Generation of Virtual Children for testing a Recommendation System for Interventions with Children with Dyslexia

J. Ignacio Mateo-Trujillo[1⋆] **Ignacio Rodríguez-Rodríguez**[1]
Diego Castillo-Barnes[1] **Andrés Ortiz**[1] **Juan Luis Luque**[2]

[1] Communications Engineering Department,
University of Málaga 29004 Málaga, Spain
[2] Department of Developmental and Educational Psychology,
University of Málaga 29004 Málaga, Spain

⋆nachomateo912@uma.es

Abstract: The LEEDUCA project has developed a recommendation system to generate intervention sessions tailored to children with dyslexia. Due to the limitations in obtaining real data for preliminary testing, the generation of in silico data, referred to as "virtual children," has been implemented. This approach allows for the simulation of a wide range of profiles and response patterns, enabling comprehensive testing of the system before its implementation with real users. The behavior of virtual readers is modeled using logistic curves, which reflect the natural evolution of users in a system that suggests words ordered by difficulty over time. By introducing variations to the model based on the coefficients that define the logistic curve, response sequences with different difficulty levels and learning rates can be simulated. To evaluate the stability of the system, multiple variations are generated from a given virtual child, creating a shadow of possible sequences. The generation of virtual children using logistic curves and the controlled introduction of variations in their responses provide a robust framework for testing the recommendation system, ensuring its reliability and adaptability to the individual needs of children with dyslexia.

keywords: data modelling; data simulation; logistic model; recommendation system.

1 Introduction

Within the framework of the LEEDUCA project (https://leeduca.es/), a Recommendation System has been developed to generate intervention sessions tailored to children with dyslexia. During these sessions, batches of 100 consecutive words and pseudo-words generated according to phonological rules from an AI generator module are serially suggested, with an intrinsic difficulty assigned by a machine learning system and validated by a team of psychologists. The main objective is to adapt the difficulty of the recommended words to the child's performance throughout multiple interventions. Before implementing the system with real users, comprehensive testing is crucial to ensure its effectiveness and reliability. However, given that the recommendation system is intended for children, there are limitations in the availability of real data for preliminary testing. For this reason, the generation of in silico data, referred to as "virtual children," has been chosen to simulate a wide range of profiles and response patterns

[3]. In addition to overcoming data scarcity, the purpose of multiple system testing is to evaluate the error in recommendations and the stability of the system under different levels of input variability.

2 Results and discussion

2.1 Generation of the model from empirical data

Using data collected from similar intervention projects and validated by the team of psychologists, the behavior of diverse profiles of virtual children has been modeled using logistic curves. These curves reflect the natural evolution of users in a system of suggestions ordered by difficulty over time. The logistic function, also known as the S-shaped curve, is a mathematical tool widely used in various fields, such as population growth, disease propagation, and diffusion in social networks. This function constitutes a refinement of the exponential model and allows for a more realistic description of the growth of a magnitude, taking into account limiting factors [2]. In general, it is observed that the user tends to answer correctly at a rapid pace until reaching their difficulty limit, at which point they stagnate during that intervention. This behavior is captured by the logistic curve, which exhibits a rapid initial growth followed by a gradual leveling off as the user approaches their maximum performance level. Readers at risk of dyslexia show poorer performance from the beginning of literacy (assumed to be the first stage); and persistently show difficulties in learning and generalising conversion rules and automating them, resulting in slow progress, with numerous setbacks, and finally not reaching the same levels as their peers (gap). These profiles can be simulated by manipulating the parameters of the logistic curve. The versatility of the logistic curve lies in its ability to generate models by adjusting different parameters, such as the initial difficulty, the final difficulty the virtual children reach, the time they remain in the initial and final phases, and the learning speed, represented by the slope of the curve. By manipulating these parameters, it is possible to create profiles of virtual children with diverse characteristics and learning rhythms, simulating a wide range of potential users and their responses to the intervention system.

2.2 Variations to the model

Once the fit of the logistic curve to the empirical data has been verified, it is possible to introduce variations to the model based on the coefficients that define the curve. In this way, response sequences that reach higher difficulty or present a faster learning rate can be simulated without the need for specific empirical data for those scenarios. The logistic model is defined by equation 1.

$$f(x) = \frac{L}{1 + e^{-\frac{k}{x - x_0}}} \tag{1}$$

Where: L: Max difficult achieved.$x_0 = \frac{n_1+n_2}{2}$: Midpoint of the curve. $n = 100$: Total number of words. n_1: Start of the rise. n_2: Stabilization point. k: Rate of change. By adjusting these parameters, sequences of in silico simulated data can be generated that follow a specific curve (Figure 1). This flexibility allows for the exploration of a wide range of possible user behaviors and the evaluation of the system's performance under different conditions.

2.3 Derived children

To evaluate the stability of the system, multiple variations are generated from a given virtual child, which deviate from the basic data sequence until reaching a specific dispersion, generating

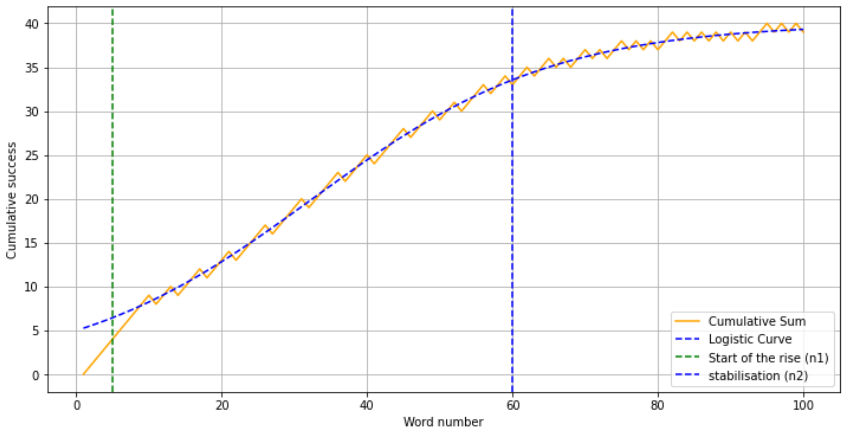

Figure 1: Construction of a virtual child.

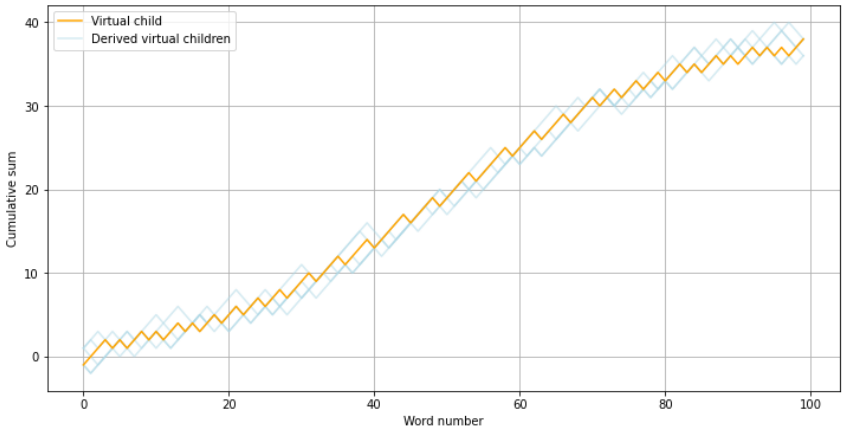

Figure 2: Derived virtual children.

a shadow of sequences whose median is the basic sequence and share the same logistic model (Figure 2). The process of creating variations involves applying a percentage of variance to the base virtual child, resulting in derived virtual children that exhibit random variations in their interactions with the system while maintaining the same overall learning trajectory. The dispersion of the difficulties recommended by the system can be observed, and it can be evaluated whether the output dispersions are contained and proportional to the input dispersions.

This evaluation is performed by analyzing the relationship between the variance introduced in the inputs (the responses of the virtual children, Figure 3) and the variance observed in the outputs (the difficulty levels proposed by the system, Figure 4). If, for a given level of input variance, the output variance does not exhibit erratic behavior or exceeds a reasonable range, it can be concluded that the system is stable and robust. A stable system should be capable of adapting to individual variations among users without compromising its ability to provide coherent and personalized recommendations [1]. The creation of virtual children using logistic curves and the generation of random variations from base profiles constitute fundamental tools to overcome the challenges of cold start and data scarcity in the development of an intervention system for children with reading difficulties. These techniques allow for the simulation of a wide range of potential users, the evaluation of system robustness, and the assurance of its reliability.