EXPLORING ENGAGEMENT IN ONLINE VIDEOS FOR LANGUAGE LEARNING THROUGH YOUTUBE'S LEARNING ANALYTICS

Maria Dolores Castrillo, Universidad Nacional de Educación a Distancia (UNED), Jorge Mañana-Rodriguez, Universidad Carlos III de Madrid, Spain

Abstract

Until a few years ago, video analytics were not accessible to learning stakeholders, mainly because online video platforms did not share the users' interactions on the system with stakeholders. However, this scenario has changed, and currently YouTube, the world's largest media sharing site, offers these data. YouTube is also the main tool for transmitting audio-visual content in Language MOOCs (massive open online courses), and its video engagement data can be monitored through the YouTube Studio channel, which provides free and open access to video analytics. In this paper we present our research based on the analysis of viewers' engagement with 35 videos of the Language MOOC entitled *Alemán para hispanohablantes: basic principles* (German for Spanish-speakers). The data provided by the YouTube Studio Learning Analytics platform has enabled new insights related to participants' watching of these videos in Language MOOCs (LMOOCs). The results of our study provide pedagogical implications for Foreign Language instructors concerning the use of videos in language learning.

Introduction

The first Language Massive Online Open Courses (LMOOCs) started to appear in October 2012. The Spanish University for Distance Education (UNED), the world pioneer in offering LMOOCs, started to offer on that date, among other courses, the LMOOC *German for Spanish-speakers*, coordinated and taught by one of the authors of the present study. The main contents of this LMOOC (which continues to be taught and updated without interruption to date) are transmitted mainly through video-pills, which according to recent investigations have turned out to be "the most powerful learning objects in this type of online courses" (Martín-Monje et al., 2018; p.4). It is specifically an xMOOC, characterized by the use of video lectures as the main tool for delivering content. In the first edition of the LMOOC, the videos were broadcast through two open channels: UNED's own channel (CanalUNED) and YouTube. Almost a decade after these 35 videos were recorded and

broadcast, at the time of this study, the videos had received more than 2,000,000 views on the YouTube channel. As research confirms, student retention is an outstanding issue in LMOOCs, and we believe that analysis of students' engagement with the videos contained in the LMOOC can help us understand why they drop out of these courses. Engagement is usually measured by video watching time, so our research begins with the assumption that there are meaningful differences in the retention of different groups of viewers, and that the retention also varies depending on the intrinsic difficulty of the videos. It is noteworthy that this study could not have been carried out only a few years ago, since video analytics were not accessible to learning stakeholders, mainly because online video platforms did not share the users' interactions on the system with stakeholders. However, this scenario has changed: YouTube, the world's largest media sharing site, now offers these data. So, the data provided by the YouTube Studio Learning Analytics platform has enabled us to gain new insights into participants' watching of these videos in our LMOOC. As stated by Ruipérez (2020; p.86) the impact on practice of the field of the Learning Analytics area, notwithstanding the popularity take-off, has been quite limited; with this study we are aiming for practical outcomes and improvements in use of video recordings for language learning. The objective of our research is, therefore, to explore if there are variables that allow to predict a certain level of retention and to define the factors that contribute to it, as a result of the Learning Analytics (LA) process. Video analytics should provide us with insights into student learning activity and their engagement. The results of our study should provide pedagogical implications for FL instructors concerning use of videos for language learning.

Online videos and language learning

Noetel et al. (2021) in their systematic review of the effects of video on learning in higher education conclude that under certain conditions videos can provide "better student learning than many other teaching methods, even when compared with face-to-face teaching" (p.227). While it is true that videos have been used in education for almost half a century (Wright, 1976), the educational use of online videos has experienced exponential growth due to the progressive incorporation of online learning by universities. Furthermore, in response to the COVID-19 pandemic in early 2020, there was a further disruptive increase of online learning and consequently also of online educational videos. Online videos have become an important content-delivery tool for Language Learning in flipped, blended, and online courses in recent years (Castrillo & Martín-Monje, 2016). They are integrated as part of face-to-face courses, serve as a base of many blended courses, and are usually the main content delivery tool in online courses (Brame, 2016). The effectiveness of using video to enhance language learning is clear and had already been established prior to the proliferation of MOOCs (Chun & Plass, 2000). As video lectures are clearly the main tool for delivering content in LMOOCs (Castrillo & Mañana-Rodríguez,

2018), video analytics should provide insights into student learning activity and their engagement. Student retention is an outstanding issue in MOOCs, and we believe that the analysis of their engagement with the videos contained in the LMOOC can help us understand why they drop out of these courses. Engagement is usually measured by the watching time of videos, thus our research begins with the assumption that there are meaningful differences in retention by different groups of viewers, and that the retention also varies depending on the intrinsic difficulty of the videos.

YouTube's Learning Analytics tool

In 2019 YouTube launched its app YouTube Studio, a new version of the old Creator Studio. It is a tool specifically designed so that content creators can access the metrics of the videos and carry out meticulous management of their channel. Among other options, it includes a video analytics page. Each video has its own analytics page, split into four tabs: Overview, Reach, Engagement, and Audience:

- The Overview tab shows the key metrics for the channel. The main graph shows watch time, views, and subscribers. There is also a graph for real-time activity, so one can see how people are interacting with the video at that very moment, and tables for likes versus dislikes, and for audience retention which is how well a video keeps people watching, i.e. do they watch it all the way to the end or leave to view something else after a few seconds.
- The Reach tab shows a similar graph to the Overview but focuses on impressions instead of views. An impression is how many times a thumbnail for a video is shown on YouTube, with a view being recorded if someone clicks on a thumbnail to visit the video page. The Reach tab also provides information about traffic sources.
- The Engagement tab has a graph that focuses on watch time in hours but also average view duration. The same tabs for audience retention and likes vs. dislikes are also provided on this tab.
- The Audience tab gives more insight into who is watching the video, including their age and gender, countries, and languages. The main graph shows returning and new viewers, unique viewers, and channel subscribers.

Methodology

The methodological approach used in this research begins with the assumption that there are meaningful differences in the retention of different groups of viewers, and that the retention also varies depending on the intrinsic difficulty of the videos. Retention can be defined as the average length of the video that the viewers watched on average. Nevertheless, it should be noted that because of the exploratory nature of this research, the identification of statistically significant differences or relationships between variables does

not have the aim of testing a specific hypothesis (which in a confirmation study would be the appropriate interpretation of the results). Its interpretation should not be understood as equivalent to that of regular statistical tests in terms of reliability or inference capacity. The growing incidence of data dredging practices makes the previous statement necessary.

Course description

The course, entitled *Alemán para hispanohablantes: nociones fundamentales* (German for Spanish-speakers) consists of a total of 35 videos which have the ultimate goal of providing the students with a basic knowledge of the German tongue, designed for the specific needs of Spanish-speakers. The channel has been available since 2012. The videos in the channel have been seen (or, at least, accessed) over 2 million times since their initial publication. It should be noted that the videos can also be viewed through the UNED's own channel with a significant number of visits not contemplated for this study. It is worth noting that this course received the first award for the best MOOC on the Telefonica MiriadaX platform in 2013 by the Spanish Education Ministry.

It is important to highlight that the videos were developed following guidelines that subsequently proved to be the most successful for retention (Guo et al., 2014), highlighting the following two characteristics: shorter videos are much more engaging (each video's duration is between 4 and 15 minutes) and informal talking-head videos are more engaging (our videos were recorded using this format). Prensky (2011) stated that Khan's videos are, above all, explanation. Our videos also basically consist of explanations of German structures in an analytical, practical and functional way, always following a principle of economy of learning: only rules that work and are proven to work are described and explained. We are convinced that the real issue for German students is that they miss important basic concepts of the German language and it is difficult for them to cover them again and progress in their learning. Our goal was not just to allow students to access the contents on demand with the video lessons but, most importantly, to give them the opportunity to replay at their own pace those contents which they may find especially complex, hence making learning much more flexible and achievable.

Data origin, description and transformations

Through YouTube Studio, a range of the most important variables in the platform was selected. These include: retention, total visualizations, total visualization time (hours), average visualization time (hours), total subscribers, total impressions, percentage of clicks of the impressions, visualizations (women), average duration of visualizations (women), percentage seen (women), visualization time (women), visualizations (men), average duration of visualizations (men), percentage seen (men), visualization time (men), and finally, the average duration of the visualizations, average percentage seen and

visualization time for the set of subjects with ages between 25 and 34 years. Although there are data available on the last three variables for other age ranges, these include large amounts of missing cases (from 37% to 100%), which has led to the exclusion of this information in the data set. It is worth noting that, although the updated YouTube Studio tool has been available since 2019, the data for the videos is available from their beginning.

Also, in order to characterize the sets of videos according to their retention, two new variables were created: "Over or equal to the median of retention" (a video with a retention value equal or over the median would have a 1 value and 0 otherwise) and "First quartile of retention", another dichotomous variable that presents the value *one* only if the retention value is over the 25th centile of the list of videos, ordered decreasingly by retention.

All the aforementioned variables have been included in the analyses carried out in this study. The initial hypothesis states that retention is not independent of the variables considered in this study.

Data analyses and tests

A battery of standard statistical techniques has been applied to the data set in order to identify the existence or otherwise of relationships between variables alongside the initial assumptions with an exploratory aim. First, the descriptive statistics were calculated for each variable (except for the two created variables, which are dichotomous). Then, an outlier analysis was performed over all continuous variables using the 1.5 inter-quartile range criterion, evidencing the lack of outliers in the variables. This was followed by the calculation of a correlation matrix of all the continuous variables using Spearman's correlation. A Kolmogorov-Smirnov test was then applied in order to test the hypothesis of normal distribution, which led to the conclusion that that a number of variables did not follow such distribution. The next step was the application of Mann-Whitney in order to identify differences between men and women regarding the continuous variables. This was followed by the application of a CHAID decision tree (minimum cases in the parent node = 2, minimum cases in child node = 1), having as dependent variable the belonging of each video to the group of videos over or equal to the median of retention, with no variable showing statistically significant differences that would lead to a segmentation of the videos. The same test was then applied having as dependent variable the belonging of the videos to the first quartile of retention distribution or otherwise.

The retention of the videos, ordered as they appear in the course, was plotted, which showed a declining tendency as the course progressed. This observation led to the application of a K-means cluster (three clusters were requested, all videos were classified) to the full set of variables, keeping the cluster of belonging of each video as a new variable,

together with the distance to each cluster's centroid for each video. Finally, in order to identify statistically significant differences between the videos belonging to each cluster in retention, a one-way ANOVA was applied to the full set of videos divided into clusters of belonging.

Results

The first results of the methodology applied are a set of correlations derived from the general correlation matrix between all the variables. Of particular interest are the correlations between the retention percentage and the rest of the variables. The most relevant correlations can be observed between the average duration of visualizations for men (Rho = .755), average percentage seen for men (Rho = .702), percentage seen for women (Rho = .660) and average duration of visualizations for women (Rho = 0.579). These correlations are, nevertheless, as expected since the retention is necessarily related to the average duration of the session and, particularly, to the percentage seen. Nevertheless, it is worth knowing if these apparent differences in visualization time are in line with the results of a Mann-Whitney U test for men and women in visualizations and visualization time.

There are statistically significant differences between men and women in the two variables. Men present a significantly greater number of visualizations (mean rank = 48.64) than women (mean rank = 24.36); U(36) = 211.000, z = -4.934, p < .00.). The same applies to the visualization time (mean rank for men = 47.88; mean rank or women = 25.13); U(36) = 238.500, z = -4,634, p < .001. It is important to mention here that these are the two variables in which statistically significant differences have been identified: the same tests were carried out for the rest of the continuous variables without significant results. In order to apply the CHAID segmentation tree, the dependent variable requires its transformation into a dichotomous variable. For this purpose, and given the limited range in the values displayed by "retention" (between 49.3 and 79.5), the videos with the first 25% accumulated retention were encoded as *1* (including 9 videos) and the rest of the videos as *0* (27 videos). The segmentation tree was then applied to the full set of variables, excluding those with an obvious relationship to retention (total visualization and percentage seen), since these would yield a segmentation tree including only those two variables.
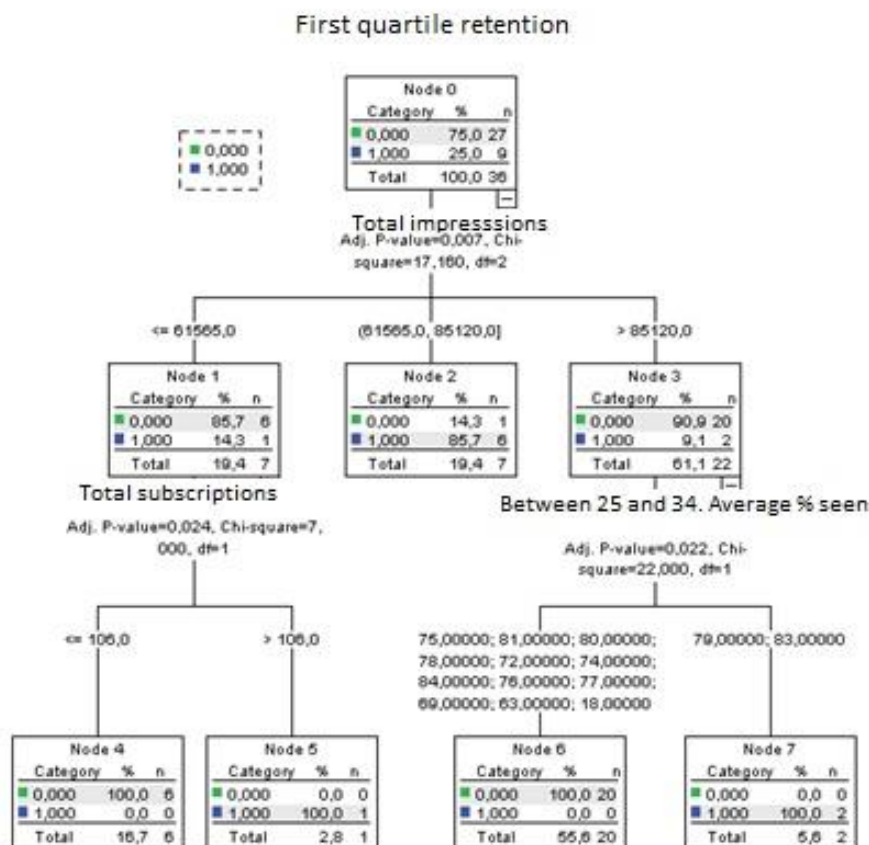
Figure 1. First quartile retention

The first split variable is the number of impressions, that is to say, the number of clicks on advertisements. The videos with an intermediate value of impressions show the greatest fraction of videos in the first quartile of retention (6). For the group with the smaller number of impressions, the most important split variable is the total number of subscribers, although the number of subjects in each node is discrete. For the group with the largest number of impressions, the videos with a percentage seen between 79 and 83%, for the segment of viewers between 25 and 34 years, are in the first quartile of retention. The overall correct classification rate of the model is 97% (96.3% for videos not in the first quartile of retention and 100% for videos in the first quartile of retention). The risk estimations provide positive results, both with the re-substitution method (Estimate = .028, Standard Error = .027) and the Cross-Validation approach (Estimate = .25, Standard Error = .072).

Retention by position of video in the MOOC sequence

Retention
Linear (Retention)

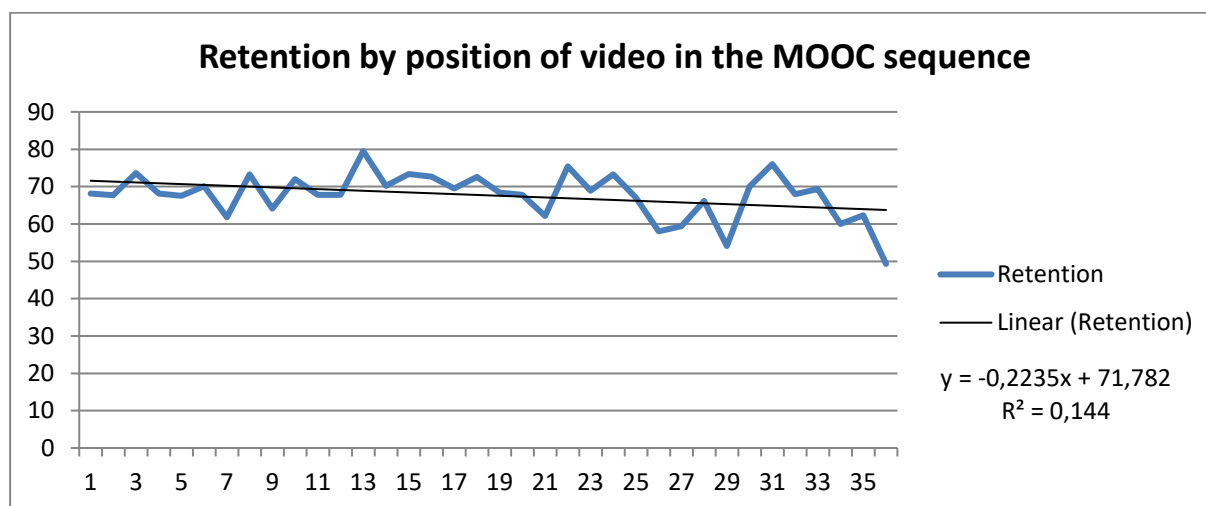$y = -0{,}2235x + 71{,}782$
$R^2 = 0{,}144$

Figure 2. Retention

The retention shows strong variations across the course (in terms of the order of the videos within the course), although the variability seems to be much greater in the second half of the series. Also, the general tendency interpolating the series by a linear equation shows a negative slope.

Discussion and conclusions

In the present study we have explored if there are variables that allow to predict a certain level of retention and what factors contribute to it, as a result of a video Learning Analytics (LA) process. Student retention is an outstanding issue in LMOOCs, and we believe that the analysis of their engagement with the videos contained in the LMOOC can help us understand why participants drop out of these courses. The results of our data analysis show that there are statistically significant differences between men and women in the total number of views and total viewing time, with men viewing significantly more than women. This result is very striking, since an intensive study of the socio-demographic analysis of the participants of this LMOOC showed that in all editions there are more women than men (Castrillo & Mañana-Rodriguez, 2017). This is corroborated by recent subsequent studies of socio-demographic profiles in LMOOCs such as that of Díez-Arcón (in press): it seems to be proven that the predominant gender of LMOOC participants is female. However, analysis of the video analytics of this study indicates that the videos are mostly followed by men. The reason for this imbalance requires further research, but, in the first instance, it could be related to the different types of intelligence and differences in cognitive processes between men and women pointed out by neuroscientists and psychologists (Halpern & LaMay, 2000; Halpern, 1997; Lim, 1994) and reinforce the theories that point to the need to create learning objects in different digital formats to adapt to the profiles of the participants.

Another result is that retention decreases, albeit with a lot of variability between videos, as the course progresses. The courses can be grouped into three sets: one, with 17 videos, which seem to be more interesting for men and for subjects aged 25 to 34, a second set (9 videos) characterized by higher user interaction (subscriptions and impressions) and higher overall interest for women, and finally, a set of 22 videos that show high levels of retention. There are differences in retention between the videos in the first set and the third, but not between the first and the second or the second and the third. The decrease in retention along as the course advances might be due to a multiplicity of factors, some of the most plausible being the increasing difficulty of the contents of each video.

The results can be incorporated into the design of future LMOOCS by splitting the course into sections with similar difficulty, potentially increasing the retention time for each section by addressing audiences with different initial levels and expectations and other variables associated with retention rate.

References

Brame, C. J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education, 15*(4), es6.

Castrillo de Larreta-Azelain, M. D., & Martín-Monje, E. (2016). Students' engagement in online language learning through short video lessons. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras, 26*, 177-186.

Castrillo, M. D., & Mañana-Rodríguez, J. (2017). Evolución del Perfil Sociodemográfico del Participante en MOOC (2013-2016): Estudio de un Caso en Lenguas Extranjeras. In EMOOCs-ES (pp. 15-24).

Chun, D., & Plass, J. L. (2000). Networked multimedia environments. In M. Warschauer, & R. Kern, R. (Eds.), *Networked-based Language Teaching. Concepts and Practice* (pp. 151-170). Cambridge: Cambridge University Press.

Díez-Arcón, P. (in press). *Perfiles del participante en cursos masivos en línea de lenguas extranjeras (LMOOC): análisis bibliográfico y estudio de caso.*

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. *Proceedings of the first ACM conference on Learning@ scale conference*, 41-50. ACM.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*(10), 1091-1102. https://doi.org/10.1037/0003-066X.52.10.1091

Halpern, D. F., & LaMay, M. L. (2000). The Smarter Sex: A Critical Review of Sex Differences in Intelligence. *Educational Psychology Review, 12*, 229-246 https://doi.org/10.1023/A:1009027516424

Khan, S. (2012). *The one world schoolhouse: Education reimagined.* London: Hodder & Stoughton.

Lim, T. K. (1994). Gender-related differences in intelligence: Application of confirmatory factor analysis. *Intelligence, 19*(2), 179-192.

Martín-Monje, E., Castrillo, M. D., & Mañana-Rodríguez, J. (2018). Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning, 31*(3), 251-272

Noetel, M., Griffith, S., Delaney, O., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2021). Video Improves Learning in Higher Education: A Systematic Review. *Review of Educational Research, 91*(2), 204-236. https://doi.org/10.3102/0034654321990713

Omnicore (2021). YouTube by the Numbers: Stats, Demographics & Fun Facts. Retrieved from https://www.omnicoreagency.com/youtube-statistics/

Prensky, M. (2011). Khan academy. *Educational Technology, 51*(5), 64.

Ruipérez-Valiente, J. A. (2020). El Proceso de Implementación de Analíticas de Aprendizaje. RIED. *Revista Iberoamericana de Educación a Distancia, 23*(2), pp. 85-101. http://dx.doi.org/10.5944/ried.23.2.26283

Wright, A. (1976). *Visual Materials for the Language Teacher.* Essex: Longman.

Acknowledgements