Best Strategies for Bilingual Education: How Can We Explain Their Success?

Claudia Cabrera[1], Olga Kosheleva[1],
Christian Servin[2], and Vladik Kreinovich[3]
[1]Department of Teacher Education
University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
olgak@utep.edu
[2]Information Technology Systems Department
El Paso Community College (EPCC)
919 Hunter Dr., El Paso, TX 79915, USA
cservin1@epcc.edu
[3]Department of Computer Science
University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
vladik@utep.edu

Abstract

When designing AI-based tools for education, it is important to take into account the experience of human teachers. In this, it is necessary to distinguish between the education features that are justified by the general features of the corresponding education task – these features should be taken into account in AI-based learning as well – and features which are specific for traditional non-AI teaching. In this paper, on the important example of bilingual education, we show that several empirically successful teaching strategies can be explained in the general context – and thus, should be implemented in AI-based teaching as well.

1 Introduction

Bilingual education is important. In the border area between domains of speakers of different languages, many people are bilingual. To function well in such an environment, it is important to train the students to be able to perform well in both languages. This is the main objective of bilingual education.

As with other education needs (see, e.g., [4]), it is desirable to use all education tools – including AI tools (see, e.g., [2]) – to help. In designing such a help, it is natural to take into account the experience of traditional – non-AI –

bilingual education. To better utilize this experience, it is important to understand which bilingual education strategies are justified by the general features of the corresponding learning tasks – features not depending on:

- whether teaching is done by human teacher only
- or by an AI-based education system.

These features need to be implemented in the AI-based system as well – while it does not necessarily make sense to implement strategies which are specific for human teachers.

From this viewpoint, it is important to analyze the existing strategies for bilingual education – which of them have deep justifications and which are specific for human teachers.

What we do in this paper. Our paper is based on the dissertation [1] that describes, in detail, the semi-empirically determined setting of a successful bilingual education program. In this paper, we provide a general theoretical explanation for the success of these particular settings.

The structure of the paper. In this section, we describe the general successful settings. In the following sections, for each of these settings, we provide a theoretical explanation.

What are the settings that we plan to explain. The first two setting relate to pre-planning of the overall teaching process.

- First, in the successful program, 50% of the time the material is taught in English, with some connections to Spanish, and 50% of the material is taught in Spanish, with some connections to English. This is different of many other bilingual programs where the proportion of languages is 80/20 or 90/10 [3].
- Second, in this program, each day, half of the day is in one language, half is in another language, with the order of the langueges changes from each day to the next.

The third setting is related to the actual teaching process: it is important *not* to correct minor mistakes all the time.

Finally, the fourth and the fifth settings are related to the assessment of students' leearning:

- The fourth setting is that it is important to regularly show, to students, their progress.
- The first setting is that it is important to grade the students you know – and if several instructors participate in grading, the grades by instructors who know the students should have more weights in the resulting grade.

2 Why 50/50 distribution between languages

Mathematical model. In this section, we provide a possible explanation of why 50/50 distribution between languages works the best. This explanation is based on the following idea.

In general, teaching any topic requires several repetitions. On each repetition i, the student's understanding is not perfect: some ideas are remembered, but, in general, the knowledge is different from ideal. In other words, there is a difference $\Delta u_i \stackrel{\text{def}}{=} \widetilde{u}_i - u$ between the student's understanding $\underline{u}_i$ of the material presented on this iteration and the ideal perfect understanding u. Based on several understandings $\widetilde{u}_1, \ldots, \widetilde{u}_n$ of the presented material, the student forms the resulting picture $\widetilde{u}$. In this formation, the student naturally wants to select $\widetilde{u}$ which is close to all previous understandings, i.e., for which the resulting vector $(\widetilde{u}, \ldots, \widetilde{u})$ is as close as possible to the perceived vector $(\widetilde{u}_1, \ldots, \widetilde{u}_n)$. By Pythagoras theorem, the square d^2 of the distance d between the two vectors is equal to the sum of the squares of the differences between the values:

$$d^2 = (\widetilde{u} - \widetilde{u}_1)^2 + \ldots + (\widetilde{u} - \widetilde{u}_n)^2. \tag{1}$$

Minimizing the distance d is, of course, equivalent to minimizing its square d^2. We can minimize the expression (1) if we simply use of the main ideas from calculus: that in the location where the minimum is attained, the derivative of the objective function is equal to 0. Differentiating the expression (1) with respect to the unknown $\widetilde{u}$ and equating the derivative to 0, we conclude that

$$2(\widetilde{u} - \widetilde{u}_1) + \ldots + 2(\widetilde{u} - \widetilde{u}_n) = 0.$$

Moving all the terms $\widetilde{u}_i$ to the right-hand side and dividing both sides by $2n$, we conclude that:

$$\widetilde{u} = \frac{\widetilde{u}_1 + \ldots + \widetilde{u}_n}{n}. \tag{2}$$

The resulting inaccuracy $\Delta u \stackrel{\text{def}}{=} \widetilde{u} - u$ of the student's understanding can be estimated if we subtract u from both sides of the equality (2). Then, we get:

$$\Delta u = \frac{\Delta u_1 + \ldots + \Delta u_n}{n}. \tag{3}$$

Ideally, this different should be as close to the ideal case 0 as possible. In other words, we want to make sure that the vector $(\Delta u^{(1)}, \ldots, \Delta u^{(K)})$ describing the understanding of all students is as close to the ideal vector $(0, \ldots, 0)$ as possible. Similarly to what we described above, this means minimizing the sum of the squares

$$\left(\Delta u^{(1)}\right)^2 + \ldots + \left(\Delta u^{(K)}\right)^2.$$

In mathematical terms, if we consider a population of students, this sum of the squares is proportional to the variance $\sigma^2 \stackrel{\text{def}}{=} E[(\Delta u)^2]$ of the value Δu. We therefore want to select a teaching strategy for which this variance is as small as possible.

In general, we have

$$\sigma^2 \stackrel{\text{def}}{=} E[(\Delta u)^2] = \frac{1}{n^2} \cdot E\left[\left(\sum_{i=1}^n \Delta u_i\right)^2\right] =$$

$$\frac{1}{n^2} \cdot \left(\sum_{i=1}^n E\left[(\Delta u_i)^2\right] + \sum_{i=1}^n \sum_{j\ne i} E\left[\Delta u_i \cdot \Delta u_j\right]\right). \tag{4}$$

Let us denote $E\left[(\Delta u_i)^2\right]$ by σ_i^2. Then, by definition of correlation r_{ij}, we have

$$E\left[(\Delta u_i)^2\right] = r_{ij} \cdot \sigma_i \cdot \sigma_j,$$

so the expression (4) takes the following form:

$$\sigma^2 = \frac{1}{n^2} \cdot \left(\sum_{i=1}^n \sigma_i^2 + \sum_{i=1}^n \sum_{j\ne i} r_{ij} \cdot \sigma_i \cdot \sigma_j\right). \tag{5}$$

Mathematical model: analysis and the resulting explanation. To estimate the variance σ^2, let us denote:

- the number of English lessons by n_E and
- the number of Spanish lessons by n_S.

Let us also denote:

- the average standard deviation of a single lesson by σ_0^2,
- the typical correlation between two lessons in the same language by r_s, and
- the typical correlation between two lessons in different languages by r_d.

Intuitively, since the two languages are different, correlation between lessons in different languages should be smaller than correlation between lessons in the same language: $r_d < r_s$.

In terms of σ_0, n_E, n_S, and correlations, the expression (5) takes the following form:

$$\sigma^2 = \frac{\sigma_0^2}{n^2} \cdot \left(n + (n_E^2 - n_E) \cdot r_s + (n_S^2 - n_S) \cdot r_s + 2 \cdot n_E \cdot n_S \cdot r_d\right). \tag{6}$$

We want to explain why it is effective to use

$$n_E = n_S = \frac{n}{2}.$$

In other words, we want to explain why it is effective to have the difference

$$\delta \overset{\text{def}}{=} n_E - \frac{n}{2}$$

equal to 0. In terms of this difference, we have

$$n_E = \frac{n}{2} + \delta$$

and

$$n_S = n - n_E = n - \left(\frac{n}{2} + \delta\right) = \frac{n}{2} - \delta.$$

Substituting these expressions for n_E and n_S into the formula (6), we get the following formula:

$$\sigma^2 = \frac{\sigma_0^2}{n^2} \cdot \left(n + r_s \cdot \left[\left(\frac{n}{2} + \delta\right)^2 - \left(\frac{n}{2} + \delta\right) + \left(\frac{n}{2} - \delta\right)^2 - \left(\frac{n}{2} - \delta\right)\right] + \right.$$

$$\left. 2r_d \cdot \left[\left(\frac{n}{2} + \delta\right) \cdot \left(\frac{n}{2} - \delta\right)\right]\right). \qquad (7)$$

Here,

$$\left(\frac{n}{2} + \delta\right)^2 - \left(\frac{n}{2} + \delta\right) + \left(\frac{n}{2} - \delta\right)^2 - \left(\frac{n}{2} - \delta\right) =$$

$$\frac{n^2}{4} + n \cdot \delta + \delta^2 - \frac{n}{2} - \delta + \frac{n^2}{4} - n \cdot \delta + \delta^2 - \frac{n}{2} + \delta = \frac{n^2}{2} + 2\delta^2 - n,$$

and

$$2 \cdot \left(\frac{n}{2} + \delta\right) \cdot \left(\frac{n}{2} - \delta\right) = 2 \cdot \left(\frac{n^2}{4} - \delta^2\right) = \frac{n^2}{2} - 2\delta^2.$$

Thus, the formula (7) takes the form

$$\sigma^2 = \frac{\sigma_0^2}{n^2} \cdot \left(n + r_s \cdot \left(\frac{n^2}{2} + 2\delta^2 - n\right) + r_d \cdot \left(\frac{n^2}{2} - 2\delta^2\right)\right) =$$

$$\frac{\sigma_0^2}{n^2} \cdot \left(n \cdot (1 - r_s) + \frac{n^2}{2} \cdot (r_s + r_d) + 2\delta^2 \cdot (r_s - r_d)\right). \qquad (8)$$

Since $r_s > r_d$, the expression (8) is an increasing function of δ^2. Thus, its smallest value is attained when δ^2 is the smallest – i.e., when $\delta = 0$ and thus,

$$n_E = n_S = \frac{n}{2}.$$

This is exactly what we tried to explain.

3 Why daily switches between languages

Empitical fact. The 50/50 arrangement – that, as we have shown, is optimal – could be, in principle, arranged by studying in one the languages one day, and in another language the next class day. However, it turned out that the more effective arrangement is when each day:

- first the teaching is in one of the languages, and
- then, in the second part of the school day, the teaching is in another language.

The language with which we start changes from one day to the next. For example:

- if the first half of Monday is in English,
- then the first half of Tuesday is in Spanish, etc.

Our explanation. Why? Here, the explanation is straightforward, does not require math. We want students:

- not only to learn the material,
- we also want them to be able to switch between languages on a short notice.

To teach this ability to the students, a natural idea is to train them on this transition every day. That is why it is useful to have language switches every day.

The same need explains the daily change the order of languages. Indeed:

- if we start with English every day,
- then students will learn how to switch from English to Spanish, but
- they will not become training in fast switching from Spanish to English.

Students need to be trained to switch in both directions equally well – this is why we alternate the languages that start the school day.

4 Why it is important not to correct minor mistakes all the time

Empirical fact. As students go through the process of acquiring new knowledge, their knowledge is not yet perfect, they make mistakes. At first glance, it seems reasonable to correct these mistakes – otherwise, how will the students know that they have made a mistake? In practice, however, the practice of always correcting all mistakes, major and minor, hinders education. It is much more effective *not* to correct mistakes all the time.

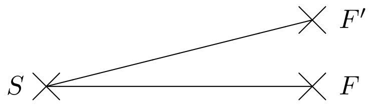
Mathematical model. In this section, we explain, on a simple mathematical model, why always correcting minor mistakes may be harmful. We will explain this phenomenon by using the following geometric representation of knowledge.

At any given moment of time, we can describe the state of a student's knowledge by a tuple $g = (g_1, g_2, \ldots)$ of relevant numbers – e.g., grades on different assignments. Each such tuple can be viewed as a point in the corresponding multi-D space, a point with coordinates g_1, g_2, etc. In these terms, the process of learning is described by a trajectory that leads the student:

- from his/her starting state S
- to as close as possible to the desired state D of perfect knowledge.

The ideal situation is when the student's trajectory brings him/her from the original state S to the desired state D by the fastest possible path. Let F denote be the first step on this ideal trajectory. In practice, students are not perfect. So:

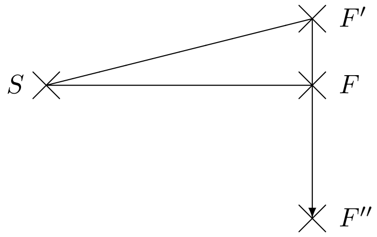
- instead of reaching the desired intermediate state F,
- students reach a somewhat different state F':



Resulting explanation. At first glance, in this situation, it makes sense to inform the student about his/her mistake. This way, the student will get to the correct intermediate state F – which will bring this student closer to the desired ideal state D.

However, the problem with this approach is that the student is not well aware of his/her location and direction: the reason why the student reached the state F' instead of the desired intermediate state F is that, at thus stage of the student's knowledge, this student is not yet able to distinguish between F and F'. So:

- even if the student understands the direction in which the correction needs to be made – namely, the direction from F' to F,
- there is a good chance that result in an overcorrection, with the new state F'' possible event further from F than F':



To avoid such an overcorrection, it is desirable to limit corrections only to *major* mistakes, mistakes for which the student is able – with a little help – to see the difference between the desired and the actual states.

5 Why it is important to regularly show, to students, their progress

Our explanation: main idea. Even when students make progress that is clear to the instructor, students – as we have mentioned in the previous section – may not be skilled enough to detect small improvement in their knowledge. As a result, by comparing their state of knowledge before and after the class – or before and after their individual discussion with an instructor or with a teaching assistant – students may not see the difference and thus, get discouraged. One of us (VK) had this experience when taking Basic Spanish for Faculty class – when his grade after the first test turned out to be similar to his grade on a similar previous quiz.

To avoid such discouragements, it is desirable to explicitly remind the students of their state of knowledge at earlier stages of the class: this difference is much larger and is, thus, has a higher possibility to be visible to students.

Let us reformulate our explanation in terms of a simple mathematical model. Let us describe the above explanation in more mathematical terms. At the intermediate stage of their studies, student cannot tell the difference between their knowledge states $K(t')$ and $K(t)$ at two consequent moments of time $t' < t$ when the distance $d(K(t'), K(t))$ between these two states is smaller than some threshold $\varepsilon > 0$.

Students naturally compare their current knowledge state $K(t)$ with the most recent state – since this is the state that happened most recently and is, thus, best remembered. When t' is close to t, the state $K(t')$ is close to $K(t)$ – and thus, it is highly probable that the distance $d(K(t'), K(t))$ will be smaller than ε.

Thus, to make the progress visible, it is important to remind the students about their earlier knowledge states, at a moment $t'' \ll t'$ – this way the distance $d(K(t''), K(t))$ will be larger and thus, have more chances to be visible to students.

6 Why it is important to grade students you know – and for committee grading, to give more weight to grades of instructors who know the student

Empirical fact. At first glance, it may seem that the most objective grading comes from outside graders: these graders will judge the level of knowledge shown on the test, without their opinion being prejudiced by the previous successes and failures of this particular student.

To some extent, this is correct. However, empirical data shows that for the most effective teaching, it is desirable to have grading done by instructors who know the student – and when the grading is done by a committee, grades by instructors who know the student should get more weight.

How we usually grade. To explain this empirical fact, let us recall that usually, the overall grade for a test is computed as an average – or a weighted average – of grades for individual problems, and the overall grade for a class is similarly computed as an average or as a weighted average of grades for all tests and graded assignments. Similarly to what we described earlier – when we talked about student's knowledge $\widetilde{u}$ after n classes, the average naturally appears if we are trying to find the overall grade that is the best fit for all the component grades.

Resulting explanation. However, the above argument does not take into account that once in a while, we have *outliers*: e.g., a student had a bad day and gets 0 on a test, not because of the lack of knowledge, but because of some stressful event unrelated to class.

Outliers occur in statistics in general: a measuring instrument can malfunction, etc.; see, e,g., [5]. If we do not detect them, they can make the resulting averages meaningless. For example, if the patient's body temperature during the day was recorded as 37.5 C, 38.0 C, 10.0 C, and 37.3 C – with 10 C being a clear outlier – then the average body temperature of this patient is a meaningless 30.7 C.

The usual way to detect outliers – and not to take them into account when making an overall estimate – is to use *expert knowledge*. For example, if a electronic thermometer shows that the patient's body temperature is 10 C, this is clearly the sensor failure, so this value should not be taken into account when estimating the average daily temperature of a patient.

In teaching, who are the experts? Teachers who know the student. They can detect (and ignore) outliers and thus, come up with grades that better reflect the student's level of knowledge.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.

It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI), and by the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany.

References

[1] C. M. Cabrera, *Teacher Language Ideologies Concerning the Reclassification of Emergent Bilingual Students in Dual Language Bilingual Education: Navigating the Levels of Power in Reclassification*, PhD Dissertation, Doctoral Program in Teaching, Learning, and Culture, University of Texas at El Paso, 2024.

[2] J. Gomez Galan, *Innovation and ICT in Education: The Diversity of the 21st Century Classroom*, River Pubishers, 2021, ISBN: 9788770221986, e-ISBN: 9788770221979.

[3] J. Medina, "Navigating competing biliteracy ideologies: ¿Qué capirotada!", *Proceedings of the October 2022 Institute of Texas Association for Bilingual Education TABE*, Houston, Texas, USA, 2022.

[4] O. A. Ponce, J. Gomez Galan, N. Pagán-Maldonado, and A. L. Canales Encarnación (eds.), *Introduction to the Philosophy of Educational Research*, River Publishers, 2021, ISBN: 9788770226370, e-ISBN: 9788770226363.

[5] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.