frontiers

Proof-of-Concept of Feasibility of Human-Machine Peer Learning for German Noun Vocabulary Acquisition

Daniel D. Hromada [1,2], **Hyungjoong Kim** [1,2]

[1] *Institute of Time-based Media, Faculty of Design, Berlin University of the Arts, Berlin, Germany*
[2] *Digital Education, Einstein Center Digital Future, Berlin, Germany*

Correspondence*:
Daniel D. Hromada
dh@udk-berlin.de

ABSTRACT

We provide the first empiric evidence that creation of human - machine peer learning (HMPL) couples can lead to increase of level of mastery of different competences in both humans and machines alike. The feasibility of the HMPL approach is demonstrated by means of an exercise whereby the human learner H gradually acquires a vocabulary of foreign language while the artificial learner m fine-tunes his ability to understand H's speech. We evaluated the feasibility of the HMPL approach in an proof-of-concept experiment composed of pre-learn assessment, 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 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, exercise is terminated by the post-learn assessment phase. In both assessment phases, number of mis-matches between expected answer and answer provided by human and recognized by machine provides the main evaluation metrics. In case of all six learners who participated in the proof-of-concept experiment, we observed increase in amount of matches between expected and predicted labels which was caused both by increase of human learner's vocabulary, as well as by increase of recognition accuracy of machine's speech-to-text model. Therefore, we consider as reasonable to postulate that curricula could be drafted and deployed for different domains of expertise whereby humans learn from AIs in the same time as AIs learn from humans.

Keywords: peer learning, machine learning, foreign language learning, vocabulary learning, automatic speech recognition, DeepSpeech, German nouns, minimization of mis-match

1 INTRODUCTION

1.1 Human-Machine Peer Learning

Human-Machine Peer Learning (HMPL) is a proposal positioned at the very frontier between educational, cognitive and computer sciences. HMPL's core precepts which we have recently introduced in (Hromada, 2022) 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 an existence of a so-called "human-machine learning parallelism". That is, of a simple fact that 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 proper understanding of our proposal - is strong preference of human learners - notably children (Golbeck, 1999; Freinet, 1990), but not only - to acquire knowledge, behaviors and competences (Cooper and Cooper, 1984) from other learners who exhibit 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" proposal observes "great learning potential inherent in teaching" (Cortese, 2005).

In HMPL, it is an artificial system - the machine m - which assumes, aside 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 creation of such human - machine couples can lead to increase of 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 (L_2) while the artificial learner m fine-tunes his ability to understand H's spoken L_2 production.

1.2 AI-assisted Vocabulary Learning

By allowing to 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 L_2 teaching practice, being only rarely

explicitly and directly addressed during L_2 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 (Alnajjar and Brick, 2017; Perea-Barberá and Bocanegra-Valle, 2014) as well as mobile devices (Hu, 2013). Often 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, such digital assistants are indeed useful mediators of L_2 acquisition.

One among the most important features of such digital systems is the ability to recognize and process learner's speech. Albeit 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 is often deployed with certain amount of success in renown products like, for example, Duolingo (Teske, 2017), the problem of accurate ASR in domain of L_2 is far from being solved, notably for students with 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 human learner with highly accurate feedback about his/her pronunciation. Such problems are further exacerbated for huge majority of all non-English languages where there is not yet enough data publicly available for induction of highly accurate acoustic models (Schlotterbeck et al., 2022).

1.3 Small data

There is little doubt that recent advances in domain of artificial intelligence (AI) & 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 "importance of starting small" (Elman, 1993) and knowing that so-called few-shot or one-shot (Vinyals et al., 2016) learning is possible and provides a viable path to increase of accuracy of 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 artificial learning systems (ALS) which adopt to one - or fairly few - specific human beings.

Or, stated otherwise, instead of aiming to provide mediocre understanding of speech of practically all humans of the planet, we are satisfied if the ALS m hereby introduced would provide 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

In order to facilitate any future communication, we adopt 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 uppercase characters, artificial agents or other learners of non-organic origin are to be denoted with lower-case characters [1]
- Each distinct skill, faculty, technique or a competence is to be denoted by a distinct symbol issued from Greek alphabet. Skills which are to be acquired by learners of organic origin are denoted with upper-case characters, skills which are to be acquired by learners of artificial generic are denoted with lower-case characters. In order to avoid ambiguous interpretations, only those characters of Greek alphabet are to be used which are graphically distinct from their Latinized counterparts.
- 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 level of mastery of σ of involved participants clearly and undeniably share Vygotskian "zone of proximal development" (Shabani et al., 2010). E.g. $T_\sigma >\sim P_\sigma$ describes an ideal didactic situation whereby the LoM of competence σ as exhibited by the human teacher T is located within zone of proximal development of the pupil P.
- Combined operator $=\sim$ (approximately same level as) denotes the situation of *a didactic equilibrium*. where levels of mastery of σ are more or less the same. E.g. in 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.

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 develop 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.) develop her/his/their Γ directly involve knowledge and competence exhibited by a (resp. b, c, etc.)
4. activities by means of which a (resp. b, c, etc.) develop her/his/their σ directly involve knowledge and competence exhibited by G (resp. H, I, etc.)

HMPL 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 mutandi, 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: $\Pi \neq \sigma$.

The notion of HMPL curricula and their most important sub-types thus introduced, we now proceed to a concrete practical example of a HMPL-curriculum labeled as Curriculum 1 (HMPL-C1).

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

Table 1. Structure of first exercise of $HMPL-C_1$. See Section 2.1 for 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	$H_\Pi =\sim m_\Pi$	$m_\sigma =\sim H_\sigma$

3 OBJECTIVES

It is important to underline that the ultimate aim of our research is not limited to 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. Create a curriculum increasing competence helping the human learner H to acquire foreign language λ_2.
2. Create a curriculum which adapts an artificial learner m to properly "understand" H's speech.
3. Evaluate how much the mutual-learning method leads to increase in amount of cases of matching vocabularies among both *learners*.

These objectives are to be attained by conducting an experiment which 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 human learner to 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 atomic units of an exercise, and thus of a curriculum.

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

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

Note the presence of word "'picture" in both H and m columns in the "Prior Knowledge" row of Table 1. This indicates that there is at least some knowledge which can considered to be "'shared" between H and m even before the learning starts.

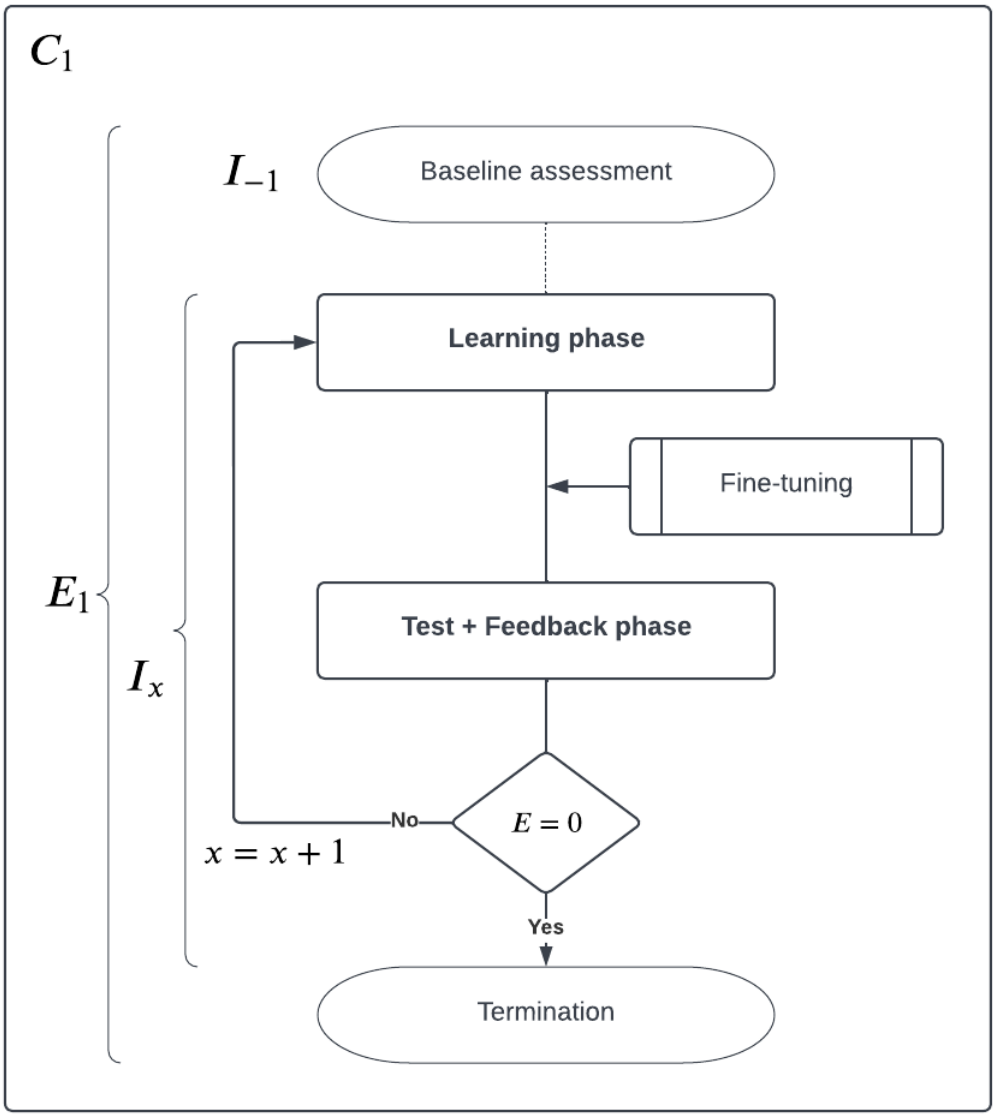

Figure 1. Diagram describing the generic structure of a $HMPL - C_1$ exercise. Curriculum is composed of exercises which are composed of iterations containing learning and testing phase. Within this article, we describe only the most simple case with one single iteration (e.g. $x < 2$).

That is, both H knows from previous experience that the picture of a book corresponds to the phonetic label /bʊk/ and, analogically, m also disposes of information associating the picture with the textual label "book".

In the context of the exercise E_1 presented in this article, such *machine's knowledge* is stored in a pre-defined dataset WL_{100} (c.f. 5.1.2).

It is thanks to - and by means of - such shared priors that communication and sharing can be established, preparing ground for subsequent informational transfer. Without such shared priors, there is nothing which could provide the base for subsequent man-machine co-development, no reference point which 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)

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

All along the test phase, information concerning matches and mismatches between expected word W and predicted label $L_{predicted}$ is collected and aggregated. In a multi-iteration exercise, such information is used to determine input into subsequent iterations. That is, 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, learning phase is composed of different tasks. During each task, m exposes H to the answer in context of "ground truth" information. For each element of the given set of words, each corresponding text and illustration is 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 do, next task is activated by showing next word on a 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 collection of H's pronunciations of all words, m uses - in 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 acquisition of German language, we used a DeepSpeech architecture (Hannun et al., 2014) model trained by (Agarwal and Zesch, 2019) on german speech data as such "generic" model. This provided sufficient but necessary starting point for further fine-tuning of often strongly-accented recordings collected during the proof-of-concept $HMPL - C_1$ exercise hereby introduced.

4.3 pre-learn and post-learn assessments

In order to facilitate entry to 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

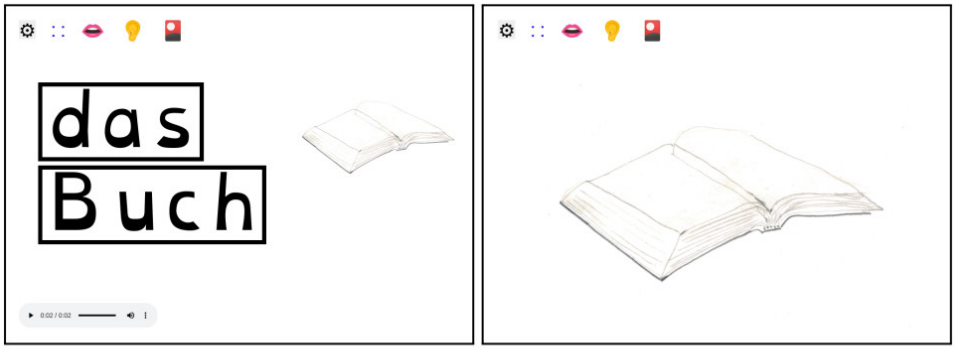

Figure 2a. **Figure 2b.**

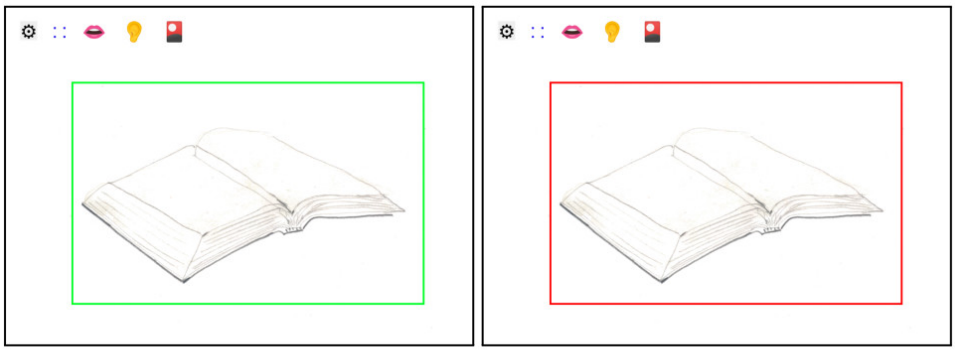

Figure 2c. **Figure 2d.**

Figure 2. Web-based Interface of HMPL-C1-E1 **(a)** Learning phase **(b)** Test+Feedback phase **(c)** Correctly recognized **(d)** Not recognized

of I_1. Under such setup, an initial "pre-learn assessment" corresponds to testing phase of iteration I_0 and "post-learn assessment" corresponds to testing phase of subsequent iteration I_1.

5 METHODOLOGY

5.1 Materials

5.1.1 Web-based Environment frontend to DeepSpeech

$HMPL-C1$ 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 being used and all DeepSpeech inferences are executed in absence of a supplementary language model (e.g. "n scorer" setup).

5.1.2 Wordlist- WL_{100}

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

[2] `https://fibel.digital`

[3] `https://github.com/hromi/lesen-mikroserver`

Labels have mostly mono- and bi- syllabic structure with 9 tri-syllabic and one tetrasyllabic (e.g. "Schokolade") item. Semantically, these hundred substantives were selected because they denote concrete objects like body parts, food or animals and can be easily and unambiguously depicted by our illustrators [4]:

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 to 67 years of age participated in the proof-of-concept experiment. All learners were in process of learning German as foreign language, with their level of mastery spanning A1 - B2 levels of Common European Framework of Reference for Languages (CoE, 2001). All participants had strong accent influenced by their mother language and all gave explicit consent for recording and further processing and publication of their voice data for the purpose of the current study.

Table 2. Information on each participant's age, gender, mother language, 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

5.3 Procedure

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 hundred naming tasks, each corresponding to one element of the WL_{100} wordlist.

After collecting 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.

Main interfaces which we implemented for this study are illustrated on Figure 2. After accessing the website, the pre-learn assessment begins by giving the first illustration to H. The audio recording process

[4] Active view of items WL_{100} including the illustrations used in the study is accessible at `https://fibel.digital/HMPL++DE++C1++E1/play`

is initiated by a tactile command - for example by H touching the given illustration - and stops when H aborts the contact.

Audio signal is sent from H's microphone to H's browser in order to be transferred by means of WebSockets protocol to back-end system running DeepSpeech models on our local instance of a lesen-mikroserver [5] engine running on CUDA-supported NVIDIA Jetson Xavier Dev Kit. This returns predicted label to H's browser and depending on whether a match occurs between the human and the machine or not, a green resp. red border appear around the figure. Subsequently, new task is given. Once $N = 100$ tasks are done, learning phase starts.

In the learning phase, a corresponding label and audio recording are provided alongside the illustration. Like 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 which one just heard).

Once H solves all 100 tasks of the learning phase (s)he H needs to wait at least 20 hours for subsequent assessment. This is to make sure that we evaluate will 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, adaptation of m's STT model to H's voice on an NVIDIA Jetson takes approx. 30 minutes.

Note that during both testing and learning phases, learners are instructed to pronounce articles - der / die / das - along with the substantive. Like this, exercise hereby described targets acquisition of both lexical as well as morpho-syntactic competence.

5.4 Minimization of Mis-Match metrics

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

In fact, we prefer to speak about "**minimization of mis-match**" (MoMM) to point out the fundamental difference between HMPL and classical signal detection theory (SDT) and machine-learning methodologies. For while in SDT one normally deals with one classification system - for example an ML algorithm - in HMPL, **we simultaneously deal with two such *cognizing systems*: the human H and the machine m.**

And since in HMPL one system encodes information into modality from which the other system decodes it - e.g. human speaks out the word W corresponding to 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 MoMM approach is that instead of one source of erroneous behaviour, one now have two potential sources of errors which - in the worst case - could result into behaviour erroneously evaluated as "valid" by an external observer. For it may happen that 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, mis-classifying 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.

[5] `https://github.com/hromi/lesen-mikroserver`

Table 3. Number of mis-matches between words whose images were displayed ($L_{expected}$) and labels predicted by generic (resp. fine-tuned) speech-to-text models. "Pre-learn" rows inform about the result of pre-learning assessment of H's vocabulary acquisition, "post-learn" rows denote state assessed not earlier than 20 hours after the "mutual learning phase". Worst result where no inference matched the displayed label is 100; best result where no mis-match between $L_{expected}$ and $L_{predicted}$ occurs is 0.

Participant $H1$		
Human	Machine	
	generic	fine-tuned
pre-learn	92	76
post-learn	92	71

Participant $H2$		
Human	Machine	
	generic	fine-tuned
pre-learn	83	68
post-learn	67	61

Participant $H3$		
Human	Machine	
	generic	fine-tuned
pre-learn	93	89
post-learn	91	88

Participant $H4$		
Human	Machine	
	generic	fine-tuned
pre-learn	97	92
post-learn	93	89

Participant $H5$		
Human	Machine	
	generic	fine-tuned
pre-learn	99	93
post-learn	90	85

Participant $H6$		
Human	Machine	
	generic	fine-tuned
pre-learn	65	45
post-learn	69	57

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

An upside, however, is that *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

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

6.1 MoMm

Summary "minimization of mismatch" results are presented in the Table 3. Decrease 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 <.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 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 case of H_1, the brief look at the "fine-tuned" column of the same participant makes it obvious that the lack of observation of vocabulary increase is not due to the fact that H_1 hadn't learn anything, but due to the fact that the "generic" model wasn't able to properly process H_1's accent.

The situation is different in case of H_6, the most German-proficient learner and the co-author of this article. To avoid any fallacy due to self-observational bias, we simply focus the attention of the reader on zero (resp. non-zero) values in the column "None" of 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 competence of both m as H before and after mutual learning phase - one observes results of statistical significance (Paired t(5) = 3.97, p = 0.01, mean of the differences = 12.5).

Table 4. Analytic overview of development of human (Π) and machine (σ) competence for all combinations of speech-recognition models and pre-, resp. post- learning assessments. On the left, human-related side,"Full" denotes 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 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 correct human input, "Valid" refers to a correct inference from an erroneous human input and "MNM" denotes a theoretically possible match resulting from combination of erroneous H-input and incorrect m-inference.

			Π = vocabulary learning				σ = accurate recognition of H's speech			
H	Assessment	Model	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

6.2 $HMPL - C_1 - E_1$ overview

Table 4. provides more detailed description of 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

One occurrence of "mutually neutralizing mistake" has been observed in case of subject H_2 whose pre-learn articulation - as annotated by the human annotator - ($L_{annotated}$='die blille') of name for an object associated to 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.

Thus, an initially only a theoretical concept of an MNM - as introduced in the penultimate paragraph of section 5 - has been empirically validated.

7 CONCLUSION

An anecdote of unknown generic 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 have provided first bits of empiric evidence supporting an insight that one of two agents (e.g. "I" and "You") does not necessarily need to be of organic or human generic. 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 in order 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 of 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 which took place at least 20 hours after the learning phase.

As our results indicate, this decrease of mis-match is both due to increase of H's vocabulary, as well as due to increase of 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 just started their trajectories towards their mutual didactic equilibrium $H_\Pi =\sim m_\Pi \wedge m_\sigma =\sim H_\sigma$.

As noteworthy is considered the empirical confirmation of occurrence of one instance of "mutually-neutralizing mistakes" spontaneously emergent after one single human-machine "mutual learning phase". We consider occurrence of such MNM phenomenon to be consistent with Nowak's information-theoretical account of emergence of 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 that of "symbiotic education" and "digital twins" (Kinsner and Saracco, 2019). Indeed, both in our as well as Kisner's approach based on so-called *symbions*, one can speak about a complementary symbiotic relation between a human individual and the corresponding digital twin (DT) system. However, concept of DT is based on

synchronization between physical and virtual object, which can be done by receiving data from physical to virtual in object's full life cycle. This is different from $HMPL$ where it is not a synchronization between the human and the digital, but mutual co-participation of development of different skills which stays in the foreground.

Results of this first empiric HMPL study may be of certain interest for both computer-scientific as well as paedagogico-didactic communities. From computer-scientific perspective, one can interpret $HMPL$ as a form of interactive supervision of a machine learning process realized by a human operator who is also learning.

Additionally, the metrics based on "minimization of mis-match" can also turn out to be of certain practical importance. This is so, because by focusing on existence or 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 what exact content does the W in between contain. Such simplification may lead to decrease of costly manual annotation and correction of one's data and may be of importance in many a scenario, including an educational one where teacher does not have time nor resources to process recordings of all her pupils.

From paedagogico-didactic perspective, one can start drafting diverse exercises and/or even wider curricula where *mutual win-win interlock* between human learning and training of artificial agents is expected to occur. For surely, the "*curriculum one (i.e. C_1=' second language acquisition') - exercise one (i.e. vocabulary learning) for German language (i.e. λ_2='DE')*" is just an introductory proof-of-concept for some more to come.

8 ACRONYMS

Acronym	Meaning
HMPL	human-machine peer learning
MDE	mutual didactic equilibrium
MNM	mutually neutralizing mistake
MLP	mutual learning phase
MoMM	minimization-of-mismatch metrics

CONFLICT OF INTEREST STATEMENT

Driven by hope that this article shall lead to increase of understanding between humans and humans, humans and machines and machines and machines, authors of this article 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.

AUTHOR CONTRIBUTIONS

DDH contributed to approx. 80% of text of this article, HJK contributed the rest (notably in sections 4 and 5). Diagrams, figures and table 2 created by HJK, tables 1, 3 and 4 by DDH. Backend and frontend code for $HMPL - C1 - E1$ was programmed by DDH. Analysis of data was performed by DDH and HJK together. Five manual annotations done by HJK, one by DDH.

FUNDING

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 Artificial Intelligence in Education Society who gave us highly useful feedback to preliminary draft of this article.

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, Germany: 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. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)* 7, 1–18

Bajorek, J. P. (2017). L2 pronunciation in call: The unrealized potential of rosetta stone, duolingo, babbel, and mango languages. *Issues and Trends in Educational Technology* 5, 24–51

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. *Computer Assisted Language Learning* 20, 209–233

CoE (2001). *Common European framework of reference for languages: Learning, teaching, assessment* (Cambridge University Press)

Cooper, C. R. and Cooper, R. G. (1984). Skill in peer learning discourse: what develops? In *Discourse development* (Springer). 77–97

Cortese, C. G. (2005). Learning through teaching. *Management Learning* 36, 87–115

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*

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition* 48, 71–99

Frager, S. and Stern, C. (1970). Learning by teaching. *The Reading Teacher* 23, 403–417

Freinet, C. (1990). *Cooperative learning & social change: Selected writings of Celestín Freinet*, vol. 15 (James Lorimer & Company)

Golbeck, S. L. (1999). Implications of piagetian theory for peer learning. *Cognitive perspectives on peer learning* , 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*

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 335–346

Hogan, D. M. and Tudge, J. R. (1999). Implications of vygotsky's theory for peer learning.

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

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

Hu, Z. (2013). Emerging vocabulary learning: From a perspective of activities facilitated by mobile devices. *English Language Teaching* 6, 44–54

Hung, H.-T. (2015). Intentional vocabulary learning using digital flashcards. *English Language Teaching* 8, 107–112

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 Journal* 39, 51–76

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

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

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 745–777

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)* (IEEE), 6229–6233

Nikoopour, J. and Kazemi, A. (2014). Vocabulary learning through digitized & non-digitized flashcards delivery. *Procedia-Social and Behavioral Sciences* 98, 1366–1373

Nowak, M. A., Plotkin, J. B., and Krakauer, D. C. (1999). The evolutionary language game. *Journal of theoretical biology* 200, 147–162

Oxford, R. and Crookall, D. (1990). Vocabulary learning: A critical analysis of techniques. *TESL Canada journal* , 09–30

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* (Springer). 129–154

Qian, D. D. and Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21, 28–52

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* (Springer), 269–280

Shabani, K., Khatib, M., and Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English language teaching* 3, 237–248

Teske, K. (2017). Duolingo. *calico journal* 34, 393–401

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems* 29