


ARTICLE

A learning perspective on the emergence of abstractions: the curious case of phone(me)s

Petar Milin¹ , Benjamin V. Tucker^{2,3} and Dagmar Divjak^{1,4}

¹Department of Modern Languages, University of Birmingham, Birmingham, UK; ²Department of Linguistics, University of Alberta, Edmonton, AB, Canada; ³Department of Communication Sciences and Disorders, Northern Arizona University, Flagstaff, AZ, USA; ⁴Department of English Language and Linguistics, University of Birmingham, Birmingham, UK

Corresponding author: Petar Milin; Email: p.milin@bham.ac.uk

(Received 17 September 2022; Revised 16 March 2023; Accepted 03 April 2023)

Abstract

In this study, we propose an operationalization of the concept of *emergence* which plays a crucial role in usage-based theories of language. The abstractions linguists operate with are assumed to emerge through a process of generalization over the data language users are exposed to. Here, we use two types of computational learning algorithms that differ in how they formalize and execute generalization and, consequently, abstraction, to probe whether a type of language knowledge that resembles linguistic abstractions could emerge from exposure to raw data only. More specifically, we investigated whether a phone, undisputedly the simplest of all linguistic abstractions, could emerge from exposure to speech sounds using two computational learning processes: memory-based learning and error-correction learning (ECL). Both models were presented with a significant amount of pre-processed speech produced by one speaker. We assessed (1) the consistency or stability of what these simple models learn and (2) their ability to approximate abstract categories. Both types of models fare differently regarding these tests. We show that only ECL models can learn abstractions and that at least part of the phone inventory and its grouping into traditional types can be reliably identified from the input.

Keywords: Error-correction learning; memory-based learning; phone; emergence; abstraction

1. Introduction

Many theories of language presuppose the existence of abstractions that aim to organize the extremely rich and varied experiences language users have. These abstractions are thought either to be innate (generative theories) or to arise from experience (emergentist, usage-based approaches). Several mechanisms or principles by which abstractions could form have been identified. The mechanisms that are of particular interest for this study include *grouping* similar experiences together,



forgetting some dimensions of experience, and *filtering* uninformative cues by not attending to certain information streams to manage the influx of information. In essence, the general assumption is that the vast amount of experience with raw language needs to be reduced to its useful and usable core. The listed mechanisms – grouping, forgetting, and filtering – should not be considered mutually exclusive, nor is this list comprehensive as it focuses mostly on principles that have appeared attractive for linguistic theorizing and research (e.g., Milin et al., 2016) and for which computational implementations exist.

Following the ‘quantitative turn’ in the early 2000s, we are currently embarking on a ‘computational turn’ in which usage-based linguists engage with computational models to put their theories to the test (Divjak & Milin, 2023). Among other things, this turn presents us with an opportunity for investigating in detail core processes by which linguistic knowledge comes about, including the process of abstracting linguistic knowledge from exposure to usage. In this study, we take the English sound system as case study and model computationally whether an abstract phone could emerge from supervised exposure to speech sounds. We simulate this process using two machine learning principles: memory-based learning (MBL) and error-correction learning (ECL); the latter being further developed into two closely related learning rules: Widrow–Hoff (WH) and temporal difference (TD) learning (a direct generalization of WH). Both algorithms are simple, and therefore their inner workings are easy to trace: they consist of two layers only, input and output, and connection weights between those layers. Hence, it is much easier to understand how these algorithms work and what they do than is the case for more sophisticated, deep(er) networks (many of which, actually, consist of layers of WH-like units; cf. Widrow & Lehr, 1990). They have also been proven successful in modeling language phenomena as disparate as allomorphy, regular/irregular forms, alternations, and the processing of morphologically complex words, including compounds (for a brief overview, see Divjak & Milin, 2023 and the references therein). Seminal work of Bybee and McClelland (2005) brings together usage-based linguistic theory and parallel distributed processing (PDP) or connectionist computational modeling, which typically relies on more complex architectures built from WH-like units (see also Bybee, 2013; McClelland et al., 2010; McClelland & Bybee, 2007).

In what follows, we summarize the notion of linguistic abstraction, and present phones, the focus of the present study, in some detail. In Section 1, we describe the opposing learning principles and their implementation. Then, in Section 2, we present the data, the methodology, and the tasks the learning algorithms are compared on. In Section 3, we present the details of the performance, and algorithms’ capability to learn the linguistic abstraction. This is followed, in Section 4, by a summary of findings and a discussion of the implications for usage-based linguistics.

1.1. Abstractions of sounds

In linguistic theory, abstractions are either innate and present at birth, as is the case in generative frameworks, or they have emerged from exposure to the ambient language, as is assumed in usage-based approaches (for an overview of these theories and the predictions they make regarding language acquisition or development, see Ambridge & Lieven, 2011). This implies that abstractions play a different role in both theoretical approaches. The difference could be said to lie in the sequence: on a generative approach, in a sense, abstractions give rise to usage, whereas on a usage-based approach, usage gives rise to abstractions.

The long-standing practice of describing language and its structures by relying on abstractions has led to the assumption that language knowledge and use must involve knowledge of the relevant abstractions (but see Ambridge, 2020). The challenge facing those who assume the existence of abstractions is one of proposing a hypothesized construct or mechanism that would give rise to these abstractions. For nativists, this means explaining how these abstractions could be present at birth, and for emergentists, this means explaining how such abstractions can emerge from exposure to usage (Pierrehumbert, 2016).

In this article, we focus on abstractions of sounds. The most commonly proposed and assumed building block or unit of abstraction in this area is the phoneme or syllable (Goldinger & Azuma, 2003). Generally, phonemes are defined as abstractions of speech sounds that ignore or ‘abstract away’ details of the speech signal that do not contribute to differentiating meaning. Yet phonemes have come under fire more than once in recent history. A recurring claim in the argument against the cognitive reality of phonemes is the richness of memories of experiences, including experiences of sounds (cf. Port, 2010). Although we do not deny the validity of the findings reported (cf. Goldinger & Azuma, 2003; Port, 2007), we do not consider these findings as evidence that there cannot possibly be any cognitive reality to phoneme-like units. Instead, we accept them as an indication that phoneme-like units are not necessarily the whole story. Phonemes are not the only abstract cognitive representation of sound either: abstractions are likely language-dependent and even speaker-dependent, and they can range from what might traditionally be referred to as features, gestures, moras, allophones, phonemes, demi-syllables, syllables, words, and sequences of words (e.g., Browman & Goldstein, 1992; Goldinger & Azuma, 2003; Liberman & Mattingly, 1985; Tremblay & Tucker, 2011). We chose not to use the term *phoneme* to refer to the speech sounds in our study to steer away from the assumptions associated with this term. Henceforth, we will use the term *phone* to refer to the speech sounds we are working with.

1.2. To learn or not to learn

For the purposes of this study, we focus on a specific subset of linguistic theories and the computational frameworks (i.e., models and algorithms) commonly used to support these theories. Linguistically, we assume that universal sound inventories are not available at birth but are learned using general cognitive mechanisms. Psychologically, we contrast models that assume veridical encoding of the input with those that allow or even encourage information sifting. In terms of machine learning approaches, this opposition would contrast *grouping* with *filtering* principles. More concretely, we position MBL against ECL. Much more sophisticated models are available, and detailed neural mechanisms for learning and inference in complex probabilistic models, claimed to be biologically and cognitively plausible, have been proposed (e.g., Haefner et al., 2016; Vértes & Sahani, 2019). However, our selection of algorithms is based on the linguistically appealing and, thus, popular choice (MBL) and the simplicity of the competitor to that popular choice (ECL). Furthermore, because both algorithms are simple, they are also transparent, which allows us to retain control over all phases of the process of interest and consequently derive knowledge regarding that process.

Technically, most (if not all) MBL approaches store past experiences as correct input–output pairs: the input is a (vector) representation of the environment and the system, whereas the output assumes an action or an entity – a desired outcome (see Haykin, 1999). Memory is organized in neighborhoods of similar or proximal inputs, which presumably elicit similar actions or entities. Thus, a memory-based algorithm consists of (1) a criterion for defining the neighborhoods (similarity and proximity) and (2) a learning principle that guides the assignment of a new exemplar to the right neighborhood. One simple yet efficient example of such a principle is that of Nearest Neighbors, where a particular type of distance (say, the Euclidean distance) is minimal in a given group of (near) neighbors. The principle appeals to linguists, and usage-based linguists in particular, as it equates learning to categorization or classification of newly encountered items into *neighborhoods of alikes*: new items (exemplars) are processed on the basis of their similarity to the existing neighborhoods, which resembles the process of analogical extension. And, for usage-based linguists, and cognitive linguists in particular, categorization is one of the foundational, general cognitive abilities that facilitate learning from input (Taylor, 1995).

ECL runs on very different principles than MBL. The algorithm gradually builds connections or associations between input information (a signal and a cue) and an outcome (a target). A well-known example is the association of the sound of a bell and the presentation of food, initially documented in Ivan Pavlov's lab (Pavlov, 1927). The process of updating the weight of the cue for an outcome gravitates toward the least erroneous identification of a desired outcome. The process unfolds in a stepwise manner, as new input information becomes available, and the weighting or evaluation of the current success is done relative to a mismatch between the *current state of learning* (i.e., the current best guess) and the *true state* (i.e., the target state or the desired outcome). That mismatch induces a small disturbance or change in the association weights which, effectively, evaluates the input data themselves to bring the current state closer to the true state. In the domain of speech production, for example, Anderson et al. (2019) established experimentally that learning is more efficient when productions have more errors. This indirectly shows that learning of phonotactic patterns specifically is triggered by (the need for) error correction.

The smallest common denominator of the MBL and ECL approaches is their ability to learn from the environment for the sake of better adaptation. Under such an assumption, rather simplified, a system learns by processing input information about the environment, and about itself in relation to that environment (Rescorla, 1988; Skinner, 1974). However, MBL and ECL could be seen as opposing principles: in the extreme, an exemplar approach does not require anything beyond organized (i.e., grouped) storage, whereas an error-correction approach is a stockpile of perpetually filtered associations.

As mentioned above, to researchers interested in linguistics, and usage-based linguistics in particular, MBL appeals because it explains linguistic knowledge in terms of categorization and analogical extension. An MBL approach also appeals as a straightforwardly suitable computational implementation of exemplar-based models of language (e.g., Daelemans & Van den Bosch, 2005; Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2006). However, ECL appears to be conceptually better suited for testing the emergentist perspective on language knowledge as it is directly revealing about the process itself – the emergence of knowledge (see Julià, 1983 about the importance of understanding the process and not just the end-state for linguistic theorizing; and see Bybee & McClelland, 2005 for a pioneering modeling attempt). In

other words, learners may disregard the aspects of the input that do not allow them to build up reliable associations, keeping only what is *useful* (sufficient and necessary) in an adaptive sense.

1.3. *This study: pursuing that which is learnable, with some caveats*

In this article, we contrast a model that stores the auditory signal in memory with one that learns to ‘find’ (i.e., filter out) phones in the auditory signal. Modeling whether an abstract phone could emerge from exposure to speech sounds by applying the principle of ECL crucially relies on the concept of unlearning, whereby the strength of the links between any cues that do not co-occur with a certain outcome are reduced (see Ramscar et al., 2013 for discussion; see also Nixon & Tomaschek, 2020, who trained a naïve discriminative learning model relying on the Rescorla–Wagner learning rule, to form expectations about a subset of phones from the surrounding speech signal). The principle is such that the model gradually tries to unlearn irrelevant variations in the speech input, to discard or filter out the dimensions of the original experience and to retain more abstract and parsimonious ‘memories’. This strategy is the radical opposite of a model that presupposes that we store our experiences veridically, that is, in their full richness and idiosyncrasy. On the latter account, learning equals encoding, storing, and organizing that storage to facilitate retrieval.

We provide our models with a common input in automatic speech recognition, mel-frequency cepstral coefficients (MFCCs), derived from the acoustic signal, and with phones as outcomes. We assign our models the task of storing vs. learning to associate the two. This implies that, in a very real sense, *our modeling is artificial*; our models learn phoneme-like units because those are the outputs that we provide. If we provided other types of units, other types of outputs might be learned. Yet this does not annul our finding that phones can be successfully learned from input. Recall also that we do not claim that phonemic variation would be something, let alone everything, that listeners identify. Skeptics might argue that an explicitly supervised learning scenario for the acquisition of phonetic abstraction is, in itself, cognitively unrealistic. We agree with this claim, in principle at least, assuming that the irrelevance is established from the point of view of cognitive (neuro) science and language learning in the wild. But that does not make the task irrelevant to our lives (see, e.g., Graves & Schmidhuber, 2005): the process of learning to read includes, for example, learning to identify phone(me)s and to syllabify the printed words (e.g., Nesdale et al., 1984; Tunmer & Rohl, 1991).

Our learning-based approach can be discussed in terms of similarities and differences to the recent mechanism-driven approach proposed in Schatz et al. (2021). Their findings suggest that distributional learning might be the right approach to explain how infants become attuned to the sounds of their native language, which is the assumption we rely on in the present study. The important difference, however, is that Schatz et al.’s (2021) proposed learning algorithm did not try to achieve any biological or cognitive realism; it focused purely on the power of generative machine learning techniques. As explained in Section 1.2, we run our simulations on much simpler learning algorithms. While using state-of-the-art learning algorithms, Schatz et al. (2021) concluded that plausible or realistic data can generate a learnable signal (i.e., in terms of the joint probability distribution,

given a problem space), we now move to make the crucial step to simultaneously test (1) plausible learning mechanisms against (2) realistic data.

As said, learnability is our minimal requirement *and* our operationalization of cognitive reality. On a usage-based, emergentist approach to language knowledge, abstractions must be learnable from the input listeners are exposed to if these abstractions are to lay claims to cognitive reality (Milin et al., 2016). Whereas earlier work explored whether abstract labels can be learned and whether they map onto distributional patterns in the input (cf. Divjak et al., 2015), more recent work has brought these strands together by modeling directly how selected abstract labels would be learned from input using a cognitively realistic learning algorithm (cf. Milin et al. 2017a).

Of course, because we use a simple algorithm, performance differences with state-of-the-art classifiers should be expected. Yet, what we trade in, in terms of performance, we hope to gain in terms of (a) interpretability of the findings relevant to natural classes of phonemes, and (b) continuity with productive language research traditions that make use of MBL and/or ECL (for a brief overview, see Divjak & Milin, 2023). Rather than maximizing performance, we minimize two critical ‘investments’ (1) in cultivating the input and (2) in algorithmic sophistication. While the former creates a problematic detachment from what exposure to realistic data ought to be, the latter increases computational intractability, for example, given the number of parameters. For these reasons, we remain as naïve as possible and engage with MFCCs directly as input while using two of the simplest yet surprisingly efficient computational principles of learning.

2. Methods

In this section, we describe the speech materials and how they were pre-processed before we move on to presenting the details of the modeling procedures. Two caveats are in order. First, our methodology consists of a carefully chosen set of worst-case scenarios, that is, computationally challenging learning situations that would assume hearing one speaker only, or learning from data with strictly random noise. In this way, we advance our understanding of both the data and the learning mechanisms. Second, while we use phones to refer to the sounds we are working with, we are manipulating phoneme-like units out of convenience and not out of any theoretical leaning. We believe that similar results would be obtained with any number of other possible levels of representation.

2.1. Speech material

The phones used for modeling are taken from the force-aligned stimuli in the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2019). This database contains 26,793 words stimuli and 9,592 pseudo-words produced by a single male speaker from Western Canada; together, this amounts to approximately 5 hours of speech. The details about the pre-processing and the data are available at <http://mald.artsrn.ualberta.ca/>. Note that, when discussing individual sounds, we use the ARPABET transcription scheme.

In order to prepare the files for modeling, we first set the sampling frequency to 16 kHz. We then windowed and extracted MFCCs for all of the words using 25 ms long

overlapping windows in 10 ms steps over the entire acoustic signal. By way of example, if there was a 100 ms signal, it would be divided into nine overlapping windows with some padding at the end to allow the windows to fit; the identity of the phone was determined by the phone occurring at the center of the window. MFCCs are summary features of the acoustic signal used commonly in automatic speech recognition, created by applying a mel transformation – a non-linear transform meant to approximate human hearing, to the frequency components of the spectral properties of the signal (Mermelstein, 1976). The mel frequencies are then submitted to a Fourier transform, so that, in essence, we are taking the spectrum of a spectrum known as the cepstrum. Following standard practice in automatic speech recognition, we used the first 13 coefficients in the cepstrum, calculated the first derivative (delta) for the next 13 coefficients, and the second derivative (delta–delta) for the final set of 13 coefficients (cf. Holmes & Holmes, 2002). This process results in each window or trial having 39 features; thus, for each word, we created an MFCC matrix of the 39 coefficients and the corresponding phone label. We chose these acoustic representations as they have been quite successful in speech technology applications; they are a convenient way to extract numeric summary information from the signal for computational modeling, but we do not believe that the brain processes the acoustic signal in this way.

2.2. Algorithms

As said, MBL and ECL can be positioned as representing opposing principles in learning and adaptation, with the former relying on storage of otherwise unaltered input–outcome experiences, and the latter using filtering of the input signal to generate a maximally accurate (minimally erroneous) approximation of the true outcome. Importantly, however, we do not see either algorithm as something the human brain would implement; rather, these algorithms have been devised to capture the functions or operations the brain *might* be carrying out. In other words, we assume that the algorithms discussed in the present study (or any other for that matter) can be considered plausible candidates for describing the workings of the brain at a higher, conceptual level (i.e., Marr’s (1982) computational level of what the system does and the algorithmic level of how it achieves that). In that same sense, our research focus contrasts with that of automatic speech recognition systems which are designed for optimal accuracy.

We applied MBL by populating neighborhoods with exemplars that are corralled by the smallest Euclidean distance, given exemplars of MFCC inputs. Then, we assigned new MFCC input to the neighborhood of least distant neighbors. This is, arguably, a rather typical MBL implementation, as Euclidean distance is among the most commonly used distance metrics, while k -nearest neighbors is a straightforward yet efficient principle for storing new exemplars (Haykin, 1999). For the purpose of the current study, all MBL trainings and tests utilized the `class` package (version 7.3-17; Venables & Ripley, 2013) in the **R** software environment (R Core Team, 2021).

ECL learning was implemented using two different models: WH (Widrow & Hoff, 1960) and TD (Sutton & Barto, 1987, 1990). Both are versions of a closely related idea: WH (also known as the Delta rule or the Least Mean Square rule; see Milin et al., 2020; Sutton & Barto, 1981 for further references therein) represents the simplest error-correction rule; TD generalizes to a rule that accounts for future rewards, too.

In non-technical terms, at each discrete time step, when new data are presented, learning makes small changes to the association weights between input cues (MFCCs) and an outcome (a phone). The change depends on the system's learning state: which outcome would be favored given the current cues and their existing association weights. The change is small to reflect the trustworthiness of 'knowledge' that is gradually accrued (in machine learning terms, this is expressed as the *principle of minimal disturbance*; cf. Haykin, 1999). Finally, the change occurs to ensure a closer match between the current state of learning and the outcome on its next encounter. In both ECL rules, mismatches lead to dissociating cues and outcomes, whereas matches strengthen cue–outcome associations (for details, see Gershman, 2015). Crucially, many cues compete to become associated with an outcome, to gain stronger weights for that outcome. In time, many cues become light-weighted, and get filtered out because they are irrelevant. Those cues that showed systematicity in their co-occurrence with an outcome weigh the most. Both ECL algorithms were developed as a combination of routines in Python (version 3.9) for data preparation, and MATLAB (version R2021a) for the trainings and tests.

For WH, the future does not exist, and the weight correction happens based on the present cues and outcomes. TD, however, looks further into the future, as the current state of learning state is *discounted* by what is yet to come. In other words, TD's present state of 'knowledge' is additionally corrected as the system also 'knows' that the current data are not all there are, but that future data are expected too. This simple idea turns into a rather elegant extension of the WH rule since, if TD would completely discount future data, the opportunity to learn from the future would also cease to exist (i.e., become void), and TD would reduce to its simpler cousin, WH. Our decision to use two closely related ECL algorithms, WH and TD, is to allow for a detailed examination of the effect that temporally overlapping data (see Section 2.1) might have on error-correction itself. If the data in the next (temporally overlapping) time step are informative for forming accurate expectations, then the TD rule, by explicitly using those future data, should show an advantage over the 'greedy' WH that does not consider anything beyond the now. In the context of the present study, we applied only the nearest future data, that is, the MFCC acoustic input that follows the current input. We also probed a two-step lookahead (i.e., two following acoustic inputs), but that did not improve the performance of the model.

2.3. Training

We ran several computational simulations of learning English phones directly from the raw MFCCs which were used as input. This constitutes a 40-way classification task, with 39 English phones and a category of silence. The full learning sample consisted of 21,061 words (1,197,862 overlapping acoustic inputs) that were split into 18,954 (1,078,200 inputs) training words and 2,107 (119,662 inputs) test words (i.e., approximately 90% and 10% split). Each learning trial represents a single MFCC acoustic signal (25 ms window extracted from the speech signal in 10 ms steps) and is matched with a phone label.

The first training regime used the training data trials as they are, that is, raw. At each trial, input cues were represented as an MFCC row vector, and the outcome was the corresponding 1-hot encoded (i.e., present) phone. The raw input, hence, consisted of MFCC inputs overlapping in time. The input was randomized at a word

level (following the argument in Harmon & Kapatsinski, 2021). The second training regime used 100 randomly selected occurrences of each phone and a silence, again after word-level randomization. As the number of MFCC inputs representing a single phone can vary across words (since a phone can have a different duration), we used all such inputs while controlling for the total number of phone occurrences ($n = 100$). The average number of MFCCs per phone was $M = 823.6$, ranging from $min = 453$ (/ah/) to $max = 1,335$ (/ay/). The total number of acoustic inputs for this random sample was 32,944 MFCC vectors, each coupled with the 1-hot encoded phone outcome. As we accounted for all within-phone inputs, they still overlapped in time, the same as for the raw sample. Finally, the third training sample started from the same 90% of training data from which we calculated mean and standard error per phone and silence. We used these two measures to generate 100 normally distributed data points (Gaussian) for each phone and silence, given a 99.98% interval of the standard error of the mean (SE). In sum, the raw training dataset consisted of 1,078,200 trials, the random training dataset of 32,944 trials, and the Gaussian data of 4,000 trials. Across all samples, a learning event was a vector of MFCC acoustic inputs and a 1-hot encoded vector of phone outcome. The sheer size of the raw training dataset could benefit the MBL model. The raw and random datasets could benefit TD, given the temporally overlapping acoustic inputs (MFCCs). In addition to that, the random dataset also kept the number of phone exemplars constant, which could benefit the ECL models as they tend to overweight token frequency (cf. Caballero & Kapatsinski, 2022; note that some connectionist PDP models applied frequency ‘flattening’ to account for this issue (see, e.g., Plunkett & Juola, 1999). Finally, the Gaussian data establish the base performance of the chosen learning rules (MBL, WH, and TD) when the data are simulated as normally distributed.

The possible combination of the types of data (raw, random, and Gaussian) and the learning models (MBL, WH, and TD) gave us a total of nine different results against the same test data (10% of all data; i.e., 2,107 words or 119,662 input–output trials). The tests were the same for all three algorithms: the prediction of an unseen phone outcome was made using the new MFCC acoustic input and then compared with the correct phone. This resulted in nine confusion matrices, for three types of data and three learning models. For the MBL model, the predictions were made using a majority vote given the k -nearest inputs (MFCC vectors) stored during the training phase. The ECL models’ (WH and TD) choices were determined given the input cue activation, that is, the weighted sum of the new input (i.e., also called total or net input). The activation was additionally pitted against the diversity of weights given the new input (the absolute length or 1-norm; for details, see Milin et al., 2017b), which indicates the amount of competition among possible outcomes. Hence, the ratio of activation over diversity for ECLs was used to put support (activation) against competition (diversity). Crucially, during the test phase, the unseen data did not lead to further weight or neighborhood updates.

2.4. Tuning the free parameters of learning

We applied a grid search to find the best values of the free parameters for the three learning models (MBL, WH, and TD). The MBL algorithm has the number of neighbors k as free parameter, and we cast the search net reasonably wide, from 3 to 40 ($k = [3, 40]$). For the highest accuracies across datasets, k -values were

as follows: 29 for raw, 30 for random, and 4 for Gaussian. The difference between k -values for raw and random vs. Gaussian is large, and the k -values of 29 and 30 neighbors might not be cognitively plausible (anecdotally, one can think of Miller's (1956) magical number 7; see also Baddeley, 1994). We will return to this issue in our general discussion.

The WH algorithm also has one free parameter: learning rate (λ). Given the considerable range of MFCC values ($min = -90.41$, $max = 68.73$, $range = 159.14$), the learning rate needed to be a small value to avoid the risk of overflow. We set the search between $6e-09$ and $5e-07$ ($\lambda = [6e-09, 5e-07]$) in miniscule increments. The best performances (accuracies) for different datasets were achieved with the λ -values of $\approx 2.5e-07$ on raw, $\approx 4.0e-08$ on random, and $\approx 6.0e-08$ on Gaussian.

Finally, the TD algorithm has two free parameters: learning rate (λ) and discount factor (γ) for future rewards (for technical details, see Appendix A at <https://doi.org/10.25500/edata.bham.00000942>). As TD represents a generalization of the WH algorithm, the learning rate in both algorithms serves the same purpose. Thus, we used the same learning rates for TD as the one found with grid search for WH. Then, we grid-searched for the discount factor from 0 to 1 ($\gamma = [0.0, 1.0]$), again in small increments. To our surprise, for both raw and Gaussian datasets, the best performance of TD was achieved when $\gamma = 0.0$, whereas the random dataset suggested $\gamma = 0.17$. This will be considered further in the general discussion.

3. Results and discussion

The following two subsections contain independent but related pieces of evidence for assessing the learnability of abstract categories, given MFCC input, and the performance of the three chosen algorithms (MBL, WH, and TD) under different learning scenarios. Even though in two out of three cases TD was formally identical to the WH algorithm (with the discount factor set to $\gamma = 0.0$, thus ignoring future data), we will report the results for all three learning models explicitly. This enables us to present the models' performance with different datasets and discuss whether they required changes in the values of their respective free parameters. For TD learning, specifically, it seems important to know if and when there are benefits from future rewards.

First, we examine horizontal generalizability, that is, whether phones are learnable can be inferred from the algorithm's success in handling new and unseen data from the same speaker (Section 3.1). We also examine whether and how well our learning models capture the assumed 'natural' classes; that is, do we see and, if so, to what extent do we see an emergence of plausible phone clusters, resembling traditional classes, directly from the learned weights (Section 3.2).

3.1. More of the same: predicting new input from the same speaker

We can consider our simulations as, essentially, a 3×3 experimental design, with two factors: learning model and type of input data. The success rates of our three algorithms (MBL, WH, and TD) are summarized in Table 1. These rates represent the achievement in predicting unseen data, given the input (raw, random, and Gaussian).

Overall, the results reveal an above chance performance, with chance assessed as either the naive or 'flat' probability of any given phone occurring by chance, which

Table 1. Success rates, in percentages, of predicting the correct phone for three different learning models (WH, TD, and MBL) using three different types of input data (raw, random, and Gaussian). The leftmost column represents the probability of a given phone in the raw test sample. Random data have a uniform probability distribution of phones (100 randomly chosen occurrences) while still allowing for difference in the durations of each sampled phone. Gaussian data have a fully uniform probability distribution of phones (100 generated datapoints)

| Phone | Occurrence in test sample | Raw | | | Random | | | Gaussian | | |
|----------------------|---------------------------|--------|-------|-------|--------|-------|-------|----------|-------|-------|
| | | MBL | WH | TD | MBL | WH | TD | MBL | WH | TD |
| aa | 1.94 | 52.00 | 69.36 | 69.36 | 43.10 | 23.20 | 22.50 | 33.32 | 48.01 | 48.01 |
| ae | 2.60 | 66.06 | 32.76 | 32.76 | 56.78 | 42.17 | 44.91 | 53.25 | 51.13 | 51.13 |
| ah | 4.93 | 48.70 | 17.33 | 17.33 | 41.07 | 8.74 | 9.23 | 18.48 | 16.15 | 16.15 |
| ao | 1.42 | 56.45 | 20.44 | 20.44 | 39.35 | 38.13 | 36.91 | 27.29 | 33.86 | 33.86 |
| aw | 0.62 | 58.05 | 5.21 | 5.21 | 14.16 | 23.57 | 23.82 | 9.21 | 22.83 | 22.83 |
| ay | 2.43 | 61.08 | 63.11 | 63.11 | 41.28 | 37.89 | 35.13 | 30.94 | 25.88 | 25.88 |
| b | 1.18 | 69.03 | 12.61 | 12.61 | 41.50 | 26.64 | 27.91 | 18.35 | 25.97 | 25.97 |
| ch | 0.74 | 37.47 | 62.13 | 62.13 | 13.32 | 49.29 | 49.42 | 9.66 | 53.18 | 53.18 |
| d | 2.48 | 63.21 | 14.83 | 14.83 | 55.87 | 7.00 | 7.14 | 16.18 | 13.04 | 13.04 |
| dh | 0.06 | 100.00 | 12.28 | 12.28 | 2.16 | 42.11 | 40.35 | 0.95 | 10.53 | 10.53 |
| eh | 2.16 | 51.01 | 9.16 | 9.16 | 30.73 | 39.19 | 38.78 | 26.06 | 30.00 | 30.00 |
| er | 3.83 | 60.87 | 28.24 | 28.24 | 44.67 | 37.28 | 39.44 | 42.71 | 38.45 | 38.45 |
| ey | 2.36 | 52.51 | 10.75 | 10.75 | 34.31 | 63.11 | 64.30 | 31.88 | 56.43 | 56.43 |
| f | 1.70 | 57.73 | 47.84 | 47.84 | 34.13 | 51.13 | 53.99 | 34.37 | 57.41 | 57.41 |
| g | 0.77 | 80.22 | 12.43 | 12.43 | 59.95 | 16.47 | 17.85 | 14.08 | 12.01 | 12.01 |
| hh | 0.49 | 59.46 | 29.25 | 29.25 | 20.11 | 26.88 | 27.31 | 7.16 | 21.08 | 21.08 |
| ih | 3.82 | 48.75 | 62.56 | 62.56 | 38.36 | 6.53 | 6.50 | 21.66 | 17.26 | 17.26 |
| iy | 4.31 | 71.02 | 54.22 | 54.22 | 67.02 | 32.08 | 34.68 | 58.18 | 44.09 | 44.09 |
| jh | 0.78 | 63.17 | 20.00 | 20.00 | 26.90 | 29.42 | 28.06 | 18.34 | 28.06 | 28.06 |
| k | 3.83 | 77.78 | 21.01 | 21.01 | 76.03 | 6.69 | 8.22 | 38.45 | 12.38 | 12.38 |
| l | 5.74 | 74.53 | 65.23 | 65.23 | 78.70 | 25.01 | 29.10 | 73.12 | 30.14 | 30.14 |
| m | 2.73 | 79.00 | 49.81 | 49.81 | 68.75 | 31.96 | 32.62 | 47.92 | 39.99 | 39.99 |
| n | 5.49 | 73.55 | 60.33 | 60.33 | 75.46 | 17.18 | 19.30 | 69.67 | 24.41 | 24.41 |
| ng | 1.63 | 75.61 | 51.10 | 51.10 | 39.71 | 36.08 | 36.20 | 35.16 | 23.49 | 23.49 |
| ow | 1.61 | 56.14 | 32.59 | 32.59 | 38.70 | 39.67 | 40.05 | 38.82 | 37.82 | 37.82 |
| oy | 0.21 | 46.67 | 0.40 | 0.40 | 6.01 | 38.55 | 38.15 | 5.53 | 30.12 | 30.12 |
| p | 2.15 | 60.96 | 2.32 | 2.32 | 36.72 | 2.45 | 1.94 | 10.43 | 3.37 | 3.37 |
| r | 3.65 | 65.73 | 33.23 | 33.23 | 61.59 | 21.85 | 21.22 | 41.65 | 19.42 | 19.42 |
| s | 9.51 | 70.75 | 63.32 | 63.32 | 73.19 | 59.19 | 60.33 | 66.62 | 64.31 | 64.31 |
| sh | 1.88 | 59.55 | 10.02 | 10.02 | 45.25 | 23.02 | 23.34 | 55.40 | 22.93 | 22.93 |
| t | 5.10 | 62.19 | 20.90 | 20.90 | 60.40 | 14.14 | 12.70 | 18.64 | 15.27 | 15.27 |
| th | 0.29 | 61.54 | 14.40 | 14.40 | 8.79 | 10.47 | 9.95 | 4.23 | 7.59 | 7.59 |
| uh | 0.18 | 37.93 | 45.45 | 45.45 | 6.42 | 10.45 | 9.09 | 2.36 | 8.64 | 8.64 |
| uw | 0.94 | 60.74 | 25.97 | 25.97 | 32.11 | 33.30 | 32.76 | 40.47 | 35.11 | 35.11 |
| v | 0.99 | 52.82 | 8.51 | 8.51 | 30.06 | 8.97 | 8.04 | 19.79 | 14.54 | 14.54 |
| w | 0.78 | 78.40 | 18.78 | 18.78 | 46.17 | 16.36 | 15.10 | 22.16 | 22.85 | 22.85 |
| y | 0.43 | 64.32 | 41.42 | 41.42 | 29.24 | 11.13 | 10.40 | 9.54 | 8.39 | 8.39 |
| z | 5.00 | 64.20 | 42.81 | 42.81 | 45.63 | 30.31 | 31.12 | 38.11 | 23.54 | 23.54 |
| zh | 0.08 | 75.00 | 0.78 | 0.78 | 6.43 | 18.60 | 18.60 | 4.23 | 11.63 | 11.63 |
| silence | 9.18 | 62.54 | 5.36 | 5.36 | 66.45 | 12.78 | 11.95 | 64.12 | 12.06 | 12.06 |
| Average success rate | | 62.92 | 29.96 | 29.96 | 40.91 | 26.73 | 26.96 | 29.46 | 26.83 | 26.83 |

Abbreviations: MBL, memory-based learning; TD, temporal difference; WH, Widrow–Hoff.

stands at 2.5% assuming an equiprobable distribution of the 40 outcome instances, or the strategy of always choosing the most probable phone (/s/) at 9.5%. The average success across datasets and algorithms reveals the superior success of MBL. However, this algorithm appears more ‘sensitive’ to input data, that is, its success is more affected by the data, its type, and, arguably, its amount. At the same time, the

performance of the two ECL algorithms shows overall worse performance that is less affected by the specifics of the data.

To examine whether there is an interaction between learning model and type of input data, we applied Bayesian quantile mixed effect modeling in the **R** software environment (R Core Team, 2021) using the **brms** package (version 2.14.0; Bürkner, 2018). A Bayesian quantile model was chosen as a robust, distribution-free alternative to linear models. The priors were set as *location* = 0.0 and *quantile* = 0.5, thus evaluating the model at *median* = 0.0 (an excellent discussion of Bayesian quantile regression can be found in Yang et al., 2016; see also Schmidtke et al., 2017; Tomaschek et al., 2018 for applications in research on language); the remaining priors, for the random effects in particular, were left at their default values. The final model is summarized in the following formula:

$$\text{Success Rate} \sim \text{Learning Model} \times \text{Input Data} + \\ (1 + \text{Learning Model} | \text{Phone}).$$

The dependent variable is the success rate, and the two fixed effect factors are the learning model (three levels: MBL, WH, and TD) and the type of input data (three levels: raw, random, and Gaussian). Finally, the 40 phone categories make up the random effect factor with an additional correction for learning model. The model summary is presented in Fig. 1.

If we compare MBL and ECL (jointly WH and TD), first, we see that MBL shows a better overall performance (*Estimate* = 16.03; Bayesian credible interval: *CrI* = [9.74, 21.97]; posterior probability: $p(\text{MBL} > \text{ECL}) = 1.0$). The MBL performance decreases significantly with a change of data, between Raw and Random (*Estimate* = 20.18; *CrI* = [15.02, 25.36]; $p(\text{MBL}_{\text{raw}} > \text{MBL}_{\text{random}}) = 1.0$),

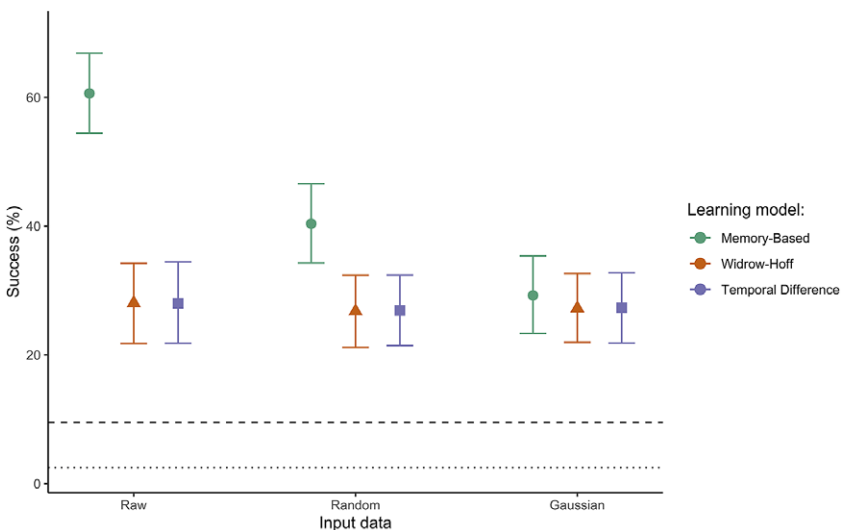


Figure 1. Conditional effect of learning model and type of data on success rates with respective 95% credible intervals. The dashed line represents the most probable phone (/s/) choice at 9.5%; the dotted line marks the flat chance level at 2.5%.

and between Random and Gaussian ($Estimate = 11.13$; $CrI = [6.06, 16.44]$; $p(MBL_{random} > MBL_{Gaussian}) = 1.0$). The difference between MBL and ECL performance on Gaussian data is not significant ($Estimate = 2.02$; $CrI = [-5.43, 9.23]$; $p(MBL_{Gaussian} > WH_{Gaussian}) = 0.72$). Finally, no differences across datasets for the two ECL algorithms are significant. Overall, it appears that MBL performs better with more data, which is to be expected as the pool of memorized exemplars grows.

The arguably ‘cleverer’ of the two ECL models, TD, does not benefit from looking into the future. Once the learning rate parameter is properly tuned, TD’s discount factor parameter suggests a value of $\gamma = 0.0$, which turns the TD algorithm into its ‘simpler cousin’, WH (for formal details, see Appendix A at <https://doi.org/10.25500/edata.bham.00000942>). In case of random data, when the number of occurrences of each phone is kept constant and when the parameter search suggests $\gamma = 0.17$, considering future data still does not improve the performance of TD, as we can infer from Fig. 1. An immediate reward strategy, as implemented in WH, does not perform worse than a strategy which considers future rewards. This could suggest that future data are not meaningful or informative for the current choice of phone.

The strength of the correlations between the probability of the phone, on the one hand, and the success rates of the three learning models, on the other hand, adds an interesting layer of information for assessing the overall efficiency of the chosen learning models: the higher the correlation, the more frequency-driven the learning model is (the correlation between probabilities in the training and the test sample was near-perfect: $r = 0.998$). To bring this out, we made use of Bayesian Kendall’s tau-b (τ_B), a non-parametric alternative to Pearson’s product-moment correlation coefficients. Table 2 contains a summary of these correlation coefficients.

There is no correlation between phone probabilities and MBL rates of success when the model is trained on raw data. The correlations of the MBL success rates, when trained on either random or Gaussian data, and the phone probabilities are, however, strong and positive. The trend is different for the two ECL models’ rates of success: they enter into a weak correlation with the probabilities when trained on the raw dataset and show no correlation when ECL models are trained on random or Gaussian data. Overall, the change in correlation given the three training datasets is small (or negligible) for the ECL algorithms, while being substantial for the MBL algorithm.

Each in its own way, the random and Gaussian datasets control for the effect of token frequency. The random dataset (randomly) samples 100 occurrences of each phone, whereas the Gaussian dataset represents generated or simulated values using empirical means and standard errors of MFCC acoustic inputs for each phone and silence. This means we can expect an effect of token frequency only for the raw dataset. Our results confirm the observation by Caballero and Kapatsinski (2022) that

Table 2. Bayesian Kendall’s tau-b between the probability of a phone in the test sample and the combination of learning (MBL, WH, and TD) and type of data (raw, random, and Gaussian)

| | Raw | | | Random | | | Gaussian | | |
|---------------------|-------|-------|-------|--------|--------|-------|----------|-------|-------|
| | MBL | WH | TD | MBL | WH | TD | MBL | WH | TD |
| Phone probabilities | 0.101 | 0.276 | 0.276 | 0.653 | -0.027 | 0.009 | 0.624 | 0.189 | 0.189 |

Abbreviations: MBL, memory-based learning; TD, temporal difference; WH, Widrow-Hoff.

ECL models tend to overweight token frequency: if allowed, as in the case of raw data, these algorithms will show such tendency, which is expressed as a weak correlation ($\tau_B = 0.276$). The correlations between MBL model performance across the dataset, and the phone probabilities, must have different reasons than the token frequency: when trained on the raw data, MBL performance shows no correlation with the phone probabilities ($\tau_B = 0.101$), while training on random and Gaussian data, which kept the phone occurrence frequency under control (a sample 100 occurrences of each phone for random, and 100 generated datapoints per phone for Gaussian), gives high correlation with phone probabilities ($\tau_B > 0.6$).

3.2. Abstracting coherent groups of phones

In this section, we examine whether the patterns in the confusion matrices show signs of meaningful groupings, that is, whether the confusions follow traditionally recognized groups of phones. If this is so in most cases, then that could be considered indirect evidence for the emergence of (simple linguistic) abstractions. In other words, since traditional classes of phones are well established in the literature, the question is whether our learning models make misclassifications *within* groups of traditionally or theoretically related phones, or whether the misclassifications show no systematicity.

We used two complementary statistical methods to depict and discuss the emergence of classes of phones from the respective confusion matrices. First, we applied hierarchical cluster analysis using the **pvclust** package (version 2.2-0; Suzuki & Shimodaira, 2013) in the **R** software environment (R Core Team, 2021) to the matrices. We used row-based Euclidean distances – distances between phones suggested by the outcome of the learning algorithm training (i.e., expected phones), with Ward's agglomerative clustering method, with 1,000 bootstrap runs. For brevity, we focus here on the confusion matrices from trainings on raw data. Other results are provided in Appendix B at <https://doi.org/10.25500/edata.bham.00000942>.

The second statistical method used a theoretical matrix of phone features, such as *sonorant*, *syllabic*, *nasal*, and *long*, which could be either present (1), absent (–1), or irrelevant (0) for each given phone (the full matrix is available on the University of Birmingham Institutional Research Archive, UBIRA at <https://doi.org/10.25500/edata.bham.00000942>). This matrix was converted to a Euclidean distance matrix between phones, which served as a criterion for comparisons with 'empirical' matrices of distances, used for cluster analysis. For each empirical matrix, we applied the Mantel test for correlation between two distance matrices (Mantel, 1967). We utilized the **ade4** package (Thioulouse et al., 2018) in **R** with efficient permutation runs to obtain simulated *p*-values.

The results of the cluster analyses are summarized in Figs. 2 and 3. The **R** package **pvclust** allows assessing the uncertainty in hierarchical cluster analysis (see Divjak & Fieller, 2014). Using bootstrap resampling, **pvclust** makes it possible to calculate *p*-values for each cluster in the hierarchical clustering. The *p*-value of a cluster ranges between 0 and 1, and indicates how strongly the cluster is supported by data. The package **pvclust** provides two types of *p*-values: the *au* (approximately unbiased) *p*-value (in red) and the *bp* (bootstrap probability) *p*-value (in green). The *au* *p*-value, which relies on multiscale bootstrap resampling, is considered a better approximation to an unbiased *p*-value than the *bp* value, which is computed by normal

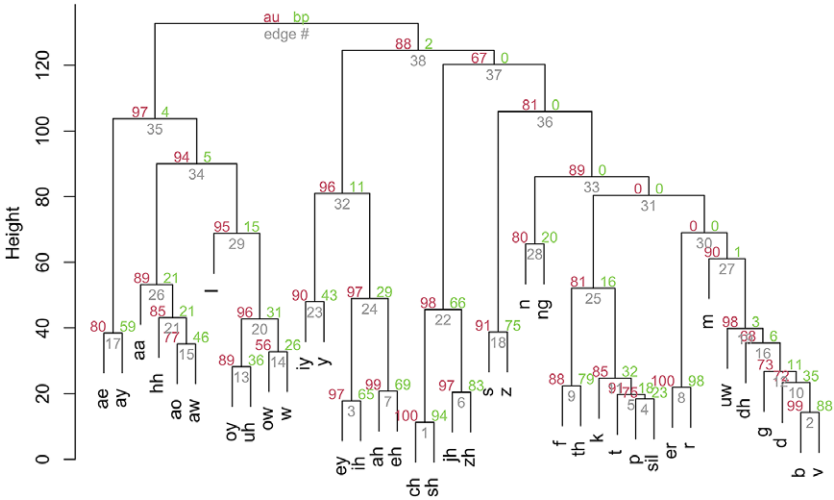


Figure 2. Dendrogram of phone clustering using Widrow-Hoff weights from training on raw data.

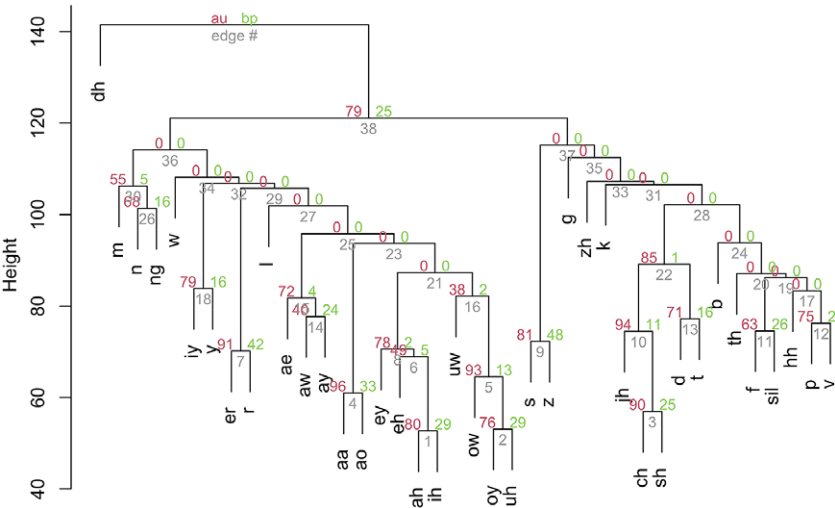


Figure 3. Dendrogram of phone clustering using memory-based probabilities from training on raw data.

bootstrap resampling. The gray values indicate the order of clustering, with smaller values signaling cluster that were formed earlier in the process, and hence contain elements that are considered more similar. For example, Fig. 2 shows an early cluster of /ch/ and /sh/ in the midsection of the plot, which is later joined by a similarly early cluster of /jh/ and /zh/. Similarly, a bit more to the left, another ‘in-between’ cluster of /ao/ and /aw/ is joined by /hh/ and then /aa/ in two steps.

The WH learning weights appear to yield relatively intuitive phone groups (Fig. 2). Specifically, the left-hand branch of the dendrogram yields clusters of vowels and approximants (with the exception of /hh/). This branch is mostly made up of low and

back vowels with the inclusion of some diphthongs along with /l/ and /w/. Acoustically, the first and second formants of these sounds are all relatively close together, at least at the beginning of the sound. There is a right-side group in this left-hand branch that is well supported and contains many of the back high and mid vowels. The right-hand branch of the dendrogram, which is less likely to replicate as a whole, has two branches. Its left side is made up of mostly high to mid front vowels in addition to /ə/ (schwa in IPA) and the post-alveolar approximant /y/ which is acoustically very similar to /iy/. Its right side is largely dominated by obstruents. They are grouped into a number of reasonable obstruent clusters. For example, the post-alveolar fricatives and affricates (as noted above), /ch/, /sh/, /jh/, and /zh/ form a strongly supported cluster, likely due to shared acoustics of the post-alveolar frication. The voiced stops /b/, /d/, and /g/ and the voiced non-sibilant fricatives /v/ and /dh/ also form a strong cluster. The /b/ and /v/ are shown to be closely tied like due to the fact that they are both labial obstruents and will share some of the same acoustic cues. There is also an unexpected phone, /uw/, in this group. The rhotacized /er/ and the rhotic approximant /r/ also form a well-supported cluster. In addition, we get a cluster of voiceless stops /k/, /t/, and /p/ along with *sil* grouped together with a branch from this group containing the voiceless non-sibilant fricatives /th/ and /f/. While not as well supported, the alveolar fricatives are grouped together, as are the alveolar and velar nasals /n/ and /ng/.

Since the confusion matrices from WH and TD training on raw data are identical, so are the results of cluster analysis (with minor differences in *au* and *bp* *p*-values, which is expected given independent bootstrap resampling). The MBL-based confusion matrix and the resulting clusters, however, paint a rather different picture.

The clusters based on MBL also seem to split into two groups. The left-hand side contains vowels, approximants, and nasals, whereas the right-hand side contains obstruents (plus *sil*). Unfortunately, the *p*-values in this dendrogram hardly ever reach an acceptable level, and none of the clusters identified can be expected to replicate.

In fact, the *p*-values for the clusters in Figs. 2 and 3 are strikingly different. Despite some quibbles with the WH phone groups (Fig. 2), many of the clusters in the WH solution are rather likely to replicate. The clusters in the MBL solution (Fig. 3), however, show the opposite tendency, where there are no clusters that receive an *au* *p*-value equal to or higher than 0.95. Higher-order clusters fare particularly poorly in this respect, with most receiving no chance at all of replicating in a different dataset. Note, also, that the values on the vertical axis are very different for ECL (i.e., WH in Fig. 2) vs. MBL: MBL clusters start forming at value 50, by which point many of the WH and TD cluster solutions have already converged.

To support our elaboration of the obtained cluster solutions with empirical evidence, we rely on the Mantel test for correlation between distance matrices. This test shows consistently higher, statistically more significant correlations between ECL-based matrices and the criterion matrix based on theoretically well-known features, as compared with correlations between MBL-based matrices and the same given criterion matrix, for the raw and random datasets; the difference between correlations for MBL and the criterion and ECL and the criterion with Gaussian data is, however, only marginal ($r_{MBL} = 0.208$ vs. $r_{ECL} = 0.268$: $t = 1.75$; $p < 0.08$). The results are summarized in Table 3.

In summarizing the results, two points appear to be particularly interesting in how the algorithms (ECL vs. MBL) compare with respect to (a) handling new input,

Table 3. Mantel test results of correlations with the criterion matrix based on the phone features. Simulated p-values are based on 1,000 replicates

| | Raw | | | Random | | | Gaussian | | |
|-------------|-------|--------|--------|--------|--------|--------|----------|--------|--------|
| | MBL | WH | TD | MBL | WH | TD | MBL | WH | TD |
| Correlation | 0.073 | 0.272 | 0.272 | 0.134 | 0.284 | 0.280 | 0.208 | 0.268 | 0.268 |
| p-value | <0.02 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |

Abbreviations: MBL, memory-based learning; TD, temporal difference; WH, Widrow–Hoff.

and (b) providing a basis for meaningful abstractions. On the first point, MBL convincingly outperforms the ECL models in assigning unheard acoustic input (MFCC) to the correct phone category. Even though the ECL models make more misclassifications, they are, however, meaningful: these models confuse phones that are more likely to be confused as they share more features and thus belong to the same or similar traditionally recognized groups of phones. This is supported by the comparisons of the respective confusion matrices and the criterion matrix of said phone features. The cluster analysis, furthermore, suggests consistency (replicability) of groupings for ECL, yet an almost complete idiosyncrasy of the MBL-based groupings.

4. General discussion

We set out to test two opposing principles regarding the development of language knowledge: MBL and ECL. MBL relies on a large pool of veridically stored experiences and a mechanism to match new ones with already stored exemplars, focusing on efficiency in storage and matching. ECL filters the input data to minimize the discrepancy between the system's current state of learning (built based on available input), on the one hand, and the true state (the actual outcome), on the other hand. Through filtering, those dimensions of the experience that are not useful for avoiding a discrepancy are removed. What gets retained, thus, changes continuously as the cycles of matching and confirmation/error-correction alternate. Such an *evolving unit* cannot be a veridical reflection of an experience, but contains only the most useful essence, distilled from many exposure events (or usage events, viz. Langacker, 1991).

A process of generalization underlies the abstractions linguists operate with, and we probed whether MBL and/or ECL could give rise to a type of generalized language knowledge that resembles linguistic abstractions. The aim is to understand what can be learned and what can be generalized from what is learned if the data are as basic and unprocessed as possible. A particular choice of data 'enrichment' might be beneficial to one algorithm more than to another.

We put all learning algorithms through two performance tests: a horizontal test that probes the quality of their learning by assessing the models' ability to predict data that are very similar to the data they were trained on, that is, the same speaker, new input; a vertical test assesses the ability of the models to give rise to abstract categories. As expected, both types of models fare differently in these tests.

In this context, it is important to keep in mind some limitations of the choices made in this study. First, all models are trained on pronunciations of isolated words, not on naturalistic speech excerpts. Next, we did not attempt the task of generalizing

to a new speaker, but only to new pronunciations by the same speaker. These limitations are, as said, consequences of conscious choices made to enable us to take reliable first steps in this area of enquiry.

4.1. *The learnability of abstractions*

MBL outperforms ECL in predicting new input from the same speaker on raw and random data. With Gaussian data, the learning models do not differ significantly in their performance. ECL may in fact overfit in case of larger datasets, essentially overcommitting to the already processed data, in particular with respect to token frequencies (Caballero & Kapatsinski, 2022). Whether such overfitting to input data is necessarily unnatural or negative is a separate question, however, as it represents known phenomena in language acquisition more generally (see, e.g., Arnon & Ramscar, 2012) as well as in the L1 and L2 learning of phonetic input specifically. For the latter, infants, at least in Western cultures, typically have one primary caregiver, which makes it likely that they, too, *overfit their mental models* to the input of that particular caregiver. There is evidence from learning in other sensory domains, that is, vision, that casts this approach in a positive light: infants use their limited world to become experts at a few things first, such as their caregiver's face(s). The same, very few faces are in the infant's visual field frequently and persistently and these extended, repeated exposures have been argued to form the starting point for all visual learning (Jayaraman & Smith, 2019). However, while overfitting may be an advantage in some learning situations, it may cause challenges in others. The use of a single speaker may indeed introduce a challenge to the model, not dissimilar to the long-known speaker normalization problem in speech perception (e.g., Ladefoged & Broadbent, 1957): the process of a listener adapting to sounds (typically vowels) produced by different speakers. L2 learners are also known to overfit their input data, and this tendency has been found to hinder their progress (for a review, see Barriuso & Hayes-Harb, 2018). While it has been shown that adult L2 learners benefit from multiple talkers and contexts (for the initial report, see Lively et al., 1993), child L2 learners do not appear to benefit more from high variability input (e.g., Barriuso & Hayes-Harb, 2018; Brekelmans et al., 2020).

We also tested to what extent the confusion matrices, which we understand as coarse proxies of (failure to) learn(ing), facilitate the meaningful groupings of phones into traditionally acknowledged groups. On that task, ECL performed reasonably well: it gave rise to a dendrogram that allocated 30 out of 40 phones to groups that resemble their traditional phonetic classification, based on their shared features (e.g., *syllabic, nasal, labial, and round*). Furthermore, the groups it identified were estimated to have a high chance of being detected in new, independent datasets. MBL, on the other hand, performed poorly. Some phone groups were intuitively recognizable, but the estimated lack of replicability of the MBL clusters is quite striking. Generalization is, of course, one of the most basic attributes of the broad notion of learning. To quote one of the pioneers of artificial neural networks, intelligent signal processing, and machine learning: 'if generalization is not needed, we can simply store the associations in a look-up table, and will have little need for a neural network', where the 'inability to realize all functions is in a sense a strength [...] because it [...] improves its ability to generalize' to new and unseen patterns (Widrow & Lehr, 1990, p. 1422). From this, by extension, we are in a position to draw several conclusions.

First, the task of correctly categorizing new acoustic input appeared to require very little, if any, generalization. The amount of training data, indeed, created an excellent ‘look-up table’ for MBL to find the right neighborhoods for a given exemplar.

One possible reason for MBL’s strong performance when predicting new data from the same speaker is the unconstrained neighborhood size, which is arguably not very plausible from a cognitive point of view: by fine-tuning its k -parameter we favored accuracy.¹ Next, a closer look at TD’s performance can also shed light on some less obvious reasons for MBL’s apparent superiority. Recall that, once the learning rate parameter was carefully tuned for WH and its value used for TD training, future data were either disregarded (raw and Gaussian data, for which best performance was achieved when the discount factor, γ , was set to 0) or did not improve the performance (random data). From this, we can conclude that, even though two out of three datasets contain temporally overlapping data (raw and random) and that TD ought to show superior performance to that of WH by explicitly factoring in TDs, future information did not contribute to the model’s performance. This observation also explains why MBL did so well by simply storing the input data as discrete exemplars and ignoring the temporality in the data, which is that model’s defaults.

4.2. Implications for a theoretical discussion of abstractions

Our findings show that we should not throw the abstractions-baby out with the theoretical bathwater. Instead, the way in which we approach the discussion of abstractions should change. Up until now, the amount of theoretical speculation about the existence of abstractions (cf. Ambridge, 2020; Ramscar, 2019) exceeds the amount of empirical work done to support these positions. This is specifically true for attempts to model how such abstractions might emerge from exposure to input. For example, Ramscar and Port (2016) discuss the impossibility of a discrete inventory of phonemes and claim that “none of the traditional discrete units of language that are thought to be the compositional parts of language – phones or phonemes, syllables, morphemes, and words – can be identified unambiguously enough to serve as the theoretical bedrock on which to build a successful theory of spoken language composition and interpretation. Surely, if these units were essential components of speech perception, they would not be so uncertainly identified” (p. 63). What our results show, however, is that at least part of the phone inventory can be reliably identified from input, even by the simple algorithms applied here – ECL and MBL. Findings reported elsewhere suggest that phoneme-like abstractions, even if not essential, do play a role (Pierrehumbert, 2016). Following a similar line of reasoning, some studies suggest a hybrid approach, with both abstractions and exemplars (Ernestus, 2014).

¹The MBL model performed considerably poorer with a reduced number of neighbors. We systematically varied the k -parameter, and ran the training and tests on a subset of the raw sample; recall that the full raw training sample contained 1,078,200 inputs (see Section 2.3), and we sampled a subset of 108,319 inputs (roughly 10%). Comparatively, the mean accuracy at, for example, $k = \{3, 10, 20\}$ for the full raw sample was 51.0%, 58.2%, and 62.3%, and for the 10% subset, it was 43.6%, 49.8%, and 54.5%. Thus, the jump in accuracy between 3 and 10 neighbours is approximately 6.7%, and between the full and the 10% sample, approximately an additional 8.0%.

The conclusions about the generalization or abstraction challenge are, however, more divisive: while we believe we have provided enough evidence that ECL principles can generalize and, thus, unveil the process of *emergence*, MBL principles seem to fail on this task, both (1) in terms of ‘mimicking’ the phone misclassifications in respect of their overapplying features (*sonorant*, *long*, etc.), (2) in providing information that would enable the appearance of natural groups of phones, and (3) in making such emerging groups replicable in a different dataset. In the end, this is exactly what Widrow’s work warned us about: memory-based exemplar models are, de facto, structured look-up tables, well suited to categorize new input, but less obviously able to generalize (see also Widrow, 1959). How can these models abstract away from said exemplars and ‘create’ representations? As such, work in phonology that requires abstractions should explore hybrid approaches to phonological representation (cf. Ernestus, 2014).

At the same time, not all phones appear to be equally learnable from exposure, at least not from exposure to one speaker. Perhaps the idiosyncrasies of our one speaker show that more diverse training is necessary to learn other groups of phone-like units. Cognitively speaking, the phoneme inventory that is traditionally assumed for English is possibly either too specific or too coarse in places, in particular when dealing with one speaker’s input. Schatz et al. (2021) likewise reported finding a different type of unit, as the ones their model learned were too brief and too variable acoustically to correspond to the traditional phonetic categories. They, too, took the view that such a result should not automatically be taken to challenge the learning mechanism, but rather to challenge what is expected to be learned, for example, phonetic categories (cf. Donegan, 2015).

Shortcomings aside, we have shown that ECL methods can learn abstractions. Furthermore, the argument for or against the relevance of units and abstractions is, as we said, void if not pitted against a particular task that may or may not require such units to be discerned and/or categorized. This point was superbly demonstrated by models of categorization, such as Love et al. (2004) and Nosofsky (1986), in which continuous representations gave rise to discrete categories, if required by the task at hand. In that same sense, as it seems, ECL offers the more versatile type of learned information, which can handle tasks at the level of exemplars and of abstractions.

Acknowledgments. We are grateful to Ben Ambridge for his seminar which sparked our interest in this topic, and to the members of the Out of Our Minds team and the Alberta Phonetics Laboratory for sharing their thoughts. We also thank Danielle Matthews, Gerardo Ortega, and Eleni Vasilaki for useful discussions and/or pointers to the literature.

Data availability statement. The data that support the findings of this study are openly available on the University of Birmingham Institutional Research Archive, UBIRA at <https://doi.org/10.25500/edata.bham.00000942>. The R-scripts for all analyses presented here are available on the GitHub repository at <https://github.com/ooominds/The-emergence-of-abstractions>.

Financial disclosure. This work was funded by a Leverhulme Trust award (RL-2016-001) which funded the first and last authors and by the Social Sciences and Humanities Research Council of Canada (4352014-0678), which funded the second author.

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5–6), 509–559.

- Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.
- Anderson, N. D., Holmes, E. W., Dell, G. S., & Middleton, E. L. (2019). Reversal shift in phonotactic learning during language production: Evidence for incremental learning. *Journal of Memory and Language*, 106, 135–149.
- Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122(3), 292–305.
- Baddeley, A. (1994). The magical number seven: Still magic after all these? *Psychological Review*, 101, 353–356.
- Barrusio, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *CATESOL Journal*, 30(1), 177–194.
- Brekelmans, G., Evans, B., & Wonnacott, E. (2020). No evidence of a high variability benefit in phonetic vowel training for children. In *Book of abstracts: 2nd workshop on speech perception and production across the lifespan (SPPL2020)*. London, UK.
- Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3–4), 155–180.
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411.
- Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, 22(2–4), 381–410.
- Bybee, J. L. (2013). Usage-based theory and exemplar representations of constructions. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 49–69). Oxford University Press.
- Caballero, G., & Kapatsinski, V. (2022). How agglutinative? Searching for cues to meaning in Choguita Rarámuri (Tarahumara) using discriminative learning. In A. D. Sims, A. Ussishkin, J. Parker, & S. Wray (Eds.), *Morphological diversity and linguistic cognition* (pp. 121–160). Cambridge University Press.
- Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press.
- Divjak, D., & Fieller, N. (2014). Cluster analysis: Finding structure in linguistic data. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (Vol. 43, pp. 405–441). John Benjamins Publishing Company.
- Divjak, D., & Milin, P. (2023). Using computational cognitive modeling in usage-based linguistics. In M. Diaz-Campos & S. Balasch (Eds.), *The handbook of usage-based linguistics*. Wiley.
- Divjak, D., Szymor, N., & Socha-Michalik, A. (2015). Less is more: possibility and necessity as centres of gravity in a usage-based classification of core modals in Polish. *Russian Linguistics*, 39(3), 327–349.
- Donegan, P. (2015). The emergence of phonological representation. In *The handbook of language emergence* (pp. 33–52). Wiley.
- Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, 142, 27–41.
- Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, 11(11), e1004567.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105(2), 251–279.
- Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, 31(3–4), 305–320.
- Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610.
- Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3), 649–660.
- Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*, 128(6), 1112–1144.
- Haykin, S. S. (1999). *Neural networks: A comprehensive foundation*, Second edition. Prentice-Hall.
- Holmes, J., & Holmes, W. (2002). *Speech Synthesis and Recognition*. CRC Press.
- Jayaraman, S., & Smith, L. B. (2019). Faces in early visual environments are persistent not just frequent. *Vision Research*, 157, 213–221.
- Johnson, K. (1997). *The auditory/perceptual basis for speech segmentation* [working paper no. 50]. Department of Linguistics, Ohio State University.

- Julià, P. (1983). *Explanatory models in linguistics: A behavioral perspective*. Princeton University Press.
- Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, 29(1), 98–104.
- Langacker, R. W. (1991). *Foundations of cognitive grammar: Descriptive application* (Vol. 2). Stanford University Press.
- Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36.
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94(3), 1242–1255.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111(2), 309–332.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2), 209–220.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman & Co.
- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356.
- McClelland, J. L., & Bybee, J. (2007). Gradience of gradience: A reply to Jackendoff. *The Linguistic Review*, 24(4), 437–455.
- Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 116, 374–388.
- Milín, P., Divjak, D., & Baayen, R. H. (2017a). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(11), 1730.
- Milín, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, 27(4), 507–526.
- Milín, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017b). Discrimination in lexical decision. *PLoS One*, 12(2), e0171935.
- Milín, P., Madabushi, H. T., Croucher, M., & Divjak, D. (2020). *Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences* [arXiv preprint]. [arXiv:2003.03813](https://arxiv.org/abs/2003.03813)
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Nesdale, A. R., Herriman, M. L., & Tunmer, W. E. (1984). Phonological awareness in children. In *Metalinguistic awareness in children: Theory, research, and implications* (pp. 56–72). Springer.
- Nixon, J. S., & Tomaschek, F. (2020). Learning from the acoustic signal: Error-driven learning of low-level acoustics discriminates vowel and consonant pairs. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 585–591). Cognitive Science Society.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39–57.
- Pavlov, I. P. (1927). *Conditioned reflexes*. Dover.
- Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, 4(34), 516–530.
- Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, 2(1), 33–52.
- Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23(4), 463–490.
- Port, R. F. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, 25(2), 143–170.
- Port, R. F. (2010). Language as a social institution: Why phonemes and words do not live in the brain. *Ecological Psychology*, 22(4), 304–326.
- R Core Team. (2021). *R: A language and environment for statistical computing* (version 4.1.1). R Foundation for Statistical Computing. <https://www.R-project.org/>

- Ramscar, M. (2019). *Source codes in human communication* [arXiv preprint]. arXiv:1904.03991
- Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, 89(4), 760–793.
- Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, 53, 58–74.
- Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151–160.
- Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7), e2001844118.
- Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(11), 1793–1820.
- Skinner, B. F. (1974). *About behaviourism*. Alfred A. Knopf Inc.
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 135–170.
- Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the ninth annual conference of the cognitive science society* (pp. 355–378). Erlbaum.
- Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). MIT Press.
- Suzuki, R., & Shimodaira, H. (2013). *Hierarchical clustering with p-values via multiscale bootstrap resampling* [R package]. R Foundation for Statistical Computing.
- Taylor, J. R. (1995). *Linguistic categorization: Prototypes in linguistic theory*. Clarendon Press.
- Thioulouse, J., Dray, S., Dufour, A.-B., Siberchicot, A., Jombart, T., & Pavoine, S. (2018). *Multivariate analysis of ecological data with ade4*. Springer.
- Tomaschek, F., Tucker, B. V., Fasiolo, M., & Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, 4(s2), 20170018.
- Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, 6(2), 302–324.
- Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, 51(3), 1187–1204.
- Tunmer, W. E., & Rohl, M. (1991). Phonological awareness and reading acquisition. In *Phonological awareness in reading: The evolution of current perspectives* (pp. 1–30). Springer.
- Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.
- Vértes, E., & Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. *Advances in Neural Information Processing Systems*, 32, 13714–13724.
- Widrow, B. (1959). *Adaptive sampled-data systems: A statistical theory of adaptation*. WESCON.
- Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits*. IRE WESCON.
- Widrow, B., & Lehr, M. A. (1990). 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9), 1415–1442.
- Yang, Y., Wang, H. J., & He, X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, 84(3), 327–344.

Cite this article: Milin, P., Tucker, B. V. & Divjak, D. (2023). A learning perspective on the emergence of abstractions: the curious case of phone(me)s *Language and Cognition* 15: 740–762. <https://doi.org/10.1017/langcog.2023.11>