frontiers | Frontiers in Education

TYPE Curriculum, Instruction, and Pedagogy
PUBLISHED 21 February 2023
DOI 10.3389/feduc.2023.1063337

OPEN ACCESS

EDITED BY
Heidi Kloos,
University of Cincinnati, United States

REVIEWED BY
Ebubekir Bozavlı,
Atatürk University, Türkiye
Charles Tijus,
Université Paris 8, France

*CORRESPONDENCE
Daniel D. Hromada
✉ dh@udk-berlin.de

SPECIALTY SECTION
This article was submitted to Digital Education, a section of the journal Frontiers in Education

RECEIVED 07 October 2022
ACCEPTED 23 January 2023
PUBLISHED 21 February 2023

CITATION
Hromada DD and Kim H (2023) Proof-of-concept of feasibility of human–machine peer learning for German noun vocabulary learning. *Front. Educ.* 8:1063337. doi: 10.3389/feduc.2023.1063337

Proof-of-concept of feasibility of human–machine peer learning for German noun vocabulary learning

Daniel D. Hromada* and Hyungjoong Kim

Digital Education, Institute of Time-Based Media, Berlin University of the Arts, Berlin, Germany

The present study provides the first empiric evidence that the creation of human–machine peer learning (HMPL) couples can lead to an increase in the level of mastery of different competences in both humans and machines alike. The feasibility of the HMPL approach is demonstrated by means of Curriculum 1 whereby the human learner H gradually acquires a vocabulary of foreign language, while the artificial learner fine-tunes its ability to understand H's speech. The present study evaluated the feasibility of the HMPL approach in a proof-of-concept experiment that is composed of a pre-learn assessment, a mutual learning phase, and post-learn assessment components. Pre-learn assessment allowed us to estimate prior knowledge of foreign language learners by asking them to name visual cues corresponding to one among 100 German nouns. In a subsequent mutual learning phase, learners are asked to repeat the audio recording containing the label of a simultaneously presented word with the visual cue. After the mutual learning phase is over, the subjacent speech-to-text (STT) neural network fine-tunes its parameters and adapts itself to peculiar properties of H's voice. Finally, the exercise is terminated by the post-learn assessment phase. In both assessment phases, the number of mismatches between the expected answer and the answer provided by human and recognized by machine provides the metrics of the main evaluation. In the case of all six learners who participated in the proof-of-concept experiment, we observed an increase in the amount of matches between expected and predicted labels, which was caused both by an increase in human learner's vocabulary as well as by an increase in the recognition accuracy of machine's speech-to-text model. Therefore, the present study considers it reasonable to postulate that curricula could be drafted and deployed for different domains of expertise, whereby humans learn from AIs at the same time as AIs learn from humans.

KEYWORDS

human-machine peer learning, foreign language learning, vocabulary learning, automatic speech recognition, DeepSpeech, small data, German nouns, minimization of mismatch

1. Introduction

1.1. Human–machine peer learning

Human–Machine Peer Learning (HMPL) is a proposal that is positioned at the very frontier between educational, cognitive, and computer sciences. HMPL's core precepts which Hromada (2022) introduced in a recent study are simple:

Humans and machines can learn together.
Humans and machines can learn from each other.

One reason which makes us postulate these two statements is the existence of a so-called "human-machine learning parallelism," that is, both processes of human and machine learning have some features in common (Hromada, 2022). Another reason–and it is this one whose understanding is crucial for a proper understanding of our proposal–is the strong preference of human learners, notably children (Freinet, 1990; Golbeck, 1999), not only–to acquire knowledge, behaviors, and competences (Cooper and Cooper, 1984) from other learners who exhibit a similar–but slightly higher–level of mastery (LoM) of such knowledge, behaviors, and competence. We label such acquisition processes between learners mutually located in their zones of proximal development (Hogan and Tudge, 1999) "peer learning" (PL).

In real life, PL often goes hand in hand with practices and situations, whereby the learner assumes the role of the teacher in the same time as the teacher assumes the role of the learner. In the article entitled "learning by teaching," Frager and Stern (1970) starts their treatise with an observation:

> *A sixth grader who reads at a first or second grade level might be rebelliously indignant if he were asked to increase his reading skills by using primers appropriate to his reading level. However, when he is asked to take on the role of teacher with a first or second grade child who needs help, the same materials become part of a program invested with status and responsibility. In this manner, the older child is given the opportunity of building up his self-confidence even as he builds his reading* (Frager and Stern, 1970).

Analogically, the author of the "learning through teaching" observes "great learning potential inherent in teaching" (Cortese, 2005).

In HMPL, it is an artificial system–the machine m–that assumes, aside from the human learner H, a simultaneous role of the one who teaches as well as the one who is being taught. In a sense that both H and m are teachers and learners at the same time, in that sense both H and m can be considered to be "peers."

Within this article, we provide the first empiric evidence that the creation of such human-machine couples can lead to an increase in LoM in both humans and machines alike. The feasibility of the HMPL approach is demonstrated by means of "Curriculum 1," whereby the human learner gradually acquires a vocabulary of foreign or second language (L2), while the artificial learner fine-tunes his ability to understand H's spoken L2 production.

1.2. AI-assisted vocabulary learning

By allowing the human learner to assimilate the fundamental units of language-word-vocabulary learning (VL) is an important component of any L_2 class. In spite of the fact that many, both theorists and practitioners of L_2 teaching, observe direct relations between VL and L2 learning (Qian and Schedl, 2004; Jun Zhang and Bin Anual, 2008), VL is often neglected in common L2 teaching practice, being only rarely explicitly and directly addressed during L2 seminars and often reduced to rote learning of a word list from a school book (Oxford and Crookall, 1990).

To fill this gap, diverse digitally assisted systems have been developed, deployed, and evaluated for computers (Perea-Barberá and Bocanegra-Valle, 2014; Alnajjar and Brick, 2017) and for mobile devices (Hu, 2013). Often, digital assistants implementing an algorithmic variant of the flashcard principle (Nikoopour and Kazemi, 2014; Hung, 2015) and exposing the learner not only to written representations of the vocabulary to be learned but also to pictures or audio recordings are indeed useful mediators of L_2 acquisition.

One of the most important features of such digital systems is the ability to recognize and process a learner's speech. Despite the fact that automatic speech recognition (ASR) and speech-to-text (STT) systems have been used in foreign language learning for almost two decades (Chiu et al., 2007; Bajorek, 2017) and are often deployed with a certain amount of success in renowned products such as, for example, Duolingo (Teske, 2017), in which the problem of accurate ASR in the domain of L_2 is far from being solved, notably for students with a strong accent (Matassoni et al., 2018) or young children (Dubey and Shah, 2022) whose voices are not accurately classified by ASR/STT systems. Additionally, in spite of impressive progress in the field of noise-robust ASR (Li et al., 2014), background sounds and other environmental factors–imagine, for example, a classroom filled with 30 simultaneously speaking children–often make it impossible to provide a human learner with a highly accurate feedback about his/her pronunciation. Such problems are further exacerbated for a huge majority of all non-English languages where there are not yet enough data publicly available for induction of the highly accurate acoustic models (Schlotterbeck et al., 2022).

1.3. Small data

There is little doubt that recent advances in the domain of artificial intelligence (AI) and machine learning (ML) have been, in great part, made possible thanks to the massive data processing aggregation of billions of users, often unaware of their role of data providers. For reasons more closely elaborated in Hromada (2022), HMPL educators ought to prioritize the "small data" paradigm over the "big data" one.

Being aware of the "importance of starting small" (Elman, 1993) and knowing that the so-called few-shot or one-shot (Vinyals et al., 2016) learning is possible and that it provides a viable path to increase one's ML systems, the paradigm adopted in this and the future HMPL curricula is simple to explain: instead of aiming to train and deploy artificial systems adapted to masses of "customers" or "users," an HMPL educator or engineer deploys the artificial learning systems (ALS) that adopt to one–or fairly few–specific human beings.

In other words, instead of aiming to provide a mediocre understanding of the speech of practically all humans on the planet,

Abbreviations: HMPL, Human–machine peer learning; MDE, Mutual didactic equilibrium; MNM, Mutually neutralizing mistake; MLP, Mutual learning phase; MoMM, Minimization-of-mismatch metrics.

we are satisfied if the ALS m hereby introduced would provide a superior understanding of its human "peer" H, on whose data it is trained and to whom it adapts.

2. Framework: HMPL curricula

2.1. HMPL convention

To facilitate any future communication, we adopt the following conventions in this–c.f. Table 1–as well as any future article addressing the topic of *HMPL*:

- Human subjects and other learners of organic origin are to be denoted with upper-case characters, whereas artificial agents or other learners of non-organic origin are to be denoted with lower-case characters.[1]
- Each distinct skill, faculty, technique, or competence is to be denoted by a distinct symbol issued from a Greek alphabet. Skills, which are to be acquired by learners of organic origin, are to be denoted with upper-case characters, whereas skills, which are to be acquired by learners of artificial origin, are to be denoted with lower-case characters. To avoid ambiguous interpretations, only those characters of Greek alphabet, which are graphically distinct from their latinized counterparts, are to be used.
- Skills are attached to their respective "carriers" as right-side subscripts: e.g., expression H_Γ denotes H's level of mastery (LoM) of Γ.
- Combined operators $>\sim$ (somewhat greater than) and $<\sim$ (somewhat smaller than) denote the situation where the level of mastery of σ of involved participants clearly and undeniably share Vygotskian "zone of proximal development" (Shabani et al., 2010). For example, $T_\sigma >\sim P_\sigma$ describes an ideal didactic situation, whereby the LoM of competence σ, as exhibited by the human teacher T, is located within the zone of proximal development of the pupil P.
- Combined operator $=\sim$ (approximately same level as) denotes the situation of *a didactic equilibrium*. where the levels of mastery of σ are more or less the same. For example in a situation where $T_\sigma =\sim P_\sigma$, the human teacher T and the human pupil P master σ at more or less same level: there is very little, resp. nothing, which P could learn about σ from T or vice versa. When it comes to observable mastery of σ, T and P are in equilibrium: the objective of the learning process was attained.

TABLE 1 Structure of the first exercise of *HMPL* – C_1. See Section 2.1 for a closer description of the employed formalism.

Curriculum 1	Human	Machine
Role	Human H	Machine m
Curricular objective	Acquisition of λ_2	Understanding H's speech
Exercise 1		
Skill	Π=vocabulary learning	σ=accurate processing of H's speech
Initial Non-Equilibrium	$H_\Pi <\sim s_\Pi$	$m_\sigma <\sim H_\sigma$
Prior knowledge	Picture-speech associations	Text-picture associations
Input	Visual representation	Speech
Output	Speech	STT model
Post-learn Equilibrium	$H_\Pi =\sim m_\Pi$	$m_\sigma =\sim H_\sigma$

2.2. Structure

A human-machine peer learning curriculum (i.e., a HMPL-C) is a planned sequence of educational instructions–i.e., a curriculum–which involves:

1. At least one human learner G, H, I, ... which gradually develops her/his/their skill Γ.
2. At least one artificial learner a, b, c, ... which gradually develops its/her/his/their skill σ.
3. Activities by means of which G (resp. H, I, etc.) develops her/his/their skill Γ, which directly involve knowledge and competence exhibited by a (resp. b, c, etc.).
4. Activities by means of which a (resp. b, c, etc.) develops her/his/their skill σ, which directly involve knowledge and competence exhibited by G (resp. H, I, etc.).

Human–machine peer learning curricula could be either convergent or divergent. In convergent HMPL curriculum, the learning objective–i.e., a competence whose LoM is to be increased–of a human learner coincides, *mutatis mutandis*, to the learning objective of an artificial learner (e.g., morality or social competence learning). That is, $\Pi = \sigma$.

On the other hand, in a divergent HMPL curriculum, the learning objective differs from the objective of a machine learner: $\Pi \neq \sigma$.

With the notion of HMPL curricula and their most important subtypes thus introduced, we then proceed to a concrete practical example of a HMPL curriculum labeled as Curriculum 1 (HMPL-C1).

3. Objectives

It is important to underline that the ultimate aim of our research is limited not only to the sole improvement in skills and knowledge of the human learner but also to provide foundations for a symbiotic co-development whereby human and machine learn from each other, and together, in a shared system of exercises.

1 Note that the choice of a purely graphemic distinction "upper-case for organic" and "lower-case for artificial" in no way intends to imply that organic learners would be classified by definition as higher, upper, greater, or superior in any other way to non-organic learners. The choice of distinction is simply motivated by the historical fact that as upper-case characters preceded lower-case characters in the evolution of script, as do organic learners precede non-organic learners in the evolution of mind.

1. Create a curriculum increasing competence by helping the human learner H to acquire foreign language λ_2.
2. Create a curriculum that adapts an artificial learner m to properly "understand" H's speech.
3. Evaluate how much the mutual-learning method leads to an increase in the amount of cases of matching vocabularies among both *learners*.

These objectives are to be attained by conducting an experiment that is both pedagogic and computer-scientific in the same time.

4. Format: HMPL curriculum 1

Curriculum 1 (C_1) is a divergent HMPL curriculum whose goal is to help the human learner acquire foreign language λ_2 while simultaneously allowing an artificial learner m to increase its ability to accurately understand H's speech.

4.1. Exercise 1: Vocabulary learning

Being a curriculum, $HMPL - C_1$ is an ordered sequence of common exercises. At its base, each exercise is composed of *tasks*, which are hereby defined as the atomic unit of an exercise and thus of a curriculum.

Within the framework of an exercise, tasks are batched into iterations that are composed of learning and test + feedback phase. Figure 1 shows the diagram of the process.

The first exercise E_1 (resp. $HMPL - C_1 - E_1$) focuses on the acquisition of most basic building blocks of λ_2: vocabulary learning. Table 1 summarizes the distinctive aspects of $HMPL - C_1 - E_1$.

The presence of word "'picture" in both H and m columns in the "prior knowledge" row in Table 1 indicates that there is at least some knowledge which can considered to be "shared" between H and m, even before the learning starts. That is, both H knows from previous experience that the picture of a book carries a phonetic label /bk/ and, analogically, *m knows* that the picture of the book is to be associated with the textual label "book." In the context of the exercise presented in this article, such *machine's knowledge* is stored in a predefined word-list dataset *WL*.

By means of such shared priors can communication and sharing be established, preparing the ground for subsequent information transfer. Without such shared priors, there is nothing which could provide the base for subsequent man–machine co-development, where no reference point could initiate the mutual symbol grounding (Harnad, 1990).

4.2. Iterations and phases

Exercises E_X of HMPL curricula are composed of multiple iterations. Each iteration I_x is composed of:

1. Test + Feedback phase
2. Mutual Learning phase (MLP)

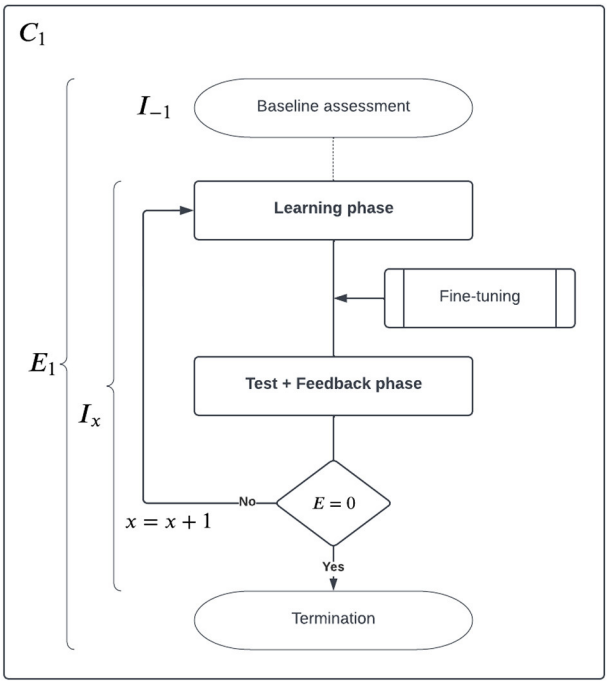

FIGURE 1
Diagram describing the generic structure of a $HMPL - C_1$ exercise. The curriculum is composed of exercises which are composed of iterations containing learning and testing phases. Within this article, we describe only the most simple case with one single iteration (e.g., $x < 2$).

4.2.1. Test + Feedback phase

Test + Feedback: In this phase, m evaluates what H already knows at the moment when the test phase is executed. Thus, during the task testing H's knowledge of word W, m displays to H the picture depicting W. No additional audio or text cues are available to H. After H names the picture he/she sees, m processes the audio signal through its speech-to-text models and obtains the predicted label $L_{predicted}$.

In case of a match between W and $L_{predicted}$, m provides H with encouraging feedback (e.g., a green rectangle). In case of absence of such a match, m provides H with corrective feedback (e.g., red rectangle + audio recording with a correct pronounciation of W). After providing the feedback, a new picture is displayed and a new task begins.

All along the test phase, information on matches and mismatches between expected word W and predicted label L_P is collected and aggregated. In a multi-iteration exercise, such information is used to determine the input into subsequent iterations. That is, it determines which tasks will be presented to H and in which order.

4.2.2. Mutual learning phase

The core of every *HMPL* iteration is the learning "phase" during which H learns and reinforces associations between what H hears, sees, reads, and speaks. Again, the learning phase is composed of different tasks. During each task, m exposes H to the answer in

the context of "ground truth" information. For each element of the given set of words, each corresponding text and illustration are displayed on the screen. At the same time, the corresponding audio file is played to aid *H* how to read. Once *H* speaks the word, *m* immediately evaluates if the expected text and the predicted text match. If they match, the next task is activated by showing the next word on the screen. Otherwise, *H* is required to speak again until *m* recognizes the word properly.

All along the learning phase, audio recordings are collected and serve as input for machine learning process, which is initiated immediately after *H* concludes all tasks batched in the learning phase. This is also a **mutual** learning phase because, after the collection of *H*'s pronunciations of all words, *m* uses–the process known as fine-tuning–the collected data to adapt parameters of its "generic" speech-to-text model to properties of *H*'s speech.

Given that we focused on the acquisition of German language, we used a DeepSpeech architecture (Hannun et al., 2014) model trained by Agarwal and Zesch (2019) on German speech data, such as the "generic" model. This development provided sufficient but necessary starting point for further fine-tuning of often strongly accented recordings collected during the proof-of-concept $HMPL_{C1}$ exercise introduced hereby.

4.3. Pre-learn and post-learn assessments

To facilitate entry to the understanding of our implementation of the *HMPL* concept, this article presents only the most simple setup composed of one full iteration I_0, followed by a subsequent test phase of I_1. Under such setup, an initial "pre-learn assessment" corresponds to the testing phase of iteration I_0 and "post-learn assessment" corresponds to the testing phase of subsequent iteration I_1.

5. Methodology

5.1. Materials

5.1.1. Web-based environment

Human-machine peer learning (HMPL) curriculum labeled as Curriculum 1 exercises are implemented as web-based components[2] of a digital primer project (Hromada, 2019). The learner communicates by means of her browser and WebSockets protocol with our own[3] open-source implementation of Mozilla's "DeepSpeech" speech-to-text system. No third-party or cloud-based platform is used.

5.1.2. Wordlist- WL_{100}

Items of WL_{100} are a subset of items that are used in the so-called Würzburger Reading Probe (Küspert and Schneider, 2000), an established tool that is used in Germany to assess the reading competence of elementary school pupils. WL_{100} contains 25 neutral, 30 masculine, and 45 feminine nouns prefixed with their determinate article (e.g., der / die / das).

Labels have mostly mono- and bi-syllabic structure with nine tri-syllabic and one tetrasyllabic (e.g., "Schokolade") items. Semantically, these 100 substantives were selected because they denote concrete objects like body parts, food, or animals and can be easily and unambigously depicted by our illustrators:

das Auge, das Auto, das Bett, das Blatt, das Brot, das Buch, das Ei, das Fahrrad, das Feuer, das Handy, das Haus, das Herz, das Kamel, das Krokodil, das Küken, das Lamm, das Mädchen, das Messer, das Netz, das Pferd, das Radio, das Schaf, das Schwein, das Tor, das Wasser, der Affe, der Apfel, der Ball, der Bär, der Baum, der Elefant, der Engel, der Fisch, der Hammer, der Hase, der Hund, der Igel, der Junge, der Käfer, der Kaktus, der Käse, der Ketchup, der Knopf, der Löffel, der Mais, der Mond, der Mund, der Pinsel, der Salat, der Schlüssel, der Schneemann, der Schnuller, der Schrank, der Schuh, der Stern, der Stift, der Stuhl, der Teller, der Tisch, der Topf, der Turm, der Wurm, der Zahn, der Zucker, der Zug, die Ampel, die Ananas, die Banane, die Biene, die Blume, die Brille, die Dose, die Ente, die Erdbeere, die Feder, die Flasche, die Gabel, die Giraffe, die Gitarre, die Gurke, die Hand, die Himbeere, die Hose, die Kartoffel, die Kuh, die Milch, die Mütze, die Nudel, die Orange, die Schere, die Schokolade, die Schule, die Seife, die Socken, die Tasse, die Tür, die Uhr, die Wurst, die Zahnbürste, die Zwiebel.

5.2. Participants

Three women and three men between 15 and 67 years of age participated in the proof-of-concept experiment. All learners were in the process of learning German as foreign language, with their level of mastery spanning A1-B2 levels of the Common European Framework of Reference for Languages (CoE, 2001). All participants had a strong accent influenced by their mother tongue and all of them gave explicit consent for recording and further processing and publication of their voice data for the purpose of the current study. Summary of participant information is displayed on Table 2.

5.3. Procedure

Before proceeding with the creation of a full-fledged, multi-iterative *HMPL* curriculum, we conducted a preliminary

TABLE 2 Information on each participant's age, gender, mother tongue, and CEFR German level.

Participant	Age	Gender	Mother lang.	CEFR Lv
H1	34	M	Turkish	B2
H2	34	F	Korean	C2
H3	30	F	Chinese	B1
H4	67	F	Slovak	A2
H5	34	M	Japanese	A2
H6	32	M	Korean	C1

2 https://fibel.digital

3 https://github.com/hromi/lesen-mikroserver

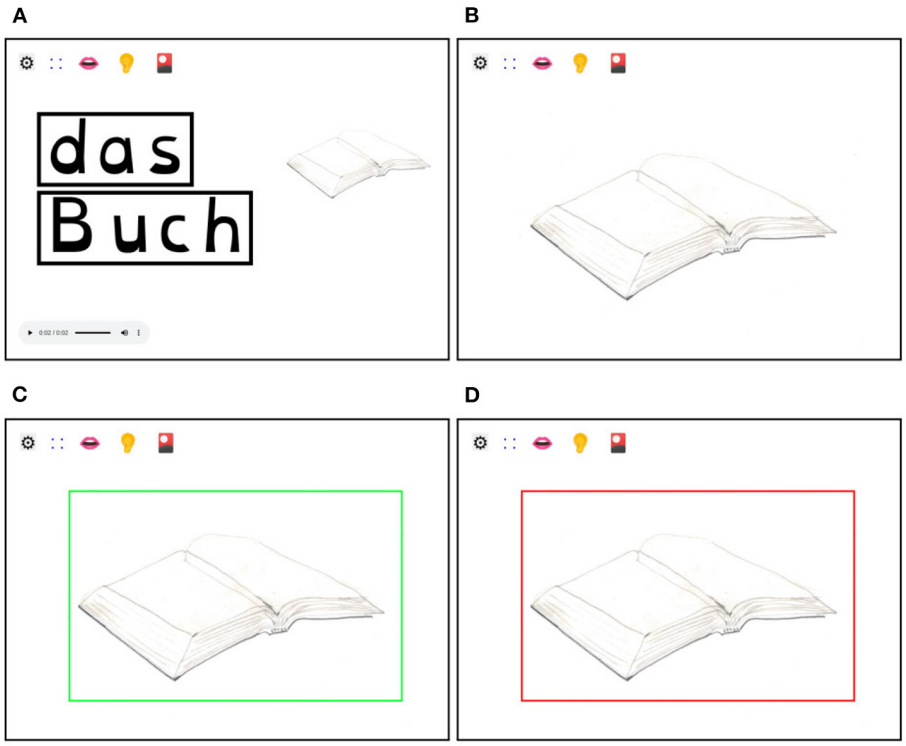

FIGURE 2
Web-based Interface of HMPL-C1-E1 **(A)** Learning phase. **(B)** Test+Feedback phase. **(C)** Correctly recognized. **(D)** Not recognized.

experiment to prove that *HMPL* is possible not only in theory, but also in practice. Thus, six learners were asked to go through the pre-learn assessment (e.g., test phase of I_0), "mutual learning phase" (MLP), and post-learn assessment (e.g., test phase of I_1). Within each phase, participants were exposed to 100 naming tasks, each corresponding to one element of the WL_{100} wordlist.

After collecting the voice samples of participant H_X during the MLP, a generic STT model is separately fine-tuned to new model M_X, which is better adapted to H_X's accent and other peculiarities of his/her voice.

The main interfaces, which we implemented for this study, are illustrated in Figure 2. After accessing the website, the pre-learn assessment begins by giving the first illustration to *H*. The audio recording process is initiated by a tactile command–for example, by *H* touching the given illustration–and is stopped when *H* aborts the contact.

An audio signal is sent from *H*'s microphone to *H*'s browser to be transferred by means of WebSockets protocol to the back-end system running DeepSpeech models on our local instance of a lesen-mikroserver[4] engine. Engine sends predicted label to H's browser and based on match between the human and the machine, a green or a red border appears around the illustration. Then, a new task is given. Once $N = 100$ tasks are done, the learning phase starts.

In the learning phase, a corresponding label and audio recording are provided alongside the illustration. Similar to this, *H*'s seeing, reading, hearing, and speaking activities are executed simultaneously (e.g., hearing while watching the picture and reading the text) or closely after each other (e.g., repeating the word that one just heard).

Once *H* solves all 100 tasks of the learning phase, (s)he *H* needs to wait at least 20 h for subsequent assessment. This is to make sure that we evaluate mid-term and long-term vocabulary extension and not some short-term memory, recency effects. In the meanwhile, fine-tuning is automatically executed on *m* once *H* terminates the learning phase: with 25 epochs and batch size 1, with an adaptation of *m*'s STT model to *H*'s voice on an NVIDIA Jetson takes cca 30 min.

During both testing and learning phases, learners are instructed to pronounce articles–der / die / das–along with the substantive. Similar to this, the exercise hereby described targets the acquisition of both lexical and morpho-syntactic competence.

5.4. Minimization of mismatch metrics

To allow for comparison with exercises of arbitrary lengths, the results are presented as the "minimization of error," whereby the ideal case corresponds to zero error.

In fact, we prefer to speak about "**minimization of mismatch**" (MoMM) to point out the fundamental difference between *HMPL* and classical signal detection theory (SDT) and machine-learning methodologies. In SDT, one normally deals with one classification system—for example an ML algorithm—in HMPL,

4 https://github.com/hromi/lesen-mikroserver

TABLE 3 The number of mismatches between words whose images were displayed ($L_{expected}$) and labels predicted by generic (resp. fine-tuned) speech-to-text models.

Participant $H1$			Participant $H2$			Participant $H3$		
Human	Machine		Human	Machine		Human	Machine	
	Generic	Fine-tuned		Generic	Fine-tuned		Generic	Fine-tuned
Pre-learn	92	76	Pre-learn	83	68	Pre-learn	93	89
Post-learn	92	71	Post-learn	67	61	Post-learn	91	88
Participant $H4$			Participant $H5$			Participant $H6$		
Human	Machine		Human	Machine		Human	Machine	
	Generic	Fine-tuned		Generic	Fine-tuned		Generic	Fine-tuned
Pre-learn	97	92	Pre-learn	99	93	Pre-learn	65	45
Post-learn	93	89	Post-learn	90	85	Post-learn	69	57

"Pre-learn" rows inform about the result of pre-learning assessment of H's vocabulary acquisition, "post-learn" rows denote the state assessed not earlier than 20 h after the "mutual learning phase." Worst result where no inference matched the displayed label is 100; best result where no mismatch between $L_{expected}$ and $L_{predicted}$ occurs is 0.

we simultaneously deal with two such *cognizing systems*: the human *H* and the machine *m*.

In addition, in *HMPL* since one system encodes information into modality from which the other system decodes it—e.g. human speaks out the word W corresponding to the expected label L_E and machine transcribes W into predicted label L_P - one can simply ask the question "does L_E match L_P?," **thus bypassing the necessity of often costly additional annotation** in order to understand the content of W. Note that in case of an ideal, oracle-like annotator, W=$L_{annotated}$ for all possible words of language λ_2.

A downside of the MoMM approach is that, instead of one source of erroneous behavior, one now has two potential sources of errors which–in the worst case–could result in a behavior erroneously evaluated as "valid" by an external observer. For when it may happen, a completely illiterate H will speak out the word "dog" when seeing "pig" and, simultaneously, a completely random speech classifier will neutralize the mistake by an own mistake, misclassifying the spoken word "pig" as "dog." Thus, mistake on both sides could result in a falsely positive result where activity as such would be evaluated as correctly resolved, while, in reality, errors happened on both sides.

Interestingly, the probability that such a "mutually neutralizing mistake" (MNM) would occur is inversely proportional to the product of a number of labels which H and m may generate and is thus relevant only in cases where classification into a finite, low amount of prespecified classes (N <20) takes place.

An upside, however, is that the *observation of a match between H and m provides simultaneous information about competences of both H and m*. When m displays an illustration of a dog setting the expected label to "dog" and when from all possible sound waves it can process and all possible inferences it can make it subsequently infers that H uttered "dog," one can be fairly confident that both H and m executed their part of the task in a correct manner.

6. Results

The most important results are presented in Table 3 (with input from H and m) and Table 4. Table 3 is truly a subset of Table 4 which can be obtained without the help of an additional external annotator.

6.1. MoMm

Summary "minimization of mismatch" results are presented in Table 3. Decreased observable within all different rows indicates that all six fine-tuned models started the process of successful adaptation to peculiarities of different voices and accents [Paired $t_{(15)} = 5.09$, p <0.001, mean of differences = 9.33].

One also observes a decrease within different in majority of columns of Table 3. This indicates that majority of human learners made less errors during post-learning test than in the pre-learning assessment: we interpret this as amelioration of each participant's vocabulary. Only cases where such amelioration is not observed are the "generic" column of the H_1 and both "generic" and "fine-tuned" columns of participant H_6.

In the case of H_1, a brief look at the "fine-tuned" column of the same participant makes it clear that the lack of observation of vocabulary increase is not due to the fact that H_1 had not learned anything, but due to the fact that the "generic" model was not able to properly process H_1's accent.

The situation is different in the case of H_6, the most German-proficient learner and the co-author of this article. To avoid any fallacy due to self-observation bias, we simply focus the attention of the reader on zero (resp. non-zero) values in the column "None" of the last four rows of Table 4.

Finally, after executing the "canonic *HMPL* analysis" and comparing the values on the main diagonal–that is, by comparing the competence of both m as H before and after mutual learning phase–one observes the results of statistical significance [Paired $t_{(5)} = 3.97$, $p = 0.01$, mean of the differences = 12.5].

TABLE 4 Analytic overview of the development of human (Π) and machine (σ) competences for all combinations of speech-recognition models and pre-, resp.

H	Assessment	Model	Π = vocabulary learning				σ = accurate recognition of H's speech			
			Knowledge				Incorrect inferences		Correct inferences	
			Full	Noun	Article	None	False	MNM	Valid	Match
H_1	Pre-learn	Generic	8	65	8	19	84	0	8	8
	Post-learn	Generic	8	77	6	9	85	0	7	8
	Pre-learn	Fine-tuned	24	49	8	19	62	0	14	24
	Post-learn	Fine-tuned	29	56	6	9	55	0	16	29
H_2	Pre-learn	Generic	17	57	2	24	61	0	22	17
	Post-learn	Generic	33	65	0	2	62	0	5	33
	Pre-learn	Fine-tuned	32	43	1	24	44	1	25	30
	Post-learn	Fine-tuned	39	59	0	2	55	0	6	39
H_3	Pre-learn	Generic	7	44	19	30	84	0	9	7
	Post-learn	Generic	9	69	4	18	87	0	4	9
	Pre-learn	Fine-tuned	11	39	20	30	77	0	12	11
	Post-learn	Fine-tuned	12	66	4	18	80	0	8	12
H_4	Pre-learn	Generic	4	31	6	59	81	0	15	4
	Post-learn	Generic	4	59	7	30	91	0	5	4
	Pre-learn	Fine-tuned	8	27	6	59	80	0	12	8
	Post-learn	Fine-tuned	10	52	7	31	86	0	4	10
H_5	Pre-learn	Generic	1	60	5	34	72	0	27	1
	Post-learn	Generic	10	73	3	14	71	0	19	10
	Pre-learn	Fine-tuned	7	54	5	34	81	0	12	7
	Post-learn	Fine-tuned	15	68	3	14	66	0	19	15
H_6	Pre-learn	Generic	35	57	4	4	55	0	10	35
	Post-learn	Generic	31	69	0	0	68	0	1	31
	Pre-learn	Fine-tuned	55	39	2	4	34	0	11	55
	Post-learn	Fine-tuned	43	57	0	0	56	0	1	43

Post-learning assessments. On the left, human-related side,"Full" denotes the full match between what H was supposed to say and what H actually said; "None" indicates that neither annotation of "Noun" nor that of "Article" component of the expected label-matched noun resp. Article component of the annotation. On the right, machine-related side, "False" refers to an invalid inference, "Match" refers to a correct inference based on the correct human input, "Valid" refers to a correct inference from an erroneous human input, and "MNM" denotes a theoretically possible match resulting from a combination of erroneous H input and an incorrect m inference.

6.2. *HMPL*-C_1-E_1 overview

Table 4 provides a more detailed description of the phenomena taking place before (pre-/generic) and after (post-/fine-tuned) a single MLP of $HMPL - C_1 - E_1$.

6.3. Presence of MNMs

A quantitative analysis revealed one occurrence of "mutually neutralizing mistake" which has been observed in the case of subject H_2 whose pre-learn articulation–as annotated by the human annotator–($L_{annotated}$="die blille") of the name for an object associated to the illustration of glasses ($L_{expected}$="die brille") has been evaluated as ($L_{predicted}$="die brille") by the DeepSpeech model fine-tuned on 200 (100 articles + 100 nouns) tokens of German language.

However, subsequent qualitative analysis of the recording by additional annotator revealed that the MNM actually had not occured and its observation was caused by error in annotation. Thus, a theoretical concept of a MNM still awaits empirical proof of its existence.

7. Conclusion

An anecdote of unknown origin states: "*If you have an apple and I have an orange and we exchange these fruits, then you and I will still each have one fruit. But if show You what I know and You will show me what You know, both of us will know two things at the end.*"

Pointing out to a fundamentally different essence of knowledge and information–as compared to matter–the proverb tacitly

illustrates how mutual learning can lead to enrichment of all parties involved.

Within this article, we provided first bits of empiric evidence supporting an insight that one of the two agents (e.g., "I" and "You") does not necessarily need to be organic or human origin. In other words, our results show that mutual co-development of human and machine competences is possible, at least within the domain of vocabulary learning on one hand, and speech recognition on the other.

More concretely, we demonstrate that one single "mutual learning phase," consisting of 100 nouns which are being learned and spoken out by human learner H to subsequently direct the fine-tuning of an artificial speech-to-text system m, is enough to induce useful mid-term and potentially long-term increase in both H's and m's skills. When compared with the pre-learning assessment, 12 more predicted labels matched the expected labels during the post-learn assessment that took place at least 20 h after the learning phase.

As our results indicate, this decrease in mismatch is both due to an increase in H's vocabulary and due to an increase in m's ability to accurately process H's voice. Thus, H's vocabulary competence Π and m's competence σ to properly process H's speech only started their trajectories toward their mutual didactic equilibrium $H_{\Pi} =\sim m_{\Pi} \wedge m_{\sigma} =\sim H_{\sigma}$.

As noteworthy, it is considered that the empirical confirmation of the occurrence of one instance of "mutually-neutralizing mistakes," turns out to be spontaneously emergent after one single human–machine "mutual learning phase." We consider the occurrence of such an MNM phenomenon to be consistent with Nowak's information-theoretical account of the emergence of a common language as a system of two co-developing *signifier* − *signified* association matrices (Nowak et al., 1999).

Additionally, it is appropriate to see certain parallels between the *HMPL* approach and those of "symbiotic education" and "digital twins" (Kinsner and Saracco, 2019). Indeed, both in our and Kisner's approach based on the so-called *symbionts*, one can speak about a complementary symbiotic relation between a human individual and the corresponding digital twin (DT) system. However, the DT concept is based on synchronization between physical and virtual object, which can be done by receiving data from physical to virtual in an object's full life cycle. This method is different from *HMPL*, where it is not a synchronization between the human and the digital but a mutual co-participation of the development of different skills, which stays in the foreground.

The results of this first empiric *HMPL* study may be of certain interest to both computer-scientific and pedagogico-didactic communities. From a computer-scientific perspective, one can interpret *HMPL* as a form of interactive supervision of a machine learning process that is realized by a human operator who is also learning.

Additionally, the metrics based on the "minimization of mismatch" can also turn out to be of certain practical importance. This is so because by focusing on the existence or the absence of a match between $L_{expected}$ and $L_{predicated}$, MoMM is in certain use-cases able to bypass the "ground truth" necessity: if one knows that $L_{expected}$ matches the $L_{predicted}$, one does not need to know of what exact content does the W in between contain. Such simplification may lead to a decrease in costly manual annotation and correction of one's data and may be of importance in many a scenario, including an educational one where the teacher does not have time or resources to process the recordings of all his/her pupils.

From the pedagogico-didactic perspective, one can start drafting diverse exercises and/or even wider curricula where a *mutual win-win interlock* between human learning and training of artificial agents is expected to occur. Surely, the "*curriculum one (i.e. C1= 'second language acquisition') - exercise one (i.e. vocabulary learning) for German language (i.e. λ_2='DE')*" is simply an introductory proof-of-concept for some more to come.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

DH contributed to cca 80% of text of this article and HK contributed the rest (notably in Sections 4, 5). Diagrams, figures, and Table 2 were created by HK and Tables 1, 3, 4 by DH. Backend and frontend codes for *HMPL* − *C*1 − *E*1 were programmed by DH. Data analysis was performed by DH and HK together. Five manual annotations were done by HK and one by DH. All authors contributed to the article and approved the submitted version.

Funding

The research presented in this article is closely related to Personal Primer collaboration between Berlin University of the Arts and Einstein Center Digital Future, which is jointly funded as a public-private partnership project by Cornelsen Verlag, Einstein Foundation, and City of Berlin.

Acknowledgments

We would like to express our gratitude to members of the Artificial Intelligence in Education Society who gave us a highly useful feedback to preliminary draft of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Agarwal, A., and Zesch, T. (2019). "German end-to-end speech recognition based on deepspeech," in *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers* (Erlangen: German Society for Computational Linguistics & Language Technology), 111–119.

Alnajjar, M., and Brick, B. (2017). Utilizing computer-assisted vocabulary learning tools in english language teaching: examining in-service teachers' perceptions of the usability of digital flashcards. *Int. J. Comput. Assist. Lang. Learn. Teach.* 7, 1–18. doi: 10.4018/IJCALLT.2017010101

Bajorek, J. P. (2017). L2 pronunciation in call: the unrealized potential of rosetta stone, duolingo, babbel, and mango languages. *Trends Educ. Technol.* 5, 24–51. doi: 10.2458/azu_itet_v5i1_bajorek

Chiu, T.-L., Liou, H.-C., and Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for efl college learning. *Comput. Assist. Lang. Learn.* 20, 209–233. doi: 10.1080/09588220701489374

Co, E. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, MA: Cambridge University Press.

Cooper, C. R., and Cooper, R. G. (1984). "Skill in peer learning discourse: what develops?" in *Discourse Development* (New York, NY: Springer), 77–97.

Cortese, C. G. (2005). Learning through teaching. *Manag. Learn.* 36, 87–115. doi: 10.1177/1350507605049905

Dubey, P., and Shah, B. (2022). Deep speech based end-to-end automated speech recognition (asr) for indian-english accents. *arXiv preprint arXiv*:2204.00977. doi: 10.48550/arXiv.2204.00977

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4

Frager, S., and Stern, C. (1970). Learning by teaching. *Read. Teach.* 23, 403–417.

Freinet, C. (1990). *Cooperative Learning and Social Change: Selected Writings of Celestín Freinet, Vol. 15*. Toronto, ON: James Lorimer & Company.

Golbeck, S. L. (1999). "Implications of piagetian theory for peer learning," in *Cognitive Perspectives on Peer Learning* (New York, NY), 3–37.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep speech: scaling up end-to-end speech recognition. *arXiv preprint* arXiv:1412.5567. doi: 10.48550/arXiv.1412.5567

Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Hogan, D. M., and Tudge, J. R. (1999). "Implications of vygotsky's theory for peer learning," in *Cognitive perspectives on peer learning*, eds A. M. O'Donnell and A. King (Lawrence Erlbaum Associates Publishers), 39–65.

Hromada, D. D. (2019). After smartphone: Towards a new digital education artefact. *Enfance* 3, 345–356. doi: 10.3917/enf2.193.0345

Hromada, D. D. (2022). "Foreword to machine didactics: on peer learning of artificial and human pupils," in *International Conference on Artificial Intelligence in Education* (New York, NY: Springer), 387–390.

Hu, Z. (2013). Emerging vocabulary learning: from a perspective of activities facilitated by mobile devices. *English Lang. Teach.* 6, 44–54. doi: 10.5539/elt.v6n5p44

Hung, H.-T. (2015). Intentional vocabulary learning using digital flashcards. *English Lang. Teach.* 8, 107–112. doi: 10.5539/elt.v8n10p107

Jun Zhang, L., and Bin Anual, S. (2008). The role of vocabulary in reading comprehension: the case of secondary school students learning english in singapore. *RELC J.* 39, 51–76. doi: 10.1177/0033688208091140

Kinsner, W., and Saracco, R. (2019). Towards evolving symbiotic education based on digital twins. *Mondo Digitale* 2, 1–14. doi: 10.1109/ICCICC46617.2019.9146095

Küspert, P., and Schneider, W. (2000). "Die würzburger leise leseprobe (wllp)," in *Diagnostik von Lese-Rechtschreibschwierigkeiten* (Göttingen), 81–89.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio, Speech Lang. Process.* 22, 745–777. doi: 10.1109/TASLP.2014.2304637

Matassoni, M., Gretter, R., Falavigna, D., and Giuliani, D. (2018). "Non-native children speech recognition through transfer learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 6229–6233.

Nikoopour, J., and Kazemi, A. (2014). Vocabulary learning through digitized and non-digitized flashcards delivery. *Procedia Soc. Behav. Sci.* 98, 1366–1373. doi: 10.1016/j.sbspro.2014.03.554

Nowak, M. A., Plotkin, J. B., and Krakauer, D. C. (1999). The evolutionary language game. *J. Theor. Biol.* 200, 147–162. doi: 10.1006/jtbi.1999.0981

Oxford, R., and Crookall, D. (1990). Vocabulary learning: a critical analysis of techniques. *TESL Canada J.* 7, 09–30. doi: 10.18806/tesl.v7i2.566

Perea-Barberá, M., and Bocanegra-Valle, A. (2014). "Promoting specialised vocabulary learning through computer-assisted instruction," in *Languages for Specific Purposes in the Digital Era* (New York, NY: Springer), 129–154.

Qian, D. D., and Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Lang. Testing* 21, 28–52. doi: 10.1191/0265532204lt273oa

Schlotterbeck, D., Jiménez, A., Araya, R., Caballero, D., Uribe, P., and Van der Molen Moris, J. (2022). ""Teacher, can you say it again?" improving automatic speech recognition performance over classroom environments with limited data," in *International Conference on Artificial Intelligence in Education* (New York, NY: Springer), 269–280.

Shabani, K., Khatib, M., and Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English Lang. Teach.* 3, 237–248. doi: 10.5539/elt.v3n4p237

Teske, K. (2017). Duolingo. *Calico J.* 34, 393–401. doi: 10.1558/cj.32509

Vinyals, O., Blundell, C., Lillicrap, T., and Wierstra, D. (2016). "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems, Vol. 29* (New York, NY).