Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate

Christo Kirov
Johns Hopkins University
Baltimore, MD
ckirov1@jhu.edu

Ryan Cotterell
Johns Hopkins University
Baltimore, MD
ryan.cotterell@jhu.edu

Abstract

Can advances in NLP help advance cognitive modeling? We examine the role of artificial neural networks, the current state of the art in many common NLP tasks, by returning to a classic case study. In 1986, Rumelhart and McClelland famously introduced a neural architecture that learned to transduce English verb stems to their past tense forms. Shortly thereafter, in 1988, Pinker and Prince presented a comprehensive rebuttal of many of Rumelhart and McClelland's claims. Much of the force of their attack centered on the empirical inadequacy of the Rumelhart and McClelland model. Today, however, that model is severely outmoded. We show that the Encoder-Decoder network architectures used in modern NLP systems obviate most of Pinker and Prince's criticisms without requiring any simplification of the past tense mapping problem. We suggest that the empirical performance of modern networks warrants a reexamination of their utility in linguistic and cognitive modeling.

1 Introduction

In their famous 1986 opus, Rumelhart and McClelland (R&M) describe a neural network capable of transducing English verb stems to their past tense. The strong cognitive claims in the article fomented a veritable brouhaha in the linguistics community and eventually led to the highly influential rebuttal of Pinker and Prince (1988) (P&P). P&P highlighted the extremely poor empirical performance of the R&M model, and pointed out a number of theoretical issues with the model, which they suggested would apply to *any* neural network, contemporarily branded connectionist approaches. Their critique was so successful that many linguists and cognitive scientists to this day do not consider neural networks a viable approach to modeling linguistic data and human cognition.

In the field of natural language processing (NLP), however, neural networks have experienced a renaissance. With novel architectures, large new data sets available for training, and access to extensive computational resources, neural networks now constitute the state of the art in many NLP tasks. However, NLP as a discipline has a distinct practical bent and more often concerns itself with the large-scale engineering applications of language technologies. As such, the field's findings are not always considered relevant to the scientific study of language (i.e., the field of linguistics). Recent work, however, has indicated that this perception is changing, with researchers, for example, probing the ability of neural networks to learn syntactic dependencies like subject–verb agreement (Linzen et al., 2016).

Moreover, in the domains of morphology and phonology, both NLP practitioners and linguists have considered virtually identical problems, seemingly unbeknownst to each other. For example, both computational and theoretical morphologists are concerned with how different inflected forms in the lexicon are related and how one can learn to generate such inflections from data. Indeed, the original R&M network focuses on such a generation task, namely, generating English past tense forms from their stems. R&M's network, however, was severely limited and did not generalize correctly to held-out data. In contrast, state-of-the art morphological generation networks used in NLP, built from the modern evolution of recurrent neural networks (RNNs) explored by Elman (1990) and others, solve the same problem almost perfectly (Cotterell et al., 2016). This level of performance on a cognitively relevant problem suggests that it is time to consider further incorporating network modeling into the study of linguistics and cognitive science.

Crucially, we wish to sidestep one of the issues that framed the original debate between P&P and R&M—whether or not neural models learn and use "rules." From our perspective, any system that

Transactions of the Association for Computational Linguistics, vol. 6, pp. 651–666, 2018. Action Editor: Sebastian Padó.
Submission batch: 2/2018; Revision batch: 5/2018; Published 12/2018.

picks up systematic, predictable patterns in data may be referred to as rule-governed. We focus instead on an empirical assessment of the ability of a modern state-of-the-art neural architecture to learn linguistic patterns, asking the following questions: (i) Does the learner induce the full set of correct generalizations about the data? Given a range of novel inputs, to what extent does it apply the correct transformations to them? (ii) Does the behavior of the learner mimic humans? Are the errors human-like?

In this work, we run new experiments examining the ability of the Encoder-Decoder architecture developed for machine translation (Bahdanau et al., 2014; Sutskever et al., 2014) to learn the English past tense. The results suggest that modern nets absolutely meet the first criterion, and often meet the second. Furthermore, they do this given limited prior knowledge of linguistic structure: The networks we consider do not have phonological features built into them and must instead learn their own representations for input phonemes. The design and performance of these networks invalidate many of the criticisms in Pinker and Prince (1988). We contend that, given the gains displayed in this case study, which is characteristic of problems in the morpho-phonological domain, researchers across linguistics and cognitive science should consider evaluating modern neural architectures as part of their modeling toolbox.

This paper is structured as follows. Section 2 describes the problem under consideration, the English past tense. Section 3 lays out the original Rumelhart and McClelland model from 1986 in modern machine-learning parlance, and compares it to a state-of-the-art Encoder-Decoder architecture. A historical perspective on alternative approaches to modeling, both neural and non-neural, is provided in Section 4. The empirical performance of the Encoder-Decoder architecture is evaluated in Section 5. Section 6 provides a summary of which of Pinker and Prince's original criticisms have effectively been resolved, and which ones still require further consideration. Concluding remarks follow.

2 The English Past Tense

Many languages mark words with syntactico-semantic distinctions. For instance, English marks the distinction between present and past tense verbs, for example, *walk* and *walked*. English verbal morphology is relatively impoverished, distinguishing maximally five forms for the copula *to be* and only three forms for most verbs. In this work, we consider learning to conjugate the English verb forms, rendered as phonological strings. As it is the focus of the original R&M study, we focus primarily on the English past tense formation.

orthographic			IPA			
stem	past	part.	stem	past	part.	infl. type
go	went	gone	goʊ	wɛnt	gon	suppletive
sing	sang	sung	sɪŋ	sæŋ	sʊŋ	ablaut
swim	swam	swum	swɪm	swæm	swʊm	ablaut
sack	sacked	sacked	sæk	sækt	sækt	[-t]
sag	sagged	sagged	sæg	sægd	sægd	[-d]
pat	patted	patted	pæt	pætɪd	pætɪd	[-ɪd]
pad	padded	padded	pæd	pædɪd	pædɪd	[-ɪd]

Table 1: Examples of inflected English verbs.

Both regular and irregular patterning exist in English. Orthographically, the canonical regular suffix is *-ed*, which, phonologically, may be rendered as one of three phonological strings: [-ɪd], [-d], or [-t]. The choice among the three is deterministic, depending only on the phonological properties of the previous segment. English selects [-ɪd] where the previous phoneme is a [t] or [d], for example, [pæt]↦[pætɪd] (*pat*↦*patted*) and [pæd]↦[pædɪd] (*pad*↦*padded*). In other cases, English enforces voicing agreement: It opts for [-d] when the proceeding phoneme is a voiced consonant or a vowel (e.g., [sæg]↦[sægd] (*sag*↦*sagged*) and [ʃoʊ]↦[ʃoʊd] (*show*↦*showed*)), and for [-t] when the proceeding phoneme is an unvoiced consonant (e.g., [sæk]↦[sækt] (*sack*↦*sacked*)). English irregulars are either suppletive (e.g., [goʊ]↦[wɛnt] (*go*↦*went*)), or exist in sub-regular islands defined by processes like ablaut (e.g., *sing*↦*sang*) that may contain several verbs (Nelson, 2010); see Table 1.

Single vs. Dual Route. A frequently discussed cognitive aspect of past tense processing concerns whether or not irregular forms have their own processing pipeline in the brain. Pinker and Prince (1988) proposed separate modules for regular and irregular verbs; regular verbs go through a general, rule-governed transduction mechanism, and exceptional irregulars are produced via simple memory look-up.[1] While some studies (e.g., Marslen-Wilson and Tyler, 1997; Ullman et al., 1997)

[1]Note that irregular look-up can simply be recast as the application of a context-specific rule.

provide corroborating evidence from speakers with selective impairments to regular or irregular verb production, others have called these results into doubt (Stockall and Marantz, 2006). From the perspective of this paper, a complete model of the English past tense should cover both regular and irregular transformations. The neural network approaches we advocate for achieve this goal, but do not clearly fall into either the single or dual-route category—internal computations performed by each network remain opaque, so we cannot at present make a claim whether two separable computation paths are present.

2.1 Acquisition of the Past Tense

The English past tense is of considerable theoretical interest because of the now well-studied acquisition patterns of children. As first shown by Berko (1958) in the so-called wug-test, knowledge of English morphology *cannot* be attributed solely to memorization. Indeed, both adults and children are fully capable of generalizing the patterns to novel words (e.g., [wʌg]↦[wʌgd] (*wug*↦*wugged*)). During acquisition, only a few types of errors are common; children rarely blend regular and irregular forms—for example, the past tense of *come* is either produced as *comed* or *came*, but rarely *camed* (Pinker, 1999).

Acquisition Patterns for Irregular Verbs. It is widely claimed that children learning the past tense forms of irregular verbs exhibit a "U-shaped" learning curve. At first, they correctly conjugate irregular forms (e.g., *come*↦*came*), then they regress during a period of **overregularization** producing the past tense as *comed* as they acquire the general past tense formation. Finally, they learn to produce both the regular and irregular forms. Plunkett and Marchman, however, observed a more nuanced form of this behavior. Rather than a *macro* U-shaped learning process that applies globally and uniformly to all irregulars, they noted that many irregulars *oscillate* between correct and overregularized productions (Marchman, 1988). These oscillations, which Plunkett and Marchman refer to as a *micro* U-shape, further apply at different rates for different verbs (Plunkett and Marchman, 1991). Interestingly, although the exact pattern of irregular acquisition may be disputed, children rarely **overirregularize**, that is, misconjugate a regular verb as if it were irregular, such as *ping*↦*pang*.

3 1986 vs. Today

In this section, we compare the original R&M architecture from 1986 to today's state-of-the-art neural architecture for morphological transduction, the Encoder-Decoder model.

3.1 Rumelhart and McClelland (1986)

For many linguists, the face of neural networks to this day remains the work of R&M. Here, we describe in detail their original architecture, using modern machine learning parlance whenever possible. Fundamentally, R&M were interested in designing a sequence-to-sequence network for variable-length input using a small feed-forward network. From an NLP perspective, this work constitutes one of the first attempts to design a network for a task reminiscent of popular NLP tasks today that require variable-length input (e.g., part-of-speech tagging, parsing, and generation).

Wickelphones and Wickelfeatures. Unfortunately, a fixed-sized feed-forward network is not immediately compatible with the goal of transducing sequences of varying lengths. Rumelhart and McClelland decided to get around this limitation by representing each string as the set of its constituent phoneme trigrams. Each trigram is termed a **Wickelphone** (Wickelgren, 1969). As a concrete example, the IPA-string [#kæt#], marked with a special beginning- and end-of-string character, contains three distinct Wickelphones: [#kæ], [kæt], [æt#]. In fact, Rumelhart and McClelland went one step further and decomposed Wickelphones into component **Wickelfeatures**, or trigrams of phonological features, one for each Wickelphone phoneme. For example, the Wickelphone [ipt] is represented by the Wickelfeatures ⟨+vowel,+unvoiced,+interrupted⟩ and ⟨+high,+stop,+stop⟩. Because there are far fewer Wickelfeatures than Wickelphones, words could be represented with fewer units (of key importance for 1986 hardware) and more shared Wickelfeatures potentially meant better generalization.

We can describe R&M's representations using the modern linear-algebraic notation standard among researchers in neural networks. First, we assume that the language under consideration contains a fixed set of phonemes Σ, plus an edge symbol # marking the beginning and end of words. Then, we construct the set of all Wickelphones Φ and the set of all Wickelfeatures $\mathcal{F}$ by enumeration. The first layer of the R&M neural

network consists of two deterministic functions: (i) $\boldsymbol{\phi} : \Sigma^* \rightarrow \mathbb{B}^{|\Phi|}$ and (ii) $\boldsymbol{f} : \mathbb{B}^{|\Phi|} \rightarrow \mathbb{B}^{|\mathcal{F}|}$, where we define $\mathbb{B} = \{-1, 1\}$. The first function $\boldsymbol{\phi}$ maps a phoneme string to the set of Wickelphones that fire, as it were, on that string; for example, $\boldsymbol{\phi}$ ([#kaet#]) = {[#kæ], [kæt], [æt#]}. The output subset of Φ may be represented by a binary vector of length $|\Phi|$, where a 1 means that the Wickelphone appears in the string and a -1 that it does not.[2] The second function $\boldsymbol{f}$ maps a set of Wickelphones to its corresponding set of Wickelfeatures.

Pattern Associator Network. Here we define the complete network of R&M. We denote strings of phonemes as $\mathbf{x} \in \Sigma^*$, where x_i is the i^{th} phoneme in a string. Given source and target phoneme strings $\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \in \Sigma^*$, R&M optimize the following objective, a sum over the individual losses for each of the $i = 1, ..., N$ training items:

$$\sum_{i=1}^{N} \left|\left| \max\left\{0, -\boldsymbol{\pi}(\mathbf{y}^{(i)}) \odot \left(W\boldsymbol{\pi}(\mathbf{x}^{(i)}) + \boldsymbol{b}\right)\right\}\right|\right|_1 \tag{1}$$

where $\max\{\cdot\}$ is taken point-wise, $\odot$ is point-wise multiplication, $W \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}|}$ is a projection matrix, $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{F}|}$ is a bias term, and $\boldsymbol{\pi} = \boldsymbol{\phi} \circ \boldsymbol{f}$ is the composition of the Wickelphone and Wickelfeature encoding functions. Using modern terminology, the architecture is a linear model for a multi-label classification problem (Tsoumakas and Katakis, 2006): The goal is to predict the *set* of Wickelfeatures in the target form $\mathbf{y}^{(i)}$ given the input form $\mathbf{x}^{(i)}$ using a point-wise perceptron loss (hinge loss without a margin); that is, a binary perceptron predicts each feature independently, but there is one set of parameters $\{W, \boldsymbol{b}\}$. The total loss incurred is the *sum* of the per-feature loss, hence the use of the L_1 norm. The model is trained with stochastic sub-gradient descent (the perceptron update rule) (Rosenblatt, 1958; Bertsekas, 2015) with a fixed learning rate.[3] Later work augmented the architecture with multiple layers and nonlinearities (Marcus, 2001, Table 3.3).

[2] We have chosen -1 instead of the more traditional 0 so that the objective function that Rumelhart and McClelland optimize may be more concisely written.

[3] Follow-up work, e.g., Plunkett and Marchman (1991), has speculated that the original experiments in R&M may not have converged. Indeed, convergence may not be guaranteed depending on the fixed learning rate chosen. As Equation (1) is jointly convex in its parameters $\{W, \boldsymbol{b}\}$, there exist convex optimization algorithms that will guarantee convergence, albeit often with a decaying learning rate.

Decoding. Decoding the R&M network necessitates solving a tricky optimization problem. Given an input phoneme string $\mathbf{x}^{(i)}$, we then must find the corresponding $\mathbf{y}' \in \Sigma^*$ that minimizes

$$\left|\left|\boldsymbol{\pi}(\mathbf{y}') - \text{threshold}\left\{W\boldsymbol{\pi}(\mathbf{x}^{(i)}) + \boldsymbol{b}\right\}\right|\right|_0 \tag{2}$$

where threshold is a step function that maps all non-positive reals to -1 and all positive reals to 1. In other words, we seek the phoneme string $\mathbf{y}'$ that shares the most features with the maximum a posteriori decoded binary vector. This problem is intractable, and so R&M provide an approximation. For each test stem, they hand-selected a set of likely past-tense candidate forms, for example, good candidates for the past tense of *break* would be {*break*, *broke*, *brake*, *braked*}, and choose the form with Wickelfeatures closest to the network's output. This manual approximate decoding procedure is not intended to be cognitively plausible.

Architectural Limitations. R&M used Wickelphones and Wickelfeatures in order to help with generalization and limit their network to a tractable size. However, this came at a significant cost to the network's ability to represent unique strings—the encoding is lossy: Two words may have the same set of Wickelphones or features. The easiest way to see this shortcoming is to consider morphological reduplication, which is common in many of the world's languages. P&P provide an example from the Australian language of Oykangand, which distinguishes between *algal* 'straight' and *algalgal* 'ramrod straight'; both of these strings have the identical Wickelphone set {[*#al*], [*alg*], [*lga*], [*gal*], [*al#*]}. Moreover, P&P point out that phonologically related words such as [slɪt] and [sɪlt] have disjoint sets of Wickelphones: {[#sl], [slɪ], [lɪt], [ɪt#]} and {[#sɪ], [sɪl], [ɪlt], [lt#]}, respectively. These two words differ only by an instance of metathesis, or swapping the order of nearby sounds. The use of Wickelphone representations imposes the strong claim that they have nothing in common phonologically, despite sharing all phonemes. P&P suggest this is unlikely to be the case. As one point of evidence, the metathesis of the kind that differentiates [slɪt] and [sɪlt] is a common diachronic change. In English, for example, [horse] evolved from [hross], and [bird] from [brid] (Jesperson, 1942).

3.2 Encoder-Decoder Architectures

The NLP community has recently developed an analogue to the past-tense generation task originally

considered by R&M: morphological paradigm completion (Durrett and DeNero, 2013; Ahlberg et al., 2015; Cotterell et al., 2015; Nicolai et al., 2015; Faruqui et al., 2016). The goal is to train a model capable of mapping the lemma (stem in the case of English) to each form in the paradigm. In the case of English, the goal would be to map a lemma, for example, *walk*, to its past-tense word *walked* as well as its gerund and third person present singular, *walking* and *walks*, respectively. This task generalizes the R&M setting in that it requires learning more mappings than simply lemma to past tense.

By definition, any system that solves the more general morphological paradigm completion task must also be able to solve the original R&M task. As we wish to highlight the strongest currently available alternative to R&M, we focus on the state of the art in morphological paradigm completion: the Encoder-Decoder network architecture (ED) (Cotterell et al., 2016). This architecture consists of two RNNs coupled together by an attention mechanism. The encoder RNN reads each symbol in the input string one at a time, first assigning it a unique embedding, then processing that embedding to produce a representation of the phoneme given the rest of the phonemes in the string. The decoder RNN produces a sequence of output phonemes one at a time, using the attention mechanism to peek back at the encoder states as needed. Decoding ends when a halt symbol is output. Formally, the ED architecture encodes the probability distribution over forms

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{N} p(y_i \mid y_1, \ldots, y_{i-1}, c_i) \quad (3)$$

$$= \prod_{i=1}^{N} g(y_{i-1}, s_i, c_i) \quad (4)$$

where g is a non-linear function (in our case it is a multi-layer perceptron), s_i is the hidden state of the decoder RNN, $\mathbf{y} = (y_1, \ldots, y_N)$ is the output sequence (a sequence of $N = |\mathbf{y}|$ characters), and finally c_i is an attention-weighted sum of the the encoder RNN hidden states h_i, using the attention weights $\alpha_k(s_{i-1})$ that are computed based on the previous decoder hidden state: $c_i = \sum_{k=1}^{|\mathbf{x}|} \alpha_k(s_{i-1}) h_k$.

In contrast to the R&M network, the ED network optimizes the log-likelihood of the training data, that is, $\sum_{i=1}^{M} \log p(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$ for $i = 1, ..., M$ training items. We refer the reader to Bahdanau et al. (2014) for the complete architectural specification of the specific ED model we apply in this paper.

Theoretical Improvements. Although there are a number of possible variants of the ED architecture (Luong et al., 2015),[4] they all share several critical features that make up for many of the theoretical shortcomings of the feed-forward R&M model. The encoder reads in each phoneme sequentially, preserving identity and order and allowing any string of arbitrary length to receive a unique representation. Despite this encoding, a flexible notion of string similarity is also maintained, as the ED model learns a fixed embedding for each phoneme that forms part of the representation of all strings that share the phoneme. When the network encodes [sɪlt] and [slɪt], it uses the same phoneme embeddings—only the order changes. Finally, the decoder permits sampling and scoring arbitrary length *fully formed strings* in polynomial time (forward sampling), so there is no need to determine which string a non-unique set of Wickelfeatures represents. However, we note that decoding the 1-best string from a sequence-to-sequence model is likely NP-hard (1-best string decoding is even hard for weighted finite-state transducers [Goodman, 1998]).

Multi-Task Capability. A single ED model is easily adapted to multi-task learning (Caruana, 1997; Collobert et al., 2011), where each task is a single transduction (e.g., stem $\mapsto$ past). Note that R&M would need a separate network for each transduction (e.g., stem $\mapsto$ gerund and stem $\mapsto$ past participle). In fact, the current state of the art in NLP is to learn all morphological transductions in a paradigm jointly. The key insight is to construct a single network $p(\mathbf{y} \mid \mathbf{x}, t)$ to predict all inflections, marking the transformation in the input string—that is, we feed the network the string "`w a l k <gerund>`" as input, augmenting the alphabet Σ to include morphological descriptors. We refer to reader to Kann and Schütze (2016) for the encoding details. Thus, one network predicts all forms; for example, $p(\mathbf{y} \mid \mathbf{x}$=*walk*, t=past) yields a distribution over past tense forms for *walk* and $p(\mathbf{y} \mid \mathbf{x}$=*walk*, t=gerund) yields a distribution over gerunds.

[4] For the experiments in this paper, we use the variant in Bahdanau et al. (2014), which has explicitly been shown to be state of the art in morphological transduction (Cotterell et al., 2016).

Type of Model	Reference	Input	Output
Feedforward Network	Rumelhart and McClelland (1986)	Wickelphones	Wickelphones
Feedforward Network	MacWhinney and Leinbach (1991)	Fixed Size Phonological Template	Fixed Size Phonological Template
Feedforward Network	Plunkett and Marchman (1991)	Fixed Size Phonological Template	Fixed Size Phonological Template
Attractor Network	Hoeffner (1992)	Semantics	Fixed Size Phonological Template
Feedforward Network	Plunkett & Marchman (1993)	Fixed Size Phonological Template	Fixed Size Phonological Template
Recurrent Neural Network	Cottrell & Plunkett (1994)	Semantics	Phonological String
Feedforward Network	Hare, Elman, & Daugherty (1995)	Fixed Size Phonological Template	Inflection Class
Feedforward Neural Network	Hare & Elman (1995)	Semantics	Fixed Size Phonological Template
Recurrent Neural Network	Westermann & Goebel (1995)	Phonological String	Phonological String
Feedforward Neural Network	Nakisa & Hahn (1996)	Fixed Size Phonological Template	Inflection Class
Convolutional Neural Network	Bullinaria (1997)	Phonological String	Phonological String
Feedforward Neural Network	Plunkett & Nakisa (1997)	Fixed Size Phonological Template	Inflection Class
Feedforward Neural Network	Plunkett & Juola (1999)	Fixed Size Phonological Template	Fixed Size Phonological Template
Feedforward Neural Network	Hahn & Nakisa (2000)	Fixed Size Phonological Template	Inflection Class

Table 2: A curated list of related work, categorized by aspects of the technique. Based on a similar list found in Marcus (2001, page 82).

4 Related Work

In this section, we first describe direct follow-ups to the original 1986 R&M model, using various neural architectures. Then we review competing non-neural systems of context-sensitive rewrite rules in the style of the Sound Pattern of English (SPE) (Halle and Chomsky, 1968), as favored by Pinker and Prince.

4.1 Follow-ups to Rumelhart and McClelland (1986) Over the Years

Following R&M, a cottage industry devoted to cognitively plausible connectionist models of inflection learning sprouted in the linguistics and cognitive science literature. We provide a summary listing of the various proposals, along with broad-brush comparisons, in Table 2.

Although many of the approaches apply more modern feed-forward architectures than R&M, introducing multiple layers connected by nonlinear transformations, most continue to use feed-forward architectures with limited ability to deal with variable-length inputs and outputs and remain unable to produce and assign probability to arbitrary output strings.

MacWhinney and Leinbach (1991), Plunkett and Marchman (1991, 1993), and Plunkett and Juola (1999) map phonological strings to phonological strings using feed-forward networks, but rather than turning to Wickelphones to imprecisely represent strings of any length, they use fixed-size input and output templates, with units representing each possible symbol at each input and output position. For example, Plunkett and Marchman (1991, 1993) simplify the past-tense mapping problem by only considering a language of artificially generated words of exactly three syllables and a limited set of constructed past-tense formation patterns. MacWhinney and Leinbach (1991) and Plunkett and Juola (1999) additionally modify the input template to include extra units marking particular transformations (e.g., *past* or *gerund*), enabling their network to learn multiple mappings.

Some proposals simplify the problem even further, mapping fixed-size inputs into a small finite set of categories, solving a classification problem rather than a transduction problem. (Nakisa and Hahn, 1996; Hahn and Nakisa, 2000) classify German noun stems into their appropriate plural inflection classes. Plunkett and Nakisa (1997) do the same for Arabic stems.

Hoeffner (1992), Hare and Elman (1995), and Cottrell and Plunkett (1994) also solve an alternative problem—mapping semantic representations (usually one-hot vectors with one unit per possible word type, and one unit per possible inflection) to phonological outputs. As these networks use unstructured semantic inputs to represent words, they must act as memories—the phonological content of any word must be memorized. This precludes generalization to word types that were not seen during training.

Of the proposals that map semantics to phonology, the architecture in Hoeffner (1992) is unique in that it uses an attractor network rather than a feed-forward network, with the main difference being training using Hebbian learning rather than the standard backpropagation algorithm. Cottrell and Plunkett (1994) present an early use of a simple recurrent network (Elman, 1990) to decode output strings, making their model capable of variable length output.

Bullinaria (1997) includes one of the few models proposed that can deal with variable length *inputs*. They use a derivative of the NETtalk pronunciation model (Sejnowski and Rosenberg, 1987) that would today be considered a convolutional network. Each input phoneme in a stem is read independently along with its left and right context

phonemes within a limited context window (i.e., a convolutional kernel). Each kernel is then mapped to zero or more output phonemes within a fixed template. Because each output fragment is independently generated, the architecture is limited to learning only *local* constraints on output string structure. Similarly, the use of a fixed context window also means that inflectional patterns that depend on long-distance dependencies between input phonemes cannot be captured.

Finally, the model of Westermann and Goebel (1995) is arguably the most similar to a modern ED architecture, relying on simple recurrent networks to both encode input strings and decode output strings. However, the model was intended to explicitly instantiate a dual route mechanism and contains an additional explicit memory component to memorize irregulars. Despite the addition of this memory, the model was unable to fully learn the mapping from German verb stems to their participle forms, failing to capture the correct form for strong training verbs, including the copular *sein* → *gewesen*. As the authors note, this may be due to the difficulty of training simple recurrent networks, which tend to converge to poor local minima. Modern RNN varieties, such as long short-term memory (LSTM) networks in the ED model, were specifically designed to overcome these training limitations (Hochreiter and Schmidhuber, 1997).

4.2 Non-neural Learners

P&P describe several basic ideas that underlie a traditional, symbolic rule learner. Such a learner produces SPE-style rewrite rules that may be applied to deterministically transform the input string into the target. Rule candidates are found by comparing the stem and the inflected form, treating the portion that changes as the rule that governs the transformation. This is typically a set of non–copy edit operations. If multiple stem–past pairs share similar changes, these may be collapsed into a single rule by generalizing over the shared phonological features involved in the changes. For example, if multiple stems are converted to the past tense by the addition of the suffix [-d], the learner may create a collapsed rule that adds the suffix to all stems that end in a [+voice] sound. Different rules may be assigned weights (e.g., probabilities) derived from how many stem–past pairs exemplify the rules. These weights decide which rules to apply to produce the past tense.

Several systems that follow this rule-based template have been developed in NLP. Although the SPE itself does not impose detailed restrictions on rule structure, these systems generate rules that can be compiled into finite-state transducers (Kaplan and Kay, 1994; Ahlberg et al., 2015). These systems generalize well, but even the most successful variants have been shown to perform significantly worse than state-of-the-art neural networks at morphological inflection, often with a >10 percentage point differential in accuracy on held-out data (Cotterell et al., 2016).

In the linguistics literature, the most straightforward, direct, machine-implemented instantiation of the P&P proposal is, arguably, the Minimal Generalization Learner (MGL) of Albright and Hayes (2003) (c.f., Allen and Becker, 2015; Taatgen and Anderson, 2002). This model takes a mapping of phonemes to phonological features and makes feature-level generalizations like the post-voice [-d] rule described earlier. For a detailed technical description, see Albright and Hayes (2002). We treat the MGL as a baseline in §5.

Unlike Taatgen and Anderson (2002), who explicitly account for dual route processing by including both memory retrieval and rule application submodules, Albright and Hayes (2003) and Allen and Becker (2015) rely on discovering and correctly weighting (using weights learned by log-linear regression) highly stem-specific rules to account for irregular transformations.

Within the context of rule-based systems, several proposals focus on the question of rule generalization, rather than rule synthesis. That is, given a set of predefined rules, the systems implement metrics to decide whether rules should generalize to novel forms, depending on the number of exceptions in the data set. Yang (2016) defines the 'tolerance principle,' a threshold for exceptionality beyond which a rule will fail to generalize. O'Donnell (2011) treats the question of whether a rule will generalize as one of optimal Bayesian inference.

5 Evaluation of the ED Learner

We evaluate the performance of the ED architecture in light of the criticisms P&P levied against the original R&M model. We show that, in most cases, these criticisms no longer apply.[5]

[5]Data sets and code for all experiments are available at https://github.com/ckirov/RevisitPinkerAndPrince.

The most potent line of attack P&P use against the R&M model is that it simply does not learn the English past tense very well. Although the non-deterministic, manual, and non-precise decoding procedure used by R&M makes it difficult to obtain exact accuracy numbers, P&P estimate that the model only prefers the correct past tense form for about 67% of English verb stems. Furthermore, many of the errors made by the R&M network are unattested in human performance. For example, the model produces blends of regular and irregular past-tense formation (e.g., *eat* $\mapsto$ *ated*) that children do not produce unless they mistake *ate* for a present stem (Pinker, 1999). Furthermore, the R&M model frequently produces irregular past tense forms when a regular formation is expected (e.g., *ping* $\mapsto$ *pang*). Humans are more likely to overregularize. These behaviors suggest that the R&M model learns the wrong kind of generalizations. As shown subsequently, the ED architecture seems to avoid these pitfalls, while outperforming a P&P-style non-neural baseline.

5.1 Experiment 1: Learning the Past Tense

In the first experiment, we seek to show: (i) the ED model successfully learns to conjugate both regular and irregular verbs in the training data, and generalizes to held-out data at convergence and (ii) the pattern of errors the model exhibits is compatible with attested speech errors.

CELEX Data Set. Our base data set consists of 4,039 verb types in the CELEX database (Baayen et al., 1993). Each verb is associated with a present tense form (stem) and past tense form, both in IPA. Each verb is also marked as regular or irregular (Albright and Hayes, 2003). A total of 168 of the 4,039 verb types were marked as irregular. We assigned verbs to train, development, and test sets according to a random 80-10-10 split. Each verb appears in exactly one of these sets once. This corresponds to a uniform distribution over types because every verb has an effective frequency of 1.

In contrast, the original R&M model was trained and tested (data was not held out) on a set of 506 stem/past pairs derived from Kučera and Francis (1967). A total of 98 of the 506 verb types were marked as irregular.

Types vs. Tokens. In real human communication, words follow a Zipfian distribution, with many irregular verbs being exponentially more common than regular verbs. Although this condition is more true to the external environment of language learning, it may not accurately represent the psychological reality of how that environment is processed. A body of psycholinguistic evidence (Bybee, 1995, 2001; Pierrehumbert, 2001) suggests that human learners generalize phonological patterns based on the count of word *types* they appear in, ignoring the *token* frequency of those types. Thus, we chose to weigh all verb types equally for training, effecting a uniform distribution over types as described above.

Hyperparameters and Other Details. Our architecture is nearly identical to that used in Bahdanau et al. (2014), with hyperparameters set following Kann and Schütze (2016, §4.1.1). Each input character has an embedding size of 300 units. The encoder consists of a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with two layers. There is a dropout value of 0.3 between the layers. The decoder is a unidirectional LSTM with two layers. Both the encoder and decoder have 100 hidden units. Training was done using the Adadelta procedure (Zeiler, 2012) with a learning rate of 1.0 and a minibatch size of 20. We train for 100 epochs to ensure that all verb forms in the training data are adequately learned. We decode the model with beam search ($k = 12$). The code for our experiments is derived from the OpenNMT package (Klein et al., 2017). We use accuracy as our metric of performance. We train the MGL as a non-neural baseline, using the code distributed with Albright and Hayes (2003) with default settings.

Results and Discussion. The non-neural MGL baseline unsurprisingly learns the regular past-tense pattern nearly perfectly, given that it is imbued with knowledge of phonological features as well as a list of phonologically illegal phoneme sequences to avoid in its output. However, in our testing of the MGL, the preferred past-tense output for all verbs was never an irregular formulation. This was true even for irregular verbs that were observed by the learner in the training set. One might say that the MGL is only intended to account for the regular route of a dual route system. However, the intended scope of the MGL seems to be wider. The model is billed as accurately learning "islands of subregularity" within the past tense system, and Albright and Hayes use the model to make predictions about which irregular forms of novel verb stems are preferable to human speakers (see the subsequent discussion of wugs).

	all			regular			irregular		
	train	dev	test	train	dev	test	train	dev	test
Single-Task (MGL)	96.0	96.0	94.5	99.9	100.0	100.0	0.0	0.0	0.0
Single-Task (Type)	99.8†	97.4	95.1	99.9	99.2	98.9	97.6†	53.3†	28.6†
Multi-Task (Type)	100.0†	96.9	95.1	100.0	99.5	99.7	99.2†	33.3†	28.6†

Table 3: Results on held-out data in English past tense prediction for single- and multi-task scenarios. The MGL achieves perfect accuracy on regular verbs, and 0 accuracy on irregular verbs. † indicates that a neural model's performance was found to be significantly different ($p < 0.05$) from the MGL.

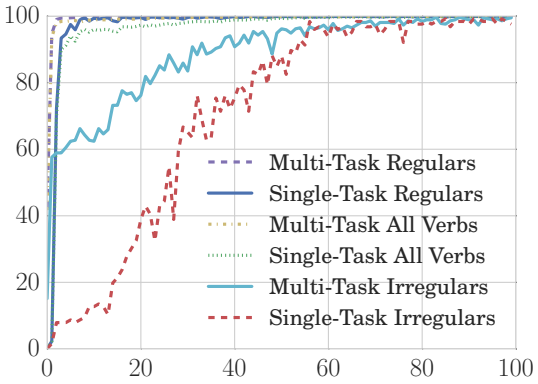

Figure 1: Single-task vs. multi-task. Learning curves for the English past tense. The x-axis is the number of epochs (one complete pass over the training data) and the y-axis is the accuracy on the training data (not the metric of optimization).

In contrast, the ED model, despite no built-in knowledge of phonology, successfully learns to conjugate nearly all the verbs in the training data, including irregulars—no reduction in scope is needed. This capacity to account for specific exceptions to the regular rule does not result in overfitting. We note similarly high accuracy on held-out regular data—98.9% to 99.2% at convergence depending on the condition. We report the full accuracy in all conditions in Table 3. The † indicates when a neural model's performance was found to be significantly different ($p < 0.05$) from the MGL according to a χ^2 test. The ED model achieves near-perfect accuracy on regular verbs, and irregular verbs seen during training, as well as substantial accuracy on irregular verbs in the dev and test sets. This behavior jointly results in better overall performance for the ED model when all verbs are considered. Figure 1 shows learning curves for regular and irregular verbs types in different conditions.

An error analysis of held-out data shows that the errors made by this network do not show any of the problems of the R&M architecture. There are no blend errors of the *eat* $\mapsto$ *ated* variety. Indeed, the *only* error the network makes on irregulars is overregularization (e.g., throw $\rightarrow$ throwed). In fact, the overregularization-caused lower accuracy that we observe for irregular verbs in development and test is expected and desirable; it matches the human tendency to treat novel words as regular, lacking knowledge of irregularity (Albright and Hayes, 2003).

Although most held-out irregulars are regularized, as expected, the ED model *does*, perhaps surprisingly, correctly conjugate a handful of irregular forms it has not seen during training—five in the test set. However, three of these are prefixed versions of irregulars that exist in the training set (retell $\rightarrow$ retold, partake $\rightarrow$ partook, withdraw $\rightarrow$ withdrew). One (sling $\rightarrow$ slung) is an analogy to similar training words (fling, cling). The final conjugation, forsake $\rightarrow$ forsook, is an interesting combination, with the prefix "for," but an untested base form "sake" that is similar to "take."[6]

From the training data, the only regular verb with an error is *compartmentalized*, whose past tense is predicted to be "compartmentalized," with a spurious vowel change that would likely be ironed out with additional training. Among the regular verbs in the development and test sets, the errors also consisted of single vowel changes (the full set of these errors was "thin" $\rightarrow$ "thun," "try" $\rightarrow$ "traud," "institutionalize" $\rightarrow$ "instititionalized," and "drawl" $\rightarrow$ "drooled").

Overall then, the ED model performs extremely well, a far cry from the $\approx$67% accuracy of the R&M model. It exceeds any reasonable standard of empirical adequacy, and shows human-like error behavior.

Acquisition Patterns. R&M made several claims that their architecture modeled the detailed acquisition of the English past tense by children. The core claim was that their model exhibited a *macro* U-shaped learning curve as in §2. Irregulars were initially produced correctly, followed by a period of overregularization preceding a final correct stage. However, P&P point out that R&M only achieve this pattern by manipulating the input distribution fed into their network. They trained only on irregulars for a number of epochs,

[6] [s] and [t] are both coronal consonants, a fricative and a stop, respectively.

CLING		MISLEAD		CATCH		FLY	
#	output	#	output	#	output	#	output
5	[klɪŋd]	8	[mɪsliːdɪd]	7	[kætʃ]	6	[flaɪd]
11	[klʌŋ]	19	[mɪslɛd]	31	[kætʃ]	31	[fluː]
13	[klɪŋ]	21	[mɪslɛd]	43	[kɔt]	40	[flaɪd]
14	[klɪŋd]	23	[mɪslɛd]	44	[kætʃ]	42	[fleɪ]
18	[klʌŋ]	24	[mɪsliːdɪd]	51	[kætʃ]	47	[flaɪd]
21	[klɪŋd]	29	[mɪslɛd]	52	[kɔt]	56	[fluː]
28	[klʌŋ]	30	[mɪsliːdɪd]	66	[kætʃ]	62	[flaɪd]
40	[klʌŋ]	41	[mɪslɛd]	73	[kɔt]	70	[fluː]

Table 4: Here we evince the oscillating development of single words in our corpus. For each stem, e.g., CLING, we show the past form that produced at change points to show the diversity of alternation. Beyond the last epoch displayed, each verb was produced correctly.

before flooding the network with regular verb forms. R&M justify this by claiming that young children's vocabulary consists disproportionately of irregular verbs early on, but P&P cite contrary evidence. A survey of child-directed speech shows that the ratio of regular to irregular verbs a child hears is constant while they are learning their language (Slobin, 1971). Furthermore, psycholinguistic results suggest that there is no early skew towards irregular verbs in the vocabulary children understand or use (Brown, 1973).

Although we do not wish to make a strong claim that the ED architecture accurately mirrors children's acquisition, only that it ultimately learns the correct generalizations, we wanted to see if it would display a child-like learning pattern *without* changing the training inputs fed into the network over time—that is, in all of our experiments, the data sets remained fixed for all epochs, unlike in R&M. We do not clearly see a macro U-shape, but we *do* observe Plukett and Marchman's predicted oscillations for irregular learning—the so-called micro U-shaped pattern. As shown in Table 4, individual verbs oscillate between correct production and overregularization before they are fully mastered.

Wug Testing. As a further test of the MGL as a cognitive model, Albright and Hayes created a set of 74 nonce English verb stems with varying levels of similarity to both regular and irregular verbs. For each stem (e.g., *rife*), they picked one regular output form (*rifed*), and one irregular output form (*rofe*). They used these stems and potential past-tense variants to perform a wug test with human participants. For each stem, they had 24 participants freely attempt to produce a past tense form. They then counted the percentage of participants who produced the pre-chosen regular and irregular forms (**production probability**). The production probabilities for each pre-chosen regular and irregular form could then be correlated with the predicted scores derived from the MGL. In Table 5, we compare the correlations based on their model scores, with correlations comparing the human scores to the output probabilities given by an ED model. As the wug data provided with Albright and Hayes (2003) use a different phonetic transcription than the one we used, we trained a separate ED model for this comparison. Model architecture, training verbs, and hyperparameters remained the same. Only the transcription used to represent input and output strings was changed to match Albright and Hayes (2003). Following the original paper, we correlate the probabilities for regular and irregular transformations separately. We apply Spearman's rank correlation, as we don't necessarily expect a linear relationship. We see that the ED model probabilities are slightly more correlated than the MGL's scores.

English	Network	MG
Regular (rife $\sim$ rifed, n=58)	**0.48**	0.35
Irregular (rife $\sim$ rofe, n=74)	**0.45**	0.36

Table 5: Spearman's ρ of human wug production probabilities with MG scores and ED probability estimates.

5.2 Experiment 2: Joint Multi-Task Learning

Another objection levied by P&P is R&M's focus on learning a single morphological transduction: stem to past tense. Many phonological patterns in a language, however, are not restricted to a single transduction—they make up a core part of the phonological system and take part in many different processes. For instance, the voicing assimilation patterns found in the past tense also apply to the third person singular: we see the affix *-s* rendered as [-s] after voiceless consonants and [-z] after voiced consonants and vowels.

P&P argue that the R&M model would not be able to take advantage of these shared generalizations. Assuming a different network would need to be trained for each transduction (e.g., stem to gerund and stem to past participle), it would be impossible to learn that they have any patterns in common. However, as discussed in §3.2, a single ED model can learn multiple types of mapping, simply by tagging each input–output pair in the training

set with the transduction it represents. A network trained in such a way shares the same weights and phoneme embeddings across tasks, and thus has the capacity to generalize patterns across all transductions, naturally capturing the overall phonology of the language. Because different transductions mutually constrain each other (e.g., English in general does not allow sequences of identical vowels), we actually expect faster learning of each individual pattern, which we test in the following experiment.

We trained a model with an architecture identical to that used in Experiment 1, but this time to jointly predict four mappings associated with English verbs (past, gerund, past participle, third-person singular).

Data. For each of the verb types in our base training set from Experiment 1, we added the three remaining mappings. The gerund, past-participle, and third-person singular forms were identified in CELEX according to their labels in Wiktionary (Sylak-Glassman et al., 2015). The network was trained on all individual *stem* $\mapsto$ *inflection* pairs in the new training set, with each input string modified with additional characters representing the current transduction (Kann and Schütze, 2016): *take* <PST> $\mapsto$ *took*, but *take* <PTCP> $\mapsto$ *taken*.[7]

Results. Table 3 and Figure 1 show the results. Overall, accuracy is >99% after convergence on train. Although the difference in final performance is never statistically significant compared to single-task learning, the learning curves are much steeper, so this level of performance is achieved much more quickly. This provides evidence for our intuition that cross-task generalization facilitates individual task learning due to shared phonological patterning (i.e., jointly generating the gerund hastens past-tense learning).

6 Summary of Resolved and Outstanding Criticisms

In this paper, we have argued that the Encoder-Decoder architecture obviates many of the criticisms P&P levied against R&M. Most importantly, the empirical performance of neural models is no longer an issue. The past tense transformation is learned nearly perfectly, compared to an approximate accuracy of 67% for R&M. Furthermore, the ED architecture solves the problem in a fully general setting. A single network can easily be trained on multiple mappings at once (and appears to generalize knowledge across them). No representational cludges such as Wickelphones are required—ED networks can map arbitrary length strings to arbitrary length strings. This permits training and evaluating the ED model on realistic data, including the ability to assign an exact probability to any arbitrary output string, rather than "representative" data designed to fit in a fixed-size neural architecture (e.g., fixed input and output templates). Evaluation shows that the ED model does not appear to display any of the degenerate error-types P&P note in the output of R&M (e.g., regular/irregular blends of the *ate* $\rightarrow$ *ated* variety).

Despite this litany of successes, some outstanding criticisms of R&M still remain to be addressed. On the trivial end, P&P correctly point out that the R&M model does not handle homophones: *write* $\mapsto$ *wrote*, but *right* $\mapsto$ *righted*. This is because it only takes the phonological make-up of the input string into account, without concern for its lexical identity. This issue affects the ED models we discuss in this paper as well—lexical disambiguation is outside of their intended scope. However, even the rule learner that P&P propose does not have such functionality. Furthermore, if lexical markings *were* available, we could incorporate them into the model just as with different transductions in the multi-task set-up (i.e., by adding the disambiguating markings to the input).

More importantly, we need to limit any claims regarding treating ED models as proxies for child language learners. P&P criticized such claims from R&M because they manipulated the input data distribution given to their network over time to effect a U-shaped learning curve, despite no evidence that the manipulation reflected children's perception or production capabilities. We avoid this criticism in our experiments, keeping the input distribution constant. We even show that the ED model captures at least one observed pattern of child language development—Plukett and Marchman's predicted oscillations for irregular learning, the micro U-shaped pattern. However, we did *not* observe a macro U-shape, nor was the micro effect consistent

[7] Without input annotation to mark the different mappings the network must learn, it would treat all input/output pairs as belonging to the same mapping, with each inflected form of a single stem as an equally likely output variant associated with that mapping. It is not within the scope of this network architecture to solve problems other than morphological transduction, such as discovering the range of morphological paradigm slots.

across all irregular verbs. More study is needed to determine the ways in which ED architectures do or do not reflect children's behavior. Even if nets do not match the development patterns of any individual, they may still be useful if they ultimately achieve a knowledge state that is comparable to that of an adult or, possibly, the aggregate usage statistics of a population of adults.

Along this vein, P&P note that the R&M model is able to learn highly unnatural patterns that do not exist in any language. For example, it is trivial to map each Wickelphone to its reverse, effectively creating a mirror-image of the input, for example, *brag*↦*garb*. Although an ED model could likely learn linguistically unattested patterns as well, some patterns may be more difficult to learn than others—for example, they might require increased time-to-convergence. It remains an open question for future research to determine which patterns RNNs prefer, and which changes are needed to account for over- and underfitting. Indeed, any sufficiently complex learning system (including rule-based learners) would have learning biases that require further study.

There are promising directions from which to approach this study. Networks are in a way analogous to animal models (McCloskey, 1991), in that they share interesting properties with human learners, as shown empirically, but are much easier and less costly to train and manipulate across multiple experiments. Initial experiments could focus on default architectures, as we do in this paper, effectively treating them as inductive baselines (Gildea and Jurafsky, 1996) and measuring their performance given limited domain knowledge. Our ED networks, for example, have no built-in knowledge of phonology or morphology. Failures of these baselines would then point the way towards the biases required to learn human language, and models modified to incorporate these biases could be tested.

7 Conclusion

We have shown that the application of the ED architecture to the problem of learning the English past tense obviates many, though not all, of the objections levied by P&P against the first neural network proposed for the task, suggesting that the criticisms do not extend to all neural models, as P&P imply. Compared with a non-neural baseline, the ED model accounts for both regular and irregular past tense formation in observed training data and generalizes to held-out verbs, all without built-in knowledge of phonology. Although not necessarily intended to act as a proxy for a child learner, the ED model also shows one of the development patterns that has been observed in children, namely, a micro U-shaped (oscillating) learning curve for irregular verbs. The accurate and substantially human-like performance of the ED model warrants consideration of its use as a research tool in theoretical linguistics and cognitive science.

References

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1024–1029, Denver, Colorado. Association for Computational Linguistics.

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the Workshop on Morphological and Phonological Learning*, volume 6, pages 58–69. Association for Computational Linguistics.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90:119–161.

Blake Allen and Michael Becker. 2015. Learning alternations from surface forms with sublexical phonology. Technical report, University of British Columbia and Stony Brook University.

R. Harald Baayen, Richard Piepenbrock, and Rijn van H. 1993. The CELEX lexical data base on CD-ROM.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14(2-3):150–177.

Dimitri P. Bertsekas. 2015. *Convex Optimization Algorithms*. Athena Scientific.

Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA.

John A Bullinaria. 1997. Modeling reading, spelling, and past tense learning with artificial neural networks. *Brain and Language*, 59(2):236–266.

Joan Bybee. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10:425–455.

Joan Bybee. 2001. *Phonology and Language Use*. Cambridge University Press, Cambridge, UK.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28:41–75.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task–morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics*, 3(1):433–447.

Garrison W Cottrell and Kim Plunkett. 1994. Acquiring the mapping from meaning to sounds. *Connection Science*, 6(4):379–412.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics*, 22(4):497–530.

Joshua Goodman. 1998. Parsing inside-out. Harvard Computer Science Group Technical Report TR-07-98.

Ulrike Hahn and Ramin Charles Nakisa. 2000. German inflection: Single route or dual route? *Cognitive Psychology*, 41(4):313–360.

Morris Halle and Noam Chomsky. 1968. *The Sound Pattern of English*. Harper & Row.

Mary Hare and Jeffrey L Elman. 1995. Learning and morphological change. *Cognition*, 56(1): 61–98.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

James Hoeffner. 1992. Are rules a thing of the past? The acquisition of verbal morphology by an attractor network. In *Proceddings of the 14th Annual Conference of the Cognitive Science Society*, pages 861–866.

Otto Jesperson. 1942. *A Modern English Grammar on Historical Principles*, volume 6. George Allen & Unwin Ltd., London, UK.

Katharina Kann and Hinrich Schütze. 2016. Med: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany. Association for Computational Linguistics.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Dartmouth Publishing Group.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics (TACL)*, 4:521–535.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Brian MacWhinney and Jared Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40(1):121–157.

Virginia Marchman. 1988. Rules and regularities in the acquisition of the English past tense. *Center for Research in Language Newsletter*, 2(4):04.

Gary Marcus. 2001. *The Algebraic Mind*. MIT Press.

William D Marslen-Wilson and Lorraine K Tyler. 1997. Dissociating types of mental computation. *Nature*, 387(6633):592.

Michael McCloskey. 1991. Networks and theories: The place of connectionism in cognitive science. *Psychological science*, 2(6):387–395.

Ramin Charles Nakisa and Ulrike Hahn. 1996. Where defaults don't help: the case of the German plural system. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 177–182. San Diego, CA.

Gerald Nelson. 2010. *English: an Essential Grammar*. Routledge.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931, Denver, Colorado. Association for Computational Linguistics.

Timothy J. O'Donnell. 2011. *Productivity and Reuse in Language*. Ph.D. thesis, Harvard University, Cambridge, MA.

Janet Pierrehumbert. 2001. Stochastic phonology. *GLOT*, 5:1–13.

Steven Pinker. 1999. *Words and Rules: The Ingredients of Language*. Basic Books.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1):73–193.

Kim Plunkett and Patrick Juola. 1999. A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23(4):463–490.

Kim Plunkett and Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102.

Kim Plunkett and Virginia Marchman. 1993. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69.

Kim Plunkett and Ramin Charles Nakisa. 1997. A connectionist model of the Arabic plural system. *Language and Cognitive processes*, 12(5-6): 807–836.

Frank Rosenblatt. 1958. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2, pages 216–271. MIT Press.

Terrence J Sejnowski and Charles R Rosenberg. 1987. Parallel networks that learn to pronounce English text. *Complex systems*, 1(1):145–168.

D.I. Slobin. 1971. On the learning of morphological rules: A reply to Palermo and Eberhart. In *The Ontogenesis of Grammar: A Theoretical Symposium*. Academic Press, New York.

Linnaea Stockall and Alec Marantz. 2006. A single route, full decomposition model of morphological complexity: MEG evidence. *The mental lexicon*, 1(1):85–123.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

Niels A Taatgen and John R Anderson. 2002. Why do children learn to say broke? a model of learning the past tense without feedback. *Cognition*, 86(2):123–155.

Grigorios Tsoumakas and Ioannis Katakis. 2006. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Michael T Ullman, Suzanne Corkin, Marie Coppola, Gregory Hickok, John H Growdon, Walter J Koroshetz, and Steven Pinker. 1997. A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of cognitive neuroscience*, 9(2):266–276.

Gert Westermann and Rainer Goebel. 1995. Connectionist rules of language. In *Proceedings of the 17th annual conference of the Cognitive Science Society*, pages 236–241.

Wayne A Wickelgren. 1969. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76(1):1.

Charles Yang. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.

Matthew D Zeiler. 2012. ADADELTA: An adaptive learning rate method. *CoRR*, abs/1212.5701.