g., evidence for increased complexity in the learner language coinciding with drop in accuracy) or access to individual psychological data (which would require investigation beyond the corpus). Though the interpretation of such results is very challenging, we believe that our findings show the need for investigating individual patterns in a more comprehensive way.

Acknowledgments

The authors gratefully acknowledge the sponsorship of EF Education First.

References

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

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

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 database: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Millar et al. (Eds.), *Selected proceedings of the 2012 Second Language Research Forum. Building bridges between disciplines* (pp. 240–254). Cascadilla Proceedings Project.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*(2), 185–207.

Huang, S. (1999). The emergence of a grammatical category definite article in spoken Chinese. *Journal of Pragmatics, 31*(1), 77–94.

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

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

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

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.

Murakami, A. (2014). *Individual variation and the role of L1 in the L2 development of English grammatical morphemes: Insights from learner corpora* (Unpublished PhD thesis). University of Cambridge, Cambridge, United Kingdom.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear L2 development: The case of English grammatical morphemes. *Language Learning*. Advance online publication.

Murakami, A., & Alexopoulou, T. (2015). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*. Advance online publication.

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*(1), 69–78.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(1–4), 209-232.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research, 1*(3), 185–211.

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.

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.

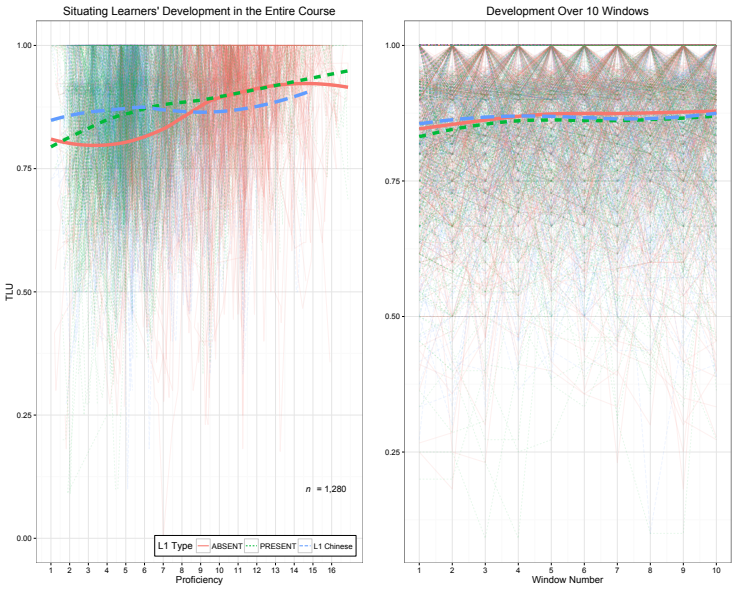

Figure 1: Longitudinal Development of Article Accuracy

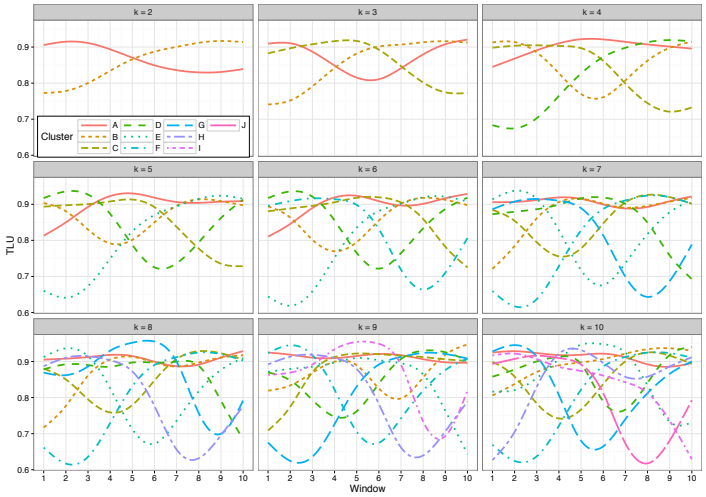

Figure 2: Developmental of Each Cluster in Varying Numbers of Clusters

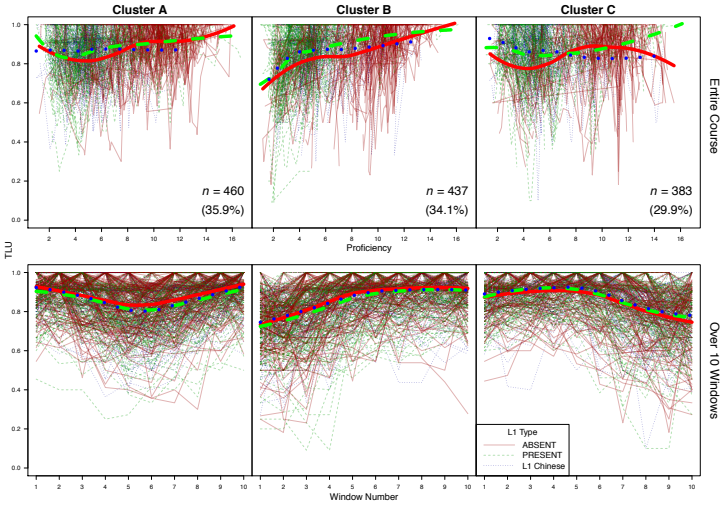

Figure 3: Results of K-Means Clustering