Kacper Łodzikowski

331069

Innovative methods for the teaching of second language metalinguistic awareness

Innowacyjne metody nauczania świadomości metajęzykowej języka drugiego

Rozprawa doktorska napisana
na Wydziale Anglistyki
Uniwersytetu im. Adama Mickiewicza w Poznaniu
pod kierunkiem prof. UAM dr hab. Piotra Gąsiorowskiego

Poznań, 2024

ACKNOWLEDGEMENTS

I would like to thank dr Jarosław Weckwerth for sparking my fascination with English pronunciation, prof. Geoffrey Schwartz for fostering my growth as a phonetician, and to prof. UAM dr hab. Piotr Gąsiorowski for trusting me (twice) with the freedom to explore the breadth of my research interests—culminating in this interdisciplinary dissertation.

I am grateful to prof. UAM dr hab. Joanna Pawelczyk for building a nurturing environment for PhD students at the Faculty of English, and to my colleagues (mgr Grzegorz Aperliński, dr Mateusz Jekiel, dr Kamil Malarski) for creating a family-like support network and long-lasting relationships that extended beyond academic corridors.

Thanks to the generations of amazing students at the Faculty of English who enthusiastically engaged in exploring and evaluating novel teaching and learning methods.

Thank you to dr John Behrens and dr Peter Foltz for reigniting my passion for academic writing and demonstrating what it means to have courage of one's convictions.

My PhD journey would not have been possible without the support of my loving parents. And while the research experience has been incredibly rewarding, the most statistically significant outcome of my years wandering the Faculty corridors has been meeting my partner. Her unwavering support and occasional not-so-gentle reminders to bring my work to completion have helped me stay focused on what truly matters in life.

Table of contents

TABLE OF CONTENTS .... **III**

LIST OF PUBLICATIONS .... **V**

FUNDING .... **VI**

LIST OF ABBREVIATIONS .... **VII**

INTRODUCTION **1**

PART 1 : BACKGROUND **3**

1.1. The evolving role of pronunciation in second language acquisition 3
1.2. Focusing on nativelikeness in the era of intelligibility 4
1.3. Factors affecting near-nativelike L2 pronunciation acquisition 5
1.4. The value of explicit pronunciation instruction 6
1.5. Research on metaphonological awareness training 7
1.6. Innovative methods for teaching metaphonological awareness 9
1.6.1. Print-and-play classroom games 9
1.6.2. Self-study transcription practice 10
1.6.3. Artificial intelligence tools 12
1.7. Knowledge gaps and research goals 14

PART 2 : RESEARCH **16**

2.1. About the publications 16
2.2. Institutional and pedagogical context for Publications 1-2 17
2.3. Publication 1 (Łodzikowski and Jekiel 2019) 18
2.4. Publication 2 (Łodzikowski 2021) 35
2.5. Publication 3 (Łodzikowski et al. 2024) 47

CONCLUSION **71**

Summary of findings 71
Limitations 72
Future directions 74

REFERENCES **76**

ABSTRACT.. 90

STRESZCZENIE... 91

APPENDIX A: AUTHOR CONTRIBUTION STATEMENTS FOR PUBLICATION 1 (ŁODZIKOWSKI AND JEKIEL 2019) 92

APPENDIX B: AUTHOR CONTRIBUTION STATEMENTS FOR PUBLICATION 3 (ŁODZIKOWSKI ET AL. 2024)... 95

List of publications

The present PhD dissertation comprises three thematically related publications:

Publication 1 (Łodzikowski and Jekiel 2019)

Łodzikowski, Kacper and Mateusz Jekiel. 2019. "Board games for teaching English prosody to advanced EFL learners", ELT Journal 73, 3: 275-285. (doi:10.1093/elt/ccy059).

Publication 2 (Łodzikowski 2021)

Łodzikowski, Kacper. 2021. "Association between allophonic transcription tool use and phonological awareness level", *Language Learning and Technology* 25, 1: 20-30. (doi:10125/44748).

Publication 3 (Łodzikowski et al. 2024)

Łodzikowski, Kacper, Peter W. Foltz and John T. Behrens. 2024. "Generative AI and Its Educational Implications", in: Dora Kourkoulou, Anastasia Tzirides, Bill Cope, and Mary Kalantzis (eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines*. Cham: Springer. (doi:10.1007/978-3-031-64487-0_2).

Funding

Two publications included in this dissertation were supported by a Preludium research grant from the National Science Centre (Poland) as part of project 2014/15/N/HS2/03867.

List of abbreviations

AI	artificial intelligence
ASR	automated speech recognition
CAPT	computer-assisted pronunciation training
ITS	intelligent tutoring system
L1	first language
L2	second language

Introduction

This PhD dissertation was created in equal measure out of curiosity and necessity. The curiosity started when, as a first-year student of English Philology at the AMU Faculty of English, I was fascinated by the ease with which my fellow students and I acquired advanced second-language (L2) vocabulary and grammar, relative to the difficulty of acquiring near-nativelike L2 pronunciation. Like many other Polish students, our aspiration to sound English was matched only by our frustration of continuing to sound Polglish despite our best efforts (Waniek-Klimczak et al. 2015).

The necessity arose when, as a PhD student, I found myself teaching the very same courses in (practical) English pronunciation and (theoretical) English phonetics and phonology that I had taken a few years earlier. The practical course was based on the emerging best practices for explicit pronunciation instruction, such as consciousness-raising activities such as corrective feedback (e.g. Saito 2011), or articulatory warm-ups and sound symbolism (e.g. Wrembel 2011a). The theoretical course was based on the accompanying body of work on L2 metaphonological awareness training (e.g. Wrembel 2005; Wrembel 2007; Dziubalska-Kołaczyk et al. 2015; Kivistö-de Souza 2015), a more advanced form of training for developing a controlled awareness of language rules.

While the practical course was reasonably effective, most learners found the theoretical course challenging. My students' sentiments echoed those of other Polish students of English Philology surveyed by Nowacka (2022). They found memorising phonetic and phonological rules difficult and boring, struggled to see the connection between theoretical concepts and practical pronunciation, and felt overwhelmed by the extensive amount of material. This is to be expected if a course in English phonetics and phonology is centred around acquiring declarative knowledge from a coursebook. Turning explicit knowledge into automatised knowledge to enable spontaneous speech production requires practice (Saito and Plonsky 2019). However, traditional course methods such as paper-based exercises in phonemic and phonetic transcription may not engage students in the type of sustained active learning required for developing applied skills (Chi and Wylie 2014). While it is evident that engaging students in metalinguistic training can enhance the effectiveness of pronunciation instruction (Kirkova-Naskova et al. 2021), finding effective strategies for the teaching of metalinguistic awareness remains a crucial challenge.

The relative successes of the broad field of computer-assisted language learning (Chapelle and Sauro 2017)—especially regarding the teaching and learning of explicit grammatical knowledge via automated writing evaluation tools—turned my attention to the subfield of computer-assisted pronunciation teaching. While the field is not new (see e.g. Chun 1998; Pennington 1999), the democratisation of access to the Internet in the 2000s brought a renewed interest and experimentation within the teaching and research community, resulting in novel applications of technologies such as spectrogram visualisaton software or automated speech recognition (see an overview in Fouz-González 2015). These tools, however, were designed primarily for mainstream pronunciation instruction for beginner and intermediate learners rather than the type of niche metalinguistic training required by advanced students. This eventually led me to prototype and pilot several innovative teaching aids for advanced students—both computer-assisted and not—some of which are presented in this dissertation.

This work has three main parts. Part 1 provides and introduction to the area of study, including a literature review and a summary of knowledge gaps and research goals. Part 2 presents a selection of my publications addressing the research goals. The first two are empirical studies investigating novel methods for the teaching of metalinguistic awareness: a pilot of classroom board games for the teaching of English prosody and a pilot of an allophonic transcription study tool. The third publication is a theoretical exploration of the future of teaching and learning enabled by artificial intelligence. Finally, the Conclusion section summarises the insights from the publications and their limitations as well as presents future research directions, including upcoming publications that were not selected for this dissertation for a number of reasons, not least brevity.

Part 1: Background

1.1. The evolving role of pronunciation in second language acquisition

The role of pronunciation in instructed second language (L2) acquisition has been shaped by shifting methodologies and priorities over the centuries. For languages such as English, pronunciation teaching before the 1850s mainly involved learners mimicking teachers and studying pronunciation from written texts, with no systematic pronunciation training (Baker 2018). The first wave of more organised pronunciation instruction (from the 1850s to the 1880s) prioritised learning through imitation methods—such as oral repetition, reading aloud, and minimal pair drills—albeit still without substantial theoretical support (Murphy and Baker 2015). A significant metalinguistic training innovation of this era was a vowel numbering system to help learners distinguish vowel sounds (Bentley 1849).

The second wave (from the 1880s to the 1980s) saw a departure from imitative-intuitive approaches towards analytic-linguistic methods. Pedagogies such as the direct method continued the tradition of imitation but incorporated accuracy drills (Wren 1912). Around the same time, the Reform Movement (e.g. Sweet 1899; Jespersen 1904) promoted a more scientific approach to phonology by introducing the International Phonetic Alphabet (IPA). Later pedagogies of the early 20th century, such as the audiolingual method, focused on oral communication and explicit pronunciation correction aimed at achieving near-nativelikeness (see an overview in Busà 2008).

The third major wave of pronunciation teaching (from the 1970s to the mid-1990s) was influenced by the early days of Communicative Language Teaching. This period was characterised by a shift from linguistic competence to communicative competence, resulting in a decreased focus on pronunciation teaching (Levis and Sonsaat 2017). This also meant that fewer new teachers received adequate pronunciation training (Derwing and Munro 2005). Over time, the need to align all aspects of language use with the ultimate goal of successful communication prompted a shift from nativelike correctness (nativeness) towards listener-oriented communicative effectiveness (intelligibility) (Munro and Derwing 1995).

The language pedagogy of the 21st century settled on a balance between meaning-focused instruction and form-focused instruction (Spada and Lightbown 2008). A renewed interest in L2 pronunciation teaching continued the focus on intelligibility over nativeness (Levis 2005; Derwing and Munro 2015) and was reinforced by an increasingly substantial body of evidence on the importance of pronunciation for communicative success (e.g. Derwing et al. 1998; Derwing and Munro 2005; Saito 2021; Suzukida and Saito 2022). In parallel, the proliferation of online resources and digital tools for pronunciation instruction (e.g. Sawala et al. 2009) lowered the entry barrier for teachers interested in introducing pronunciation instruction in their curricula.

1.2. Focusing on nativelikeness in the era of intelligibility

This PhD dissertation focuses on the acquisition of near-nativelike L2 pronunciation, defined as the ability to produce segmentals (vowels and consonants) and suprasegmentals (e.g. word stress, intonation) with accuracy close to that of a monolingual native speaker (Saito 2018). Although the research community largely agrees that mainstream pronunciation instruction should be grounded in intelligibility (Pennington 2021), some advanced adult learners may still want to strive for near-nativelikeness (usually interpreted as striving for the General British or General American pronunciation model) for personal or professional reasons (Pennington and Rogerson-Revell 2019). This is especially predominant in students enrolled in university-level English Philology programmes, mainly due to perceived attractiveness, cultural identity, or resulting confidence (Nowacka 2012; Buczek-Zawiła 2018; Lacabex and Roothooft 2023). Interestingly, later-stage students may exhibit more welcoming attitudes towards non-native speech (e.g. Waniek-Klimczak et al. 2015; Lintunen and Mäkilähde 2018). Moreover, many students understand that communicative success depends on a focus on fluency over accuracy (Waniek-Klimczak 2011). They also appreciate the diversity of regional English varieties and non-native pronunciation (Krzysik and Lewandowska, 2017). That being said, there is clearly still a need to explore methods of facilitating the acquisition of near-nativelike L2 pronunciation for those learners who want it.

1.3. Factors affecting near-nativelike L2 pronunciation acquisition

Acquiring a near-nativelike command of non-native language in adulthood presents numerous challenges. First, as posited by the Revised Speech Learning Model (SLM-r) (Flege and Bohn 2021), even though adult L2 learners use the same implicit speech learning mechanism that is successfully used by infants and toddlers, it does not work as well for adults because they start with the baggage of their L1 phonetic categories. When learners first perceive L2 speech, a latent process of interlingual identification creates perceptual links between L2 and L1 sounds, leading learners to default to producing L1 sounds instead of L2 sounds. This is not an issue if the sounds are so similar that such a substitution is unobtrusive. For other L2 sounds, however, learners need to form dedicated phonetic categories. This requires learners to process more tokens of sounds from a statistical distribution and subconsciously group them into auditorily-similar sets called equivalence classes, which in turn can be used for creating L2 phonetic categories used for the perception and production of L2 sounds. However, the formation of L2 categories is likely to be obstructed by the L1 categories that already exist in the learner's phonetic space. Specifically, the SLM-r predicts that the most difficult L2 categories to form are the ones that are somewhat similar—as opposed to very similar or very different—to existing L1 categories. Part of the challenge is that adult learners do not receive the same quantity and quality of spoken input as young learners. Even migrant adult learners—who may benefit from a naturalistic setting, acculturation, and close relationships with native speakers, such as marriage—are unlikely to get enough input to break the L2-L1 perceptual links.

This brings us to the second challenge, namely individual variability that affects pronunciation acquisition. Learner-extrinsic variability includes such aspects as quality of spoken input and pronunciation instruction methods (see an overview in Lee et al. 2015; Saito 2019). Learner-intrinsic factors include the effects of ageing on neural plasticity (Birdsong 2018) as well as differences in cognitive and socio-psychological characteristics, such as language learning aptitude, anxiety, musical hearing, and motivation (see an overview in Suzukida 2021). For example, new evidence highlights the potential role of musical hearing and musical experience on the acquisition of L2 vowels (Jekiel and Malarski 2021), L2 rhythm (Jekiel 2022), and L2 intonation (Jekiel and Malarski 2023). The main challenge of studying such factors, however, is their interplay with other variables, be it extrinsic (e.g. explicit pronunciation instruction) or intrinsic (e.g. pre-existing pronunciation proficiency). This was recently highlighted by Krzysik (2022), who found that individual differences such as phonological memory do not significantly influence the acquisition of L2 phonological perception and production. Perhaps some clarity will be provided by the emerging area of research on grit, understood as maintaining an enduring interest in particular goals over long periods and persevering in working towards achieving them. Domain-general grit has consistently been associated with achieving general life goals, such as secondary school graduation (Eskreis-Winkler et al. 2014). In the relatively new field of L2-specific grit, researchers observed that high scores on domain-specific grit assessments strongly predict language learning achievement (Teimouri et al. 2020), with the consistency-of-interest facet of grit being the most significant predictor (Sudina and Plonsky 2021), including among English Philology students (Zawodniak et al. 2021).

While the above-mentioned factors impact all learners from beginner to advanced, those who aim to acquire near-nativelike pronunciation face a third challenge, namely the time and effort needed to master segmental accuracy. Building on Munro and Derwing's (1995) framework that distinguishes between accentedness (i.e. how much a speaker's pronunciation deviates from native-like pronunciation) and comprehensibility (i.e. the ease of understanding the speaker by the listener), Saito (2021) showed that native listeners associate accentedness with segmental accuracy and comprehensibility with prosody and fluency. Unfortunately for adult L2 learners, reducing accentedness requires a larger time investment than improving comprehensibility (Saito 2015). This is favourable for learners with communicative goals, but unfavourable for learners striving for near-nativelikeness. And while advanced learners may experience an advantage due to being less impacted by lexical frequency effects (Saito 2018), they will still find attaining near-nativelike articulation challenging.

1.4. The value of explicit pronunciation instruction

While some of the above-mentioned factors are beyond the control of learners and teachers, learner-individual cognitive processing can be facilitated with instructional design focusing on explicit pronunciation instruction, i.e. providing learners with guidance on how to perceive and produce the segmental and suprasegmental features of the target language (Saito 2012; Suzukida 2021). Explicit pronunciation instruction involves a variety of approaches, ranging from classroom tactics such as corrective feedback to more strategic interventions such as articulatory training, auditory training, and metaphonological awareness training (see an overview of methods in Saito and Plonsky 2019).

Explicit pronunciation instruction assumes that L2 learning involves developing implicit knowledge (i.e. unconscious procedural speech processing) and that the process can be facilitated by explicit knowledge (i.e. declarative knowledge of the L2 consciously accessed in a controlled way) (Ellis 2005). This hypothesised weak interface assumes that explicit L2 knowledge facilitates key implicit learning mechanisms, especially input noticing and output monitoring (Schmidt 1990). For example, if learners receive instruction about the auditory characteristics of L2 vowels, they should be able to start noticing such features in their input.

In the first major meta-analysis of explicit pronunciation instruction studies, Lee et al. (2015) showed that—in within-group observational studies of mostly controlled production—explicit instruction improved L2 learners' pronunciation by 0.89 standard deviation units. They noted, however, that the studies have exhibited patterns of publication bias. Additionally, Thomson and Derwing (2015) commented that the analysed studies varied considerably regarding which aspects of speech production were measured and how, which decreased the generalisability of the findings. In another meta-analysis, Sakai and Moorman (2018) showed that perception-focused instruction increases productive pronunciation ability by 0.54 standard deviation units. Subsequently, Saito and Plonsky (2019) showed how the effect of pronunciation instruction indeed changes according to outcome definition. For example, a relatively large effect was observed for the controlled production of variables measured instrumentally (e.g. vowel formants), compared to a relatively small effect for spontaneous production thereof.

1.5. Research on metaphonological awareness training

Traditionally, phonological awareness is categorised into epilinguistic and metalinguistic (Ellis 2004). The former is an unconscious awareness of language rules, measurable through self-repair or mimicry. The latter is a conscious awareness of language rules, measurable through declarative knowledge or attention to form during production. Previous studies showed a positive relationship between L2 phonological awareness and L2 pronunciation, regardless of whether they measured its epilinguistic aspect (e.g. Mora et al. 2014), its metalinguistic aspect (e.g. Peltola et al. 2014; Kennedy and Trofimovich 2010; Saito 2019), or both (e.g. Venkatagiri and Levis 2007).

For the purpose of this dissertation, we will focus on recent research on instructed L2 phonological awareness acquisition, understood as equipping learners with explicit declarative knowledge of the L2 phonological system and its phonetic properties. Following Wrembel (2011a), we will refer to this as metaphonological awareness training. The main assumption is that the acquisition of explicit metalinguistic knowledge enables controlled speech production, which through sustained practice leads to proceduralisation and automatisation of the explicit knowledge—resulting in spontaneous speech production (Saito and Plonsky 2019). These interventions indeed demonstrated positive benefits for both production (Alves and Magro 2011; Lee et al. 2020; Saito 2011; Saito 2013; Wrembel 2005; Zhang and Yuan 2020) and perception (Carlet and Kivistö-de Souza 2018; Couper 2022).

One could argue that most explicit pronunciation instruction interventions described in the section 1.4 are by definition metaphonological, and the need for introducing a new term such as metaphonological awareness is superficial. This dissertation, however, treats metaphonological awareness interventions as a separate, though closely connected, area of study. The type of metaphonological awareness training researched here focuses on advanced learners aiming to achieve near-nativelike pronunciation, while research on explicit PI tends to focus on mainstream learners for whom achieving such pronunciation proficiency is unnecessary.

That being said, some researchers combine metaphonological awareness instruction with other consciousness-raising activities, such as corrective feedback (e.g. Saito 2011) or computer-assisted visualisations of phonetic features, such as pitch (e.g. Ramírez Verdugo 2006). Moreover, some studies use terms such as 'explicit phonetic instruction' for instruction involving elements of declarative knowledge such as articulatory description of L2 sounds (e.g. Saito 2011), while others use the same term for perception activities without the declarative aspect (e.g. Lacabex and Gallardo-del-Puerto 2020). Conversely, studies such as Lee et al. (2020) use the broad term 'pronunciation instruction' while incorporating elements of metaphonological awareness instruction.

1.6. Innovative methods for teaching metaphonological awareness

In our exploration of novel methods for the teaching of metaphonological awareness, it is important to acknowledge a broad range of techniques that, while significant, fall outside the scope of this review. The broad category of innovation in pronunciation teaching techniques (see an overview in Brinton 2017) includes, among others, kinesthetic approaches, such as using body movements or clapping to indicate stress patterns (Acton 1984; Murphy 2013), as well as drama techniques to raise awareness of suprasegmental features such as rhythm and intonation (Galante and Thomson 2016). Additionally, this dissertation excludes tools stemming from the metaphonological awareness teaching tradition, such as think-aloud protocols (Wrembel 2011b); questionnaires (Lintunen 2013), or self-reflective journals (Kennedy and Blanchet 2013; Inceoglu 2021)—since these have been already researched in some detail. Finally, this review does not include novel data collection instruments for analysing metaphonological awareness, such as the structured task for self-reflection on pronunciation deviations by Kivistö-de Souza and Lintunen (2023). Instead, this discussion centres on understudied methods and tools whose full potential is yet to be realised, namely print-and-play classroom games, self-study transcription practice tools, and artificial intelligence (AI) systems, with a particular focus on intelligent tutoring.

1.6.1. Print-and-play classroom games

Even before the advent of explicit pronunciation instruction, language teachers and learners relied on playful verbal drills and game-like exercises (Baker 2018). Historically, the syllable has been particularly suited to these activities, as demonstrated by a long history of tongue twisters addressing challenging phonotactic and articulatory combinations as well as variations of Pig Latin games, in which speakers systematically modify words by shifting their onsets or adding fabricated suffixes (Cardoso 2017).

During the 1970s and 1980s, the shift towards Communicative Language Teaching encouraged the use of role-plays, simulations, and games to practice language in context. Over time, this resulted in the inclusion of a broad spectrum of pronunciation-related language play, such as jokes or riddles (Wong 1987; Bell 2012; Bell and Pomerantz 2015).

This focus naturally extended to classroom games. The pivotal coursebook by Hancock (1995) included a variety of verbal and pen-and-paper pronunciation games, ranging from puzzles to mini board games. Hancock's foundational contributions capitalised on the peak popularity of photocopiable teacher's resource books, which made it easier and more affordable for teachers to customise class materials by freely combining paper-based handouts. Despite some notable contributions like those of Nixon and Tomlinson (2005) and Hancock (2017), the field of L2 pronunciation classroom games lagged behind the more established L2 research domains (Levis and Sonsaat 2017). Instead on pronunciation, most game-oriented coursebooks focused on grammar and related metalinguistic aspects (e.g. Zaorob and Chin 2001; Hadfield 2003).

Similarly, L2 research on puzzles and games focused mainly on their usefulness for enhancing the acquisition of grammar, vocabulary, and broad communicative skills (Danesi and Mollica 1994; Treher 2011; Wu et al. 2014). It is reasonable to expect, however, that at least some of the benefits of improved cognitive engagement and metalinguistic knowledge observed in other areas of L2 teaching and learning would extend to the domain of pronunciation.

1.6.2. Self-study transcription practice

Perhaps the longest-standing method of teaching aspects of metaphonological awareness has been via IPA-based phonemic and phonetic transcription. Transcription practice aids in forming precise mental representations of sounds and enhances learner autonomy by enabling self-guided learning and feedback (Mompeán and Fouz-González 2021). Pen-and-paper-based transcription is a readily available, affordable, and flexible tool that can be used for working with any type of speech, ranging from isolated words to longer passages of connected speech, both scripted and spontaneous (Mompeán 2017).

While IPA transcription was present in general-purpose English language learning textbooks already in the first half of the 20th century, its role diminished in the latter half of the century with the broader shift in thinking about explicit pronunciation instruction (Sobkowiak 2012). In the early 2000s, transcription practice has been successfully used to help learners to achieve better pronunciation learning outcomes (e.g. Lintunen 2004). L2 learners appreciate the way in which transcription helps them visually notice and remember sounds, especially for languages with irregular sound-to-grapheme correspondence, such as English (Mompeán and Lintunen 2015). Even though both teachers and learners can sometimes perceive phonetic notation as too theoretical and detached from practical pronunciation practice (Mompeán and Fouz-González 2021; Nowacka 2022), when applied correctly (i.e. as a means to an end) phonetic notation can serve as a flexible metalanguage that facilitates systematic reasoning about L2 speech features.

Despite these benefits, transcription remains underused. This is largely due to the significant time required to prepare engaging exercises and the need for teachers to provide learners with feedback manually, at least in the initial stages of acquiring the skill (García Lecumberri and Cooke et al. 2003). A handful of researchers and practitioners proposed addressing these challenges by dedicated computer software that automated the process of scoring learners' work products (e.g. García Lecumberri and Maidment et al. 2003; Bates et al. 2010; Bruijn et al. 2011). Interestingly, many such tools were developed outside of the domain of L2 pronunciation teaching, especially for the purpose of training speech pathologists and other L1 clinicians. The reusability of these systems into adjacent fields (e.g. from L1 speech therapy to L2 speech acquisition) only reinforces the versatile nature of transcription as a metalinguistic enabler for practical downstream skills. Speech clinician training is also the only domain that produced notable studies of digital transcription tools. For example, Titterington and Bates (2018) found a positive relationship between engagement in auto-graded transcription tasks and metaphonological awareness in language therapy students.

While self-study transcription practice offers the potential to boost learner autonomy and accelerate the development of metalinguistic awareness without overburdening teachers, the adoption of tools developed specifically for L2 professionals by their peers and researchers remains limited. This situation may change with the relatively recent advent of websites providing free (though often ad-sponsored) automated text-to-transcription services. Typically, these sites only offer text conversion based on fixed rule sets. The more advanced tools are based on curated databases of phonemically-transcribed words (e.g. PhoTransEdit 2023) and can even extrapolate patterns from such databases to previously-unseen words (Brondsted 2020). And while some websites claim to offer allophonic transcription, the range of phonetic processes they cover is usually not exhaustive and their reliability remains to be verified. Nevertheless, the landscape of such online tools is evolving rapidly and warrants careful observation.

1.6.3. Artificial intelligence tools

The definition of AI itself is subject to debate. This dissertation adopts a recently updated definition by OECD (2024), endorsed by regulatory bodies worldwide following the surge in generative AI. According to the definition, an AI system is "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." The authors acknowledge that AI systems differ in their level of autonomy and adaptiveness.

It is useful to begin this review of AI-enabled metalinguistic awareness teaching with an overview of computer-assisted pronunciation training (CAPT). This term encompasses a range of methods, some of which align more closely with the domain of AI than others. In the late 1990s and early 2000s, the main premise of CAPT revolved around involving students in instrumental speech analysis using signal processing software, such as Praat (Boersma and Weenink 2023) (see example curricula in Chun 1998; Pennington 1999; Vaissière 2003). One could argue that such systems not fit the OECD's decision-oriented definition of AI. They do, however, employ algorithms to make some decisions on how to process and visualise speech signals, and therefore could be put on the fringe of the natural language processing subfield of AI. A meta-analysis by Lee et al. (2015) showed that 39% of explicit pronunciation instruction studies relied on CAPT understood mainly as speech signal visualisations, especially spectrograms and pitch contours. They also found that human-led instruction was still more effective than computer-assisted instruction. One explanation for this is that spectrogram-based CAPT exposes learners to noisy raw data that can be difficult to interpret without expert guidance, potentially leading to incorrect self-corrections and increased frustration.

A separate category of CAPT research focuses on a type of AI known as automated speech recognition (ASR). In the 2000s, such systems were used for automated scoring of L2 pronunciation in high-stakes summative assessments (Bernstein 1999). In the 2010s, they were adapted to commercial mobile applications, which prompted a new wave of research into their use in L2 pronunciation acquisition (Rogerson-Revell 2021). The main pedagogical assumption is that ASR tools provide learners with an opportunity to notice errors in their own speech. For spontaneous speech, this usually manifests by outputting an incorrect speech-to-text transcription. For controlled speech, an ASR application usually explicitly highlights the deviation from the expected output (often modelled after a native speaker target) via a percentage score (see an overview in Walesiak 2020; see a detailed analysis of a particular application in Becker and Edalatishams 2019).

One meta-analysis of CAPT that included ASR studies (Mahdi and Al Khateeb 2019) reported a moderate effect size (Cohen's $d = 0.66$) on L2 pronunciation acquisition. Note, however, that this analysis covered a diverse array of papers, ranging from commercial language learning applications to bespoke systems not accessible to the general public as well as dictation systems that simply transcribe speech without offering explicit feedback. Interestingly, another meta-analysis focusing strictly on ASR systems (Ngo et al. 2023) reported a similarly moderate effect size (Hedges' g = 0.69).

While ASR systems have shown some success for explicit pronunciation training, their usefulness for advanced metaphonological awareness training may be limited because they usually do not provide explicit phonetic feedback to help the learner understand why they received a particular score. Many of them use methods that can confuse non-native English learners, such as providing transcriptions in the form of respelling (Coulange 2023). And while there are noteworthy examples of ASR-based metaphonological awareness training applications grounded in L2 research (see excellent example by Tejedor-García 2020), they tend to focus on the beginner or intermediate learner.

Finally, the type of AI that may be particularly relevant to metaphonological awareness training is the intelligent tutoring system (ITS) (Graesser et al. 2012). These chatbot-like (usually text-based) systems aim to replicate the nuanced dynamics of one-on-one tutor sessions. They can respond dynamically to student inputs and provide personalised feedback and guidance, much like a human tutor would. An early example of metalinguistic awareness training was a system developed by Tasso et al. (1992), which helped students understand English verb structures. This and other such systems attempted to understand a learner's misconception and provide specific feedback to correct errors and reinforce correct usage (Swartz and Yazdani 1992; Tafazoli et al. 2019.).

ITSs have been extensively tested and found successful in educational domains other than language learning (esp. math and computer science), showing an average increase in domain-specific assessments from the 50th to the 75th percentile (Kulik and Fletcher 2016). There has been less rigorous research in the field of L2 language learning, not to mention its subdomains such as pronunciation. Similarly to L2 CAPT, meta-analyses of L2-focused ITSs such as Lee and Lee (2024) encompass a diverse array of systems, ranging from commercial applications (e.g. Duolingo) to learning management systems (e.g. Moodle), which decreases the generalisability of the findings. A review of actual language ITSs by Slavuj et al. (2015) showed that—while innovative and promising—they have yet to show significant effects on learning outcomes. This is primarily due to their limitations in handling the nuanced and context-sensitive aspects of language use (Tafazoli et al. 2019).

One reason AI tutors have struggled in language teaching is that, historically, these systems primarily used basic methods to process and understand human language. They operated on relatively fixed rules—specified manually or learned from corpora—and lacked a genuine understanding of the flow of a conversation. To simplify, they worked by consulting a vast database filled with facts, rules, and previous interactions, which they then applied to new input from learners to generate appropriate responses. This method allowed the systems to detect patterns in student responses and provide seemingly insightful feedback. However, these systems often failed when faced with ambiguous or unusual situations that did not match their programmed instructions.

1.7. Knowledge gaps and research goals

While significant strides have been made in researching the teaching of L2 pronunciation and metaphonological awareness, several critical gaps remain unaddressed. These gaps, along with the research goals addressed by this dissertation, can be outlined as follows.

First, past interventions in metaphonological awareness studies have predominantly focused on beginner to intermediate learners in communication-oriented classes, where the primary goal was intelligibility. Addressing the need to broaden research to include advanced learners (Saito 2019), this dissertation focuses on interventions targeting advanced English Philology students for whom near-nativeness is often a curriculum requirement.

Second, alongside proposing novel tools for the teaching of metaphonological awareness, there is a need for a deeper understanding of how learners engage with such tools over the long term. Digital tools presented in this dissertation offer an opportunity to investigate study patterns that could shed light on data-driven pedagogy improvements as well as individual behavioural differences in acquiring metaphonological awareness.

Third, the field of computer-assisted language learning—and perhaps L2 pedagogy as a whole—struggles with transparency and reproducibility of research. Many studies still fail to provide the data and code required for replication, which hampers progress in the field. The empirical studies in this dissertation implement reproducible research practices to foster a more robust scientific dialogue.

Finally, previous-generation AI applications for pronunciation teaching have faced several limitations, especially in the realm of conversational ITSs. The rapidly evolving landscape of next-generation AI technologies—especially large language models—presents an opportunity to explore their applications in education. This dissertation provides a theoretical investigation of generative AI technologies in teaching and learning, providing a starting point for developing future ITSs targeting metalinguistic awareness training.

Part 2: Research

2.1. About the publications

To address the research questions posed in Part 1, this dissertation presents a selection of three most impactful publications from a wider body of work by the present author. Other publications on the topic, especially two detailed explorations of AI for metalinguistic awareness (in press and in review), are mentioned in the Conclusion as part of future directions.

Publication 1 (Łodzikowski and Jekiel 2019) was published in ELT Journal (140 points in the classification of the Polish Minister of Science). It investigates the use of board games to enhance the teaching of English prosody—specifically word stress, intonation, and rhythm—to English Philology students. Drawing on theories from cognitive psychology and pedagogy that highlight the benefits of game-based learning, the research implemented three custom-designed board games within an English phonetics and phonology course. While only correlational, the findings suggest that such games may increase learners' classroom engagement and performance in post-class assessments (declarative measures of metaphonological awareness).

Publication 2 (Łodzikowski 2021) was published in Language Learning and Technology (200 points). It describes an observational study of English Philology students using a tool developed by Łodzikowski and Aperliński (2016). The usage of the tool was associated with modest improvements in the students' metaphonological awareness levels, particularly when usage was spaced over time. The study was also replicated by Foung and Kohnke (2023), who not only confirmed the original findings via an exact replication but also addressed potential methodological limitations, mainly the relatively low sample size, which could affect the generalisability of the results. Their analysis further highlighted the benefits of the self-study transcription tool and reinforced the importance of replication in computer-assisted language learning research.

Publication 3 (Łodzikowski et al. 2024) is a theoretical exploration of the impact of generative AI on education, to be included in a Springer volume on AI-mediated education due in September 2024. Co-authored with leading figures in AI-enabled assessment and learning—Peter W. Foltz and John T. Behrens—this chapter presents an expansive overview of the opportunities and challenges for educators and institutions at the onset of the generative AI era. It deviates from the narrow focus of the first two publications on metalinguistic awareness to address broader pedagogical themes—a shift that was strategically chosen to engage a wider audience. The publication serves as a frame of reference for upcoming works by the present author that explore AI in language teaching (Łodzikowski in press) and report pilot results of a custom-built metaphonological awareness tutoring system powered by a large language model (Łodzikowski et al. in preparation)—both outlined at the end of this dissertation as part of future directions.

2.2. Institutional and pedagogical context for Publications 1-2

The pilot studies described in Publication 1 and Publication 2 were integrated into the same institutional context (albeit in different years). The subjects were English Philology students at Adam Mickiewicz University in Poznań (Poland), taking a 9-month course in English phonetics and phonology. The course was organised around in-person classes (90 minutes per week) and mandatory homework assignments on Moodle. The course syllabus covered over 20 topics, including phonemic and allophonic transcription, articulatory description of vowels and consonants, connected speech processes, word stress and weak forms, intonation, and regional varieties of English. The training focused on those aspects of English pronunciation with which Polish EFL learners are known to struggle due to cross-linguistic influence and low spelling-to-pronunciation correspondence (Nowacka 2016; Sawala et al. 2009; Sobkowiak 2004; Szpyra-Kozłowska 2015; Rojczyk 2010; Rojczyk and Porzuczek 2012; Rojczyk and Porzuczek 2017; Rojczyk and Porzuczek 2019). At the segmental level, this included, for example, vowel quality contrasts (challenging because General British has almost twice as many vowels as Polish) and features non-existent in Polish, such as pre-fortis clipping or non-rhoticity. At the suprasegmental level, the training similarly focused on those features which do not exist in L1, or are markedly different in L2, for example word stress in compound nouns, weak forms of function words, connected speech processes, and intonation patterns. Additionally, the training expanded beyond traditional phonological representations (phonemes as definite categories) and included elements of acoustic phonetics training and activities to help attune learners to perceptual cues (Schwartz 2005).

2.3. Publication 1 (Łodzikowski and Jekiel 2019)

Łodzikowski, Kacper and Mateusz Jekiel. 2019. "Board games for teaching English prosody to advanced EFL learners", ELT Journal 73, 3: 275-285. (doi:10.1093/elt/ccy059).

Board games for teaching English prosody to advanced EFL learners

Kacper Łodzikowski and Mateusz Jekiel

This exploratory study fills the gap in research on using print board games to teach English prosody to advanced EFL learners at university level. We developed three in-class print-and-play board games that accompanied three prosody-related topics in a course in English phonetics and phonology at a Polish university. For those topics, compared to topics without any board games, learners reported higher in-class engagement and obtained higher post-class quiz scores. At the end of the course, learners rated board games as equally or more useful than some of the other teaching aids. Although traditional printed worksheets were still rated as the most useful teaching aid, learners expressed their preference for using extra classroom time for playing board games instead of completing extra worksheet exercises. We hope these promising results will encourage teachers to experiment with implementing these and other board games in their advanced curricula.

Theoretical support for board games in ELT

Board games have been the topic of many scientific studies, primarily in psychology, covering such research areas as memory, perception, decision making, problem solving, motivation, intelligence, and neuroscience (Gobet, de Voogt, and Retschitzki 2004). According to Danesi and Mollica (1994: 13–4), recreational mental play in the form of board games can be the most memory-enhancing way in which L2 learners develop new linguistic concepts. In the ELT classroom, puzzleological techniques (i.e. board games, crosswords, word searches, etc.) are commonly used for reinforcing communicative skills as well as reviewing structural and lexical knowledge (Treher 2011). Moreover, playing board games is a social experience that can boost the development of social and emotional skills (Hromek and Roffey 2008). Additionally, since classroom engagement is partly based on peer interactions, playing board games may contribute to increasing this engagement.

Use of board games in higher education

Although we know of no research that focuses on ELT board games at higher education level, board games have been used in other higher-education contexts to teach, organize, and connect learners from different educational backgrounds (Holmes and Gee 2016). Over the last decade, the application of *game-based teaching and learning* (GBTL) in higher education has become a legitimate field of study and an accepted form of

ELT Journal Volume 73/3 July 2019; doi:10.1093/elt/ccy059

Advance Access publication February 28, 2019

university-level instruction. A case study by Cochran (2012) showed that the use of board games in the classroom can significantly improve learners' comprehension and retention. Smith (2013) asked his history students to research historical events while playing an actual board wargame, 1776 by Avalon Hill, boosting their in-class engagement by promoting a dynamic and interactive learning environment. Despite these promising results, board games are still quite uncommon in higher education. This is probably because faculty guidelines rarely mention them as standard teaching methods and not every teacher may want to experiment with them.

Board games for advanced EFL learners

Recent research on ELT games tends to focus on *digital* games (e.g. Hong, Han, Kim, and Bae 2017). However, print board games are still easier to implement in classrooms without computers or internet connectivity. The use of such tabletop board games in ELT is usually limited to grammar or vocabulary (Paris and Yussof 2012; Bakhsh 2016). The scarce research on print ELT board games focuses on young learners. This is probably because all phonetic games available on the market are aimed at either primary learners (Nixon and Tomlinson 2005) or beginner to intermediate learners of all ages (Hancock 1995, 2017). Perhaps the only category of games suitable for the more advanced adult students of English phonetics and phonology are adaptations of existing games for practising IPA symbols, e.g. the IPA versions of Scrabble and Bingo published by Cascadilla Press. Because we could not find board games dealing specifically with English prosody[1] at university level, we designed our own games.

Why board games for teaching prosody?

The decision to create board games for teaching English prosody to advanced Polish learners of English at university level was motivated by the fact that English prosody is one of the most difficult aspects of English pronunciation to teach to Polish students (Sobkowiak 2008). Over the years of teaching, our students often reported that prosody-related topics were among the most challenging ones.

Throughout our teaching, we follow Wrembel's (2007) suggestion that improving learners' performance on such challenging pronunciation topics could be achieved by increasing their metacompetence, i.e. explicitly teaching them the 'rules' of English prosody before they start their pronunciation drills. Historically, most of our learners showed little engagement in the classes focusing on prosody, so we decided to experiment with board games to see if these could help learners engage more with the topic, and, as a result, learn it better.

The study

Aim

This paper examines the usefulness of board games for learning *about* English prosody. We expect that playing board games during classes that discuss prosody will be associated with increased learners' in-class engagement and post-class assessment performance, compared to classes without board games.

Institutional context

We implemented our games in a two-semester course in English phonetics and phonology. The course is obligatory for all first-year students of English Studies at the Faculty of English, Adam Mickiewicz University, Poznan, Poland. It aims to supplement the obligatory four-semester practical pronunciation course by making students aware of how English speech sounds are produced, transcribed, and how they function in real-life situations. And since many of the students become teachers, the course also aims to help them predict and correct the pronunciation errors of their potential future students. Broadly, the first semester focuses on the phonetic description of English sounds and connected speech processes, whereas the second expands on that by introducing prosody-related topics and regional varieties of English.

The course was offered in a flipped-classroom[3] model with one 90-minute class per week. Each week covered a different topic. Before each class, the learners were required to complete an online pre-class preparation module that we created on Moodle. The module contained a short video lecture on the topic, a few close-ended ungraded activities with pre-scripted feedback, and links to one or two supplementary readings (chapters from English phonetics and phonology textbooks). The 90 minutes of classroom time was devoted to completing worksheet activities (more advanced versions of the online pre-class activities) and clarifying any confusing concepts. By the end of each course week, the learners took an online graded quiz. They also took a longer test midway through each semester and the final exam at the end of the second semester.

Game design

We created three board games that focus on three key prosody topics in our course: word stress (*Stress Run*), weak forms (*Stress Maze*), and phonotactics[2] (*Phono Tactics*). The games use components that can be printed in greyscale on A4 paper. The players only need to add dice and counters, such as coins. Each game was designed for at least two players and about 30 minutes of play. All three games are freely available in a print-and-play format at bit.ly/phongames. Due to space constraints, here we include an illustration of only one game.

Stress Run, shown in Figure 1, was designed for two or more players to review English stress patterns in compound words. In order to get from the campus to the library, players take turns clockwise to throw the dice. Players compare their dice rolls with a table that instructs them to move their counters to the nearest square with a specific stress pattern. For example, if a player rolls a two, then he or she has to move to the nearest square containing a compound in which the second element carries the primary stress (counting from the start, the first such square is ˌ*Abbey* ˈ*Road*). The player to the right checks the validity of the move in the answer key (an alphabetic index that makes it difficult to cheat). Rolling a 6 means that another player draws a challenge card with a question that the rolling player needs to answer for a bonus or a penalty.

Stress Maze helps two players to review English weak forms. Players start at opposite ends of a grid with 96 squares, each of which contains a short phrase. Some of the phrases contain a function word in its weak form, while most contain either a function word in its strong form (e.g. an auxiliary verb that occurs at the end of a phrase) or a word that does not have a weak form at all, such as ‘may’. Each player moves their counter towards the finish square located in the centre of the board by placing their counter on the nearest square containing a function word in its

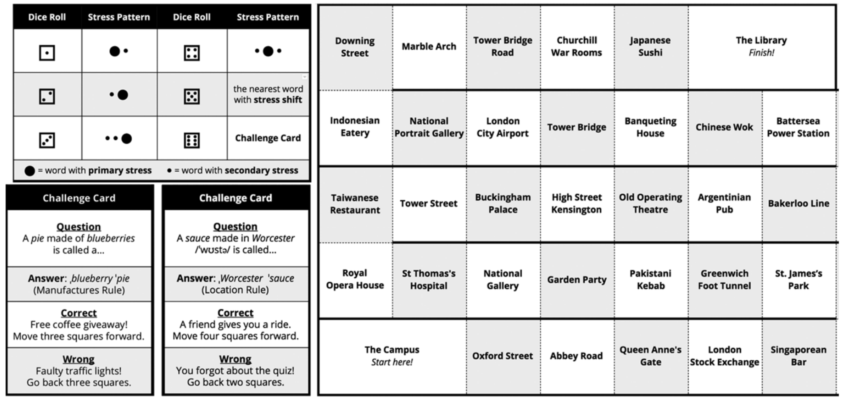

FIGURE 1
Top left: reference table for dice rolls. Right: a fragment of the board. Bottom left: two example challenge cards. Not shown are the instructions page and the answer key.

weak form. The other player confirms whether the move was correct by referring to their answer key. Each answer key was designed so that a player can only see the opponent's correct answers. After making the move, the player makes a dice roll to receive a bonus or a penalty, depending on the success or failure of their move.

Phono Tactics challenges two players to review English phonotactic constraints. Players start at opposite ends of a grid with 159 hexagonal fields. They take turns to move their counters towards the finish located in the centre of the grid by making a dice roll. A player needs to move to the nearest licit consonant cluster appearing in syllable onsets (after rolling 1 or 2), codas (after rolling 3 or 4), or onsets or codas (after rolling 5 or 6). After the first player makes a move, the other player consults their answer key to see if the move was valid.

Study design

This study focuses on 29 Polish learners of English who took our course: 25 females and 4 males (average age around 20 years old, average time spent learning English around 12 years). Although a total of 50 learners played at least one of the three board games we piloted, only those 29 played all three games, and took the final exam and the course evaluation survey. The remaining learners either dropped out of university midway through the second semester (when the games appeared) or did not attempt the final exam and the course evaluation survey. The learners belonged to three groups taught by the present authors. The groups shared the same materials and the teachers followed the same lesson plans. The classes were held on different days of the week and at different times of the day.

Each group of learners played one board game per topic. The first 45 minutes of the class were devoted to completing worksheet activities and clarifying concepts. Then, the teacher spent five minutes on distributing the game sets, and about ten minutes on reading the rules of play aloud in English, demonstrating the first couple of rounds, and answering any questions. This usually left about 30 minutes for at least one full playthrough with little teacher supervision. After the class, the learners could either take their in-class copy of the game with them or print a new copy at home.

In this observational study, the main explanatory variable is whether a topic was supplemented by a board game. The response variables are:

- **Learners' in-class engagement:** At the end of each course week, having taken the online post-class quiz, learners answered the same set of survey questions. Two of the questions are of interest here. The first one asked the learners to rate their in-class engagement on a five-point scale, where 1 meant 'Very disengaged' and 5 meant 'Very engaged' (they received longer definitions of both terms). Another question asked them to decide whether or not they think a given course topic would help them in mastering their English pronunciation.
- **Learners' performance on assessment:** These are the scores that learners received on the relevant questions from the weekly post-class quizzes, the midterm test, and the final exam. All course content and assessment was aligned in terms of learning objectives and question types. Therefore, learners' score on assessment questions for a given topic should reflect their mastery of the learning objectives for that topic.
- **Learners' perceived usefulness of games:** At the end of the course, the learners completed a course evaluation survey. Among other questions, the learners were asked to rate the usefulness of each teaching aid used in the course on a five-point scale, where 1 meant 'Not useful at all' and 5 meant 'Very useful'. They were also asked other close-ended and open-ended questions about the usefulness of the board games which we will describe in the Results section.

Additionally, we controlled for the learners' sex, their prior achievement (written and spoken secondary school final exam results), and the student group to which they belonged.

Results and discussion

Learners' self-reported in-class engagement

Figure 2 shows that learners reported a similar level of in-class engagement throughout the academic year, centred around 'somewhat engaged' (3.8 on average). There is little variation in the data: 75 per cent of all ratings fall between 3.5 and 4. For each topic in which learners played a board game, they rated their in-class engagement slightly higher than the average (Word Stress rated as 4; Weak Forms rated as 3.9; Phonotactics rated as 4). But only for the first board game topic, Word Stress, the engagement level was significantly higher than for the neighbouring topics not accompanied by board games (Connected speech processes rated as about 3.5 on average and Rhythm rated as about 3.6 on average). The other two topics with board games were rated similarly to neighbouring topics without board games (Weak Forms rated similarly to Intonation 1, and Phonotactics rated similarly to The syllable and General British vs General American).

A multiple regression analysis[4] shows that a learners' in-class engagement increases by 0.29 of a point if that learner perceives a topic as useful ($P < 0.09$), after controlling for the score on the online pre-class preparation module, the use of a board game in the class, sex, and prior achievement. Although this is a weak association, the three topics accompanied by board games were indeed similarly or more helpful in mastering English pronunciation than other course topics. Figure 3 shows that, on average, about 94 per cent of learners rated the topics covered in both semesters as helpful in mastering their English pronunciation, compared to about 93 per cent for Word Stress and 100 per cent for both Weak Forms and Phonotactics.

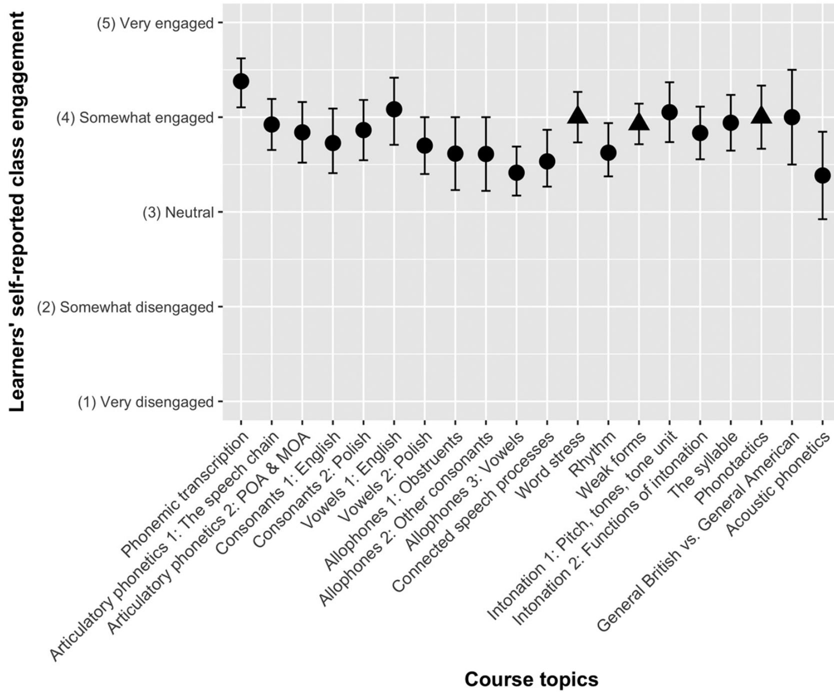

FIGURE 2
Mean in-class engagement self-reported by learners after each class (topic). Topics with board games are shown as triangles. Whiskers show 95 per cent confidence intervals.

Another explanation of the differences in in-class engagement could be more trivial, namely that learners are less engaged if a class takes place in the morning. Unsurprisingly, we observed that learners who took the class on Monday at 8 a.m. seemed less engaged than learners who took the class on Tuesday at 3 p.m. In the survey for the second-lowest rated topic (Vowels 2: Polish), one learner explained their 'somewhat disengaged' rating with a comment 'My disengagement is result of early hour. I was a bit sleepy. [The] lesson was all right.'

Although we did not systematically record the learners' interactions during the playthroughs, we observed that most learners were fully immersed in the games, as if they were regular (i.e. non-educational) games. This was probably because learners were responsible for checking each other's answers, so they needed to collaborate and stay focused from start to finish. Their high engagement may have also been influenced by switching to their native language for the playthrough. Although classes and materials for all courses in the programme are in English, some learners switch to Polish during prolonged clarifications of confusing concepts. During the playthrough, the players were required to provide each other with corrective feedback, for which they usually switched to Polish. The use of Polish may have also been prompted by the more relaxed atmosphere that more closely resembled a casual board game night than a formal class.

Learners' performance on assessment

We now move from learners' perceptions to their actual performance by looking at how playing board games is associated with scores on quizzes that learners took after each class. A multiple regression analysis shows that playing an in-class board game is associated with an increase in the expected post-class quiz score of about 8 percentage points ($P < 0.03$),

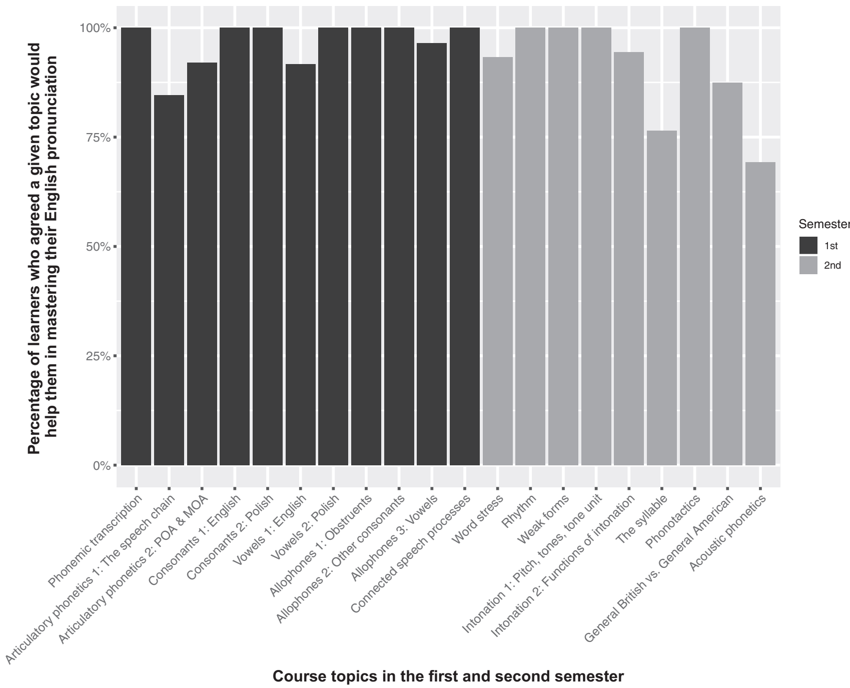

FIGURE 3
Learners' perceived usefulness of a given phonetics and phonology topic for mastering English pronunciation.

after controlling for student group, score on online pre-class preparation module, learners' perceived usefulness of a given topic, learners' self-reported in-class engagement, sex, and prior achievement.

Expectedly, three of the control variables show even stronger associations. First, we estimate an expected 0.13 percentage point increase in post-class quiz score for every 1 percentage point increase in pre-class preparation score ($P < 0.0002$). Second, we estimate an expected 5.93 percentage point increase in post-class quiz score for every 1 point increase in self-reported in-class engagement ($P < 0.0007$). Figure 4 visualizes how the association between in-class engagement and post-class quiz score was observed to increase in the presence of board games. Third, we estimate an expected 0.47 percentage point increase in post-class quiz score for every 1 percentage point increase in written (but not spoken) secondary school final exam result ($P < 0.006$). This supports our observations that learners with a better command of English seemed to have less difficulty in understanding the course material.

We built similar multiple regression models for the associations between learners' performance on (and the perception of) topics with board games and relevant tasks on the midterm test and the final exam but these results lacked both statistical and practical significance.

Learners' perceived usefulness of games

We will now look at the results of the questions asked in the course evaluation survey that learners completed at the end of the course. Figure 5 shows how learners perceived the usefulness of board games compared to other teaching aids offered throughout the course. The horizontal axis lists the teaching aids in the order they were meant to be used each week (but note the games only appeared in three weeks of the

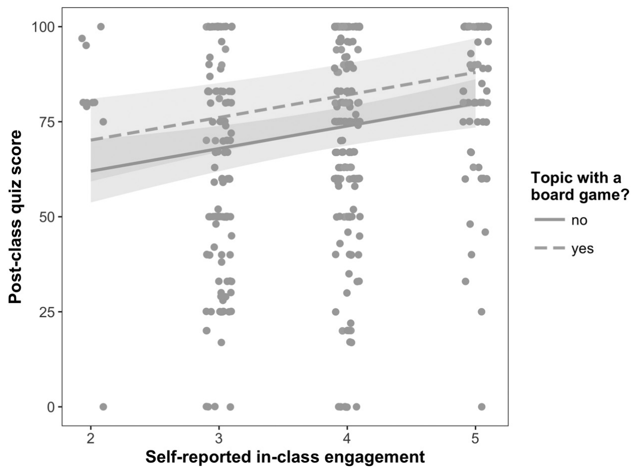

FIGURE 4
The association between in-class engagement and post-class quiz score for topics with and without a board game. Each dot represents one learner–topic pair. Shaded areas show 95 per cent confidence intervals.

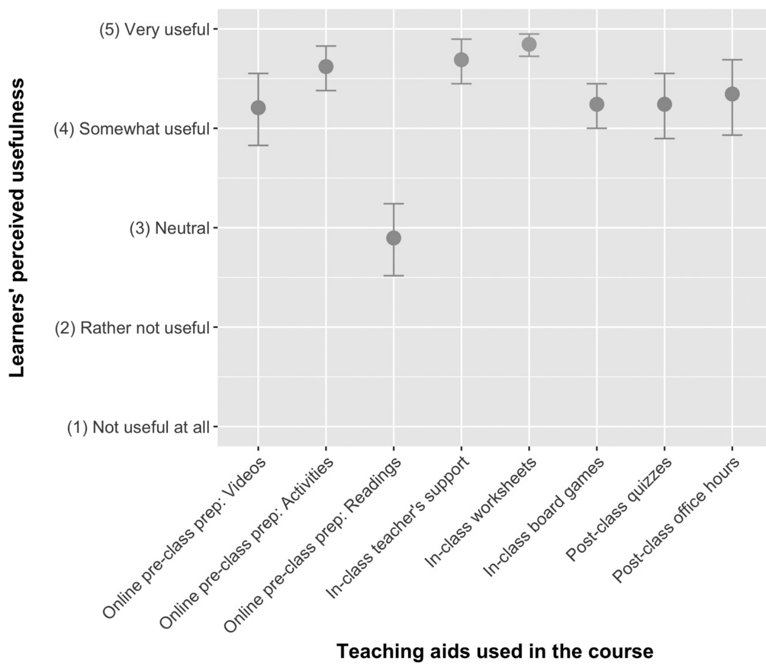

FIGURE 5
How learners rated the overall usefulness of board games compared to other teaching aids. Whiskers show 95 per cent confidence intervals.

course). The vertical axis shows the perceived usefulness of a teaching aid on a five-point scale, where 1 meant 'Not useful at all' and 5 meant 'Very useful'. On average, the learners rated all three board games collectively as about 4.2. This means that learners found the games to be, on average, as useful as pre-class video lectures and post-class quizzes (both rated as about 4.2 on average) and about 45 per cent more useful than traditional pre-class readings (about 2.9 on average) but about 13 per cent less useful than in-class worksheets.

Although learners perceive worksheets as slightly more useful in-class teaching aids, they also think board games are a good supplement. When asked if they could go back in time and choose between spending 30 minutes (about one-third of total class time) on playing the games or doing extra worksheet exercises, 62 per cent of learners said they would prefer board games over extra exercises (Figure 6).

Moreover, about 35 per cent of learners reported replaying at least one game after the class and 10 per cent of learners said they replayed at least one game while preparing with peers for the final exam (not shown). When asked to describe how exactly the board games were useful, learners said that games were '[an extra] chance to practice' that provided 'immediate feedback' and helped to 'memorize the rules' and 'remember [the learners'] mistakes' through 'fun' and 'competition'.

Conclusion

Summary of findings

To our knowledge, this is the first study on using print board games to teach English prosody to advanced EFL learners at university level. First, we showed that learners reported slightly higher in-class engagement for the three prosody topics accompanied by board games. Second, we observed a moderate increase in the expected post-class quiz score for those topics. Third, learners reported that board games can be a fun alternative to the more typically used in-class worksheets. And since these are actual games that learners can enjoy when they socialize after hours, they are also a stealthy way of introducing some extra study time in first-year students' busy schedules. We acknowledge that the study has limitations typical of an observational study conducted throughout an academic year at a national higher-education institution, especially due to learners dropping out midway through the semester.

Teaching implications

Board games for advanced EFL learners are a promising teaching tool because they provide solid instructional scaffolding that fosters collaboration and allows for precise corrective feedback. Once the teacher walks the learners through the rules of play, he or she then transfers the ownership

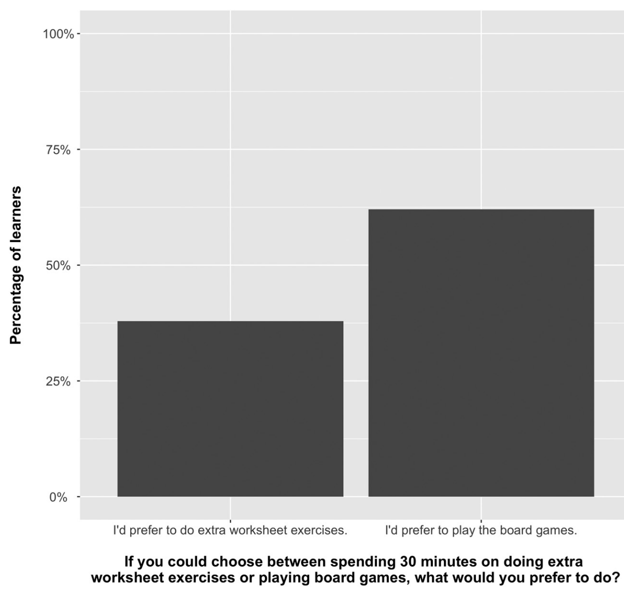

FIGURE 6
Learners' preference for using classroom time for completing extra worksheet exercises or playing board games instead.

of the learning event to the learners. The frequent turn-taking imposes a shared responsibility to pay attention to each other's choices. In our games, the learners must collaborate because each player has access only to the other player's answer key. Moreover, the answer keys allow the learners to provide each other with explicit correction, which allows the teacher to focus on observing the learning and supplementing it with metalinguistic feedback. In an example intervention, the teacher could say 'Correct, the compound ˌ*Park* ˈ*Road* has the primary stress on its final element because it is a location name. But why is ˈ*Park* ˌ*Street* stressed on the first element?' Perhaps the only major disadvantage of a board game is the time needed to assemble the game components, introduce the rules, and play the game.

We encourage teachers to experiment with implementing prosody board games in their phonetics and phonology curricula, either by printing the games we designed for this study, or by creating their own games. Designing your own game can be a valuable instructional experience in itself. In our case, it forced us to re-evaluate our assessment criteria for each topic. For those teachers who would like to design their own game, we recommend limiting the scope of the game to a single learning outcome, creating a rough prototype in up to two hours, and testing it as soon as possible with another person to spot any potential loopholes.

Future directions

The next step for ELT researchers could be to use board games for recording learners' interactions while they are playing such pronunciation games. One could even involve learners in the process of creating their own board games and ask them to record their reflections. A more ambitious opportunity is to create technology-enhanced board games with pre-scripted feedback. Although our board games proved engaging, they required teacher supervision to introduce the rules and provide feedback. In the board games industry, some publishers are now bypassing the need to consult the manual by building free mobile apps to accompany their games. In a phonetics and phonology board game, a simple companion app could replace the printed answer keys, provide instantaneous feedback to learners who keep losing points, and possibly also increase replayability by introducing modified game rules. This would bring us one step closer to what could be the holy grail of educational board games—fun games that learners can take home and continue learning from, without teacher supervision.

Acknowledgements

The authors received support from the Narodowe Centrum, Nauki; project 2014/15/N/HS2/03867.

Final version received August 2018

Notes

1 Prosody (or suprasegmental phonetics) refers to such speech features as word stress or intonation. It is a key part of pronunciation alongside segmental phonetics that focuses on single sounds.

2 Phonotactics refers to sound sequences that can occur in a syllable. For example, a Polish learner of English would pronounce *gnome* as */gnəʊm/ because /gn/ is an acceptable Polish cluster.

3 In the flipped-classroom model, learners are usually expected to watch an instructional video before coming to class, so that class time is devoted to practice and clarification.

4 This analysis is fully reproducible. The raw data and code that generated the findings are available at bit.ly/phongames_code.

References

Bakhsh, S. A. 2016. 'Using games as a tool in teaching vocabulary to young learners'. *English Language Teaching* 9/7: 120–8.

Cochran, J. J. 2012. 'Can you *really* learn basic probability by playing a sports board game?'. *American Statistician* 59/3: 266–72.

Danesi, M. and **A. Mollica**. 1994. 'Games and puzzles in the second-language classroom: a second look'. *Mosaic* 2/2: 13–22.

Gobet, F., A. de Voogt, and **J. Retschitzki**. 2004. *Moves in Mind: The Psychology of Board Games*. New York: Psychology Press.

Hancock, M. 1995. *Pronunciation Games*. Cambridge: Cambridge University Press.

Hancock, M. 2017. *PronPack 2: Pronunciation Puzzles*. Chester: Hancock McDonald ELT.

Holmes, J. B. and **E. R. Gee**. 2016. 'A framework for understanding game-based teaching and learning'. *On the Horizon* 24/1: 1–16.

Hong, J. S., D. H. Han, Y. I. Kim, and **S. J. Bae.** 2017. 'English language education online game and brain connectivity'. *ReCALL* 29/1: 3–21.

Hromek, R. and **S. Roffey**. 2008. 'Promoting social and emotional learning with games'. *Simulation and Gaming* 40/5: 626–44.

Nixon, C. and **M. Tomlinson**. 2005. *Primary Pronunciation Box: Pronunciation Games and Activities for Younger Learners*. Cambridge: Cambridge University Press.

Paris, T. N. S. T. and **R. L. Yussof**. 2012. 'Enhancing grammar using board game'. *Procedia—Social and Behavioral Sciences* 68: 213–21.

Smith, S. K. 2013. 'Pounding dice into musket balls: using wargames to teach the American revolution'. *The History Teacher* 46/4: 561–76.

Sobkowiak, W. 2008. *English Phonetics for Poles* (Third edition). Poznań: Wydawnictwo Poznańskie.

Treher, E. N. 2011. *Learning with Board Games: Tools for Learning and Retention*. Minnesota: The Learning Key, Inc.

Wrembel, M. 2007. 'Metacompetence-based approach to the teaching of L2 prosody: practical implications' in J. Trouvain and U. Gut (eds.). *Non-Native Prosody: Phonetic Description and Teaching Practice*. Berlin: Mouton de Gruyter.

The authors

Kacper Łodzikowski is a PhD student at the Faculty of English, Adam Mickiewicz University in Poznan, Poland. He is interested in educational technology and his research focuses on computer-assisted pronunciation acquisition.
Email: klodzikowski@wa.amu.edu.pl

Mateusz Jekiel is a lecturer and researcher at the Faculty of English, Adam Mickiewicz University in Poznan, Poland. His research interests include second-language acquisition, EFL pronunciation teaching, phonetics, and phonology. His current research project focuses on the role of musical hearing in the acquisition of EFL pronunciation.
Email: mjekiel@wa.amu.edu.pl

2.4. Publication 2 (Łodzikowski 2021)

Łodzikowski, Kacper. 2021. "Association between allophonic transcription tool use and phonological awareness level", *Language Learning and Technology* 25, 1: 20-30. (doi:10125/44748).

Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND
February 2021, Volume 25, Issue 1
pp. 20–30

LANGUAGE TEACHING AND TECHNOLOGY FORUM

Association between allophonic transcription tool use and phonological awareness level

Kacper Łodzikowski, Adam Mickiewicz University, Poznań

Abstract

This is the first paper that provides correlational evidence about how interacting with an online allophonic transcription tool helps learners of English as a Second Language (ESL) to improve their phonological awareness. The study investigates 55 advanced ESL learners at a Polish university enrolled in a course in English phonetics and phonology. The course placed heavy emphasis on reading and writing allophonic transcription based on the International Phonetic Alphabet. Apart from obligatory practice with traditional pen and paper worksheets, learners could also practise with a supplementary custom-designed web application that allowed them to enter the phonemic transcription of any word in order to receive its allophonic transcription. The results show that using this tool at least once during the course is associated with an expected increase in midterm test score of 5.03 percentage points, 95% CI [-10.61, 20.67]. The estimated benefit is higher for learners who space their usage of the tool; each additional distinct day of usage is associated with an additional increase in the expected midterm test score of 2.62 percentage points, 95% CI [-3.25, 8.49]. Additionally, some learners practised transcription on non-words, and these learners were observed to perform better on phonotactics-related assessment.

Keywords: *Computer Assisted Pronunciation Teaching, IPA Transcription, Learner Autonomy, Learning Analytics*

Language(s) Learned in This Study: *English*

APA Citation: Łodzikowski, K. (2021). Association between allophonic transcription tool use and phonological awareness level. *Language Learning & Technology*, *25*(1), 20–30. https://hdl.handle.net/10125/44748

Introduction

Challenges in Pronunciation Instruction

Discredited since the 1970s, pronunciation instruction in the context of English as a Second Language (ESL) acquisition has returned to favour towards the end of the century (Pennington, 2015). Over the following two decades, a growing body of research has shown which pronunciation features should be prioritised at the segmental level (e.g. Jenkins, 2000; Munro & Derwing, 2006; Gao & Weinberger, 2018) and at the suprasegmental level (e.g. Munro, 1995; Rogerson-Revell, 2012, 2014).

Despite these advancements, the teaching of pronunciation to ESL learners is still of secondary importance because the curriculum is traditionally dominated by other aspects of the English language and instructors do not have the proper training and resources (Pennington & Rogerson-Revell, 2019). This is especially a challenge for non-native ESL instructors who lack the confidence in their pronunciation skills (Bai & Yuan, 2019). And even when non-native instructors report that they are comfortable with teaching pronunciation, observations show that their teaching is rather simplistic and reactive (Buss, 2016). This is echoed by research into the instruction of other languages, such as French and Spanish (Huensch, 2018).

Role of Phonological Awareness in Facilitating Pronunciation Instruction

Traditionally, pronunciation instruction would encompass implicit exposure to speech and explicit pronunciation drills. Most recently, a new strand of research and practice has focused on supplementing that with training in phonological awareness, such as the knowledge of how the phonological system of English as the target language is different from the phonological system of the learner's native language (Wrembel, 2005). The underpinnings of this approach can be traced to theories of second-language perception, such as Best (1995), in which the central element is the interference between the two systems, and the resulting challenges in the perception, and therefore also the pronunciation, of target language sounds.

One way of developing phonological awareness is by practising transcription based on the International Phonetic Alphabet (IPA). It is a skill that helps to consolidate various aspects of declarative phonological awareness and practical pronunciation ability (Mompean, 2017). Perhaps its main benefits are that it can help learners better understand the English phonological system (in comparison to their native system) and make them more aware of issues with spelling-to-pronunciation correspondence (Mompean & Lintunen 2015). The less-detailed phonemic transcription seems to offer the best return on investment for beginner and intermediate learners who would like to work on segmental errors that impact intelligibility, for example, substituting consonantal phonemes (Gao & Weinberger, 2018). The more complex allophonic transcription is a better choice for advanced learners, such as the participants in this study, who want to work on the less salient features, such as aspiration of fortis plosives or word-initial epenthesis, in order to sound more native-like on top of being intelligible.

The link between transcription and ESL learners' phonological awareness is substantiated by research on native speakers of English in the context of communication disorders. For example, Robinson et al. (2011) showed that a pre-test of phonological awareness can predict native speakers' difficulties in learning transcription, and Werfel (2017) showed that native speakers improved their phonological awareness after completing a transcription course.

Computer-Assisted Transcription Training

Despite all its benefits, transcription is still a rare teaching technique. This is partly because many teachers do not have the time needed to prepare engaging transcription exercises. Moreover, many learners get bored by the lack of variety or frustrated by the lack of feedback (García Lecumberri, Maidment, et al., 2003). As such, transcription is a great candidate for benefitting from automation through the use of computer software.

The best example of an automated transcription tool designed for second-language acquisition is the Web Transcription Tool prototyped as a desktop application by Cooke et al. (2001). It was subsequently redesigned as a web application by García Lecumberri, Cooke, and Maidment (2003) and further improved in García Lecumberri, Maidment, et al. (2003). The tool contains phonemic transcription exercises for English, Spanish, Swedish and Romanian. It also supports selected connected speech processes and tries to provide learners with relevant feedback. The complexity of the tool requires the instructor to provide a reference transcription for each exercise. A notable mention is Jensen (2005); while it is not a transcription tool per se, it includes a variety of freely available simple transcription-related activities.

Teachers and learners who would like to use allophonic transcription instead of phonemic transcription do not have much choice. Several free transcription tools are available online, but they only offer transcribing spelling into phonemic transcription. Some of them offer a limited selection of allophonic processes, for example, checking the transcription of 'salt' would show velarisation but not pre-fortis clipping. One could again turn to the field of communication disorders, where several systems were prototyped that combine perception tasks and allophonic transcription (e.g. Bruijn et al., 2011; Bates et al., 2010). However, the former tool is not freely available, and the latter covers only selected allophonic processes. Since none of the freely available tools were appropriate for automated practice of single-word allophonic transcription, this study relies on a custom tool (Łodzikowski & Aperliński, 2016).

The Study

Purpose

This study answers two research questions. First, how do advanced ESL learners use a supplementary allophonic transcription tool? Second, how is the usage of the tool associated with learners' level of declarative phonological awareness?

Tool Design

Figure 1 shows the transcription tool used in this study. It is a simple web application written in HTML5 and hosted at a university server. The logic of the application is based on conditional statements that reflect the rules of selected allophonic processes (e.g. a sound categorised as a vowel should undergo pre-fortis clipping if followed by a sound categorised as a fortis consonant). The descriptions of those rules were obtained from English phonetics and phonology textbooks, such as Cruttenden (2014), that formed the curriculum of the present course. Understandably, these textbooks are meant to give a broad descriptive overview of English speech, so the rules may be overgeneralised.

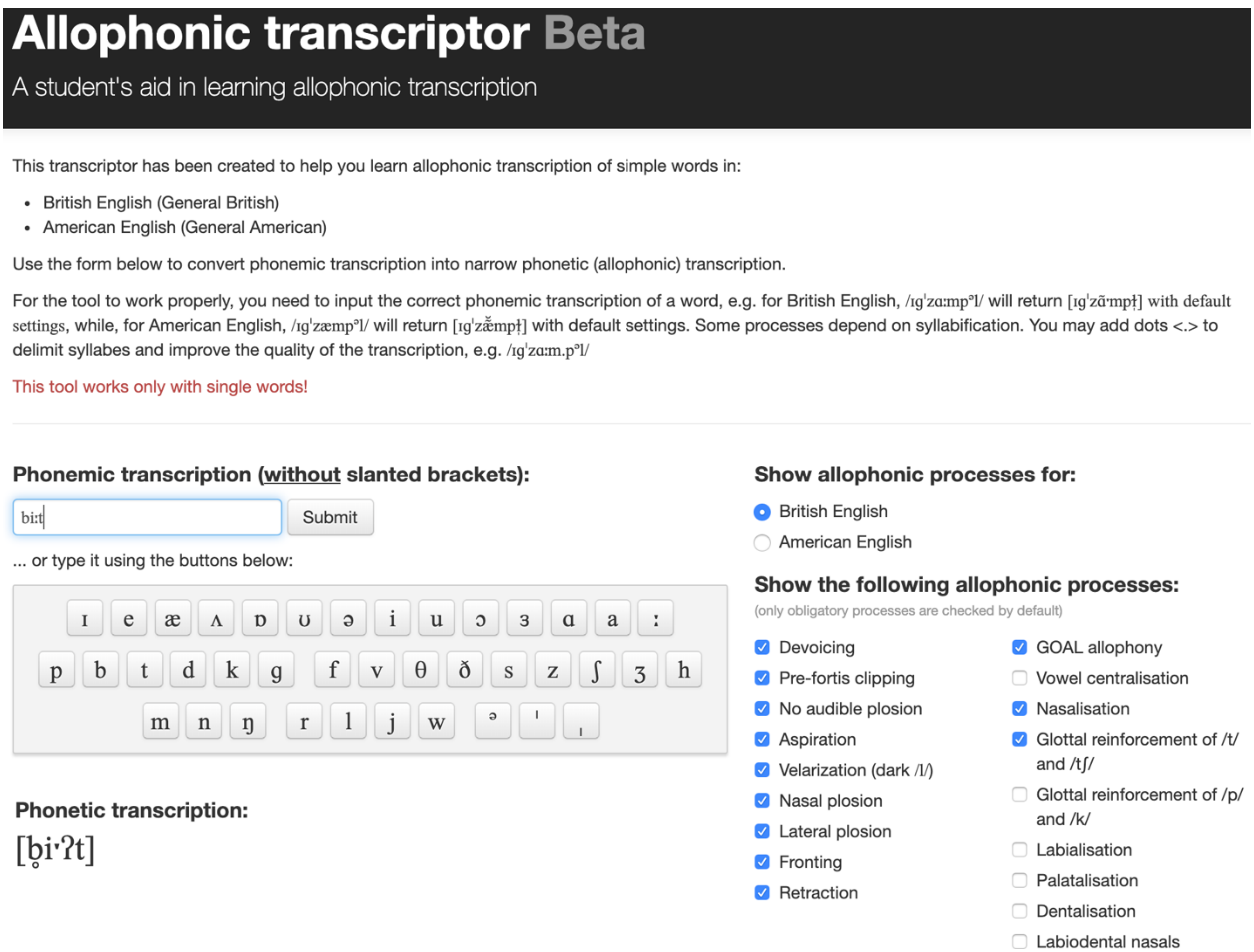

Figure 1. Example input entered into the transcription tool.

Note that the tool does not evaluate if the input is correct. If a learner enters a sequence of symbols that does not correspond to an actual word, for example <ðpt>, the tool will still provide an output that observes the phonetic rules for actual English words, for example [ð̥p̚t]. Similarly, if the learner enters the spelling form of a word, for example <neat>, the tool will incorrectly treat it as the phonemic form and provide [neaʔt] instead of [ni·ʔt].

Institutional Context

The transcription tool was implemented as part of a blended-learning course in English phonetics and phonology taken by first-year students of English Studies at a Polish university. This obligatory course, which aims to increase learners' phonological awareness, was offered alongside a four-semester practical pronunciation course that focuses on implicit exposure and drilling. The programme curriculum goes beyond intelligibility and focuses on achieving near-native pronunciation.

The English phonetics and phonology course was conducted in a flipped classroom manner (i.e. the learners were asked to watch an instructional video and complete close-ended activities before showing up to class). Each week of the course covers a different topic. The first semester focused on the speech chain, the English phonological system and allophonic variation, and the second semester focused on connected speech processes and prosody. The course started in early October and ended in mid-June. The learners received access to the transcription tool in December, when the topic of allophonic transcription was first introduced. While transcription homework was obligatory, the use of the tool was voluntary. And while the learners were encouraged to use it, no particular pedagogical guidance was given, other than the instructions in Figure 1.

While 70 learners were given access to the tool, the study focuses on the 55 learners who did not drop out before reaching the midterm test in the second semester. The learners were on average 20 years old, and 41 of them were females.

Data Collection and Cleansing

The transcription tool contains a built-in data collection mechanism: A snippet of JavaScript tracking code executed by the user's browser. Whenever someone visits the tool or clicks on the Submit button to request allophonic transcription, their browser sends a tracking message to a server running a free analytics application Piwik (Aubry, 2014). The Piwik database logs an event containing the following metadata: visit timestamp and duration, entered text (input), device information (e.g. screen resolution), and additional user information (e.g. approximate geographical information). Because the tracking code is executed by the browser, a small amount of errors is to be expected, for example, a visit is not logged if someone visits the tool and the browser window is closed because the browser crashes before the tracking script is executed.

A separate process was used to identify learners across visits. This required an orchestrated onboarding of learners, so that their first visit to the tool was made from Moodle. A link to the tool was put on the Moodle course homepage, and a message was sent to each learner encouraging them to start using the tool. When a learner clicked on the link, Moodle passed that learner's unique identification (Learner ID) to the tracking code sent to Piwik, which assigned that Learner ID to its own unique tracking identification (Visitor ID), and then asked the browser to store that Visitor ID in a cookie. As a result, even if that learner made subsequent visits to the transcription tool directly (e.g. by bookmarking the address), he or she had the same Visitor ID. If the learner cleared browser cookies, the relationship was re-established the next time the learner visited the tool from Moodle. In the meantime, Piwik guessed the visitor by device fingerprint (e.g. device type, operating system, screen resolution, browser type and plugins, IP address, etc). The risk of cookie loss means this data collection method is not infallible. However, it provides a good balance of reliability and ease of use, compared to adding an extra sign in screen. Moreover, an exploratory analysis of the log data did not reveal any patterns suggesting that learners were misidentified. For example, everyone who accessed the tool from a new device did it from Moodle, which established the relationship between Learner ID and Visitor ID, and which created the device fingerprint. Furthermore, in the accompanying demographic survey, all learners stated that they owned a personal computer or mobile device, so it is unlikely they used a public computer for homework. It is similarly unlikely that they frequently cleared browser cookies because that would require them to repeatedly sign into Moodle and possibly other web services like social media or email. One unavoidable scenario of misidentified transcriptions would be if one learner shared a personal device with another learner, for example during classroom pair work.

Expectedly, the online log data collected for this study required some pre-processing. Based on summary statistics, a number of outlier entries were identified and examined. For example, 16 visits with legitimate inputs showed a duration of 0 seconds. After a careful inspection, these tokens were interpreted as legitimate, and they were included in the analysis. Other outliers were entries that contained phrases instead of single words. This is probably because some learners wanted to use the transcription tool for practising connected speech processes, which it does not yet support. These entries were excluded from the analysis. The data cleansing and subsequent modelling was done in R (R Core Team, 2018) and RStudio (RStudio Team, 2016). The data and code that generated the findings are available at bit.ly/phontrans_analysis. The source code for the transcription tool is available at bit.ly/phontrans_webapp. Please note that deploying the app to a server requires software engineering skills, so most teachers will need to ask their school IT administrator for assistance.

Findings

Transcription Study Patterns

Of learners who were given access to the transcription tool, 91% visited it at least once. Of those learners, about 51% visited it five or more times, and about 16% more than ten times. The learners entered a total of 3,119 inputs over the course of 312 visits. The median number of inputs per visit was four (1^{st} quartile = one; 3^{rd} quartile = 12). The median duration of a visit was about 8 minutes (1^{st} quartile about 1 minute; 3^{rd} quartile about 23 minutes).

Figure 2 shows the distribution of visits across the period from when the transcription tool was made available midway through the first semester (December) until the end of the second semester (June). Overlaid are the dates of selected assessments relevant to this study. In the first three post-class quizzes and the midterm test, phonemic transcriptions of words were provided, and learners needed to mark the relevant allophonic processes. The activity type and rubric were the same as in classroom practice, but the examples were different. The midterm contained additional transcription activities (e.g. the learners needed to choose the correct allophonic transcription out of three provided—example: ‘crude’ [kɹ̥uːd̥], *[kʰɹuˑd̥], *[kɹ̥ʉˑd̥])—or the learners needed to correct the provided phonemic transcription of a word—example: ‘throughout’ given as */θruwˈɑʊt/). The words did not appear in previous practice or quizzes. The final assessment relevant to this study is the post-class quiz on phonotactics. Learners needed to decide if five non-words were acceptable from the point of view of English phonotactics, and justify their decision (e.g. /skrɔː/ is a licit onset because it occurs in a word such as ‘screen’)

By examining Figure 2 from the left, we see spikes in visits to the tool in December and January, around the dates of the three quizzes. We then see that learners continued practising with the tool between late February and early March, even though the topics covered during that period (connected speech processes and weak forms) required only phonemic transcription. This could be because allophonic practice helped the learners understand the connected speech concepts or, more likely, because they wanted to continue practising allophonic transcription before the midterm test in April. The midterm was also preceded by a spike in visits. Most of the longer periods of inactivity are due to holidays.

Figure 3 shows that about 9% of visits were made from mobile devices (of which just one was from a tablet), and that these usually occurred during the day. Most likely, this activity was generated by learners who used the tool in the classroom. Desktop visits were made mainly in the evening, and some learners studied well into the night. We do not show another interesting segmentation, namely that some learners visited the tool relatively regularly, while others visited it only a couple of days before assessment.

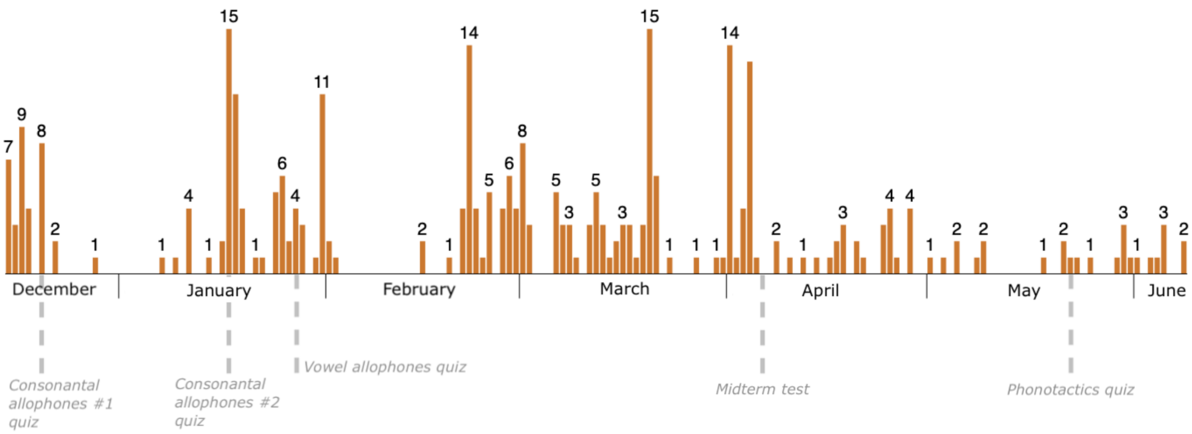

Figure 2. Course-wide distribution of learners' 312 visits to the transcription tool. Note that one learner could make more than one visit per day, so the first bar on the left represents a total of seven visits made by a total of six learners (not shown).

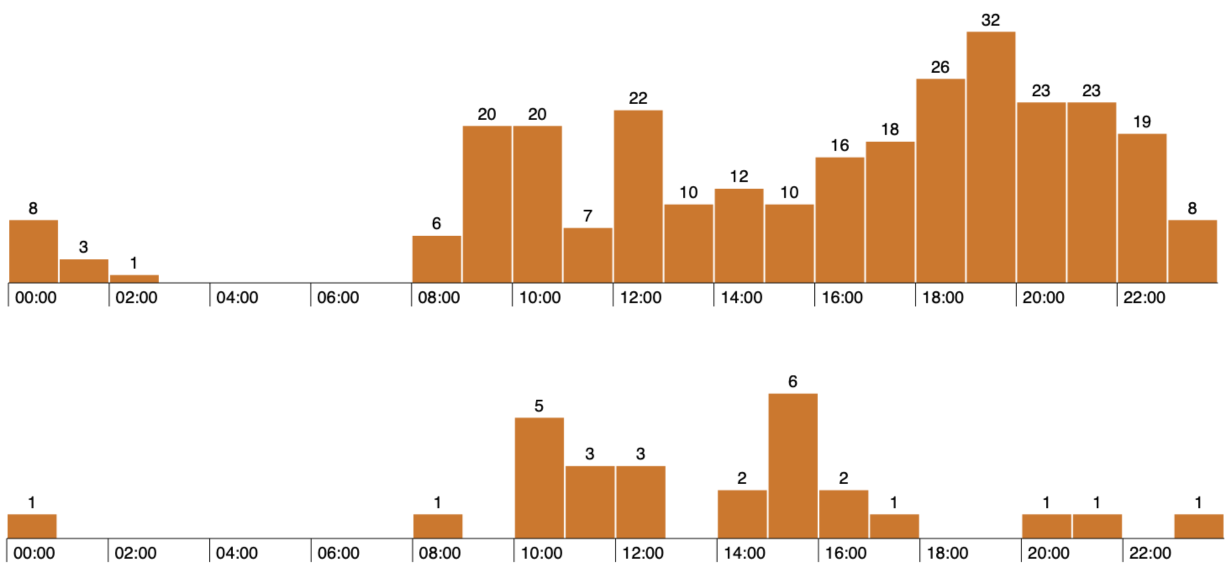

Figure 3. Distribution of times of day at which learners visited the transcription tool from desktop browsers (top) and mobile device browsers (bottom). For clarity, the figure shows all 312 visits overlaid on one chart, i.e. as if they occurred on one day.

Of the 3,119 inputs, 68 were non-words written in phonemic form. The remaining inputs were mainly words written in their phonemic forms, and some words written in their spelling forms. Words in their phonemic forms were manually mapped to their spelling forms, so that summary statistics could be calculated. This was done for two reasons. First, some words were over-represented because learners entered the same word multiple times, either by accident or to see how the resulting allophonic transcription looks with different settings enabled (e.g. with or without glottal reinforcement of /t/ and /tʃ/). Second, some words were written using the standard (Gimson's) IPA scheme for English (so that 'bet' is transcribed as /bet/) and some using a modified (Upton's) IPA scheme (so that 'bet' is transcribed as /bɛt/). While the former scheme is required of the learners in the phonetics course, some of the materials used by the learners in the practical pronunciation course may have used the latter scheme. The mapping to spelling showed that out of the 3,051 word inputs, 1,105 were distinct words.

The 15 most frequently entered words account for 10.29% of all entered words. These are: 'potential' (1.57%), 'cute' (1.21%), 'twelfth' (0.72%), 'grandchild' (0.69%), 'alcohol' (0.66%), 'bead' (0.66%), 'love' (0.59%), 'be' (0.56%), 'beat' (0.56%), 'guilty' (0.56%), 'spoilt' (0.52%), 'supermarket' (0.52%), 'pat' (0.49%), 'rescue' (0.49%), 'try' (0.49%). Many of the 50 most frequently entered words appeared on in-class transcription worksheets. On the one hand, this is a positive surprise because it shows that the learners used the tool for the reason it was designed (to supplement out-of-class practice in the phonetics course). On the other hand, after each class the learners always received the worksheet answer key anyway, so using the transcription tool mainly to check answers to worksheets is a rather limited use.

Association Between Tool Use and Phonological Awareness Level

We will construct a linear regression model that predicts a learner's midterm test score based on that learner's usage of the transcription tool. Observations following the test date (April) were excluded. The control variables included sex and prior achievement (scores from written and oral Matura, i.e. secondary school final exam). Initially, group was added as a random variable but then it was removed because it did not explain any more variance. Figure 4 visualises the model's fixed effects. While all associations are rather weak ($R^2 = 0.56$), three of them are worth noting. First, learners who visited the transcription tool at least once were observed to score higher on the midterm test by 5.03 percentage points, 95% confidence intervals [-10.61, 20.67], after controlling for sex and prior achievement. Second, while the total number of visits to the tool showed, unexpectedly, a slightly negative association with midterm scores, it seems that the spacing of the visits was more important. This was measured by looking at distinct days of visit, e.g. visiting the tool once on Monday and once on Tuesday yields a higher number of distinct visit days than visiting it twice on Monday. We see that each additional distinct visit day is associated with an increase in the expected midterm test score of 2.62 percentage points, 95% CI [-3.25, 8.49]. Third, each additional second spent during an average visit is associated with an increase in the expected midterm test score by 0.01 percentage points, 95% CI [-0.00, 0.01]. While this may seem small, note the average visit duration was 14 minutes and 34 seconds. Another explanation for the rather weak model fit is that some learners used the transcription tool in the expected way (entering legitimate transcriptions from worksheet examples) but did not show the expected improvement because their overall course performance was at the bottom quartile of the class, possibly due to their overall low ability or aptitude.

We will now discuss one unexpected finding. The analysis of the transcription tool logs showed a creative use of the tool, namely entering non-words. The context for this is that when the learners first started studying allophonic processes, they were advised to practise marking those processes on non-words. The reason for that was that such practice could help reinforce the learners' intuition about the fact that the presence of a phonetic process depends on the phonetic composition of a word (e.g. a consonant cluster will likely involve a change in how the consonants are released). It is unclear, however, why some learners wanted to transcribe the non-words. The transcription tool is agnostic of phonotactic constraints, and it does not give any feedback regarding licit and illicit onsets and codas. Most of the non-words came from the practice worksheet for the phonotactics course topic.

To further investigate this behaviour, a linear model similar to the one in the previous paragraph was built, with the difference that the outcome variable is the phonotactics quiz score, and a new variable was added indicating whether a learner entered three or more non-words over the course of using the transcription tool. The threshold was chosen arbitrarily based on the distribution of the data. While this model showed a weak fit ($R^2 = 0.3$), it predicted an increase in the phonotactics quiz score of 31.15 percentage points, 95% CI [-13.06, 75.36], for those learners who entered a total of three or more non-words. This could mean that those learners who obtain an understanding of allophonic processes have a more intuitive understanding of which combinations of sounds are licit and which are not. This incidental finding is important in the light of Gao and Weinberger (2018) who showed that syllable-level errors such as illicit elision of plosive in phrases such as ‘ask her’ are an important source of accented speech.

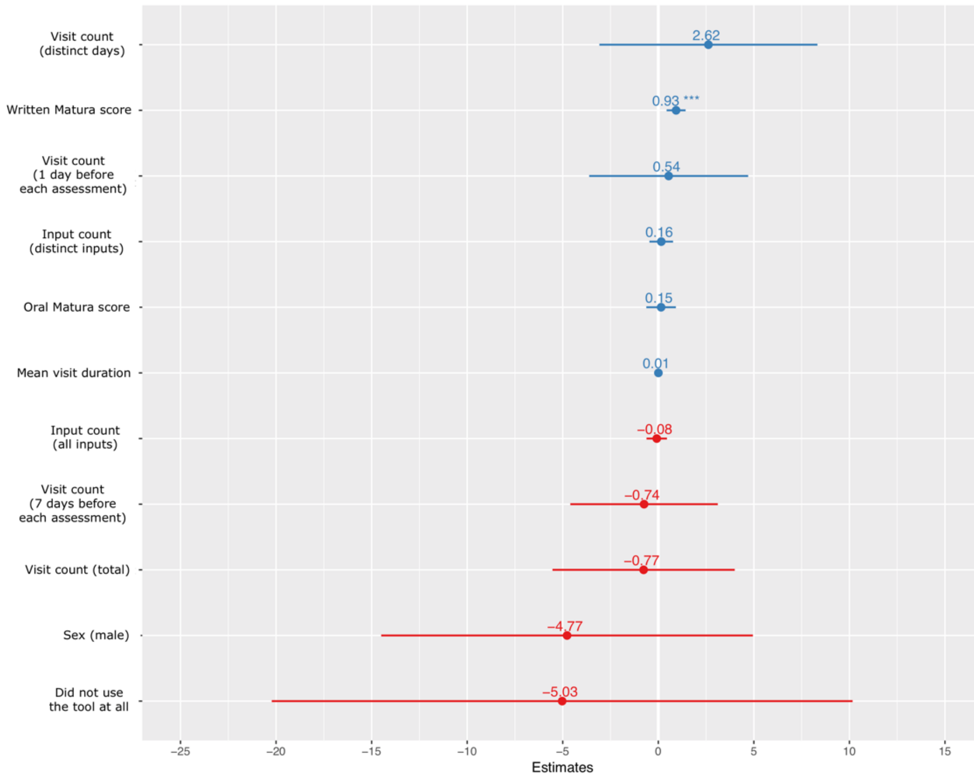

Figure 4. The association between midterm test score and transcription tool usage. Whiskers show 95% confidence intervals. The bottom fixed effect shows the estimated decrease in midterm test score for learners who did not use the tool.

Conclusion

Summary of Findings

This work is a practical investigation of ideas posed by previous researchers who pointed to studying IPA-based transcription as one of the best ways to improve learners’ phonological awareness. Undergraduate students of English studies at a Polish university were given access to a custom-designed IPA-based allophonic transcription tool to supplement a two-semester course in English phonetics and phonology. Based on a linear regression model, for the 91% of learners who visited the transcription tool at least once, we estimate an expected increase in the midterm test score of 5.03 percentage point, 95% CI [-10.61, 20.67]. Moreover, we observed that the total number of visits to the tool is less important than the self-regulated spacing of those visits; each additional distinct day with a visit was associated with an increase in the expected midterm test score of 2.62 percentage points, 95% CI [-3.25, 8.49). Additionally, we saw that some learners used the transcription tool for entering non-words, which could help them in grasping the rules of English phonotactics, although this association is weak and needs further investigation. Due to the limitations of this observational study, the claims presented here are associative and directional rather than causal and definitive.

Implications

While the transcription tool described here was borne out of necessity, this research was conducted out of curiosity, and it is reported here in the spirit of highlighting the role of IPA-based transcription in ESL pronunciation acquisition. It is hoped this study will inspire instructors to implement transcription exercises in their curricula, even if they are in the form of simple paper and pencil activities. In fact, instructors willing to start this practice do not need to create time-consuming worksheets with answer keys. The simplest solution is to seek example worksheets online or in dedicated textbooks such as Tench (2011), or to find sources of annotated authentic speech in such corpora such as The Speech Accent Archive (Weinberger, 2015).

Those instructors who would like to leverage such automated transcription tools as those described in this study are encouraged to review the References section. Some solutions are publicly available and can be easily implemented in an existing course. At the moment, the allophonic transcription tool described here is not available publicly. However, the code repository is available to anyone who would like to host their own instance of the application, or to modify it.

Regardless of whether the instructor chooses manual or automated transcription practice, the effort will be worth it. As Pennington and Rogerson-Revell (2019 p. 202) said:

> Learners and teachers need to be aware that developing pronunciation skills, from individual sounds to discourse-level intonation patterns, is a gradual process of acquisition involving all of these subskills, rather than just correcting the odd individual pronunciation error in an isolated listen-and-repeat session. The ultimate aim is for learners to be able to recognize and correct their own errors rather than rely on the teacher to do so, thus developing learner autonomy.

IPA-based transcription has the potential to increase such autonomy by equipping learners with a framework that facilitates the identification and correction of errors.

Future Research and Practice

Regarding future research and development of such transcription tools, a welcome addition to this and similar tools would be simple ear-training activities along the lines of those proposed by Ashby et al. (2009). A good example of how these ideas could be implemented is the freely available WebFon web application (Bates et al., 2010), which allows its users to listen to authentic recordings of speakers with developmental speech disorders and juxtapose them against transcriptions.

Once such online transcription tools become more popular, instructors should be able to leverage the power of data to inform their decisions. Just like with learning management systems such as Moodle, the immediate use case for tracking real-time learner performance data is identifying learners who struggle with a given part of the material (in this case, particular phonetic processes). The long-term use case is the ability to inform curriculum design by reviewing aggregated data on the most common errors and (a)typical usage patterns.

Acknowledgements

The author received support from the National Science Centre, Poland; project 2014/15/N/HS2/03867. Thank you to Przemysław Kaszubski and Michał Remiszewski for enabling the data collection effort.

References

Ashby, M., Yanagisawa, K., Kim, Y. S., Maidment, J., Przedlacka, J. (2009, August 6–8). *Achieving interactivity in online learning of phonetic skills* [Paper presentation]. Phonetics Teaching and Learning Conference: PTLC 2009, London, United Kingdom. https://www.phon.ucl.ac.uk/ptlc/ptlc2009/ptlc2009-proceedings/ptlc2009_ashby_015_ed.pdf

Aubry, M. (2014). Piwik (Version 2.17) [Computer software]. https://matomo.org

Bai, B., & Yuan, R. (2019). EFL teachers' beliefs and practices about pronunciation teaching. *ELT Journal*, *73*(2), 134–143. https://doi.org/10.1093/elt/ccy040

Bates, S., Matthews, B., & Eagles, A. (2010). *Phonetic transcription self-study programme*. Webfon. http://elearning.marjon.ac.uk/ptsp/

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 171-204). York Press.

Bruijn, C. G., Nunes, J. M., Fang, L., Pathak, R., & Zhou, J. (2011). A system for independent e-learning of practical phonetics. In L. Wai-Sum & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 368–371). Hong Kong: City University of Hong Kong.

Buss, L. (2016). Beliefs and practices of Brazilian EFL teachers regarding pronunciation. *Language Teaching Research*, *20*(5), 619–637. https://doi.org/10.1177/1362168815574145

Cooke, M. P., García Lecumberri, M. L., & Maidment, J. A. (2001). A tool for automatic feedback on phonemic transcription. In P. Dalsgaard, B. Lindberg, H. Benner, & Z. Tan (Eds.), *Proceedings: Eurospeech 2001, Scandinavia, 7th European Conference on Speech Communication and Technology* (pp. 2795–2798). Aalborg.

Cruttenden, A. (2014) *Gimson's pronunciation of English* (8th ed.). Arnold.

Gao, Z., & Weinberger, S. (2018). Which phonetic features should pronunciation instructions focus on? An evaluation on the accentedness of segmental/syllable errors in L2 speech. *Research in Language*, *16*(2), 135–154. https://doi.org/10.2478/rela-2018-0012

García Lecumberri, M. L., Cooke, M. P., & Maidment, J. A. (2003). Transcripción fonémica en Internet. In L. Ruiz Miyares, C. E. Álvarez Moreno, & M. R. Álvarez Silva (Eds.), *VIII Simposio Internacional de Comunicacion Social, Actas II* (pp. 1043–1047). Santiago: Centro de Lingüística Aplicada, Ministerio Ciencia.

García Lecumberri, M. L., Maidment, J., Cooke, M., Ericsson, A., & Giurgiu, M. (2003). A web-based transcription tool. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 981–984). Barcelona: Universitat Autònoma de Barcelona.

Huensch A. (2018) The pronunciation teaching practices of university-level graduate teaching assistants of French and Spanish introductory language courses. *Foreign Language Annals*, *52*(1), 13–31. https://doi.org/10.1111/flan.12372

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Jensen, C. (2005, July 27–30). *Online training and testing in phonetics* [Paper presentation]. Phonetics Teaching and Learning Conference: PTLC 2005, London, United Kingdom.

Łodzikowski, K, & Aperliński, G. (2016, December 1–3). *Usage patterns of an online allophonic transcriptor* [Paper presentation]. 10th International Conference on Native and Non-native Accents of English, Łódź, Poland.

Mompean, J. A. (2017). Doing phonetic transcription in a modern language degree. In J. A. Cutillas Espinosa, J. M. Hernández Campoy, R. M. Manchón Ruiz, & F. Mena Martínez (Eds.), *Estudios de Filología Inglesa: Homenaje a D. Rafael Monroy* (pp. 479–505). Editum.

Mompean, J. A., & Lintunen, P. (2015). Phonetic notation in foreign language teaching and learning: Potential advantages and learners' views. *Research in Language*, *13*(3), 292–314. https://doi.org/10.1515/rela-2015-0026

Munro, M. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, *17*(1), 17–34. https://doi.org/10.1017/S0272263100013735

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System: An International Journal of Educational Technology and Applied Linguistics*, *34*(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

Pennington, M. C. (2015). Research, theory and practice in L2 phonology: A review and directions for the future. In: J. A. Mompean & J. Fouz-González (Eds.), *Investigating English pronunciation* (pp. 149–173). Palgrave Macmillan, London. https://doi.org/10.1057/9781137509437_7

Pennington, M. C., & Rogerson-Revell, P. (2019). *English pronunciation teaching and research: Contemporary perspectives*. Palgrave Macmillan, London. https://doi.org/10.1057/978-1-137-47677-7_4

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.2;) [Computer software]. Vienna: R Foundation for Statistical Computing. Retrieved from www.r-project.org

Robinson, G. C., Mahurin, S. L., & Justus, B. (2011). Predicting difficulties in learning phonetic transcription: Phonemic awareness screening for beginning speech-language pathology students. *Contemporary Issues in Communication Science and Disorders*, *38*, 87–95. https://doi.org/10.1044/cicsd_38_S_87

Rogerson-Revell, P. (2012). Can or should we teach intonation? *Speak out! IATEFL Pronunciation SIG Newsletter*, *47,* 16–20.

Rogerson-Revell, P. (2014). Pronunciation matters: Using English for international business communication. In R. van den Doel & L. Rupp (Eds.), *Pronunciation matters: Accents of English in the Netherlands and elsewhere* (pp. 137–159). VU Uitgeverij.

RStudio Team. (2016). RStudio: Integrated development for R (Version 1.1.463;) [Computer software]. Boston, MA: RStudio, Inc. www.rstudio.com

Tench, P. (2011). *Transcribing the sound of English: A phonetics workbook for words and discourse*. Cambridge University Press. https://doi.org/10.1017/CBO9780511698361.001

Weinberger, S. (2015). *Speech Accent Archive*. George Mason University. http://accent.gmu.edu

Werfel, K. L. (2017). Phonetic transcription training improves adults' explicit phonemic awareness: Evidence from undergraduate students. *Communication Disorders Quarterly*, *39*(1), 281–287. https://doi.org/10.1177/1525740117702456

Wrembel, M. (2005). Metacompetence-oriented model of phonological acquisition: Implications for the teaching and learning of second language pronunciation. In J. A. Maidment (Ed.), *Proceedings of the Phonetics Teaching and Learning Conference 2005* (pp. 1–5). London: University College London.

About the Author

Kacper Łodzikowski's academic and professional interest is in educational technology. His academic research focuses on computer-assisted language learning (esp. pronunciation acquisition) and educational data mining. His professional work focuses on artificial intelligence in education, especially natural language processing, including automated speech recognition, natural language understanding, and natural language generation.

Email: klodziko@amu.edu.pl

2.5. Publication 3 (Łodzikowski et al. 2024)

Łodzikowski, Kacper, Peter W. Foltz and John T. Behrens. 2024. "Generative AI and Its Educational Implications", in: Dora Kourkoulou, Anastasia Tzirides, Bill Cope, and Mary Kalantzis (eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines*. Cham: Springer. (doi:10.1007/978-3-031-64487-0_2).

Chapter 2
Generative AI and Its Educational Implications

Kacper Łodzikowski, Peter W. Foltz, and John T. Behrens

2.1 Introduction

Over the last 50 years, each decade can roughly align with the application of transformational technology that dramatically impacted daily life across societies. The 1970s brought the semiconductor, and the 80s the personal computer. The 1990s brought the World Wide Web, the commercial use of the Internet, open-source software sharing models, and open standards for technology diffusion. The 2000s brought online search, e-commerce, and the scaling of the ideas and tools from the previous decade, fueling the evolution of the passive consumption-oriented Web 1.0 into the increasingly interactive and participatory Web 2.0. The 2010s witnessed the proliferation of mobile devices and the growth of social media platforms, both of which contributed to the unprecedented accumulation of digital data. That decade also brought dramatic advances in data science and computing concerning the application of Artificial Intelligence (AI) for prediction and classification. This has led to a broad range of commercial applications of AI, such as virtual assistants enhanced by natural language processing (e.g., Amazon's Alexa), nearly autonomously driving cars, as well as decision support systems for healthcare, finance, and many other industries.

K. Łodzikowski (✉)
Adam Mickiewicz University, Poznań, Poland
e-mail: klodziko@amu.edu.pl

P. W. Foltz
University of Colorado, Boulder, CO, USA
e-mail: Peter.Foltz@Colorado.edu

J. T. Behrens
University of Notre Dame, Notre Dame, IN, USA
e-mail: jbehrens@nd.edu

D. Kourkoulou et al. (eds.), *Trust and Inclusion in AI-Mediated Education*, Postdigital Science and Education ,
https://doi.org/10.1007/978-3-031-64487-0_2

Writing this chapter in the spring of 2023, we believe we are on the cusp of the next wave of transformational technology that, like previous waves, appears both fantastical and naturally progressing. We have little doubt that when the 2020s are characterized in the future, they will be described as the 'decade of AI'. While AI has been evolving for decades, the field has taken a significant leap forward in the last year with the general availability of generative AI (GAI). This new area of AI is a collection of technologies in which computer systems use AI techniques and large amounts of data to generate texts, images, sounds, videos, or their combination. The most well-known product from this class of technologies is ChatGPT, which became a near overnight sensation upon its release on the last day of November 2022 (Brockman 2022). It reached 100 million monthly active users in just 2 months (Hu 2023). Nearly simultaneously, its underlying functionality was implemented in a wide range of technology products produced by Microsoft with similar functionality soon followed by Google and others. While less widely covered, technical capabilities for image generation from text instructions have also made dramatic improvements. Tools such as Midjourney allow users to generate photorealistic images that are sometimes impossible to distinguish from actual photographs, creating economic opportunities and social challenges.

This current wave of GAI tools is different from prior waves in a number of practical ways. First, because it is building on top of prior technological waves, the rate of functional improvement is dramatically faster than we have seen in the past. The number of new articles referring to ChatGPT in the open-access paper repository arXiv grew from 25 in January 2023 to 772 in June 2023. This growth is enabled by the ability of researchers to (1) easily access state-of-the-art AI tools, (2) communicate with others widely and rapidly regarding findings, and (3) benefit from prior advances in open-source software characteristics, including common libraries, languages, and systems for sharing (e.g., the computer code repository GitHub). While we believe the impact of GAI as an underlying platform for many activities will match that of the Internet in the next decade, the rollout and impact will be dramatically faster as the AI evolution benefits from the existence of the Internet and related technologies, which the Internet itself could not benefit from.

Second, this wave supports the specialized behavior of text generation (i.e., writing) and image generation (i.e., visual communication) in ways previously only attributable to humans and useful in the daily life of a great swath of society. Historically, automation technologies have largely affected blue-collar workers, especially in such industries as agriculture, manufacturing, and administration—and this was also the case with computerisation in the second half of the twentieth century (Frey and Osborne 2017). And while the past two decades of AI advancements have not negatively impacted the overall job market due to AI adding some jobs on top of replacing others (Handel 2022), GAI is different in that it is also expected to also impact white-collar jobs by automating such activities as sales and marketing content creation, customer service, or software development (McKinsey 2023).

Third, this widespread impact brings numerous social conflicts and confusions between the activities of humans and machines that we have not seen before. For example, while autonomous vehicles have been available for years, they are experienced by few and their impact is not perceived widely, not least due to their cost or regulation. However, the ability for virtually everyone to use freely-available, unregulated AI tools to generate writing or images whose provenance may be indistinguishable from human artifacts upends numerous social expectations and norms. This has already led to discussion about academic integrity (Cotton et al. 2023) and AI literacy (Anders 2023), confusion over the use of some tools in highly sensitive legal situations (Weiser 2023), and even concerns about the long-term impact on modern societies writ large (Lukpat 2023).

In light of these rapid and significant changes, this chapter aims to provide the reader with an overview of text-related GAI technologies in four sections. First, we provide an overview of how AI relevant to education has evolved and a gentle introduction to how current technologies work. In the second section, we discuss how such systems can be, and are being, applied in learning contexts, followed by a third section in which we note a number of the larger societal issues that will impact AI in education. We conclude the chapter with recommendations for educational researchers. While GAI is a class of software that includes a broad range of systems including text-to-text generation, text-to-image generation, text-to-video and other combinations, in this paper we focus on text-text generation (such as ChatGPT) as that is the area with which most readers will be familiar and which, we anticipate, will have the broadest base of use in education in the near future.

2.2 Understanding Language-Based GAI Systems

To help the reader appreciate the complexities of GAI systems, we start with an overview of how the design of AI systems has evolved over the years. Since its birth after World War II, the field of AI has seen several cycles of growth and stagnation, often described in terms of technological breakthroughs and funding for commercial and academic endeavors (see e.g., Russell and Norvig 2022).

2.2.1 Previous-Generation AI: From Rule-Based to Data-Driven

* See note on preprint error below.

An early example of this approach is MYCIN (Shortliffe 1976), a dialogue-based system designed to aid physicians in diagnosing and treating bacterial infections. By posing close-ended questions to the physician, MYCIN simulated human expert decision-making using explicit rules (e.g., ‘If the patient is febrile, apply drug X’) and experience-based heuristics (e.g., ‘If the blood test shows X and Y, then it is moderately suggestive that the bacteria is Z’). Although such expert systems performed satisfactorily in well-defined domains and straightforward cases, they

Note: In this preprint version, section 2.2.1 is missing the first sentence from the submitted manuscript, which reads “The first wave of AI systems, between the 1950s and the late 1980s, relied on domain experts to manually encode knowledge into a set of rules for logic-based decision-making.” The Editorial team will correct this for the published version of record.

struggled with complex real-world scenarios that fell outside their predefined rules and included elements of uncertainty. Moreover, the labor-intensive process of encoding the vast knowledge of human experts into hundreds of rules was a bottleneck.

In the realm of education, this wave introduced intelligent tutoring systems that could dynamically track student knowledge, apply contextual tutoring strategies, and provide scaffolded support (e.g., Anderson et al. 1995; Sleeman and Brown 1982). For example, a variant of MYCIN called GUIDON (Clancey 1984) could engage a student in a mock dialogue about a patient's condition and give feedback on the chosen treatment. The system was proficient in remembering and applying rules and could understand and analyze the learner's input within the boundaries of its hand-programmed knowledge. However, it struggled with evaluating ambiguous cases and was unable to create novel solutions not encoded in its database.

In the late 1980s, the second wave of AI introduced a shift from rule-based systems to data-driven machine learning systems. In this approach, the system identifies meaningful patterns in historical data and uses those patterns to generate rules for automated decision-making in the future. The role of the domain expert shifts from specifying the logical beliefs of experts to collecting relevant real-life data and pre-selecting (or 'engineering') data features (also called 'variables' or 'attributes' in the social sciences) likely to predict the outcome in question. Then, an algorithm ingests the features and goes through a cycle of 'learning' (also known as 'training') to produce a model which is used for prediction (also known as 'inference').

In the education space, this wave brought commercial-grade systems for automated formative and summative assessment. For example, the Intelligent Essay Assessor (Foltz et al. 1999) removed the burden of manual essay grading from teachers and provided learners with just-in-time feedback. Compared to a first-wave AI system, such a system was more robust in analyzing and evaluating learners' work products. This is because it no longer relied on manually crafted expert rules that could not cover the wealth of real-life situations.

The third wave of AI (from around 2011) introduced deep learning as a subfield of machine learning. Deep learning systems learn from data without requiring explicit feature engineering by domain experts. For example, to recognize handwriting, a deep learning system only needs a sufficiently large number of labeled pictures representing each handwritten character. This is made possible by a family of algorithms known as artificial neural networks, which are roughly inspired by the interconnected neurons in the human brain. In the 2010s, the capabilities of these deep learning models catalyzed the development of consumer-grade AI systems, such as automated labeling of photos on social networking platforms, speech recognition on mobile devices, and automated translation across multiple languages. As these AI-enabled tools became widely available, they propelled AI into the public's consciousness. Within the education industry, a new breed of self-study AI companions incorporated deep learning to offer more natural ways for the human to interact with the machine. For example, using Aida Calculus (Pearson Education 2019), learners could take pictures of their handwritten math problems to get step-by-step feedback on the process rather than just the final answer.

2.2.2 Educational Applications of Previous-Generation AI

At this point, we can see that attempts at applying AI in education are not new. Previous-generation AI-based educational systems have been deployed for large-scale assessment, classroom and individual tutoring systems, and teacher support. Studies have shown that intelligent tutoring systems can raise student test scores 0.66 standard deviations over conventional classroom training and be as effective as expert tutors (e.g., D'Mello and Graesser 2023; Kulik and Fletcher 2016; VanLehn 2011). AI-based automated essay scoring has been used operationally since the early 2000s to grade high-stakes exams as well as provide students with instant feedback (e.g., Yan et al. 2020).

While successful, however, these AI-based educational approaches have had strong limitations. First, they have often been rule-based or trained on specific topics which can be inflexible and difficult to adapt to new situations. Second, they had limited natural language processing capabilities, resulting in more stilted, non-human-like language interactions. This has also limited systems to focusing less on higher-order thinking skills such as reasoning, argumentation, or collaboration. Third, while there has been research on multimodal processing, few incorporated modalities such as eye-gaze, gestures, facial reactions, or emotion detection, which can provide a deeper understanding of the learning context. Finally, automated educational systems have been expensive to build, often requiring collecting hundreds or thousands of hours of student interaction data to train models for specific domains or have needed content experts to code knowledge and design the interactions. Thus, while much has already been achieved in AI-driven educational systems, it is important to consider what GAI enables that can accelerate the advancement.

2.2.3 Pathway to GAI

The current wave of GAI (originating around 2017) can be considered an extension of deep learning. For the past two decades, the AI research community has demonstrated that the quality of machine learning models, including deep learning models, tends to improve with their 'scale', defined as the amount of data available for model training and the computational resources required to process that data (e.g., Halevy et al. 2009; Sun et al. 2017). However, researchers soon encountered a time, cost, and quality bottleneck in the form of data collection and labeling. For example, building a neural network for detecting toxic social network posts required a dedicated data team to meticulously label thousands of historical posts as toxic or non-toxic, so that the AI could mine those examples for patterns and create generalized rules for classifying future posts. This forced a shift in approach: instead of relying on painstakingly curated datasets, researchers began experimenting with large, unstructured, and unlabeled datasets. They quickly turned to Internet-derived text, such as web pages, online encyclopedias, discussion forums, or digital books. This coincided with the development of a new neural network architecture called the transformer (Vaswani et al. 2017), which allowed for faster processing of large text files. Coupled together, these two developments paved the way for the 'large' language models we see today.

At the training stage, the goal of a language model is to find patterns in texts in order to learn language patterns. In the 2000s, a language model could predict and correct words typed in a text message. In the 2010s, a deep learning-based language model could use its capability to predict language to generate grammatically correct passages in the style of Shakespeare plays (Karpathy 2015)—albeit coherence degraded with longer generated texts. A breakthrough came with Generative Pre-trained Transformer-2, or GPT-2 (Radford et al. 2019), which was able to not only generate plausible language but also perform a wide range of tasks, such as document summarization, question answering, or translation. We should note that, even though GPT-2 could carry out a surprisingly wide array of tasks, its performance did not exceed that of humans or specialized AI systems of the time. For example, its cross-language translations were of lower quality than those of a specialized AI-based translation system developed for years using dedicated translation datasets. However, it demonstrated the feasibility of moving away from the established practice of developing multiple narrow-domain language processing systems towards a new paradigm of developing a single general-purpose system, or a 'foundation model'.

GPT-2 was followed by GPT-3 (Brown et al. 2020), which approximated human-level performance on certain tasks and which was used, with modifications, in ChatGPT (OpenAI 2022). In the spring of 2023, ChatGPT was upgraded with GPT-4 (OpenAI et al. 2023), which exceeded human-level performance on certain tasks (Bubeck et al. 2023), including beating specialized translation systems (Jiao et al. 2023).

2.2.4 Capabilities of Current-Generation Large Language Models

While the capabilities of present-day large language models are still being explored, there are two main characteristics that distinguish them from previous-generation language models. The first is that they learn tasks from training data without supervision, that is, without humans specifying the things the models should learn apart from their basic goal to learn to generate sentences word-by-word. This unsupervised task learning capability emerged because the Internet is a treasure trove of real-life task demonstrations, and feeding a sufficiently large amount of such data into a transformer model allows it to learn not only the structure of language(s), but also the characteristics of the featured tasks.

For example, if the training dataset includes text sources that feature the same sentences in English and French (e.g., language learning textbooks, fan translation websites, multilingual versions of the same document), then the language model will not only learn how to generate plausible English and French, but also what humans mean when they ask for a 'translation'. The unsupervised nature of training means that these models have learned to pick up more nuanced features of human language uses and contexts manifested in the data, such as sarcasm, sentiment, or cultural references. Consequently, current-generation language models not only complete tasks, but do so in a more human-like and contextually-appropriate way than ever before.

The second foundational capability is a human-like ability to learn how to perform tasks according to text-based commands, or 'prompts'. For example, while training an earlier-generation AI model to identify toxic social media posts would have required showing it thousands of examples, one now can provide a large language model with just a few example posts (real or fabricated) and it should be able to classify future posts based on those examples. This capability emerged because large language models were trained on increasingly larger and more varied text datasets, which allowed them to create representations of human reasoning and behavior demonstrated in the data, including learning from instructions.

A trained model holds information that allows the production of language writ large, along with information on how to carry out tasks. When this general capability is combined with new, more specific information, such as examples of toxic posts, the system behaves in a manner that looks as if the system has integrated the principles of broad reasoning with specific information. The user can provide supporting instructions on how to execute the task and provide the model with feedback on how it should adjust its outputs. This practice of 'prompt engineering' enables people without extensive programming skills or computational resources to effectively 'program' their own copy of an AI model. In a way, this brings us back to the era of humans encoding their knowledge and preferences into AI systems, albeit not through manual programming of handcrafted rules, but by providing instructions and examples in natural, everyday language.

2.2.5 *From Research to Application*

As the general-purpose capabilities of large language models grew, AI developers began to adapt them to more specific uses, such as chatbots. For example, ChatGPT (OpenAI 2022) does not use the base GPT-3 model alone, but combines it with additional software and modified model layers focused on conversational interactions. The modification was needed because GPT-3, while capable, is not fully aligned to societal expectations due to its propensity for bias, toxicity, and misinformation (e.g. Lucy and Bamman 2021; Weidinger et al. 2022). This is because the model does not actually understand the text it generates in the same way that humans do—it mainly mimics previously seen texts. The performance of the model varies significantly according to the type of task given and its relationship to the training data. When it receives a prompt about a topic that was only briefly mentioned in its training data, it may generate text that is not aligned with reality. In contexts where creativity or novelty are valued, this may be a valuable characteristic. In contexts where facts are involved, however, such errors are called 'hallucinations' and are typically disregarded—though it is incumbent upon the human end-user to make the distinction. Accordingly, in many contexts the systems require targeted fine tuning to teach the system the most important information until it reaches the necessary threshold of accuracy or sustained human (or other computing agents) in the loop for risk mitigation.

To align models with human expectations of task performance, AI developers employ a few key techniques at the intersection of computer science and data science. One such technique is instruction tuning, which involves training the model on smaller curated dataset that consist of prompts and corresponding desired outputs, such as examples of how to give helpful and safe relationship advice (e.g., Zhou et al. 2023). Another technique is reinforcement learning from human feedback (Ouyang et al. 2022). In this approach, the model is trained to adapt its responses based on feedback it receives over time. For example, if the model provides a response that is factually incorrect or inappropriate, it can be corrected, and that correction is factored into its future interactions. A more fundamental approach to minimizing harmful model behavior is to train it on a dataset of higher quality, such as websites known for factual accuracy (e.g., Touvron et al. 2023).

Other strands of research focus on finding strategies for interacting with models to reliably obtain truthful outputs (e.g., Bubeck et al. 2023). For example, when consulting a generative chatbot for answering fact-based multiple-choice questions, it may be more beneficial to ask the model to explain its reasoning before trusting its correct answer choice (Bowman et al. 2022). Another emerging practice involves prompt engineering templates, that is, proven strategies that maximize the chance of obtaining desired outputs, such as asking the model to generate probing questions until it collects adequate information to deliver a relevant response (White et al. 2023). The field is also witnessing developments that enhance the capability of GAI systems to retrieve information from trusted sources. For example, systems such as Toolformer (Schick et al. 2023) can call upon other models or databases to provide factual information or perform complex calculations.

As we look forward, the horizon of large language model research is widening beyond just natural language. Models trained on computer code datasets, such as Codex (Chen et al. 2021), can be prompted to generate entire computer programs. Multimodal models, such as Kosmos-1 (Huang et al. 2023) or GPT-4 (OpenAI et al. 2023), can process and generate more than one type of data. For example, the user can upload an image and the model can describe it. This broadens the ways in which they can understand and interact with the world, addressing one of the fundamental concerns behind text-only language models, namely that their understanding of reality is only grounded in what can be represented in text form.

2.3 Opportunities and Applications

2.3.1 Interaction and Assessment in Education

To examine where advancements can be made from GAI, we consider two key facets that comprise education and work together to create an effective educational experience: *interaction* and *assessment*. Much of education comprises a multiway multi-modal *interaction* between learners and agents (e.g., other learners, instructors, or responsive educational artifacts such as ITSs). An agent can be conceived as a 'system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future' (Franklin and Graesser 1997). For example, in a learner-instructor face-to-face dialogue, an instructor can question the learner, dynamically adapt their responses to the level of the learner, and provide visual, auditory, gestural, and emotive responses. Thus, the agent can sense the state of a learner and choose responses that can be most effective for impacting a student's learning.

On the other hand, much of education is an interaction between learners and static materials (e.g., books, manuals, web pages). For example, a book is carefully crafted by the author so that each paragraph follows coherently from the next with an organized structure that is oriented to providing new information at a rate that can be absorbed by the reader within their zone of proximal development (e.g., Vygotsky 1978). However, it does not adapt itself to differing learning contexts or learner levels.

In order to be effective in interacting with a learner, an agent must be able to perform an *assessment*. Assessment in education means being able to infer attributes of the learner through observation of their performances and activities in natural or controlled contexts (e.g., Behrens and DiCerbo 2014). It, therefore, provides the means for an agent to sense the environment (e.g., the state of the learner in relation to the learning situation) and guide how best to act upon the learner's state. Assessment is critical for evaluating learning, guiding instruction, knowing when to provide feedback, tracking progress, as well as measuring accountability of educational systems. Within digital environments, assessments can be embedded and integrated as part of the natural learning experience.

Furthermore, assessments deriving from a variety of digital experiences can be combined to make inferences about student ability over longer time frames (e.g., DiCerbo and Behrens 2012). To accomplish this, the types of assessments must be aligned with the tasks being performed by the learners. While multiple choice and fill-in-the-blank type assessment items have been widely used and are easy to automatically score within digital learning environments, they often reduce the complexity of the activity to match the scoring format rather than considering how richer inferences can be extracted from more complex tasks and responses.

Analyzing information from complex performances, such as speaking, writing, and the logging of process data is difficult for both humans and computers. However, automating these assessments enables the integration of more complex performances within learning environments (e.g., Behrens et al. 2019). Over the past 30 years, there has been great advancement in applying AI for assessing writing, analyzing spontaneous speech within tutoring contexts, and mining process data (see Koedinger et al. 2015; Yan et al. 2020; Zechner and Evanini 2019 for reviews). These advances have allowed the development of more interactive learning systems in which the assessments are embedded as part of the performance. These systems include interactive dialogue-based tutoring, automated assessment of writing with instant formative feedback, and tracking and feedback on teams performing collaborative tasks. Yet, the AI-based assessment techniques that are used typically require collecting a large number of samples of student performance, hand-scoring them, and then using machine learning techniques to train an AI model to learn to score them automatically. This limitation has confined the applicability of AI to areas where data collection is straightforward, interactions can be hand-designed, and human coders can easily characterize performance.

The advent of GAI, however, promises to greatly transform assessment methodologies, addressing many of the limitations currently faced in the field. Whereas automating assessment has required handcrafted models and training data, now, with its broad domain knowledge and ability to generate learning experiences through prompting, GAI can be easily implemented by teachers and developers without advanced AI training and can be used in many domains. For one, when provided with spoken or written language from a learner as input, it can characterize multiple qualities of a learner's language and cognitive abilities. It can also integrate multimodal data, such as speaking, writing, facial emotions allowing for a more personalized understanding of a student's strengths and weaknesses.

Additionally, GAI can be instructed to adhere to a particular rubric, providing an objective and standardized means of evaluation through prompt engineering. By writing a carefully crafted prompt, an educational designer can instruct the AI to assess consistently, thereby reducing inconsistencies in grading that may arise due to limitations in human assessment capabilities, such as the need for training, and the requirement of continued human attention. Furthermore, assessment through GAI not only provides a measurement but can also give meaningful explanations for each assessment, fostering understanding and transparency in the evaluation process. Lastly, different types of language models can be applied across different written and spoken languages as well as software code, making it a versatile tool in multilingual education environments as well as for learning programming skills.

2.3.2 GAI for Complex Performances

The advent of GAI presents an opportunity to overcome the above-mentioned hurdles and provide agency for interactivity and assessment in educational technologies. The nature of the prior training of GAI means that the automation of digital interactive learning experiences does not have to be as hand-crafted or developed through collection of large amounts of prior training data that is specific to the contexts. Thus, the foundation models in GAI provide a means to jump off into new educational innovations, much in the same way that the Internet suddenly allowed data interchange, which resulted in many new forms of knowledge sharing which have become the primary means for communicating and collaborating. For example, prior generations of AI question/answering systems would have required painstaking training of the dialog system with numerous specific examples of acceptable or partially acceptable responses. The new large language models come pre-built with that language assessment functionality built in, thereby greatly accelerating the speed of development for many new systems.

Interactivity in digital environments will change through GAI allowing the creation of more engaging, realistic learning experiences. First, as an agent, AI can take on roles that are much more like human-human interactions (see Office of Educational Technology 2023). While learners have formerly mostly written and made simple click responses with online systems, AI will allow in-depth interactive conversations through speaking and drawing, with the system able to respond conversationally with the dialogue adapted to the appropriate level of knowledge and language ability of the learner. Second, these agents further have the ability to assume different roles, such as a mentor, tutor, coach, peer teammate, as a student that needs teaching, or as an embedded simulation (e.g., Mollick and Mollick 2023). Each role may be optimized for different learning situations. For example, learners working with an AI teammate on a collaborative problem solving task can learn strategies such as how to construct shared knowledge and maintain team functions (see Graesser et al. 2018). By participating as a teammate, the AI-agent can both support the team of learners by serving as an example, but also monitor and adapt its responses to help improve the functioning of the team.

Third, GAI has the ability to generate information on the fly that is adapted to the learner's needs. Rather than choosing a static textbook that is written at the level of the learner, a learner can choose to read about a topic and interact with a system that generates content adapted to the learner. For a learner that needs to study mitosis in biology, an AI system can generate text explanations adapted to the student's background knowledge and reading level. Moreover, it can generate images, movies and simulations to further explicate the examples. It can further respond to various forms of communication including spoken language, written texts, and even facial emotions to continually adapt based on how well the learner is grasping the material.

Thus, GAI opens the field of education to novel approaches to creating learning assessment contexts, evaluating the quality of responses and generating contextually appropriate feedback. We summarize this potential in Table 2.1, showing different types of multimodal language models, how they can provide interactive and/or assessment, the kinds of educational methods that can be applied, and potential educational applications that can result from them.

Table 2.1 A selection of multimodal language model types and their potential applications in education

Language model type	Interactivity and assessment	Sample methods	Sample application
Text-to-text	Create	Instructional material generated on the fly adjusted to learner level	Personalized textbooks
		Generate contextual assessment activities e.g., multiple choice questions	Practice items with difficulty adjusted to student learning level
	Evaluate and feedback	Compare student written response to domain content and generate feedback	Writing practice for content areas with instant formative feedback
	Create and adapt	Act as roleplay participant, adapting character based on student prompts	Dynamic learning environments that facilitate integrated development of critical analysis skills
Text-to-code	Create	Generate contextual assessment activities, e.g., logically-correct computer code with syntax errors	Software debugging practice
Code-to-text	Evaluate and feedback	Assess quality of student computer code and convert to description of errors	Instant assessment and deep conceptual feedback and training
Text-to-image	Create materials in new modality	Generate illustrations/diagrams based on textual descriptions	Visual aids for complex theoretical concepts
Image-to-text	Evaluate activity in new modality	Recognize handwritten math to provide step-by-step feedback	Pinpoint diagnosis of gaps for remediation
Text-to-speech	Create materials in new modality	Generate speech from automatically generated training materials	Interactive speech-based tutors for content domains
Speech-to-text	Evaluate activity in new modality	Interpret quality and accuracy of speech signal	Interactive dialogue for language learning and practice

2.3.3 Towards More Engaging Real-World Learning Experiences

By combining interactivity and assessment, GAI enables more engaging, natural learning experiences with a higher level of fidelity in measuring learner performance. This advancement not only facilitates a deeper understanding of students' abilities but also opens avenues for a broader range of real-world experiences through realistic simulations and embedded games. Moreover, the integration of GAI holds great potential to increase the relevance of training in schools and workforce development programs. It can cater to a variety of higher-order thinking skills and domains, such as coding, critical thinking through writing, and teamwork, particularly in genres where learners have limited experience. For instance, GAI can support learners in tasks such as crafting a compelling argument after reading multiple documents, effectively collaborating with team members, or simulating realistic interview scenarios. By immersing students in these practical performances, GAI fosters skill acquisition and prepares learners for real-world challenges.

The approach also changes how we conceive of assessment and allows us to move towards a model of continuous assessment and learning. Instead of treating the educational process as a set of separated learning experiences and summative assessments, all assessments are embedded in the learning activities with real-time feedback. These kinds of instantaneous feedback loops using AI have proved advantageous for learning higher-order skills. For instance, in learning to write in content domains, AI-based feedback on learners' content knowledge and writing skills allows learners to iterate with the computer to refine their essays before submitting them to teachers. This has resulted in faster learning of the domain knowledge and writing skills as well as providing a function for assessing thousands of drafts and alerting the teacher to students who are struggling (e.g., Foltz et al. 2013).

AI-based tools open new models of education for both students and teachers. For students, they can iterate with agents, practicing and learning. Teachers can rely more on formative assessment loops where they are still the guides of the learning process, directing when and how the AI will be used, but able to be continually informed about the state of student learning, and able to intervene and engage with students. As such, the goal of AI in the classroom is not to replace the teacher, but to empower them with tools that increase their effectiveness.

2.4 Challenges to Implementing, Deploying, and Using GAI-Based Educational Tools

2.4.1 Choosing the Right Tool for the Job

The application of large language models in education is a nascent field, and many aspects of their behavior have yet to be sufficiently explored. The availability of large language models such as the GPT family and the ease of prompt engineering allow people to rapidly develop systems for assessment and learning, such as chatbots. However, just because such a system is built on a GAI model that has proven effective for some online tasks does not mean that the approach will be effective as an educational tool. As more and more companies release new models, there arises a concern about determining which one is the most effective for a particular educational experience. This is because the models differ in fundamental assumptions, such as the quantity and quality of training data or any additional alignment of the model's capability to human expectations in a given context and domain. Another set of considerations involves the learning context, the specific needs and preferences of the students, and the objectives of the course or program. Therefore, there are a number of challenges that still need to be addressed for implementing, deploying, and using GAI-based educational tools at scale.

2.4.2 Data Bias and Design Transparency

One of the principal concerns of using AI-based models is that they reflect the data they are trained upon. The quality of the data that is used in training is crucial. As machine learning models learn from data, any inaccuracies, omissions, or biases within the data can be reproduced and magnified in the AI's behavior. Bias in AI can take many forms and can result in unfair or inequitable outcomes (e.g., Baker and Hawn 2022). For example, if algorithms are not trained on a diverse set of student responses, they may disproportionately penalize or reward certain ways of communicating, thinking, or problem-solving (e.g., Kwako et al. 2023). This can have different kinds of effects on students from various cultural, linguistic, or socio-economic backgrounds.

However, ensuring the accuracy, diversity, and breadth of data is a significant challenge. As of today, educational technology developers do not have control over how large language models are trained or what kind of data was used in the training. Moreover, most model providers are not transparent regarding the design of their systems and do not provide guarantees against bias. This is a major concern in educational contexts, which require fair and equal opportunities for all students. Addressing this challenge requires continuous efforts in bias detection and mitigation in both the data and algorithms used in AI systems. Indeed, developers of educational systems may still need to test and certify their systems across wide ranges of inputs to assure that biases are mitigated, or at least are known so that the system is only used in contexts for which the models are appropriate.

2.4.3 Algorithmic Explainability and Propensity for Misinformation

Assuring the quality of GAI systems and their outputs is another significant challenge. As AI often functions as a 'black box', it is difficult to understand precisely why a model is making a certain decision or prediction. This is particularly problematic when we consider the psychometric properties of an assessment, especially validity and reliability. Traditional methods of evaluating those characteristics may not directly apply to GAI systems. Most likely, new methods are being and will have to be developed that consider changes in the assessments and the nature of the learning environments (e.g., Liu et al. 2024; von Davier et al. 2021).

The issue of model quality is related to another key challenge, namely the potential for large language models to 'hallucinate', that is, generate information that seems plausible but is incorrect or misleading. This can be especially harmful in educational contexts, where accuracy of information is paramount. And while hallucination seems to be caused primarily by the word-by-word nature of text generated by transformer models, addressing it systemically requires not only better models but also robust systems for verifying and validating AI outputs.

2.4.4 *Introducing and Maintaining Standards for GAI in Education*

These challenges illustrate that GAI-based systems cannot be deployed on learners without a significant amount of research, testing, validation, and human oversight. Indeed, the field will need to internally police itself with standards that espouse transparency and explainability around the methods. This includes being open about how the models were developed, tested and validated, and providing information on their intended use and limitations in their educational context (e.g., Mitchell et al. 2019). Concurrently, the field will need to continually incorporate external guidance to help steer ethics in this field, such as the European Union's ethical guidelines on the use of AI and data in teaching and learning and education (European Commission 2022). In certain areas of the world, such guidelines will be reinforced by formal transparency regulations (e.g., European Parliament 2023).

While AI may provide greater autonomy for learners and instructors, it need not take human instructors out of the educational process. It may change how they interact with learners and computational systems. It may change how they select educational material or structure their courses. It may change the kinds of information they receive about learners and allow them to focus more on those learners in need. But humans will still play a critical role in orchestrating how the AI is applied to best impact learners. Indeed, we see that our future world will require a human-AI partnership in which each provides their specialized capabilities and collaborate, resulting in something more educationally effective than either working individually (e.g., Hellman et al. 2019). Thus, we advocate for a human-in-the-loop approach throughout the development and use of AI-based systems in education, as it is the human teacher who will act as the ultimate regulator.

2.5 Challenges for Educational Ecosystems

In the previous sections, we have primarily focused on technological advances and their impacts on providing educational support through technology. In this section, we aim to highlight a few areas in which education itself will experience second-order effects stemming from the larger societal changes that these technologies will bring, independent of their educational applications. We will briefly discuss this in relation to the evolution of the workforce and daily life activities, the transformation of modes of communication and its impact on issues of trust, and finally, the evolution of social norms. These represent significant societal changes that will undoubtedly have a profound impact on education on a grand scale.

2.5.1 New Work, New Curricula

Perhaps the clearest signal from the new AI is that digital technologies can now perform many tasks that only humans could do just a very short time ago. For example, systems such as ChatGPT can write quality computer code at a scale that is already changing the landscape and best practices of software development. When large language models are fine-tuned for specific tasks, such as writing computer code or answering legal questions, their outputs are often sufficient as first-draft work products to be inserted into production workflows. While the prior phases of AI and related technologies such as robotics replaced mainly certain physical tasks, such as elements of the automobile assembly process, the current wave is squarely focused on language-based, and therefore cognitive, activity. In fact, higher-order thinking skills required for such jobs as accounting may be more impacted by AI in the foreseeable future than sensorimotor skills required for such jobs as housekeeping.

This is backed up by historical data—despite their proliferation, autonomous robotic vacuum cleaners have not impacted housekeeping jobs (Handel 2022)—as well as the challenges of building AI-powered robots that will walk, sense, and act with the same dexterity as humans (Deranty and Corbin 2022). Even though predictions vary on how much cognitive work will be augmented by GAI versus replaced by it, it is clear that the workforce of the future will have to master GAI tools. They will also need to evolve their skills and foci, so that they complement rather than compete with the new capabilities of AI.

In the education realm, there will be many impacts at the administrative levels. First, curricula must change to prepare students for a rapidly evolving world. Otherwise, we risk preparing the next generation for a world that no longer exists. At present, many educational institutions are not well positioned to evolve their curriculum quickly, typically having well established curricula and faculty incentivized for long-term concerns. This leads to a second concern, that the manner in which faculty operate will need to change with regard to both the production of research and the conduct of instruction. In the same way that the shift from print to digital representation of knowledge via the Internet led to a sea change in the speed and quality of academic research, we expect to see a similar explosion of productivity. In instruction, not only must the curricula change, but also the modes of instruction must, and will change, as we discussed above. While tools to support this will unfold over the next few years, we are likely to see a period of disconnect where digitally-native students are out of sync with the understandings and practices of less technically-oriented faculty.

2.5.2 New Modes of Communication and Trust

Aside from the pragmatic issues of workforce evolution and appropriate skill development, the availability of GAI presents a fundamental challenge. Until now, while we have seen an ongoing increase in automation and the use of information technologies, it has always been, by and large, straightforward to distinguish between human-generated and machine-generated products. For instance, in the realm of images, we have a long history of understanding print and digital images as having a high verisimilitude to physical reality because photographs were designed for that purpose. However, current image generation technology, available to everyone through the Internet, can generate photo realistic images that are extremely difficult to distinguish from photographs. This suggests that the epistemic and social assumptions we bring to interpreting images need to be rethought. For example, while anticipating Donald Trump's indictment in 2023, a journalist used a widely-available AI tool to generate images imagining Trump's arrest (Belanger 2023). These images were widely reposted, often with viewers believing they were photographs, rather than computer-generated images. Similarly, in July of 2022, a generated picture of the Pentagon on fire was passed around the Internet with some attributing a drop in the stock market to the perceived 'news' (Polus 2023).

There is a fundamental issue that whereas earlier we could assume with high (though not perfect) confidence that images reflected physical realities, that assumption can no longer be held without question. We expect to see similar issues in text generation. Among the many interesting qualities of large language models are their ability to generate or regenerate text 'in the voice of…'. For example, ChatGPT produces the following opening sentences when asked to rewrite the previous paragraph in the voice of Thomas Jefferson:

> In the realm of workforce evolution and the cultivation of suitable skills, there arises a profound quandary concerning the advent of GAI. Hitherto, while witnessing the continuous rise of automation and the integration of information technologies, we have generally been able to discriminate with ease between human-crafted endeavors and those wrought by machines.

While this is an interesting linguistic and historical exercise that may have curricular implications for historical analysis, it could also lead to widespread fakery and political misinformation for historical figures as well as for current events. This is a watershed moment in how our societies will understand and react to attribution and provenance going forward, and how educational institutions will evolve to support them.

2.5.3 The Collaboration Boundary and Social Norms

The rapid introduction of GAI into daily routines raises numerous questions about appropriate use which are not always straightforward to answer. As the fundamental differences blur between what humans and computers are capable of, fundamental questions of attribution and provenance are raised as well. At present, people are generally not required to cite the version of a grammar checker they use to manipulate the text of a passage for increased clarity. However, the additional cognitive-like functionality of large language models requires a more precise language about what 'what I have done' and 'what the computer has done', which is not evident at present.

There is clarity on extremes, such as when an AI system writes an entire paper, or when it is used only as a research tool similar to searching the Internet. But in the middle, for example when the software has synthesized ideas or provided novel formulations, should the software be cited as a co-author or a tool in the same way statistical software may be cited in a quantitative analysis? We do not yet have answers to such questions but advise patience and generosity. There will be many perceived social transgressions and mistakes while social and professional societies evolve their understandings and practices.

2.6 Conclusion

We are at an inflection point in the relationship between computers and humans that has only been previously suggested in science fiction, with both utopian and dystopian implications. The behavioral capabilities of large language models and other forms of GAI are evolving so rapidly that the technical leaders in the field frequently express surprise at these systems' behaviors; a fact that led some to sign a letter requesting a slow down in GAI-related product development (Bengio 2023; Seetharaman 2023). Whether educational researchers are focused on the utopian or dystopian implications, building or using the technologies, or focusing on the social and ethical critiques, these technologies are impacting our societies and educational systems and must be actively engaged. Several recommendations follow.

First, educational researchers should start using the freely-available text-to-text generation tools as part of their ongoing personal or professional activities. Both the use of these tools and the tools themselves are rapidly evolving as products. For example, in March of 2023, ChatGPT was updated with the latest large language model from the GPT family, GPT-4, which allowed the general public to experience the improvements in model output quality. Almost immediately, another feature called plug-ins was introduced that allowed the system to connect to other software, such as Internet search engines, mathematical problem solvers, and travel databases, thereby uniting the 'large language brain' with access to real time data. These innovations will continue and these systems will evolve. The speed of technological evolution puts numerous social practices at risk and we encourage the scholarly community to engage with the technologies to help guide social evolution.

Second, educational researchers need to become conversant with the fundamental logics we introduce here. We can think of the role of the Internet in the evolution of organizations over the last 20 years. While not every organization became an ‘Internet company’, almost all organizations have become Internet-dependent. Similarly, while not all educational researchers need to become AI researchers, all researchers must know enough to evolve their research and teaching on the new AI platforms as appropriate. To help society in its social evolution with technology, and to take advantage of its benefits, we must achieve the required level of understanding and engagement.

Third, researchers must rethink their relationship to technology and its use. For many in education, technology is a niche topic for others to consider. We hope we have sufficiently communicated that the inflection point of technology infusion we are facing will change how society interacts with technology and how we as educators relate to technology. The concerns and opportunities we are facing involve curricular issues, psychological and social issues, computational and media issues and so forth. Technologies such as data science and machine learning are no longer topics but substrates to our daily lives and our educational research should reflect it.

Fourth, for those interested in computational aspects of education, this is both an exciting and challenging time. The speed of change in both research results and industrial applications is remarkable. The fact that large language models perform well at computer coding means that new support is available for those who want to enter the world of computer science. At the same time, for those with prior software experience, there are a flood of support tools for learning how to use and adapt open-source or proprietary AI software. A notable example is the model repository and cloud computing environment offered by Hugging Face.[1]

For both good and ill, the biggest limitation that educational researchers will face is their imagination. The new AI systems act, and are interacted with, in such novel ways that students and researchers with limited computational background may find it difficult to appreciate the opportunity to build something new. At the same time, those with engineering background may be limited by old conceptualisations of learning and assessment instead of reimagining what might be possible. Either way, we encourage all researchers to learn, experiment, and integrate their domain knowledge with these new developments.

References

Anders, B. A. (2023). Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns*, *4*(3), 100694. https://doi.org/10.1016/j.patter.2023.100694.

[1] See https://www.huggingface.co. Accessed 14 April 2024.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9.

Behrens, J. T., & Dicerbo, K. E. (2014). Technological Implications for Assessment Ecosystems: Opportunities for Digital Technology to Advance Assessment. *Teachers College Record*, *116*(11), 1–22. https://doi.org/10.1177/016146811411601112.

Belanger, A. (2023). AI-faked images of Donald Trump's imagined arrest swirl on Twitter. Ars Technica, 21 March. https://arstechnica.com/tech-policy/2023/03/fake-ai-generated-images-imagining-donald-trumps-arrest-circulate-on-twitter/. Accessed 30 June 2023.

Bengio, Y. (2023). How rogue AIs may arise. Yoshua Bengio, 22 May. https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise. Accessed 30 June 2023.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). Measuring progress on scalable oversight for large language models. arXiv:2211.03540. https://doi.org/10.48550/arxiv.2211.03540.

Brockman, G. (2022). @gdb. 5 December. https://twitter.com/gdb/status/1599683104142430208. Accessed 2 May 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165. https://doi.org/10.48550/arxiv.2005.14165.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. arXiv:2107.03374. https://doi.org/10.48550/arxiv.2107.03374.

Clancey, W. J. (1984). Use of MYCIN's rules for tutoring. In B. G. Buchanan & E. H. Shortliffe (Eds.), *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project* (pp. 464–489). Reading, MA: Addison-Wesley Publishing Company.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148.

Deranty, J.-P., & Corbin, T. (2022). Artificial intelligence and work: a critical review of recent research from the social sciences. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01496-x.

DiCerbo, K., & Behrens J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

D'Mello, S. K., & Graesser, A. C. (2023). Intelligent tutoring systems: How computers achieve learning gains that rival human tutors. In P. A. Schutz & K. R. Muis (Eds.), *Handbook of educational psychology.* 4th Ed. New York: Routledge. https://doi.org/10.4324/9780429433726.

European Commission, Directorate-General for Education, Youth, Sport and Culture. (2022). Final report of the Commission expert group on artificial intelligence and data in education and training: Executive summary. Luxembourg: Publications Office of the European Union. https://doi.org/10.2766/65087.

European Parliament. (2023). Report on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial

Intelligence Act) and amending certain Union Legislative Acts. https://www.europarl.europa.eu/doceo/document/A-9-2023-0188_EN.html#_section1. Accessed 30 June 2023.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2), 939–944.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (68–88). New York: Routledge. https://doi.org/10.4324/9780203122761.ch5.

Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, & N. R. Jennings (Eds.), *Lecture notes in computer science: Vol. 1193. Intelligent agents III: Agent theories, architectures, and languages* (pp. 21–35). Berlin: Springer. https://doi.org/10.1007/BFb0013570.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36.

Handel, M. J. (2022). Growth trends for selected occupations considered at risk from automation. *Monthly Labor Review*. https://www.jstor.org/stable/48716837. Accessed 14 April 2024.

Hellman, S., Rosenstein, M., Gorman, A., Murray, W., Becker, L., Baikadi, A., Budden, J., & Foltz, P. W. (2019). Scaling up writing in the curriculum: Batch mode active learning for automated essay scoring. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (pp. 1–10). New York: Association for Computing Machinery. https://doi.org/10.1145/3330430.3333629.

Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. Reuters, 2 February. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed 2 May 2023.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., & Wei, F. (2023). Language is not all you need: Aligning perception with language models. arXiv:2302.14045. https://doi.org/10.48550/arxiv.2302.14045.

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv:2301.08745. https://doi.org/10.48550/arxiv.2301.08745.

Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networks. Andrej Karpathy blog, 21 May. http://karpathy.github.io/2015/05/21/rnn-effectiveness. Accessed 2 May 2023.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 333–353. https://doi.org/10.1002/wcs.1350.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420.

Kwako, A., Wan, Y., Zhao, J., Hansen, M., Chang, K. W., & Cai, L. (2023). Does BERT exacerbate gender or L1 biases in automated English speaking assessment? In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 668–681). Toronto: Association for Computational Linguistics. https://aclanthology.org/2023.bea-1.54. Accessed 14 April 2024.

Liu, Q., Pinto, J. D., & Paquette, L. (2024). Applications of Explainable AI (XAI) in Education. In D. Kourkoulou, A.-O. Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and Inclusion in AI-mediated Education: Where Human Learning Meets Learning Machines.* Cham: Springer.

Lucy, L., & Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, & L. J. Martin (Eds.), *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48–55). Toronto: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nuse-1.5.

Lukpat, A. (2023). AI poses 'risk of extinction' on par with pandemics and nuclear war, tech executives warn. The Wall Street Journal, 30 May. https://www.wsj.com/articles/ai-threat-is-on-par-with-pandemics-nuclear-war-tech-executives-warn-39105eeb. Accessed 30 June 2023.

McKinsey. (2023). The economic potential of generative AI: The next productivity frontier. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights. Accessed 30 June 2023.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596.

Mollick, E. R., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4475995.

Office of Educational Technology. (2023). Artificial intelligence and future of teaching and learning: Insights and recommendations. Washington, DC: Office of Educational Technology.

OpenAI. (2022). ChatGPT. https://chat.openai.com/chat. Accessed 2 May 2023.

OpenAI, Achiam, J., Adler, S., Agarwal, S., ... Zoph, B. (2023). GPT-4 Technical Report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155. https://doi.org/10.48550/arxiv.2203.02155.

Pearson Education. (2019). Aida Calculus. https://apps.apple.com. Accessed 1 June 2023.

Polus, S. (2023). Market dips briefly after AI image of fake explosion near Pentagon goes viral. The Hill, 22 May. https://thehill.com/policy/technology/4015817-market-dips-briefly-after-ai-image-of-fake-explosion-near-pentagon-goes-viral/. Accessed 30 June 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 14 February. https://openai.com/research/better-language-models. Accessed 2 May 2022.

Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach.* 4th Ed. Harlow: Pearson Education Limited.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. arXiv:2302.04761. https://doi.org/10.48550/arxiv.2302.04761.

Seetharaman, D. (2023). Elon Musk, other AI experts call for pause in technology's development. The Wall Street Journal, 29 March. https://www.wsj.com/articles/elon-musk-other-ai-bigwigs-call-for-pause-in-technologys-development-56327f. Accessed 30 June 2023.

Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN*. New York: Elsevier. https://doi.org/10.1016/b978-0-444-00179-5.x5001-x.

Sleeman, D., & Brown, J. S. (Eds.). (1982). *Intelligent tutoring systems*. London: Academic.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 843–852). IEEE. https://doi.org/10.1109/ICCV.2017.97.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. https://doi.org/10.48550/arxiv.2307.09288.

von Davier, A. A., Mislevy, R. J., & Hao, J. (Eds.). (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With Examples in R and Python*. Cham: Springer. https://doi.org/10.1007/978-3-030-74394-9.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762. https://doi.org/10.48550/arxiv.1706.03762.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 214–229). New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088.

Weiser, B. (2023). Here's what happens when your lawyer uses ChatGPT. The New York Times, 27 May. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html. Accessed 30 June 2023.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382. https://doi.org/10.48550/arXiv.2302.11382.

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781351264808-19.

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. New York: Routledge. https://doi.org/10.4324/9781315165103.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: Less is more for alignment. arXiv:2305.11206. https://doi.org/10.48550/arxiv.2305.11206.

Conclusion

Summary of findings

The publications presented in this PhD dissertation contributed to four research goals. First, Publication 1 (Łodzikowski and Jekiel 2019) and Publication 2 (Łodzikowski 2021) contributed to the need to focus on metaphonological awareness training for advanced learners of English who aim at achieving near-nativelike pronunciation proficiency. The first study demonstrated that print-and-play board games can serve as a readily available tool for enhancing in-class (and post-class) student engagement with prosody topics challenging for Polish learners of English (word stress, intonation, and rhythm) and potentially improving declarative metaphonological knowledge on those topics. Specifically, playing these games was associated with an increase in post-class quiz scores (low-stakes measures of metaphonological awareness) by about 8 percentage points. When asked to choose between 30 minutes of board games or extra worksheet exercises, 62% of learners preferred board games. Additionally, 35% of learners replayed games after class, and 10% used them to prepare for the final exam, citing benefits like extra practice, immediate feedback, and learning through fun and competition. Similarly, the second study demonstrated that encouraging learners to autonomously engage with an IPA-based allophonic transcription tool can improve their metaphonological awareness; see paragraph below for details. Similarly to Jekiel and Malarski (2021), greater gains were observed for learners with a higher overall level of English proficiency, suggesting the tools' design (or the course matter in general) needs to become more accessible to various learners, perhaps with the help of generative AI.

Second, Publication 2 (Łodzikowski 2021) showed how a digital tool for metaphonological awareness training can provide insights into behavioural self-study patterns, such as the frequency and extent of practice sessions. Such details are typically unobservable in conventional classroom settings and may provide more objective evidence of learner activities compared to subjective self-reports. In this study, the analysis of the log data revealed that regular and spaced usage of the tool was a more reliable predictor of improved metaphonological awareness than merely the frequency of tool usage. Specifically, learners who visited the transcription tool at least once were observed to score higher on the midterm test (mid-stakes measures of metaphonological awareness) by about 5 percentage points. Moreover, each additional distinct visit day was associated with an increase in the expected test score of about 2.6 percentage points on the assessment. Overall, this publication has successfully demonstrated the idea by Chun (2016) that digital pronunciation teaching tools can be more than just (potentially) helpful teaching aids. They can also be indispensable mechanisms of collecting data for future studies as well as pedagogical practice improvements.

Third, both publications contributed to promoting transparency in reporting research in the field of CAPT by providing open access to the developed tools (print-and-play board games and transcription tool code) as well as the necessary data and methods to replicate the studies. In fact, Publication 2 was replicated by Foung and Kohnke (2023), which not only largely reaffirmed the original findings but also enhanced their visibility in the broader academic community.

Finally, Publication 3 explored the integration of next-generation AI technologies, specifically large language models such as GPT-4, into the broad education domain. Due to their advanced capabilities in understanding and generating natural language, generative AI systems can offer dynamic and personalised learning experiences that enhance the validity of assessments, thus addressing prior criticisms of AI-enabled (language) education. And while this publication did not focus on metalinguistic awareness per se, it addresses the call by Rogerson-Revell (2021) for the next wave of computer-assisted teaching and learning to be pedagogy-led rather than technology-led.

Limitations

Publications 1 and 2 have contributed valuable insights but also face inherent limitations typical of in situ research, and as a result they do not provide definitive answers to the research questions.

First, the research setting influenced the sample composition, resulting in biases such as an underrepresentation of male participants and issues with student dropout. To improve the robustness of future studies, researchers should collect larger and more diverse participant samples to strengthen the validity and reliability of the statistical analysis.

Second, the exploratory nature of these studies means that the findings about the relationships between variables remain tentative. While past research in the field often relied on controlled experimental designs that provided incomplete or subjective behavioural data, this work used an observational approach that captures more detailed behavioural insights. However, fully randomised control trials may not be feasible in institutional settings because teachers and students must adhere to a fixed curriculum. If organising an experimental study is not an option, then perhaps a viable alternative could be to extend the scope of behavioural data collection to new datasets. See, for example, the use of physiological data by Giannakos et al. (2019).

Third, some lurking variables could not have been excluded. For instance, some learners might have used other study tools alongside those being evaluated by the present author, potentially skewing the results. Conversely, some learners may have used the website-based transcription practice tool less than traditional pen-and-paper transcription exercises not because they found the digital tool unhelpful but because they preferred to keep the web browser window closed to avoid distractions.

Fourth, even though digital tools offer a rich data source for standardising measurements of educational interventions (e.g. site visits and time spent), interpreting computer-aided interventions involves some complexities. The effectiveness of an intervention can be significantly impacted by challenges such as onboarding, troubleshooting, and cognitive overload (Shadiev and Yang 2020). For example, relatively unpredictable issues such as outdated browsers could lead to an improper display or operation of the transcription practice tool, potentially frustrating students and reducing their engagement.

Finally, and perhaps most importantly, the two empirical studies could have been enhanced by juxtaposing learners' interactions with the metaphonological awareness tools against instrumental and impressionistic analyses of segmental and suprasegmental pronunciation features at the start and end of the course.

Despite these challenges, the findings from these studies offer directional insights valuable for researchers undertaking confirmatory or registered studies, as well as for pronunciation instructors seeking to refine their course curricula based on the observed usage patterns.

Future directions

One noticeable trend in the history of CAPT tools is that industry professionals often develop them for specific use cases (such as speech transcription system for clinicians in Bailey et al. 2022) and that they are subsequently reimplemented in industry-specific training contexts (just like Speights Atkins et al. 2023 implemented the above work and evolved it into automated transcription practice for speech pathology trainees). Previously, developing such tools required niche programming skills and substantial time investments. However, the emergence of generative AI now allows researchers and practitioners to create novel tools effortlessly using simple natural language commands.

Among the various ways in which language teachers can harness generative AI, a few opportunities stand out immediately (Łodzikowski in press). First, teachers can rapidly generate diverse and relevant classroom or digital content—textual, visual, and auditory—tailored to exact needs of each learner and resulting in greater engagement. Moreover, this allows the teacher to focus on more meaningful activities, such as one-to-one tutoring with at-risk learners. Another significant application is in designing authentic assessments that simulate real-world challenges. For instance, teachers might use AI to create scenarios where students must employ diplomatic language and problem-solving skills in dealing with a simulated angry customer.

Expanding on these applications, Łodzikowski et al. (in preparation) built and piloted an ITS to support English Philology students in acquiring L2 metaphonological awareness. The application leveraged a state-of-the-art large language model that was tweaked for accuracy by consuming English phonetics and phonology coursebooks. The authors designed a set of custom homework assignments that leveraged the capabilities of the ITS (such as the ability to invent helpful analogies), while exposing its shortcomings (esp. the propensity towards misinformation). These tasks allowed students to interact with the AI for discussions and analyses of phonetic concepts and transcriptions. While the ITS generally provided helpful feedback and explanations, it occasionally presented incorrect information, challenging students to critically assess and correct the errors, thereby deepening their engagement and learning. Overall, the study showed that there are some aspects of metaphonological training for which generative AI could serve as a natural extension of the teacher, and many others for which it could not.

The above-mentioned strategies are not confined to teaching phonological awareness. They are applicable across various domains of language learning and L2 awareness. Regardless of the domain, integrating AI into the teaching process will require educators to cultivate three key areas of AI literacy. The first area, technological literacy, requires obtaining an intuitive understanding of the technology to be able to determine its strengths and weaknesses for various teaching and learning use cases. The second area is pedagogical literacy, namely opening to a new paradigm of shared agency between humans and AI (Godwin-Jones, 2024) and critically rethinking outdated conceptualisations of assessment and skill acquisition (Moorhouse and Kohnke 2024). The third area is governance literacy, that is, ensuring responsible stewardship of student data and intellectual property, safeguarding against inappropriate content, and ensuring compliance with institutional or legal requirements (Department for Education 2024). As educators begin to experiment with AI tools within their institutional learning management systems such as Moodle, ensuring stringent data privacy through will be non-negotiable (see a discussion and guidelines in Dondorf 2022).

References

Acton, William. 1984. “Changing fossilized pronunciation”, *TESOL Quarterly* 18, 1, 71-85. (doi: 10.2307/3586336).

Alves, Ubiratã Kickhöfel and Vivian Magro. 2011. “Raising awareness of L2 phonology: Explicit instruction and the acquisition of aspirated /p/ by Brazilian Portuguese speakers”, *Letras de Hoje* 46, 3: 71-80.

Bailey, Dallin J., Marisha Speights Atkins, Ishaan Mishra, Sicheng Li, Yaoxuan Luan and Cheryl Seals. 2022. “An automated tool for comparing phonetic transcriptions”, *Clinical Linguistics & Phonetics* 36, 6: 495-514. (doi:10.1080/02699206.2021.1896783).

Baker, Amanda A. 2018. “Pronunciation teaching in the pre-CLT era”, in: Okim Kang, Ron I. Thomson and John M. Murphy (eds.), *The Routledge handbook of contemporary English pronunciation*. London: Routledge, 249-266.

Bates, Sally, Bruce Matthews and Alan Eagles. 2010. *Phonetic transcription self-study programme: Webfon*. (http://elearning.marjon.ac.uk/ptsp/) (date of access: 31 Aug. 2017).

Bell, Nancy. 2012. “Formulaic language, creativity, and language play in a second language”, *Annual Review of Applied Linguistics* 32: 189-205. (doi:10.1017/S0267190512000013).

Bell, Nancy and Anne Pomerantz. 2015. *Humor in the classroom: A guide for language teachers and educational researchers*. London: Routledge.

Bentley, Rensselaer. 1849. *The pictorial spelling book: Containing an improved method of teaching the alphabet, and likewise spelling and pronunciation, by the use of pictures*. New York, NY: Pratt, Woodford & Co.

Boersma, Paul and David Weenink. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.10. (http://www.praat.org/) (date of access: 1 May 2023).

Brinton, Donna M. 2017. “Innovations in pronunciation teaching”, in: Okim Kang, Ron I. Thomson and John M. Murphy (eds.), *The Routledge handbook of contemporary English pronunciation*. London: Routledge, 449-461.

Brondsted, Tom. 2020. Automatic phonemic transcriber. (http://tom.brondsted.dk/text2phoneme/) (date of access: 1 Dec. 2023).

Bruijn, Christel de, Miguel Baptista Nunes, Linhao Fang, Rigved Pathak and Jingchao Zhou. 2011. A system for independent e-learning of practical phonetics. (Paper presented at the 17th International Congress of Phonetic Sciences, 17-21 Aug. 2011.).

Buczek-Zawiła, Anita. 2018. "English pronunciation standard preferences among students of Polish universities—self-reports", *Anglica Wratislaviensia* 56: 251-271. (doi:10.19195/0301-7966.56.16).

Busà, Maria Grazia. 2008. "New perspectives in teaching pronunciation", in: Anthony Baldry, Maria Pavesi, Carol Taylor Torsello and Christopher Taylor (eds.), *From Didactas to ecolingua. An ongoing research project on translation and corpus linguistics*. Trieste: Università degli Studi di Trieste, 165-182.

Cardoso, Walcir. 2017. "English syllable structure", in: Okim Kang, Ron I. Thomson and John M. Murphy (eds.), *The Routledge handbook of contemporary English pronunciation*. London: Routledge, 122-136.

Carlet, Angélica and Hanna Kivistö-de Souza. 2018. "Improving L2 pronunciation inside and outside the classroom: Perception, production and autonomous learning of L2 vowels", *Ilha do Desterro* 71, 3: 99-123. (doi:10.5007/2175-8026.2018v71n3p99).

Chapelle, Carol A. and Shannon Sauro (eds.). 2017. *The handbook of technology and second language teaching and learning*. Hoboken, NJ: John Wiley & Sons, Inc.

Chi, Michelene T. H. and Ruth Wylie. 2014. "The ICAP framework: Linking cognitive engagement to active learning outcomes", *Educational Psychologist* 49, 4: 219-243. (doi:10.1080/00461520.2014.965823).

Chun, Dorothy. 1998. "Signal analysis software for teaching discourse intonation", *Language Learning & Technology* 2, 1: 74-93. (doi:10125/25033).

Chun, Dorothy M. 2016. "The role of technology in SLA research", *Language Learning & Technology* 20 2: 98-115. (doi:10125/44463).

Coulange, Sylvain. 2023. "Computer-aided pronunciation training in 2022: When pedagogy struggles to catch up", in: Alice Henderson and Anastazija Kirkova-Naskova (eds.), *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices*. Grenoble: Université Grenoble-Alp (https://hal.science/hal-04159763/document) (date of access: 15 Jan. 2024). (doi: 10.5281/zenodo.8137754).

Couper, Graeme. 2022. "Teaching and testing perception of word stress: many shades of perception", in: John Levis and Aleksandra Guskaroska (eds.), *Proceedings of the 12th Pronunciation in Second Language Learning and Teaching Conference*. (doi:10.31274/psllt.13266).

Danesi, Marcel and Anthony Mollica. 1994. "Games and puzzles in the second-language classroom: A second look", *Mosaic* 2, 2: 13-22.

Department for Education. 2024. *Generative AI in education: Educator and expert views*. (https://www.gov.uk/government/publications/generative-ai-in-education-educator-and-expert-views) (date of access: 1 Jun. 2024).

Derwing, Tracey M. and Murray J. Munro. 2005. "Pragmatic perspectives on the preparation of teachers of English as a second language: Putting the NS/NNS debate in context", in: Enric Llurda (ed.), *Non-native language teachers: Perceptions, challenges and contributions to the profession*. Philadelphia, PA: Springer, 179-191.

Derwing, Tracey M. and Murray J. Munro. 2015. *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins. (doi:10.1075/lllt.42).

Derwing, Tracey M., Murray J. Munro and Grace Wiebe. 1998. "Evidence in favor of a broad framework for pronunciation instruction", *Language Learning* 48, 3: 393-410. (doi:10.1111/0023-8333.00047).

Dondorf, Thomas. 2022. Learning analytics for Moodle: facilitating the adoption of data privacy friendly learning analytics in higher education. [Unpublished doctoral dissertation, RWTH Aachen University].

Dziubalska-Kołaczyk, Katarzyna, Anna Balas, Geoffrey Schwartz, Arkadiusz Rojczyk, and Magdalena Wrembel. 2015. "Teaching to suppress Polglish processes", in: Ewa Waniek-Klimczak and Mirosław Pawlak (eds.), *Teaching and researching the pronunciation of English*. Cham: Springer, 241-256. (doi:10.1007/978-3-319-11092-9_15).

Ellis, Rod. 2004. "The definition and measurement of L2 explicit knowledge", *Language Learning* 54, 2: 227-275. (doi:10.1111/j.1467-9922.2004.00255.x).

Ellis, Rod. 2005. “Principles of instructed language learning”, *System* 33, 2: 209-224. (doi:10.1016/j.system.2004.12.006).

Eskreis-Winkler, Lauren, Elizabeth P. Shulman, Scott A. Beal and Angela L. Duckworth. 2014. “The grit effect: predicting retention in the military, the workplace, school and marriage”, *Frontiers in Psychology* 5. (doi:10.3389/fpsyg.2014.00036).

Flege, James Emil and Ocke-Schwen Bohn. 2021. “The revised speech learning model (SLM-r)”, in: Ratree Wayland (ed.), *Second language speech learning: Theoretical and empirical progress*. Cambridge: Cambridge University Press, 3-83.

Foung, Dennis and Lucas Kohnke. 2023. “Beyond replication: An exact replication study of Łodzikowski (2021)”, *ReCALL* 35, 2: 225-238. (doi:10.1017/S0958344023000071).

Fouz-González, Jonás. 2015. “Trends and directions in computer-assisted pronunciation training”, in: Juan Antonio Mompean and Jonás Fouz-González (eds.), *Investigating English pronunciation*. London: Palgrave Macmillan, 314-342. (doi:10.1057/9781137509437_14).

Galante, Angelica and Ron I. Thomson. 2016. “The effectiveness of drama as an instructional approach for the development of second language oral fluency, comprehensibility, and accentedness”, *TESOL Quarterly* 51, 1: 115-142. (doi:10.1002/tesq.290).

García Lecumberri, María Luisa, Martin Cooke and John Maidment. 2003. Transcripción fonémica en Internet. (Paper presented at the VIII Simposio Internacional de Comunicación Social, 20 Jan. 2003).

García Lecumberri, María Luisa, John Maidment, Martin Cooke, Anders Ericsson and Mircea Giurgiu. 2003. A web-based transcription tool. (Paper presented at the 15th International Congress of Phonetic Sciences, 3-9 Aug. 2003.

Giannakos, Michail N., Kshitij Sharma, Ilias O. Pappas, Vassilis Kostakos and Eduardo Velloso. 2019. “Multimodal data as a means to understand the learning experience”, *International Journal of Information Management* 48: 108-119. (doi:10.1016/j.ijinfomgt.2019.02.003).

Godwin-Jones, Robert. 2024. “Distributed agency in language learning and teaching through generative AI”, *Language Learning & Technology* 28, 2: 5-31. (doi:10125/73570).

Graesser, Arthur C., Mark W. Conley and Andrew Olney. 2012. “ITSs”, in: Karen R. Harris, Steve Graham, Timothy Urdan, Ann G. Bus, Sandra Major and H. Lee Swanson (eds.), *APA educational psychology handbook. Vol. 3: Application to learning and teaching*. American Psychological Association, 451–473. (doi:10.1037/13275-018).

Hadfield, Jill. 2003. *Intermediate grammar games*. London: Longman.

Hancock, Mark. 1995. *Pronunciation games*. Cambridge:Cambridge University Press

Hancock, Mark. 2017. *PronPack 2: Pronunciation puzzles*. Chester: Hancock McDonald ELT.

Inceoglu, Solène. 2021. “Exploring the relationship between explicit instruction, pronunciation awareness, and the development of L2 French connected speech processes”, *Language Awareness* 30, 4: 336-354. (doi:10.1080/09658416.2021.1881527).

Jekiel, Mateusz. 2022. “L2 rhythm production and musical rhythm perception in advanced learners of English”, *Poznań Studies in Contemporary Linguistics* 58, 2: 315-340.

Jekiel, Mateusz and Kamil Malarski. 2021. “Musical hearing and musical experience in second language English vowel acquisition”, *Journal of Speech, Language, and Hearing Research* 64 5: 1666-1682.

Jekiel, Mateusz and Kamil Malarski. 2023. “Musical hearing and the acquisition of foreign-language intonation”, *Studies in Second Language Learning and Teaching* 13, 1: 151-178.

Jespersen, Otto. 1904. *How to teach a foreign language*. (Translated by Sophia Yhlen-Olsen Bertelsen.) London: Swan Sonnenschein & Co. Ltd.

Kennedy, Sara and Josée Blanchet. 2013. “Language awareness and perception of connected speech in a second language”, *Language Awareness* 23, 1-2: 92-106. (doi:10.1080/09658416.2013.863904).

Kennedy, Sara and Pavel Trofimovich. 2010. “Language awareness and second language pronunciation: a classroom study”, *Language Awareness* 19, 3: 171-185. (doi:10.1080/09658416.2010.486439).

Kirkova-Naskova, Anastazija, Alice Henderson and Jonás Fouz-González (eds.). 2021. *English pronunciation instruction: Research-based insights*. Amsterdam: John Benjamins. (doi:10.1075/aals.19).

Kivistö-de Souza, Hanna. 2015. Phonological awareness and pronunciation in a second language. [Unpublished Ph.D. dissertation, University of Barcelona].

Kivistö-de Souza, Hanna and Pekka Lintunen. 2023. "Thinking about your pronunciation: Examining phonological selfawareness with a novel task", in: Alice Henderson and Anastazija Kirkova-Naskova (eds.), *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices*. Grenoble: Université Grenoble-Alp (https://hal.science/hal-04178953) (date of access: 4 Apr. 2024). (doi: 10.5281/zenodo.8225603).

Kohnke, Lucas, Benjamin Luke Moorhouse and Di Zou. 2023. "Exploring generative artificial intelligence preparedness among university language instructors: A case study", *Computers and Education: Artificial Intelligence* 5: 100156. (doi:10.1016/j.caeai.2023.100156).

Krzysik, Iga. 2022. Individual differences in multilingualacquisition of phonology: A longitudinalstudy of adolescents. [Unpublished Ph.D. dissertation, Adam Mickiewicz University].

Krzysik, Iga and Halina Lewandowska. 2017. "English pronunciation training through the eyes of university graduates", *Research in Language* 15, 3: 299-312. (doi:10.1515/rela-2017-0017).

Kulik, James A. and J. D. Fletcher. 2016. "Effectiveness of ITSs: A Meta-Analytic Review", *Review of Educational Research* 86, 1: 42-78. (doi:10.3102/0034654315581420).

Lacabex, Esther Gomez and Francisco Gallardo-del-Puerto. 2020. "Explicit phonetic instruction vs. implicit attention to native exposure: phonological awareness of English schwa in CLIL", *International Review of Applied Linguistics in Language Teaching* 58, 4: 419-442. (doi:10.1515/iral-2017-0079).

Lacabex, Esther Gomez and Hanne Roothooft. 2023. "Pronunciation anxiety, pronunciation-related views and pronunciation learning actions of EMI and English major students", *Research in Language* 21, 4: 333-356. (doi:10.18778/1731-7533.21.4.01).

Lee, Hansol and Jang Ho Lee. 2024. "The effects of AI-guided individualized language learning: A meta-analysis", *Language Learning & Technology* 28, 2: 134-162. (https://hdl.handle.net/10125/73575).

Lee, Bradford, Luke Plonsky and Kazuya Saito. 2020. “The effects of perception- vs. production-based pronunciation instruction”, *System* 88: 102185. (doi:10.1016/j.system.2019.102185).

Lee, Junkyu, Juhyun Jang and Luke Plonsky. 2015. “The effectiveness of second language pronunciation instruction: A meta-analysis”, *Applied Linguistics* 36, 3: 345-366. (doi:10.1093/applin/amu040).

Levis, John. 2005. “Changing contexts and shifting paradigms in pronunciation teaching”, *TESOL Quarterly* 39, 3, 369-377. (doi:10.2307/3588485).

Levis, John, and Sinem Sonsaat. 2017. “Pronunciation teaching in the early CLT era”, in: Okim Kang, Ron I. Thomson and John M. Murphy (eds.), *The Routledge handbook of contemporary English pronunciation*. London: Routledge, 267-283.

Lintunen, Pekka. 2004. *Pronunciation and phonemic transcription: A study of advanced Finnish learners of English*. Turku: University of Turku.

Lintunen, Pekka. 2013. The effect of phonetic knowledge on evaluated pronunciation problems. (Paper presented at PTLC2013, 8-10 Aug. 2013.)

Lintunen, Pekka and Aleksi Mäkilähde. 2018. “Short- and long-term effects of pronunciation teaching: EFL learners’ views”, In: J. Volín and R. Skarnitzl (Eds.), The pronunciation of English by speakers of other languages (pp. 46–72). Cambridge Scholars Publishing.

Łodzikowski, Kacper. 2021. “Association between allophonic transcription tool use and phonological awareness level”, *Language Learning and Technology* 25, 1: 20-30. (doi:10125/44748).

Łodzikowski, Kacper. In press. “Harnessing the potential of generative AI in second language teaching”, in: Marta Strukowska and Katarzyna Matuszak (eds.), *Proceedings of Looking ahead: Developing academics’ and students’ linguistic and intercultural competence for a globalised world*. Poznań: Poznań University of Technology Publishing House.

Łodzikowski, Kacper and Grzegorz Aperliński. 2016. Usage patterns of an online allophonic transcriptor. (Paper presented at the 10th International Conference on Native and Non-native Accents of English, 1-3 Dec. 2016.).

Łodzikowski, Kacper and Mateusz Jekiel. 2019. “Board games for teaching English prosody to advanced EFL learners”, *ELT Journal* 73, 3: 275-285. (doi:10.1093/elt/ccy059).

Łodzikowski, Kacper, Peter W. Foltz and John T. Behrens. 2024. "Generative AI and Its Educational Implications", in: Dora Kourkoulou, Anastasia Tzirides, Bill Cope, and Mary Kalantzis (eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines.* Cham: Springer. (doi:10.1007/978-3-031-64487-0_2).

Łodzikowski, Kacper, Jarosław Weckwerth and Kamil Malarski. In preparation. Exploring large language models for L2 metaphonological awareness tutoring.

Mahdi, Hassan Saleh and Ahmed Abdulateef Al Khateeb. 2019. "The effectiveness of computer-assisted pronunciation training: A meta-analysis", *Review of Education* 7, 3: 733-753. (doi:10.1002/rev3.3165).

Mompeán, Jose A. 2017. "Doing phonetic transcription in a modern language degree", in: Juan Antonio Cutillas Espinosa, Juan Manuel Hernández Campoy, Rosa María Manchón Ruiz and Florentina Mena Martínez (eds.), *Estudios de filología inglesa: Homenaje a D. Rafael Monroy*. Murcia: Editum, 479-505.

Mompeán, Jose A. and Pekka Lintunen. 2015. "Phonetic notation in foreign language teaching and learning: Potential advantages and learners' views", *Research in Language* 13, 3: 292-314. (doi: 10.1515/rela-2015-0026).

Mompeán, Jose A. and Jonás Fouz-González. 2021. "Phonetic Symbols in Contemporary Pronunciation Instruction", *RELC Journal* 52, 1: 155-168. (doi:10.1177/0033688220943431).

Moorhouse, Benjamin Luke and Lucas Kohnke. 2024. "The effects of generative AI on initial language teacher education: The perceptions of teacher educators", *Syste*, 122, 103290. (doi:10.1016/j.system.2024.103290).

Mora, Joan C., Youssef Rochdi and Hanna Kivistö-de Souza. 2014. "Mimicking accented speech as L2 phonological awareness", *Language Awareness* 23, 1-2: 57-75. (doi:10.1080/09658416.2013.863898).

Murphy, John. 2013. *Teaching pronunciation*. Alexandria, VA: TESOL Publications.

Murphy, John and Amanda Ann Baker. 2015. "History of ESL pronunciation teaching", in: John M. Levis and Mary Reed (eds.), *Handbook of English pronunciation*. New York, NY: Wiley-Blackwell, 36-65.

Munro, Murray J., and Tracey M. Derwing. 1995. "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners", *Language Learning* 45, 1: 73–97. (doi:10.1111/j.1467-1770.1995.tb00963.x).

Ngo, Thuy Thi-Nhu, Howard Hao-Jan Chen and Kyle Kuo-Wei Lai. 2023. "The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis", *ReCALL* 36, 1: 4-21. (doi:10.1017/s0958344023000113).

Nixon, Caroline and Michael Tomlinson. 2005. *Primary pronunciation box: Pronunciation games and activities*. Cambridge: Cambridge University Press.

Nowacka, Marta. 2012. "Questionnaire-based pronunciation studies: Italian, Spanish and Polish students' views on their English pronunciation", *Research in Language* 10, 1: 43-61. (doi:10.2478/v10015-011-0048-3).

Nowacka, Marta. 2016. "English spelling among the top priorities in pronunciation teaching: Polglish local versus global(ised) errors in the production and perception of words commonly mispronounced", *Research in Language* 14, 2: 123-148. (doi:10.1515/rela-2016-0002).

Nowacka, Marta. 2022. "English phonetics course: University students' preferences and expectations", *Research in Language* 20, 1: 71-84. (doi:10.18778/1731-7533.20.1.05).

OECD. 2024. Explanatory memorandum on the updated OECD definition of an AI system. Paris: OECD Publishing. (doi:10.1787/623da898-en).

Peltola, Maija S., Pekka Lintunen and Henna Tamminen. 2014. "Advanced English learners benefit from explicit pronunciation teaching: An experiment with vowel duration and quality", *AFinLA-teema* 6: 86-98.

Pennington, Martha C. 1999. "Computer-aided pronunciation pedagogy: Promise, limitations, directions", *Computer-Assisted Language Learning* 12, 5:427-440.

Pennington, Martha C. 2021. "Teaching pronunciation: The state of the art 2021", *RELC Journal* 52, 1: 3-21. (doi:10.1177/00336882211002283).

Pennington, Martha C. and Pamela Rogerson-Revell. 2019. *English pronunciation teaching and research: Contemporary perspectives*. London: Palgrave Macmillan. (doi:10.1057/978-1-137-47677-7).

PhoTransEdit. 2023. PhoTransEdit. (http://photransedit.com) (date of access: 2 Dec. 2023).

Ramírez Verdugo, Dolores. 2006. "A study of intonation awareness and learning in non-native speakers of English", *Language Awareness* 15, 3: 141-159. (doi:10.2167/la404.0).

Rogerson-Revell, Pamela M. 2021. Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal* 52, 1: 189-205. (doi:10.1177/0033688220977406).

Rojczyk, Arkadiusz. 2010. “Forming new vowel categories in second language speech: The case of Polish learners’ production of English /ɪ/ and /e/”, *Research in Language* 8: 85-97.

Rojczyk, Arkadiusz and Andrzej Porzuczek. 2012. “Selected aspects in the acquisition of English phonology by Polish learners – Segments and prosody”, in: Danuta Gabryś-Barker (ed.), *Readings in second language acquisition.* Katowice: Uniwersytet Śląski, 93-120.

Rojczyk Arkadiusz and Andrzej Porzuczek. 2017. “Discrimination of English tone contours by Polish learners”, *Linguistica Silesiana* 38: 53-66.

Rojczyk, Arkadiusz and Andrzej Porzuczek. 2019. “EFL pronunciation teaching to Polish English studies majors,” in: Danuta Gabryś-Barker (ed.), *Challenges of foreign language instruction in the university context*. Katowice: Uniwersytet Śląski, 9-25.

Saito, Kazuya. 2011. “Examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: An instructed SLA approach to L2 phonology”, *Language Awareness* 20, 1: 45-59. (doi:10.1080/09658416.2010.540326).

Saito, Kazuya. 2012. “Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies”. *TESOL Quarterly* 46, 4: 842-854.

Saito, Kazuya. 2013. “Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information”, *Studies in Second Language Acquisition* 35, 1: 1-29. (doi:10.1017/S0272263112000666).

Saito, Kazuya. 2019. “Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners’ English /ɹ/ pronunciation”, *Second Language Research* 35, 2: 149-172. (doi:10.1177/0267658318768342).

Saito, Kazuya. 2021. “What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors”, *TESOL Quarterly* 55, 3: 866-900. (doi:10.1002/tesq.3027).

Saito, Kazuya and Luke Plonsky. 2019. “Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis”, *Language Learning* 69, 3: 652-708. (doi:10.1111/lang.12345).

Sakai, Mari and Colleen Moorman. 2018. “Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research”, *Applied Psycholinguistics* 39, 1: 187-224. (doi:10.1017/S0142716417000418).

Sawala, Krzysztof, Tomasz Szczegóła, Michał Jankowski and Jarosław Weckwerth. 2009. *Say It Right: A multimedia course in English pronunciation and vocabulary*. Poznań: Oficyna Wydawnicza Atena.

Schmidt, Richard W. 1990. “The role of consciousness in second language learning”, *Applied Linguistics* 11, 2: 129-158. (doi:10.1093/applin/11.2.129).

Schwartz, Geoffrey. 2005. “The lingua franca core and the phonetics-phonology interface”, in: Katarzyna Dziubalska-Kołaczyk and Joanna Przedlacka (eds.), *English pronunciation models: A changing scene.* Frankfurt: Peter Lang, 177-198.

Schwartz, Geoffrey and Kamil Kaźmierski. 2020. “Vowel dynamics in the acquisition of L2 English – an acoustic study of L1 Polish learners”, *Language Acquisition* 27, 3: 227-254. (doi:10.1080/10489223.2019.1707204).

Shadiev, Rustam and Mengke Yang. 2020. “Review of studies on technology-enhanced language learning and teaching”, *Sustainability* 12, 2: 524. (doi:10.3390/su12020524).

Slavuj, Vesna, Branko Kovačić and Ivana Jugo. 2015. ITSs for language learning. (Paper presented at the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, 25-29 May 2015.).

Sobkowiak, Włodzimierz. 2004. *English phonetics for Poles* (3rd ed.). Poznań: Wydawnictwo Poznańskie.

Sobkowiak, Włodzimierz. 2012. “This is tom = /ZYZYS’tom/ pronunciation in beginners’ EFL textbooks then and now”, *Research in Language* 10, 1: 111-122. (doi:10.2478/v10015-011-0028-7).

Spada, Nina and Patsy M. Lightbown. 2008. “Form-focused instruction: Isolated or integrated?”, *TESOL Quarterl* 42: 181-207. (doi:10.1002/j.1545-7249.2008.tb00115.x).

Speights Atkins, Marisha, Dallin J. Bailey, & Cheryl D. Seals. 2023. “Implementation of an automated grading tool for phonetic transcription training”, *Clinical Linguistics & Phonetics* 37, 3: 242-257. (doi:10.1080/02699206.2022.2048314).

Swartz, Merryanna L. and Masoud Yazdani (eds.). 1992. *ITSs for foreign language learning: The bridge to international communication*. Berlin: Springer-Verlag. (doi:10.1007/978-3-642-77202-3).

Sweet, Henry. 1899. *The practical study of languages*. New York, NY: Henry Holt.

Szpyra-Kozłowska, Jolanta. 2015. *Pronunciation in EFL instruction: A research-based approach*. Bristol: Multilingual Matters. (doi:10.21832/9781783092628).

Suzukida, Yui and Kazuya Saito. 2022. “What is second language pronunciation proficiency? An empirical study”, *System* 106: 102754. (doi:10.1016/j.system.2022.102754).

Tafazoli, Dara, Cristina Aránzazu Huertas Abril and María Elena Gómez Parra. 2019. “Intelligent language tutoring system: Integrating intelligent computer-assisted language learning into language education”, *International Journal of Information and Communication Technology Education* 15, 3: 60-73. (doi:10.4018/IJICTE.2019070105).

Tasso, Carlo, Danilo Fum and Paolo Giangrandi. 1992. “The use of explanation-based learning for modelling student behavior in foreign language tutoring”, in: Carlo Tasso and Danilo Fum (eds.), *ITSs for foreign language learning*. Berlin: Springer, 151-170. (doi:10.1007/978-3-642-77202-3_10).

Tejedor-García, Cristian. (2020). Design and evaluation of mobile computer-assisted pronunciation training tools for second language learning. [Unpublished Ph.D. dissertation, Universidad de Valladolid].

Thomson, Ron I. and Tracey M. Derwing. 2015. “The effectiveness of L2 pronunciation instruction: A narrative review”, *Applied Linguistics* 36, 3: 326-344. (doi:10.1093/applin/amu076).

Titterington, Jill and Sally Bates. 2018. “Practice makes perfect? The pedagogic value of online independent phonetic transcription practice for speech and language therapy students”, *Clinical Linguistics & Phonetics* 32, 3: 249-266. (doi:10.1080/02699206.2017.1350882).

Treher, Elizabeth N. 2011. *Learning with board games: Tools for learning and retention*. Minnesota: The Learning Key, Inc.

Vaissière, Jacqueline. 2003. New tools for teaching phonetics. (Paper presented at the 15th International Congress of Phonetic Sciences, 3-9 Aug. 2003.).

Venkatagiri, H. S. and John Levis. 2007. “Phonological awareness and speech comprehensibility: An exploratory study”, *Language Awareness* 16, 4: 263-277. (doi:10.2167/la417.0).

Walesiak, Beata. 2020. Technology for pronunciation training and accents? (Paper presented at the PronSIG Online Conference, 4 Oct. 2020).

Waniek-Klimczak, Ewa. 2011. “‘I am good at speaking, but I failed my phonetics class’ – Pronunciation and speaking in advanced learners of English”, in: Miroslaw Pawlak, Ewa Waniek-Klimczak and Jan Majer (eds.), *Speaking and instructed foreign language acquisition*. Bristol: Multilingual Matters, 117-130. (doi:10.21832/9781847694126-010).

Waniek-Klimczak, Ewa, Arkadiusz Rojczyk and Andrzej Porzuczek. 2015. “‘Polglish’ in Polish eyes: What English studies majors think about their pronunciation in English”, in: Ewa Waniek-Klimczak and Mirosław Pawlak (eds), *Teaching and researching the pronunciation of English. Second language learning and teaching.* Springer: Cham, 23-23. (doi:10.1007/978-3-319-11092-9_2).

Wong, Rita. 1987. *Teaching pronunciation: Focus on English rhythm and intonation*. London: Prentice-Hall.

Wrembel, Magdalena. 2005. Phonological metacompetence in the acquisition of second language phonetics [Unpublished doctoral dissertation, Adam Mickiewicz University].

Wrembel, Magdalena. 2007. “Metacompetence-based approach to the teaching of L2 prosody: Practical implications”, in: Jürgen Trouvain and Ulrike Gut (eds.), *Non-native prosody: Phonetic description and teaching practice*. Berlin: Mouton de Gruyter, 189-209.

Wrembel, Magdalena. 2011a. “Cross-modal reinforcements in phonetics teaching and learning: An overview of innovative trends in pronunciation pedagogy”, in: Wai-

Sum Lee and Eric Zee (eds.), *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong: City University of Hong Kong, 104-107.

Wrembel, Magdalena. 2011b. "Metaphonetic awareness in the production of speech", in: Miroslaw Pawlak, Ewa Waniek-Klimczak and Jan Majer (eds.), *Speaking and instructed foreign language acquisition*. Bristol: Multilingual Matters, 169-182. (doi:10.21832/9781847694126-013).

Wrembel, Magdalena. 2015. *In search of a new perspective: Cross-linguistic influence in the acquisition of third language phonology*. Poznań: Adam Mickiewicz University Press.

Wren, Percival Christopher. 1912. *The 'direct' teaching of English in Indian schools.* Bombay: Longmans, Green.

Wu, Chia-Jung, Gwo-Dong Chen and Chi-Wen Huang. 2014. "Using digital board games for genuine communication in EFL classrooms", *Educational Technology Research and Development* 62, 2: 209-226. (doi:10.1007/s11423-013-9329-y).

Zaorob, Maria Lucia and Elizabeth Chin. 2001. *Games for grammar practice: A resource book of grammar games and interactive activities*. Cambridge: Cambridge University Press.

Zawodniak, Joanna, Mirosław Pawlak, and Mariusz Kruk. 2021. "The role of grit among Polish EFL majors: A comparative study of 1st-, 2nd-, and 3rd-year university students", *Journal for the Psychology of Language Learning* 3, 2: 118-132. (doi:0.52598/jpll/3/2/8).

Zhang, Runhan and Zhou-min Yuan. 2020. "Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation", *Studies in Second Language Acquisition* 42, 4: 905-918. (doi:10.1017/S0272263120000121).

Abstract

This PhD dissertation investigates innovative methods for teaching metalinguistic awareness in second language (L2) learners of English, with a focus on pronunciation. The research is framed within the historical context of explicit pronunciation teaching (Baker 2018) and the revised Speech Learning Model by Flege and Bohn (2021). The studies focus on near-nativelike pronunciation, which—despite the need to focus on intelligibility over nativelike accuracy in general-purpose pronunciation instruction (Munro and Derwing 1995)—is often pursued by advanced learners at university level for personal or professional reasons (Pennington and Rogerson-Revell 2019).

The research is structured around three publications. The first publication (Łodzikowski and Jekiel 2019) shows that print-and-play board games can enhance classroom engagement and improve declarative knowledge of prosody. Learners reported increased engagement and a rise in quiz scores (formative metaphonological awareness assessment) by about 8 percentage points. When given a choice, 62% of learners preferred board games over extra exercises, and many replayed the games outside class, citing benefits such as extra practice, immediate feedback, and fun. The second publication (Łodzikowski 2021) demonstrates that an IPA-based transcription tool can improve phonological awareness, with regular use associated with a 5 percentage point increase on a summative metaphonological awareness assessment. The third publication (Łodzikowski, Foltz, and Behrens 2024) explores the integration of next-generation AI technologies like GPT-4, which support personalised learning experiences and deepen students' engagement through tailored feedback, despite occasional inaccuracies.

Publications 1 and 2 faced limitations such as sample biases and challenges in fully capturing variable relationships. Future research directions include developing more robust and diverse participant samples, extending behavioural data collection, and leveraging next-generation AI technologies to create innovative educational tools. This work underscores the potential of digital tools and AI in transforming L2 pronunciation instruction and metalinguistic awareness training. Moreover, it argues for a systematic increase in L2 teachers' AI literacy to fully harness these technologies' potential. Overall, this research highlights the potential of innovative tools to enhance language learning, providing insights for future studies and educational practices.

Streszczenie

Niniejsza rozprawa doktorska bada innowacyjne metody nauczania świadomości metajęzykowej u osób uczących się języka angielskiego jako drugiego języka (L2), ze szczególnym uwzględnieniem wymowy. Praca jest osadzona w historycznym kontekście bezpośredniego nauczania wymowy (Baker 2018) oraz modelu przyswajania mowy Flege i Bohn (2021). Badania koncentrują się na wymowie zbliżonej do rodzimej, do której dąży sporo zaawansowanych osób uczących się języka na poziomie uniwersyteckim z uwagi na osobiste lub zawodowe motywacje (Pennington i Rogerson-Revell 2019)—pomimo konieczności skupienia się na zrozumiałości wypowiedzi zamiast na dokładności zbliżonej do rodzimej (Munro i Derwing 1995).

Badania są zorganizowane wokół trzech publikacji. Pierwsza publikacja (Łodzikowski i Jekiel 2019) pokazuje, że gry planszowe typu wydrukuj-i-graj mogą zwiększyć zaangażowanie na zajęciach i poprawić deklaratywną wiedzę na temat prozodii. Zaobserwowano podwyższenie wyników testów świadomości metajęzykowej średnio o około 8 punktów procentowych. Co więcej, prawie dwie trzecie uczestników wskazało, że wolało gry planszowe od innych dodatkowych ćwiczeń, a wiele z nich grało w nie również poza zajęciami. Druga publikacja (Łodzikowski 2021) pokazuje jak narzędzie do transkrypcji oparte na międzynarodowym alfabecie fonetycznym (IPA) może poprawić świadomość metafonologiczną. Regularne korzystanie z narzędzia wiązało się z wzrostem wyników testów świadomości metajęzykowej średnio o 5 punktów procentowych. Trzecia publikacja (Łodzikowski, Foltz i Behrens 2024) bada możliwości zastosowania sztucznej inteligencji (SI) takiej jak GPT-4 do tworzenia spersonalizowanych treści edukacyjnych.

Publikacje 1 i 2 miały pewne ograniczenia, takie jak stronniczość próby i obserwacyjna formuła badań. Przyszłe badania powinny opierać się na większej próbie, bogatszym zbiorze danych oraz wykorzystaniu generatywnej SI do tworzenia innowacyjnych narzędzi edukacyjnych. Praca ta podkreśla potencjał cyfrowych narzędzi i SI w nauczaniu wymowy języka drugiego i treningu świadomości metajęzykowej. Ponadto, wzywa do systematycznego zwiększania kompetencji nauczycieli języka drugiego w zakresie SI, aby w pełni wykorzystać potencjał tych technologii. Podsumowując, rozprawa podkreśla potencjał innowacyjnych narzędzi w nauczaniu języków obcych, dostarczając wgląd w przyszłe badań i praktyki edukacyjne.

Appendix A: Author contribution statements for Publication 1 (Łodzikowski and Jekiel 2019)

Poznań, 6 July 2024

Author contribution statement

In reference to the publication:

Łodzikowski, Kacper and Mateusz Jekiel. 2019. "Board games for teaching English prosody to advanced EFL learners", *ELT Journal* 73, 3: 275–285. (doi:10.1093/elt/ccy059).

The co-authors hereby declare their contributions according to the CRediT (Contributor Roles Taxonomy) framework as follows:

- **Kacper Łodzikowski:** Conceptualization (equal); Writing – Original Draft Preparation (lead); Writing – Review & Editing.
- **Mateusz Jekiel:** Conceptualization (equal); Writing – Original Draft Preparation.

Authors

mgr Kacper Łodzikowski

..........Łodzikowski..........

dr Mateusz Jekiel

..........Jekiel..........

PhD Supervisor

prof. UAM dr hab. Piotr Gąsiorowski

....

Poznań, 6 lipca 2024 r.

Oświadczenie o wkładzie autora

W odniesieniu do publikacji:

> Łodzikowski, Kacper and Mateusz Jekiel. 2019. "Board games for teaching English prosody to advanced EFL learners", *ELT Journal* 73, 3: 275–285. (doi:10.1093/elt/ccy059).

Współautorzy niniejszym oświadczają swój wkład zgodnie z CRediT (Contributor Roles Taxonomy) w następujący sposób:

- **Kacper Łodzikowski:** Konceptualizacja (równy); Pisanie – opracowanie manuskryptu (główny); Pisanie – redakcja.
- **Mateusz Jekiel:** Konceptualizacja (równy); Pisanie – opracowanie manuskryptu.

Autorzy

mgr Kacper Łodzikowski

..........Łodzikowski..........

dr Mateusz Jekiel

..........Jekiel..........

Promotor pracy doktorskiej

prof. UAM dr hab. Piotr Gąsiorowski

....

Appendix B: Author contribution statements for Publication 3 (Łodzikowski et al. 2024)

Poznań, 6 July 2024

Author contribution statement

In reference to the publication:

> Łodzikowski, Kacper, Peter W. Foltz, and John T. Behrens. 2024. “Generative AI and its educational implications”, in: Dora Kourkoulou, Anastasia Tzirides, Bill Cope, and Mary Kalantzis (eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines*. Postdigital Science and Education. Cham: Springer.

The co-authors hereby declare their contributions according to the CRediT (Contributor Roles Taxonomy) framework as follows:

- **Kacper Łodzikowski:** Conceptualization (equal); Writing – Original Draft Preparation (lead); Writing – Review & Editing (lead).
- **Peter W. Foltz:** Conceptualization (equal); Writing – Original Draft Preparation (equal); Writing – Review & Editing (equal).
- **John T. Behrens:** Conceptualization (equal); Writing – Original Draft Preparation (equal); Writing – Review & Editing (equal).

Authors

mgr Kacper Łodzikowski

....

Peter W. Foltz, Ph.D.

....

John T. Behrens, Ph.D.

....

PhD Supervisor

prof. UAM dr hab. Piotr Gąsiorowski

....

Poznań, 6 lipca 2024 r.

Oświadczenie o wkładzie autora

W odniesieniu do publikacji:

> Łodzikowski, Kacper, Peter W. Foltz, and John T. Behrens. 2024. "Generative AI and its educational implications", in: Dora Kourkoulou, Anastasia Tzirides, Bill Cope, and Mary Kalantzis (eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines*. Postdigital Science and Education. Cham: Springer.

Współautorzy niniejszym oświadczają swój wkład zgodnie z CRediT (Contributor Roles Taxonomy) w następujący sposób:

- **Kacper Łodzikowski:** Konceptualizacja (równy); Pisanie – opracowanie manuskryptu (główny); Pisanie – redakcja (główny).
- **Peter W. Foltz:** Konceptualizacja (równy); Pisanie – opracowanie manuskryptu (równy); Pisanie – redakcja (równy).
- **John T. Behrens:** Konceptualizacja (równy); Pisanie – opracowanie manuskryptu (równy); Pisanie – redakcja (równy).

Authors

mgr Kacper Łodzikowski

.....Łodzikowski.....

Peter W. Foltz, Ph.D.

.....Peter W. Foltz.....

John T. Behrens, Ph.D.

....

PhD Supervisor

prof. UAM dr hab. Piotr Gąsiorowski

....