Part II

What Role Does Cue Informativity Play in Learning and How the Lexicon Evolves Over Time?

PROOFS

PROOFS

5 How Agglutinative? Searching for Cues to Meaning in Choguita Rarámuri (Tarahumara) Using Discriminative Learning

Gabriela Caballero and Vsevolod Kapatsinski

A canonically agglutinative language or morphological pattern is traditionally analyzed as building words out of independent morphemes. Using data from Choguita Rarámuri (Uto-Aztecan), we attempt to quantify this notion by examining the extent to which meanings are predictable from their exponents without reference to context. We show that two-layer connectionist networks, computational models that map form onto meaning directly, can be used for this purpose. We also show that learning the meanings of morphemes can pose significant challenges to such models and constrains the design of the learning algorithm. In particular, models trained to equilibrium tend to focus on unreliable cues to the meanings they try to predict, especially when trained on a small corpus typical of underresourced languages. Some of these issues can be alleviated by a slow learning rate. However, one issue – which we call the problem of spurious excitement – is shown to be inherent to the learning algorithm, and always arises by the time the model achieves equilibrium. Spurious excitement means that a cue becomes associated with a meaning that it does not co-occur with, simply because of co-occurring with cues that disfavor the meaning. This case raises larger implications with respect to the type of learning mechanism involved in the acquisition of natural languages. Solutions to spurious excitement are discussed. The logistic activation function is shown to improve the performance of the model in detecting reliable cues to meanings that recur across many word types (i.e., cues of high type frequency), as well as eliminating spurious excitement.

PROOF

1 Introduction

In both morphological theory and morphological processing, researchers have long debated the importance of morphemes as a reliable unit for analysis and/or learning (see Blevins et al. 2016, Kapatsinski 2018a for recent reviews). The challenges faced by the canonical notion of a morpheme as a one-to-one association between meaning and a category of forms (morphs) have long been recognized in morphological theory (Anderson 1992, Aronoff 1976, Feldman and Moscoso del Prado Martín this volume, Matthews 1972), leading to an emerging consensus that words often cannot be exhaustively

decomposed into morphemes and, moreover, that the goal of morphological analysis should not be the identification of minimal morphological pieces (see also Sims et al. this volume). Instead, the extent to which decomposition is possible varies from word to word, morph to morph, and language to language. In particular, the processing literature suggests that "phonologically opaque" morphemes that trigger or undergo somewhat lexicalized phonology are less likely to be parsed out of the signal than more phonologically transparent morphemes (Hay and Baayen 2005). Similarly, morphemes that have constant and unambiguous meanings are more likely to be parsed out than those that are more semantically opaque (Schreuder et al. 2003). Both phonological and semantic opacity make the opaque morpheme an inferior cue to the meaning(s) it co-occurs with. Furthermore, the two kinds of opacity reinforce each other: the more opaque a morpheme, in either sense, the less likely it is to be parsed out of the signal in acquisition and to be relied on in processing. The more opaque the morpheme, the more likely words containing it are therefore to be accessed directly and the more likely they are to drift apart in semantic or phonetic space, further reducing the similarity between the forms and meanings of different instances of the same morpheme and increasing the morpheme's opacity (e.g., Bybee 2001, Hay 2003). This opaque-gets-more-opaque loop means that we should expect morphemes to cluster at the ends of the opaque–transparent continuum, which makes it a promising parameter for typological classification.

Intuitively, the degree to which the morphs of a language are good, transparent cues to the corresponding meanings is potentially useful as a typological parameter discriminating agglutinative and isolating languages from flexive ones. However, how does one measure transparency? The present chapter develops one possible way to do so, by utilizing a model that allows both morphs and cross-boundary phoneme sequences to predict morpheme meanings. The model evaluates the extent to which each form cue is predictive of various semantic features of words. When a morph is a good cue to the corresponding meaning, the association between the two is expected to be strong. Furthermore, when the morpheme is phonologically transparent (i.e., neither affected by nor affecting its phonological context), phonological cues spanning morph boundaries should be relatively weak predictors of meaning. In this chapter, we evaluate the degree to which morphs and cross-boundary sequences are predictive of morpheme meanings using a corpus of Choguita Rarámuri (CR, also known as Tarahumara; Uto-Aztecan; ISO code [tar]), an underresourced language that displays both agglutinative and flexive typological characteristics.

To estimate the weights of form–meaning connections, we use three versions of a two-layer connectionist network model (perceptron). All three versions use error-driven learning to predict meanings from forms. The resulting weights

PROOFS

of form–meaning associations are intended to represent how reliably the individual bits of form cue the various meanings (see also Bates and MacWhinney 1987, MacWhinney 1987 for a related approach at the syntactic level).

$$\Delta V_{C\to O} = \Lambda\left(1 - a_O\right) \tag{1}$$

$$\Delta V_{C\to O} = \Lambda\left(0 - a_O\right) \tag{2}$$

The model updates cue–outcome associations based on Equations (1)–(2), from Rescorla and Wagner (1972). In the present study, cues are chunks of form (morphs and diphones that cross morph boundaries), while outcomes are meanings of morphemes. The model learns cue–outcome associations by updating their weights. When a set of cues occurs together with a set of outcomes, the weight of the association from a present cue to a present outcome ($V_{C\to O}$) is incremented by (1). The weight of an association from a present cue to a known absent outcome ($V_{C\to O}$) is decremented by (2). Nothing is learned about absent cues.

For example, CR has several causative morphs, including /ti/, /ri/, and /r/. When the CR causative morph /ti/ occurs in a word with a Causative meaning in the gloss, as in *rimée-n-ti-ma* 'make.tortillas-APPL-CAUS-FUT.SG,' the association between the two is increased by (1), as are the Causative associations of the phoneme sequences that span *ti*'s boundaries, *n-t* and *i-m*.[1] The Causative associations of all other morphs and cross-morph sequences in the same word (the "present" cues; here, *rimée*, *n*, *ma*, and *e-n*) are also incremented. Similarly, all of the morphs and boundaries increase their associations with Make Tortillas, Applicative and Future Singular. Conversely, their associations with absent meanings (like Desiderative here) are decremented by (2).

$$a_O = \sum_C V_{C\to O} \tag{3}$$

In these equations, Λ is the learning rate, and a_O is activation of an outcome, which is the total amount of activation it receives from the present cues. As shown in (3), activation is defined as the sum of the weights of the associations from the present cues to the outcome. The RW learning rule is error-driven: the

[1] CR data are provided with a broad phonetic transcription using the IPA, except for stress, which is marked with an acute accent. Tone is left unmarked. Abbreviations used include the following: APPL – applicative; CAUS – causative; CER – certainty; CL – end of clause particle; CONJ – conjunct; DESID – desiderative; FUT – future; IMP – imperative; INCH – inchoative; INT – intensifier; OBJ – object; PASS – passive; PL – plural; POSS – possessive; POT – potential; PROG – progressive; PROX – proximal; PST – past; PTCP – participle; REV – reversive; SG – singular; SBJ – subject; TR – transitive; VBLZ – verbalizer.

PROOFS

weights change in proportion to the difference between the "correct" activation of an outcome (1 for present outcomes, 0 for absent ones) and the extent to which it is currently activated by the whole set of cues encountered (a_O). Because a_O is the sum of all cues rather than the weight of a particular cue, RW predicts cue competition effects such as blocking and overshadowing. Thus, when Causative occurs without /ti/, but in the presence of some cue that co-occurs with /ti/, like /r/, Causative becomes harder to associate with /ti/ because its activation in the presence of /ti/ and /r/ is closer to one in the absence of a ti~Causative association. CR is a particularly interesting language in which to examine cue competition because it has an abundance of multiple exponence patterns. In particular, /r/ and /ti/ are causative morphs that often occur together in a verb (Caballero 2008, Caballero and Kapatsinski 2015).

We compare three different versions of this model. One version is simply the RW model in (1)–(3). Another, the Naïve Discriminative Learner (NDL), defines the equilibrium state of the RW model. This is the state to which the model will eventually converge after encountering the training data, in the sense that seeing the training data again will not change its mind about the association weights. Danks (2003) shows that, at equilibrium, the weight of an association from a cue *j* to an outcome is given by (4), which states that the weight increases as the conditional probability of the outcome given the cue increases and decreases if the occurrence of the cue is predictable from other cues. Note that, importantly, type frequency plays no role in this formula.

$$w_{C_j \to O} = \frac{p(O \mid C_j)}{\sum_{j=0}^{n} p\left(C_j \mid C_i\right)} \tag{4}$$

The third, logistic version of the perceptron retains Equations (1)–(2) but replaces the definition of outcome activation in (3) with the definition in (5). According to this equation, activation is passed through an S-shaped logistic function that asymptotically approaches the limits, zero and one, but never reaches or overshoots them.

$$a_O = logit^{-1}(\sum\nolimits_c V_{C \to O}) \tag{5}$$

NDL has been argued to successfully model recognition of morphologically complex words in languages at both ends of the isolated-to-flexive spectrum, from Vietnamese (Pham and Baayen 2015) to Serbian (Baayen et al. 2011, Filipović Đurđević and Milin 2019). In the present study, we apply this tool to Choguita Rarámuri, which has a more ambiguous typological status. It displays many characteristics of agglutinative languages but departs from canonical agglutination by having a significant amount of phonological cohesion

PROOFS

between exponents. In this language, some of the morphemes are more transparent than others (Caballero 2008). We show that these differences in transparency can be recovered from NDL cue weights.

That said, this does not mean that NDL performs well in CR word recognition. In fact, the reason that transparency is reflected in NDL is that transparency correlates with the extent to which NDL succeeds in discovering reliable cues to the morphological meanings. While NDL has impressively replicated many psycholinguistic effects without the use of morphemic cues (see Feldman and Moscoso del Prado Martín this volume; Sims et al. this volume, for examples), the plausibility of cue weights derived from an NDL model has not been analyzed in detail. The present study is a first step in this direction.

We show that, especially when transparency is low, NDL faces at least two problems, both of which involve rare cues. First, the model is subject to what Albright and Hayes (2006) called the "problem of accidentally exceptionless generalizations": the cues the model considers most reliable are often cues that always co-occur with the meaning they are associated with but occur only once or twice in the dataset. Even more problematically, we show that the "best" cues to a certain meaning may never actually co-occur with that meaning in the data the model is trained on (examples are discussed below). We call this the problem of spurious excitement. Following Kruschke (1992), we argue that this problem is inherent to the "strict teacher" approach to error-driven learning in RW. The strict teacher approach says that outcome (here, semantic) activations should not exceed certain limits (here, one and zero are the *limiting values* for activations), and that activations overshooting the limits are wrong and need to be ~~increased~~. For this reason, when a stimulus contains more than one cue that inhibits an outcome, and a novel cue, that novel cue is learned to be an exciter to prevent the sum of the two cue weights from being negative. For example, if Singulars are more common than Plurals, and morphemes *x1* and *x2* are plural suffixes, they come to inhibit Singular. Suppose that the language has multiple exponence (like CR), so *x1* and *x2* can occur together. Now suppose that there is an infrequent stem, *x3*, that occurs in the corpus together with *x1* and *x2*, forming the word *x3x1x2*. This word is of course Plural, but after encountering it NDL will learn that *x3* is a cue to the Singular meaning, sometimes even stronger than a Singulative suffix would be (Kapatsinski 2021a).

These problems can be minimized by training NDL on a large corpus in which there is more opportunity for unlearning any spurious associations that are true in only a few forms. For this reason, the problems are not easily apparent when NDL is applied to languages for which such corpora are available, like Serbian or Vietnamese, though Ez-Zizi et al. (2021) have also recently reported spurious excitement in applying the model to Polish. However, these problems loom large in applying NDL to discover the patterns in an underdocumented language like CR or to simulate the learning of a miniature artificial language in the laboratory.

PROOFS

For such applications, the problems can be alleviated by deriving cue weights using the Rescorla-Wagner (RW) model, for which NDL represents the equilibrium state, using a slow learning rate that ensures that the model does not approach equilibrium and the meaning activations do not approach their limiting values (see also Heitmeier et al. 2021). A slow learning rate increases the contribution of type frequency and reduces the contribution of conditional probability to the cue weights, which solves the problem of accidentally exceptionless generalizations. However, slow learning rates do not solve the problem of spurious excitement because the problem will always arise as the model approaches equilibrium (see Kapatsinski 2021a for simulation results). Although a researcher can set the learning rate to be slow enough not to see spurious excitement in applying the model to a particular corpus, it is only possible to know that a certain learning rate is too fast by evaluating the associations learned at that rate and determining them to be spurious. It therefore requires the researcher to know what the results "should be," i.e., what the strongest cues to each particular meaning are, which defeats the purpose of using a computational model to discover the best cues to a meaning.

Overall, the results suggest that examining the ability of models implementing a learning rule to discover morphological structure can help evaluate the cognitive plausibility of the learning rule. In particular, the problem of spurious excitement arises because activations of outcomes can overshoot the activation limits, and because this overshoot is counted as an error. Consequently, it can be solved by (1) not having activation limits, as in Hebbian models (e.g., Kapatsinski and Harmon 2017, McMurray et al. 2012), (2) not counting overshooting the activation limit as an error (the *humble teacher* proposal, Kruschke 1992), or (3) making the activation limits unreachable, as in the logistic perceptron, by using Equation (5) to define activation instead of (3) (Rumelhart et al. 1986).

Kapatsinski (2021a) argues for the logistic perceptron solution based on cue competition effects in associative learning and language acquisition. Cue competition is well documented in language acquisition (Arnon and Ramscar 2012, Nixon 2020, Ramscar et al. 2010) but can be overcome with additional experience (Ellis and Sagarra 2010). Hebbian models do not show cue competition effects, while the humble teacher (like RW) fails to explain how cue competition can be overcome (Kapatsinski 2021a). In this chapter, we show that the logistic perceptron also performs better than RW in discovering reliable and type-frequent cues to morpheme meanings.

2 Morphological Typology and Processing

The "agglutinative ideal," the one-to-one correspondence between meaning and form in morphological expression, has been a central assumption

PROOFS

in morpheme-based morphological theory (e.g., Distributed Morphology [Halle and Marantz 1993, 1994]). The notion of agglutination has also played a central role in morphological typology as originally conceived in the nineteenth century. Specifically, morphological typology has been concerned with the classification of entire languages along the familiar scale of agglutination–flexion, with isolating languages on one end of the spectrum and introflexive (or nonlinear) languages at the other (isolating > agglutinative > flexive > non-concatenative [or introflexive]). This single scalar hierarchy results from conflating the parameters of phonological fusion and flexivity, as defined in (6):

(6) Phonological fusion and flexivity (Bickel and Nichols 2007)

a. Fusion: the degree to which individual exponents are phonologically fused to their host (isolating > concatenative > non-concatenative)
b. Flexivity: characterization of individual exponents (or stems) in terms of whether or not they exhibit lexically conditioned variance (suppletive allomorphy)

Agglutination in this classical sense thus involves morphological patterns that are concatenative and non-flexive. In contrast to the "agglutinative" type, in "flexive" or "(in)flecting" languages, the parameters of fusion and flexivity are also conflated to refer typically to languages with a predominance of concatenative-flexive exponence, i.e., those with relatively segmentable affixes and a high degree of lexical (suppletive) allomorphy of stems and/or affixes. For example, in Russian (and other Slavic languages), relatively segmentable case desinences are largely dependent on declension classes (Bickel and Nichols 2007).

The original motivation for conflating the parameters of "fusion" and "flexion" in traditional morphological typology stems from the fact that non-concatenative morphological patterns are (arguably) less segmentable than concatenative ones, and flexive exponents are also less segmentable than non-flexive ones: in a language with flexive affixes, those affixes and a stem that selects for them are more difficult to parse out of the signal than stem–affix combinations in a language where allomorphy is morphophonologically regular. In other words, in a language where a stem lexically selects for an affix, that stem may be analyzed as co-indexing the value encoded by the affix since it occurs only in combination with this value (Bickel and Nichols 2007: 18).

The measures of morphemic cue reliability we derive from NDL are most clearly related to the flexivity dimension. In particular, lexically or morphologically conditioned variation involving a particular morpheme means that the context in which a morpheme occurs will be somewhat predictive of its meaning. For example, if the causative /ti/ causes preceding consonants to change into /r/, the phonological sequence /rt/ will be somewhat predictive

PROOFS

Table 5.1 *Agglutination versus flexion (Plank 1999).*

Parameter	Agglutinative patterns	Flexive patterns
Separative exponence	✓	X
No flexivity	✓	X
Zero exponence	✓	X
No (or little) homonymous exponence	✓	X
Multiple exponence (via multiple affixation)	✓	X
Large paradigms	✓	X
Transparent morpheme boundaries	✓	X
Low degree of phonological cohesion	✓	X
Loose morphological bonding[2]	✓	X
Optional morphological marking	✓	X

of the Causative meaning. In contrast, the fusion dimension is not directly reflected in cue reliability measures. Instead, we consider the degree of fusion to be a correlate of the degree of flexivity. As morphological patterns age, they accumulate semantic and phonological idiosyncrasies and fuse with the surrounding context (Bybee 2001, 2003, 2008).

From this perspective, fusion and flexion are distinct – though correlated – dimensions. As a result, although some combinations of the values of the two parameters are more frequently attested than others crosslinguistically – with the "flexive" or "(in)flecting" (flexive-concatenative) type being the most common crosslinguistically – all possible combinations are attested (Bickel and Nichols 2007).

Recently, morphological typology has moved away from coarse-grained dimensions like flexivity and towards increasingly fine-grained variables (Bickel and Nichols 2007, Plank 1999). A crucial question that remains to be addressed within this new framework is this: To what extent do these more fine-grained variables exhibit a greater or lesser dependency between them? For example, Plank (1999) proposes that a typological profile of "agglutinative" versus "flexive" (inflecting) patterns may be better examined as resulting from a clustering of properties from a larger set of parameters. These parameters are summarized in Table 5.1.

Each of these fine-grained parameters may be assessed individually as canonically identifying agglutinative patterns versus flexive ones. However, many of these parameters are expected to influence morphemic cue reliability

[2] Plank (1999: 283) defines loose morphological bonding as the possibility of deletion/omission of morphological marking under identity (e.g., omission of case and number marking in one of two nouns in a coordinate construction in Turkish).

PROOFS

in similar ways. In particular, along with flexivity, zero exponence requires the listener to make use of the context surrounding the morpheme to detect the morpheme's meaning in the speech signal. Homonymous exponence likewise makes context essential for identifying a morpheme's meaning and reduces the reliability of morphemic cues. If /ri/ sometimes occurs with Causative and sometimes with Future, because it can mean either, its association with Causative will weaken any time /ri/ means Future, thus the /ri/→CAUS association will be weaker than if /ri/ always had a causative meaning. Furthermore, the context might help predict whether the word is causative or future, acquiring an association with the corresponding meaning.

On the other hand, some parameter values characteristic of agglutinative processes and languages do not align with lack of flexivity in increasing the reliability of morphemic cues to meaning and reducing reliance on context. In particular, multiple exponence means that some morphemic cues may be redundant. In an error-driven model like the perceptron, cue redundancy results in cue competition, reducing the weights of the redundant cues to a meaning. Similarly, optionality reduces the reliability of the optional cue. While it intuitively seems likely that morphemic cues play a larger role in the processing of agglutinative languages compared to flexive ones, not all characteristics of a canonically agglutinative language favor such cues. Rather, contextual cues are likely important for processing both agglutinative and flexive patterns. As a result, it is difficult to predict how context-independent morphemic cues will be in a particular language or a particular morphological domain within a language. Furthermore, the answer to this question will depend on how word recognition is thought to operate. The present study evaluates this question for one particular view of morphological processing, in which word recognition involves the direct use of formal cues to predict meanings and one particular view of learning the relevant cue–outcome mappings.

The grammar of a language contains a large number of morphological processes that can vary widely on the dimensions above (Anderson 1992: 328–329). This applies to even the parade examples of morphological types, such as agglutinative Turkish: while exhibiting predominantly concatenative morphology, this language also features productive non-concatenative processes, such as stress shifts to morphologically derive place names (e.g., *Bébek* from *bebék* 'baby') and emphatic reduplication of adjectival bases (e.g., *eski* 'ancient' *ep-eski* 'very ancient,' *temiz* 'clean' *ter-temiz* 'spotlessly clean'; Lewis 1967; see also Inkelas and Orgun 2003).[3] Characterizing a language as "agglutinative" or "flexive" thus becomes a generalization about the *degree* to which an individual system exhibits

[3] Though Lewis (1967) asserts that the only instances of non-concatenative morphology in Turkish are nonproductive traces of constructions borrowed from Arabic (Bickel and Nichols 2007: 183).

PROOFS

concatenative-non-flexive exponence across its morphological constructions. The measures of cue reliability we propose are most clearly applicable to individual morphological processes, with language classification necessitating aggregation across these processes, which raises issues beyond the scope of this chapter.

Note that the implications of typological characteristics for human language processing are not model-independent. They depend on one's theory of processing as well as one's theory of learning to process a language. For example, the prediction of competition between multiple exponents crucially depends on our assumption that forms are cues to meanings. Some previous work in the error-driven paradigm has argued strenuously that meanings serve as cues to forms and not vice versa (Arnon and Ramscar 2012, Ramscar et al. 2010; see Kapatsinski 2018b, 2021b for discussion). That is, comprehension works like production: both involve predicting upcoming forms from semantic context and preceding forms. If this is true, then multiple exponents of the same meaning would not compete with each other.

The present study develops measures of morphemic cue weights using a particular model of processing coupled with a particular model of learning (error-driven cue weighting based on predicting meanings from forms). We use the weights as a measure of the statistical structure of the CR lexicon, a reflection of the extent to which morphemic cues can be used to discriminate the corresponding meanings. We also use the results to draw implications for the underlying model as a way to learn morphological structure.

We quantify the extent to which semantic discrimination can be accomplished on the basis of context-independent morpheme representations. The intuition is that, in an ideally morpheme-based ("agglutinative") language, knowing the context of a morpheme would not help predict the morpheme's meaning. We use three versions of the same model, which differ in whether the model is trained to equilibrium, reaching the point at which another run through the lexicon would not change cue weights, and in how activation is defined.

At equilibrium, a form is an ideal cue to a meaning if (1) it has no homophones, (2) it is not redundant (i.e., there is no multiple exponence), and (3) its probability of occurrence is independent of the surrounding lexical and phonological context. Homophony directly reduces the form's cue weight by making the form unreliable as a cue to meaning. Multiple exponence reduces its weight via cue competition: the learner divides the cue weight among the cues that predict the same outcome. Synonymy among non-co-occurring morphemes is not generally a problem. Each synonym is learned to be a cue to the shared meaning. Only co-occurring and therefore (partially) redundant cues compete to predict the meaning. Therefore, free variation – if it existed – would not affect morpheme cue weights. However, conditioned allomorphy does matter because it means that allomorph occurrence probabilities are not context-independent. Non-independence means that aspects of the context – to the extent that they

PROOFS

are predictive of and specific to the morpheme – can become associated with the meaning instead of the morpheme, reducing the morpheme's cue weight in the process. This version of the model is the one we propose to use to quantify flexivity and transparency.

With a limited learning rate, form variation does matter even if unconditioned – the more allomorphs a morpheme has, the fewer opportunities there are to strengthen the cue weight of each allomorph (a related discussion in terms of affix parsability is found in Hay 2003, Hay and Baayen 2005, Hay and Plag 2004; see also Caballero and Inkelas 2013). More generally, high frequency makes a morph a better cue to its meaning because the cue weight increases by only a small amount from one exposure to a form–meaning pairing. As in humans (Bybee 1995a, Perfors et al. 2014), type frequency matters more than token frequency because an additional exposure to a morpheme in a constant context increases cue weights of both the morpheme and the surrounding context (though not necessarily equally). We show that being sensitive to type frequency is crucial for learning the right cues to meanings in CR.

3 The Language

3.1 Choguita Rarámuri Morphological Structure

Choguita Rarámuri is an Uto-Aztecan (UA) language of the Taracahitan branch spoken in northern Mexico by approximately 1,000 speakers (Casaus 2008). The data addressed in this study was obtained from an ongoing language description and documentation project carried out together with CR speakers since 2003, which has produced a documentary corpus that includes both elicited and naturalistic data (Caballero 2009, 2017; Chaparro Gardea et al. 2019).

Uto-Aztecan languages have been described as prototypically agglutinative, with complex verbal morphological systems, a high degree of synthesis, a low degree of phonological cohesion between largely concatenative exponents, and a low degree of cumulation in morphological exponence (Langacker 1977: 158). Based on the criteria in Table 5.1, CR morphology, which is also highly synthetic and almost exclusively suffixing,[4] displays the following agglutinative-like properties:

(7) Agglutinative-like properties of the CR verb:

a. Mostly concatenative, separative exponence
b. Limited flexive exponence

[4] Synthesis is defined as a high degree of semantic density at the level of the word; synthetic languages are those exhibiting a moderate number of formatives together with one root within single words.

PROOFS

Table 5.2 *Suffix positions and categories of the Choguita Rarámuri verb.*

Position	Type	Categories
S1	Derivation	Inchoative
S2	Derivation	Transitive
S3	Derivation	Applicative
S4	Derivation	Causative
S5	Derivation	Applicative
S6	Modality	Desiderative
S7	Derivation	Associated Motion
S8	Modality	Auditory Evidential
S9	Inflection	TAM
S10	Inflection	TAM
S11	Inflection	TAM, indirect causative
S12	Subordination	Deverbal morphology

Table 5.3 *Stem levels of the Choguita Rarámuri verb.*

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Root	INCH	TR	APPL	CAUS	APPL	DESID	MOT	EV	TAM	TAM	TAM, CAUS.I	SUB
Inner Stem	Derived Stem		Syntactic Stem			Aspectual Stem			Finite Verb			Subord. Verb

c. Zero exponence
d. Moderate homonymous exponence
e. Large derivational paradigms
f. Multiple exponence (through multiple affixation)
g. Optional marking

While CR shares several morphological and morphophonological properties and phenomena with other morphologically complex languages that have been characterized as agglutinative, it also crucially departs from the "agglutinative" type in that it has less transparent morpheme boundaries, due to a fair amount of phonological cohesion between exponents closer to the stem, a pattern more frequently attested in morphological systems traditionally characterized as "flexive."

The details are as follows: there is evidence for twelve suffix positions that are grouped into six concentric verbal zones or layers that are semantically, morphotactically, and morphophonologically motivated (Caballero 2008). The suffix positions and categories expressed in the CR verbal structure are schematized in Table 5.2, and their arrangement within stem levels is shown in Table 5.3.

PROOFS

Table 5.4 *Position of applicative, causative, desiderative, and future singular within the CR verb structure.*

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Root	INCH	TR	**APPL** *-ni* *-si* *-wi*	**CAUS** *-ti*	**APPL** *-ki*	**DESID** *-nale*	MOT	EV	**TAM** *-ma*	TAM	TAM	SUB
Inner Stem	Derived Stem		Syntactic Stem			Aspectual Stem			Finite Verb			Subord. Verb

The phonological cohesion of concatenative markers in the CR verb is gradient and results from the asymmetric application of regular phonological processes (including vowel harmony, compensatory lengthening, morphologically conditioned vowel lengthening, and stress shifts) in the verbal structure. Markers closer to the stem in this language are less productive and undergo more phonological alternations, making them potentially difficult to parse out of the word (Caballero 2008). Thus, the different CR constructions are expected to be more or less agglutinative-like depending on their position within the morphological structure.

3.2 Morphological Constructions

From the large set of morphological constructions in CR, we focus here on applicative, causative, desiderative, and future singular constructions. These constructions, all suffixes, belong to different stem levels of the CR verbal morphological structure as outlined in Section 3.1 above. Their position within the CR verb is highlighted in Table 5.4.

We provide details on each of these constructions below.

3.2.1 Applicatives CR has several applicative suffixes, all of which add a benefactive/malefactive object argument to the base predicate ('to do X for/against Y'). A subset of the applicative suffixes (*-ni*, *-si*, and *-wi*) occupy a suffix slot close to the stem in the morphological verb structure (S3) and are of limited productivity and lexically conditioned by the roots to which they attach. The contrast between basic predicates and their applicative counterparts is shown in (8):

(8) Inner applicative suffixes

a. pá-ka
throw-IMP.SG
'Throw it!'
[BFL 06 5:147/el]

PROOFS

b. tamí ku pá-**ʃi**-ri pelóta
1SG.OBJ REV throw-**APPL**-IMP.SG ball
'Throw the ball back at me!'
[BFL 06 5:147/el]

c. ne ma wí-ma sunú
1SG.SBJ now harvest-FUT.SG corn
'I'll harvest corn now'
[LEL 06 4:151/el]

d. wí-**ni**-mo=n olá ne jé-ra sunú
harvest-**APPL**-FUT.SG=1SG.SBJ CER 1SG.SBJ mom-POSS corn
'I will harvest the corn for my mom'
[BFL 06 5:146/el]

e. waʔlú na atʃá bilé aʔpéri=ti ané
big PROX SIT.TR one lump=1PL.SBJ say
'They put (lit. sit) a lot of what we call an *a'péri* (a lump)'
[LEL tx19:33/Text]

f. mi=n napítʃi atʃí-**w**-mo la towí
/mi=ni napítʃi atʃí-**wi**-ma olá towí/
2SG.OBJ=1SG.SBJ fire sit.APPL-**APPL**-FUT.SG CER boy
'I will sit your boy down next to the fire'
[BFL 06 6:146/el]

These inner applicative markers are always unstressed and may be the target of round vowel harmony or phonological reduction, with the suffix vowel reduced to a schwa or deleted (e.g., [8f]).[5]

In addition to these inner applicative suffixes, there is an outer applicative suffix *-ki* (occupying suffix slot S5) that is fully productive and displays no restrictions in its distribution. This suffix does not undergo stress-based vowel reduction or allomorphy, though it may undergo round vowel harmony. This suffix introduces an additional argument to one-place or two-place predicates. Like the other applicatives, the argument introduced is a benefactive or malefactive argument. This suffix is exemplified in (9):

(9) Outer applicative *-ki* suffix

a. ma=n rata-bá-tʃi-ki koʔwá-ami
already=1SG.SBJ heat-INCH-TR.PL-CONJ eat-PTCP
'I already heated up the food'
[BFL 08 1:20/el]

[5] The surface form of the /-si/ suffix in (8b) is due to a general process of palatalization of fricatives before high, front vowels.

PROOFS

Table 5.5 *CR valence stem allomorphy.*

	Intransitive	Transitive	Applicative	Gloss
a.	suwi	suwá	suwé	'run out/finish up'
b.	sawi	–	sawé	'cure, heal'
c.	–	rará	raré	'buy'
d.	noko	–	noké	'move'
e.	–	itʃ͡á	itʃ͡í	'plant'
f.	uku	–	uké	'rain'
g.	wiri	wirá	wiré	'stand'
h.	tʃ͡oʔi	tʃ͡oʔá	tʃ͡oʔí	'extinguish'
i.	–	osá	osé	'write'
j.	–	kimá	kimé	'cover with blanket'

b. ne mi baʔwí rata-bá-tʃ͡-**ki**-ra
1SG.SUBJ 2SG.OBJ water heat-INCH-TR-**APPL**-POT
'Shall I heat the water for you?'
[BFL 08 1:21/el]

Finally, CR also encodes the applicative through a valence stem allomorphy system, with distinct intransitive, transitive, and applicative stems. In this system, summarized in Table 5.5, intransitive stems end in an unstressed vowel, transitive stems end in a stressed, low mid vowel, and applicative stems end in a stressed front vowel.

Applicative stems marked through this stem allomorphy system are part of the Inner Stem domain within the CR verbal structure, the innermost domain of verbal morphological structure. We do not attempt to predict applicativeness of stems here, treating it as an inseparable part of the stem's meaning rather than the same as the meaning of the applicative suffix. This leaves the following surface allomorphs as expressions of the Applicative, as transcribed in the database: *e*$_{1/1}$, *é*$_{36/43}$, *ée*$_{5/5}$, *yé*$_{3/3}$, *u*$_{2/5}$, *i*$_{2/12}$, *í*$_{1/1}$, *k*$_{4/8}$, *ki*$_{44/81}$, *ko*$_{1/5}$, *li*$_{23/90}$, *n*$_{17/27}$, *nə*$_{1/1}$, *ni*$_{23/44}$, *nu*$_{2/2}$, *ri*$_{21/72}$, *s*$_{3/4}$, *si*$_{4/10}$, *sh*$_{2/10}$, *shi*$_{4/28}$. Subscripts show type frequency in applicatives over total type frequency of the morph (counting only the instances in which it is parsed out as a morph). The subscripts illustrate the challenging nature of identifying the applicative meaning using only an applicative exponent. All of the frequent cues to an applicative are at best imperfect in their reliability (only *é* and *ki* make the applicative meaning more probable than not; *ki* barely so). Conversely, all of the perfectly reliable cues are of very low type frequency, occurring in only one to three verbs.

PROOFS

3.2.2 Causative The causative *-ti* suffix is a suffix that introduces an agent (causer) argument to the argument structure of a predicate. Causativization applies to both intransitive and transitive verbs. In the causative construction exemplified in (10b), the object corresponds to the subject argument of its basic, non-causative counterpart. The introduced agent argument causes the undergoer to perform the activity described by the verbal root.

(10) Causative suffix

a. ne mi rimé-ni-ra
1SG.SBJ 2SG.OBJ make.tortillas-APPL-POT
'I can make you tortillas'
[BFL 08 1:161/el]

b. mi=n ne ono-rá
/mi=ni ne ono-ra/
2SG.OBJ=1SG.SBJ 1SG.SBJ father-POSS
rimée-n-**ti**-ma
/rime-ni-**ti**-ma/
make.tortillas-APPL-**CAUS**-FUT.SG
'I will make you make tortillas for my dad'
[BFL 08 1:161/el]

The causative suffix has two lexically determined allomorphs: *-ti* and *-ri*. The allomorphy is also partially phonologically determined, since there is a phonological process that devoices voiced stops after another consonant (a derived environment stemming from stress-conditioned syncope). Examples of the distribution of allomorph *-ti* after consonant-final bases are provided in (11).

(11) Phonological distribution of causative allomorph *-ti*

	Form	*Unattested*
a.	láan-**ti**-ki bleed-**CAUS**-CONJ 'I made him bleed' [SFH 05 1:102/el]	*lán-**r**i-ki
b.	sikirép-**ti**-ki 'cut-**CAUS**-CONJ' 'I made him cut it' [BFL 05 1:113/el]	*sikirép-**r**i-ki

A word can also have two causative exponents, which results in the sequence *r-ti*. The causative suffix is extremely productive, displaying no restrictions as

PROOFS

to the bases to which it can attach. Like the applicative suffixes, the causative suffix is never stressed and may undergo round vowel harmony as well as vowel reduction. Thus, on the surface, the following allomorphs are observed: $t_{1/1}$, $ti_{97/105}$, $tə_{1/1}$, $ta_{3/4}$, $r_{76/98}$, $ri_{49/72}$, $rə_{2/2}$, $ra_{5/26}$. The challenge of detecting a causative meaning from its exponent appears to be easier than detecting an applicative meaning from its exponent, because most of the time the causative is expressed by a frequent and reliable cue (*ti*, *r* or *ri*), or by both *r* and *ti* together.

3.2.3 Desiderative The desiderative suffix *-nale* is a suffix that encodes agent-oriented modality, with the meaning 'X wants to/feels like doing Y,' where the argument experiencing the "wanting" and the subject of the desideratum predication are co-referent. This suffix, which may be stressed or unstressed, occupies suffix slot S6. It is exemplified in (12).

(12) Desiderative suffix

a. ne bilé nijúrka sebá-**nale** ba
INT one stubbornly reach-**DESID** CL
'He really wanted to reach it'
[BFL 07 tr191/Text]

b. t͡ʃukúri-li t͡ʃapi-**nál**-a
/t͡ʃukúri-li t͡ʃapi-**nále**-a/
go.around-PST grab-**DESID**-PROG
'He was going around wanting to get him'
[LEL 06 tr5/Text]

As shown in (12b), the final vowel of the desiderative suffix may be replaced by a vocalic t[illegible]uffix (in this case the progressive *-a* suffix). Like other disyllabic suffixes in [illegible] language, the desiderative suffix has a "short" monosyllabic suppletive allomorph (*/-na/*). Both disyllabic and monosyllabic allomorphs undergo general phonological reduction processes that target vowels in unstressed syllables as well as round harmony. The following variants are observed in the database: $nari_{7/7}$, $nár_{26/27}$, $nər_{1/1}$, $nári_{35/36}$, $nir_{8/8}$, $nil_{3/3}$, $niri_{20/20}$, $n_{10/27}$, $ni_{17/44}$, $náari_{1/1}$, $náal_{3/3}$, $nál_{8/8}$, $náro_{6/6}$, $n@ro_{1/1}$, $nəra_{2/10}$. Desideratives are expressed by the second widest variety of morphs after applicatives. Compared to applicatives, the frequent morphs tend to be quite reliable cues to the meaning.

3.2.4 Future Singular The future singular suffix is the outermost suffix of the constructions considered for this study, occupying suffix position S9. This suffix has an unstressed allomorph *-ma* and a stressed allomorph *-méa*. This is a case of suppletive allomorphy, given that the vocalic alternation exhibited in

PROOFS

these allomorphs is not attested anywhere else in the language due to presence/absence of stress. Both the unstressed and stressed allomorphs of the future singular suffix are exemplified in (13).

(13) Future singular suffix

a. he ná=ni sipútʃa sipu-tá-**mo** lá
/he ná=ni sipútʃa sipu-tá-**ma** olá/
it PROX=1SG.SBJ skirt skirt-VBLZ-**FUT.SG** CER
'I will wear this skirt'
[BFL 07 Sept 6/el]

b. ma muku-**méa** rajénali
already die-**FUT.SG** sun
'There will be an eclipse' (lit. 'The sun will die')
[SFH 05 2:63/el]

Verbs inflected for future tense are generally followed by epistemic modality markers that indicate the degree of certainty speakers have about the actuality of an event (e.g., [11a]).[6] This example also illustrates the post-lexical phonological effect that these particles have on the inflected verb's final vowel, which is replaced by the first vowel of the epistemic modality marker (/V-m**a** oˈla/ -> [V-m**o** ˈla]). This suffix does not undergo any phonological reduction processes in the lexical phonology. As a result, only three allomorphs are observed: *ma*$_{212/220}$, *méa*$_{27/27}$, and *mo*$_{42/42}$. Detecting a future singular is clearly a much easier task than detecting any of the other meanings: there are only three exponents, all reliable and frequent.

3.2.5 Summary Based on the distribution of type frequency and reliability, the future singular appears to be easiest to detect, followed by the causative, followed by the desiderative, followed by the applicative. Both desideratives and applicatives ought to be fairly challenging to detect from their suffixal exponents, without relying on context, because so many of their exponents have low type frequency. Applicatives would appear to be particularly challenging because the type-frequent exponents are also relatively unreliable.

Table 5.6 summarizes the CR constructions surveyed in this section in terms of their flexivity (whether they exhibit suppletive allomorphy), their productivity, and how much morphophonologically regular surface variation they exhibit, quantified in terms of three regular morphophonological processes: round harmony, stress-based vowel reduction, and stress-based vowel deletion. In this table "✓" indicates that a given suppletive allomorph is a

[6] Forms lacking such particles have a neutral interpretation with respect to the speaker's commitment to the expectation that the event encoded by the predicate will take place or not in the future.

PROOFS

Table 5.6 *Surveyed CR morphological constructions with suppletive allomorphs and regular morphophonological changes.*

	Suppletive Allomorphs	Round harmony	Stress-based V reduction	Stress-based V deletion	Productivity
Applicative	*-ni*	✓	✓	✓	Unproductive
	-si	✓	✓	✓	Unproductive
	-wi	✓	✓	✓	Unproductive
	-ki	✓	✓	–	Productive
Causative	*-ti*	✓	✓	–	Productive
	-ri	✓	✓	–	Productive
Desiderative	*-nale*	✓	✓	–	Productive
	-nále	–	–	–	Productive
	-na	–	✓	–	Productive
	-ná	–	–	–	Productive
Future Singular	*-ma*	–	–	–	Productive
	-méa	–	–	–	Productive

candidate for undergoing round harmony or stress-based vowel reduction or deletion.[7]

While all suffixing constructions surveyed exhibit flexivity (through suppletive allomorphy), there is a greater degree of dependence on the phonological environment of the inner suffixes, exhibiting a greater amount of allomorphic variance overall in the morphological structure of the verb. As mentioned above, inner suffixes are also less productive than outer ones.

Finally, it should be noted that vowel reduction (mostly vowel height neutralization) results in homophony (e.g., the allomorph of the monosyllabic desiderative allomorph is rendered homophonous in its surface form to the applicative suffix after vowel reduction [desiderative /-na/ → [-ni] vs. applicative /-ni/]). Homonymous morphemes can often be disambiguated using phonotactic or morphotactic context. For example, in the surface word *ri'mee**n**tima* the only interpretation available for formative *-n* (reduced after post-tonic vowel deletion) is that of an applicative (in suffix position S3), since it precedes the causative *-ti* suffix (in suffix position S4); in this case, a desiderative reading (in suffix position S6) is not available. The dependence of interpretation on context is expected to reduce the extent to which NDL relies on morphemic cues relative to cross-boundary cues that identify the morpheme's local context.

[7] Excluded from this table are processes of lexical final vowel replacement and post-lexical final vowel replacement, which are restricted to the disyllabic desiderative allomorph and the unstressed future singular allomorph, respectively. We also exclude applicative stems formed through stem allomorphy and consider only constructions that involve suffixation outside of the inner stem domain.

PROOFS

In other cases, it is not possible to disambiguate surface homophonous suffixes using lexical context, such as immediately following a verb stem and preceding an inflectional suffix: the wordform ˈ*wi**ni**ma*, for instance, could be interpreted as containing an applicative suffix ('S/he will harvest for her') or a desiderative suffix ('S/he will want to harvest'). Such cases reduce the reliability of morphemic cues but do not boost the reliability of cross-boundary cues. They therefore contribute to flexivity based on our measure but not as much as cases of disambiguated homonymy.

Homonymy between CR morphological constructions occurs only in certain contexts, contrary to what is expected in a language with predominance of flexive morphological patterns, such as Latin, Russian, or Serbian, where associations between affixes and their selecting stems is a lexical matter. That is, the "flexive-like" properties of CR stem from phonological processes that contribute to the blurring of boundaries between morphological constructions. These processes can in principle be "undone" via "phonological inference" (e.g., Darcy et al. 2007, Gaskell and Marslen-Wilson 1996). However, this requires proposing intermediate representations, which goes against the basic assumptions of the "wide learning" approach embodied by NDL/RW and other single-layer perceptrons, where forms and meanings are connected directly (e.g., Arnold et al. 2017). Thus, in the present application of this model, forms directly cue semantic features. Without intermediate representations, boundary-blurring phonology of this kind reduces the reliability of context-independent morphemic cues. We now describe the specifics of the modeling approach.

4 Computational Approach

As mentioned above, the Rescorla-Wagner (1972) model seeks to identify the cues that are most predictive of particular outcomes. In other words, the model learns to discriminate among cue sets that are paired with distinct outcomes. The weight from a cue C to a present outcome O at time $t+1$ is increased via Equation (1), while the weight from a cue C to an absent outcome is decreased using Equation (2). The learning rate Λ is determined by the salience of the cue and outcome in question. Activation is defined as in (3) in RW and as in (4) in the Logistic perceptron, which is the only difference between these two models (Dawson 2008). Danks (2003) showed that, at equilibrium, the weights settle on (4), which is how the weights are defined in the NDL model (Baayen et al. 2011).

In our application, the outcomes are morpheme meanings such as future, desiderative, applicative, and causative. Whereas previous applications of NDL to morphology used exclusively diphones or triphones as cues, we used morphemes and cross-boundary diphones. (Simulations using diphones and triphones were also attempted, but only to evaluate whether the identified shortcomings of the NDL model for learning CR morphology were due to ~~this~~

PROOFS

choice of input encoding. They were not.) Previous applications of NDL have eschewed providing morpheme boundaries to the model because they are not available to the human learner. The model was therefore left free to discover the most predictive cues to meaning. With this approach, diphones and triphones that cross morpheme boundaries have been identified as predictive of meanings that cannot be assigned to a morpheme (Baayen et al. 2011). For example, the sequence *kbo* is highly predictive of that part of the meaning of the word *blackboard* that is not part of either the meaning of *black* or the meaning of *board* (e.g., the fact that it is for writing on and may not in fact be black).

Our goal in this study is different: we aim to determine whether non-morphemic cues are helpful for predicting meanings that *are* assignable to a morpheme, in order to discover how fused the morpheme is with the surrounding phonological context. When a morpheme occurs only in certain phonological and morphological contexts, whether because of (morpho)phonology, partial productivity, or lexically conditioned allomorphy, transitions into and out of the morpheme will be predictive of the morpheme's meaning. As shown in (4), the more predictable are morphemes given contextual cues, the weaker their associations with their own meanings should be. Even if a meaning always occurs whenever a morpheme occurs, the more predictable that morpheme's occurrence is (given the other co-occurring cues), the weaker its association with its own meaning should be. Therefore, the more agglutinative a language is, the stronger the morphemic cues should be. This prediction is a simple consequence of cue competition, which is the major difference in prediction between the RW model and earlier, Hebbian approaches to learning (see Bouton 2007, for a recent review).

We applied the model to a corpus of 814 glossed verbs, which have been elicited by the first author. We do not believe that it is realistic to present the model with our corpus of CR and hope that it would discover the best cues to meaning, superseding a morphological analysis, because the corpus is relatively small and unrepresentative compared to some other languages, like English. Even English corpora have been criticized for drastically underestimating the amount of linguistic input available to the child (e.g., Lieven and Behrens 2012, Tomasello and Stahl 2004). This is even more true of understudied languages like CR. Our corpus is vastly smaller and is also composed of elicited data. Therefore, its utterance composition is unlike that of the input to a child acquiring CR morphology. Given that the model cannot be trained on the same input as a human CR learner, we cannot hold it responsible for learning the same system of form–meaning mappings. The limited nature of the dataset is also an argument for not using the Danks equilibrium equation for the RW model (4) implemented in NDL, and instead updating the weights using Equations (1)–(2) with a conservative, slow error rate where each individual verb contributes little to the beliefs of the model.

PROOFS

Providing the model with morphemes as cues is maximally conservative with respect to identifying informative cross-boundary sequences. The morphemes have been previously identified by linguistic analysis as good cues to the meanings in question. As the model enforces cue competition, the presence of morphemes in the cue set makes it maximally difficult for cross-boundary cues to develop strong associations with the morphemes' meanings. Therefore, finding such associations would provide strong evidence for a contribution of cross-boundary sequences to meaning identification and also provide evidence for the language departing from the agglutinative ideal.

We compared future, causative, desiderative, and applicative constructions. The data examined were transcribed using a broad phonetic transcription, representing vowel harmony and stress-based vowel reduction and deletion. We first trained the model on the full set of form–meaning mappings, where meanings consisted of all morpheme meanings and formal cues consisted of morphemes and phone bigrams spanning morpheme boundaries. The training used either the Danks Equation (4) as implemented in the estimateWeights() function of the NDL package in R (Arppe et al. 2015) or a custom function implementing the RW Equations (1)–(2) kindly provided to us by Harald Baayen, which was extended to implement activation either as (3), as in the original RW model, or (5), as in the logistic perceptron ("Logistic"). Learning rate was set to 0.04 for the RW model, and to 0.12 for the Logistic. The greater learning rate in the Logistic model was chosen because the learning rate is defined on the logistic scale for the Logistic and on the probability scale for RW. The results of the training were the weights used for the rest of the modeling.

We then constructed new datasets, each of which had a binary dependent variable coding the presence/absence of a particular meaning of interest. For example, the future dataset would code each word as having the future meaning or not. This new variable was then predicted from all the cues that were associated with the presence or absence of the meaning in question above a particular threshold, using the ndlClassify() function in the NDL package. The threshold was gradually increased so that progressively fewer cues were used to predict the meaning. To do this, we took the maximum cue weight in the network and set the threshold to progressively increasing proportions of this weight, so that, for example, all weights were included, then all weights with strength above 5 percent of the maximum were included and so on.[8] As the threshold rises,

[8] An anonymous reviewer points out that a network could be badly damaged by this procedure if its accurate performance relies on negative weights and suggests ablating both positive and negative connections based on absolute value. In the present networks, however, negative connections reduce accuracy. As seen in Figure 5.1, ablation of all negative and weak positive weights increases misses rather than false alarms. Negative connections help with false alarms, where help is not needed, but exacerbate the miss rate ~~further~~. In addition, negatively weighted connections link the meaning of interest to cues to other meanings. Including them would therefore make this measure not a measure of how agglutinative the expression of a particular meaning (like "causative") is.

PROOFS

fewer and fewer cues are included until none are left. At that point, the lines in the graphs below stop. Based on (4), we suspected that this point would come earlier for meanings that are expressed in less agglutinative ways.

The ndlClassify() function outputs a number of measures of the model's performance. In Figures 5.1–5.3, we concentrate on misses (not detecting the meaning when it is there) and false alarms (falsely detecting the meaning when it is not there).

The order of observations (verbs) in the database is of course arbitrary. It does not reflect the order in which human learners encounter the verbs, and the analyst could enter verbs into a database in any order. In addition, we do not know the token frequency distribution over the verbs, aside from suspecting that it is Zipfian, which appears to be true of almost any corpus (Baayen 2001). Because both of these variables matter for what the models learn, we created 100 replication runs of each model. For each run, we randomly reordered the verbs, and imposed a new Zipfian frequency distribution over them using the rzipfman() function in the tolerance package (Young 2010) in R (with N = 10 and s = 1). All three models were applied to the same orders and frequency distributions. We report results that are robust to order and frequency distribution differences (Table 5.7), and the variability across these differences (Figures 5.1–5.3).

The data and code necessary to reproduce Figures 5.1–5.3 and Tables 5.7–5.8, or to run these models on other data, are available online: https://osf.io/ctexa/.

5 Results

Table 5.7 compares the best cues to the meanings according to each model. As mentioned above, we reran each model 100 times, changing the order of observations in the corpus and the distribution of token frequencies over words every time. For each rerun, the five strongest cues to each meaning were extracted. Table 5.7 reports the cues that were among the five best in 50 percent or more of these reruns. These are cues that the models reliably extract from the database regardless of the order in which the verbs appear in the database and their relative token frequencies.

Recall that the NDL model is the equilibrium state of the RW model: what would happen if the RW model were trained on the present dataset until its weights ~~would~~ no longer change after another run through the training data. The table shows that NDL and RW are largely in agreement with respect to the morphemic cues to the future singular meaning but disagree quite radically on the cues to the other meanings. The difference between the models is particularly clear for the causatives and applicatives. Here, RW is successful at identifying some of the best morphemic cues to the two meanings, whereas NDL is not. In fact, none of the morphemic cues to either meaning are reliably detected by NDL.

PROOFS

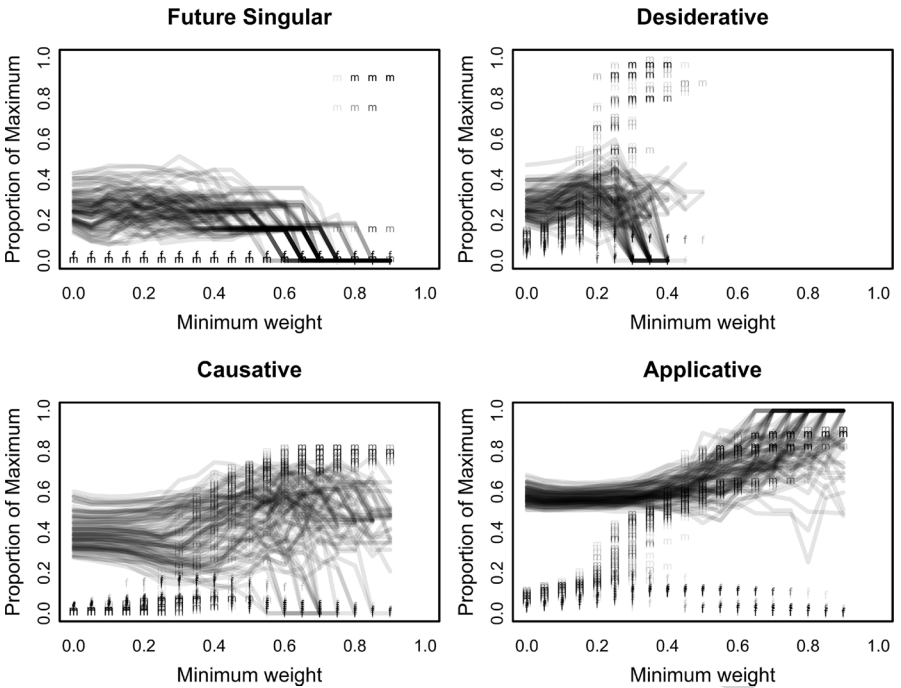

Figure 5.1 The equilibrium equation model (NDL). The x-axis shows the minimum cue weight required for the cue to be included in the model. The y-axis shows miss and false alarm rates, and the proportion of cues that are non-morphemic (lines, with each line corresponding to one run of the model). As one excludes more cues, by imposing a stricter cue weight cutoff, the proportion of "misses" – failures to detect the meaning when it is there – increases dramatically for causatives and applicatives ('m' points). The proportion of false alarms ('f' points) – detections of the meaning when it is not present in the signal – is generally lower.

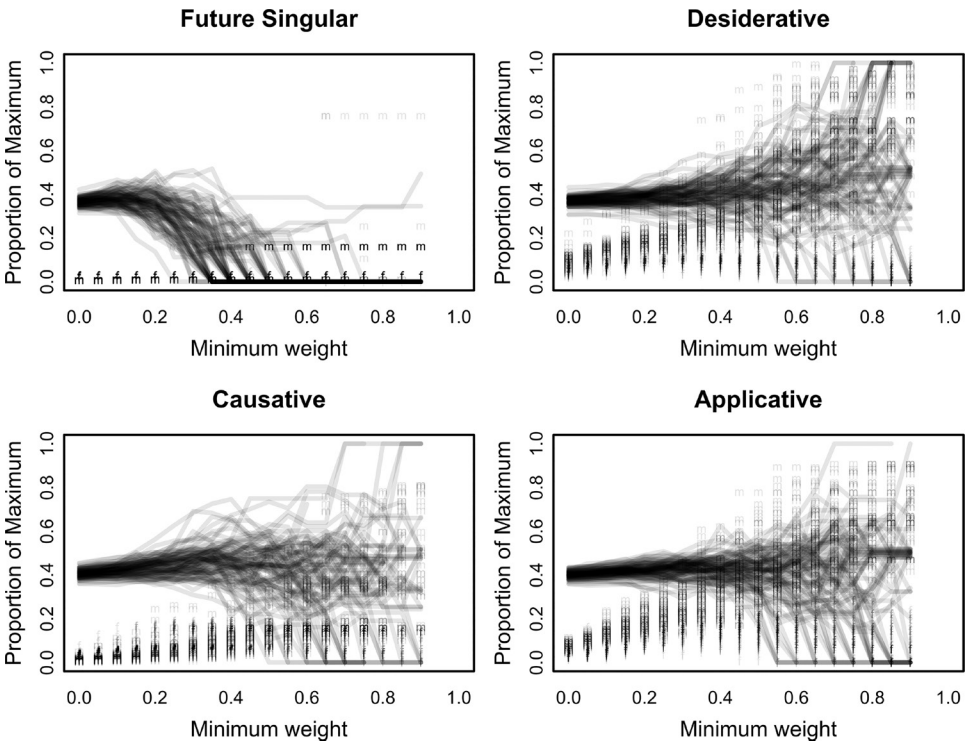

Figure 5.2 Rescorla-Wagner training with a slow learning rate. Misses (m points), false alarms (f points), and the proportion of cues that are morphemic (lines) as a function of how strong a cue has to be in order to be included as a predictor of meaning (Minimum weight).

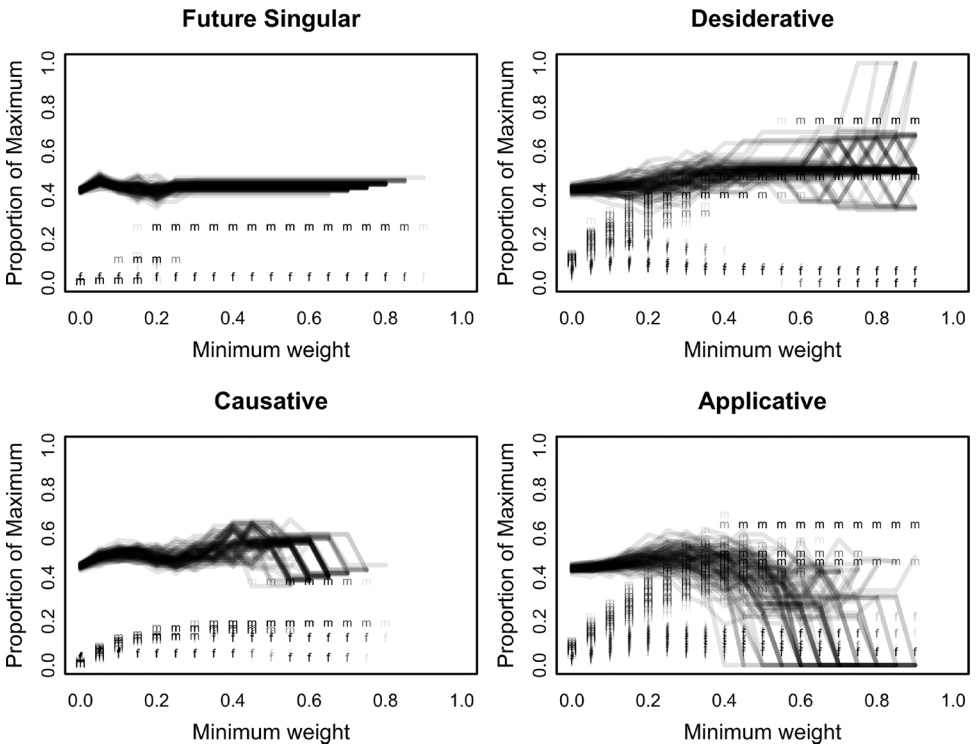

Figure 5.3 Logistic perceptron training. Misses (m points), false alarms (f points), and the proportion of cues that are morphemic (lines) as a function of how strong a cue has to be in order to be included as a predictor of meaning (Minimum weight).

Table 5.7 *The best cues to each meaning according to each model.*

	NDL	RW	Logistic
Future Singular	*méo $(100\%)_{1/1}$ *méa $(100\%)_{27/27}$ *ma $(100\%)_{212/220}$ *mo $(100\%)_{42/42}$	*mo $(100\%)_{42/42}$ *ma $(100\%)_{212/220}$ *méa $(100\%)_{27/27}$	*mo $(100\%)_{42/42}$ *ma $(100\%)_{212/220}$ i-m $(100\%)_{177/194}$ o-r $(59\%)_{38/43}$
Desiderative	kúcha $(99\%)_{1/1}$ *nər $(98\%)_{1/1}$ *nir $(88\%)_{8/8}$ *nari $(58\%)_{7/7}$	*niri $(94\%)_{20/20}$ *nari $(65\%)_{7/7}$	*nári $(100\%)_{35/36}$ ʔ-n $(100\%)_{38/38}$ koʔ $(100\%)_{38/40}$ i-n $(98\%)_{38/46}$ *nár $(73\%)_{26/27}$
Causative	o-n $(90\%)_{3/21}$ nər $(54\%)_{0/1}$ niúr $(54\%)_{2/2}$ simér $(52\%)_{1/1}$	*ti $(80\%)_{97/105}$ *ri $(75\%)_{49/72}$	*ti $(100\%)_{97/105}$ *r $(100\%)_{76/98}$ r-t $(100\%)_{61/64}$ á-r $(100\%)_{73/93}$ i-m $(63\%)_{106/194}$
Applicative	r-p $(100\%)_{1/4}$ á-n $(95\%)_{1/32}$ ó-k $(68\%)_{3/4}$ r-n $(64\%)_{0/8}$ u-á $(64\%)_{0/6}$	*é $(62\%)_{36/43}$	*é $(100\%)_{36/43}$ *ki $(99\%)_{44/81}$ ú-n $(75\%)_{22/33}$ é-k $(72\%)_{23/31}$ *ni $(64\%)_{23/44}$

Note: Proportion of times in which the cue appeared in the top five cues across replications in parentheses. Subscript: type frequency/scope, i.e., the number of words in which the form has the associated meaning over the number of words in which the form is present. Starred cues favor the meaning in the corpus, i.e., the meaning is more likely in their presence than in their absence. Underlined cues make the occurrence of the meaning they cue more likely, i.e., the probability of the meaning in the presence of the cue is higher than its probability in the absence of the cue.[9]

The cues considered best by the NDL model are often of low type frequency. For example, *méo* occurs only once in the dataset, with the future singular meaning; it is consistently considered among the five best cues to the meaning by NDL but not by RW. As indicated by (4), the weights of the NDL model reflect the conditional probabilities of meanings given forms. A conditional probability of a meaning given a form is the frequency of that form–meaning pairing divided by the frequency of the form. As a result, forms that occur only once, and happen to co-occur with a particular meaning on that occasion, are considered to be the best cues to that meaning. Albright and Hayes (2006) call this "the problem of accidentally exceptionless generalizations" and suggest that type frequency of a mapping should be taken into account. As long argued by Bybee (1985, 1995a, 2001), a generalization that holds for many observed words is likely to hold for new words as well, whereas one that holds for only a few words may not (see also Kapatsinski 2018c, Perfors et al. 2014). The RW and Logistic models naturally incorporate

[9] The data in these tables are represented using a Spanish-based orthography for the broad phonetic transcription (e.g., <ch> = [t͡ʃ]).

PROOFS

Table 5.8 *The best cues to each meaning according to each model when all verbs are assumed to have the same token frequency (n = 1), rather than having a Zipfian token frequency distribution, but the order of verbs varies across runs of a model.*

	NDL	RW	Logistic
Future Singular	*méo (100%)$_{1/1}$ *méa (100%)$_{27/27}$ *ma (100%)$_{212/220}$ *mo (100%)$_{42/42}$	*mo (100%)$_{42/42}$ *ma (100%)$_{212/220}$ i-m (100%)$_{177/194}$ o-r (100%)$_{38/43}$ *méa (100%)$_{27/27}$	*mo (100%)$_{42/42}$ *ma (100%)$_{212/220}$ i-m (100%)$_{177/194}$ o-r (100%)$_{38/43}$ rá (100%)$_{52/77}$
Desiderative	kúcha (100%)$_{1/1}$ *nər (100%)$_{1/1}$ *nir (100%)$_{8/8}$ *nari (100%)$_{7/7}$ *n@ro (100%)$_{1/1}$	*nári (100%)$_{35/36}$ *niri (100%)$_{20/20}$ i-n (100%)$_{38/46}$ ko? (83%)$_{38/40}$	*nári (100%)$_{35/36}$?-n (100%)$_{38/38}$ i-n (100%)$_{38/46}$ ko? (100%)$_{38/40}$ *nár (100%)$_{26/27}$
Causative	o-n (100%)$_{3/21}$ o-m (100%)$_{4/11}$ nər (100%)$_{0/1}$ niúr (100%)$_{2/2}$ siméŕ (100%)$_{1/1}$	*ti (100%)$_{97/105}$ *ri (100%)$_{49/72}$ á-r (100%)$_{73/93}$ á-t (100%)$_{5/19}$ r (93%)$_{76/98}$	*ti (100%)$_{97/105}$ *r (100%)$_{76/98}$ r-t (100%)$_{61/64}$ á-r (100%)$_{73/93}$ i-m (63%)$_{106/194}$
Applicative	r-p (100%)$_{1/4}$ á-n (100%)$_{1/32}$ ó-k (100%)$_{3/4}$ r-n (100%)$_{0/8}$ u-á (100%)$_{0/6}$	*é (100%)$_{36/43}$ *ni (100%)$_{23/44}$ ú-n (100%)$_{22/33}$ *ki (92%)$_{44/81}$ pá (57%)$_{6/17}$	*é (100%)$_{36/43}$ *ki (100%)$_{44/81}$ ú-n (100%)$_{22/33}$ é-k (100%)$_{23/31}$ *ni (100%)$_{23/44}$

Note: Grayed-out forms are shared with Table 5.7.

type frequency. Because the learning rate in these models is limited, exposure to a single word results in only a small change to the form–meaning weights constituting the model's beliefs. Only by repeatedly experiencing the same mapping will the model increase the weight of the association representing that mapping.

Another problem afflicting the NDL cues is spurious excitement. It can be illustrated by *nər*, which is considered to be a strong cue to the Causative meaning by NDL but is actually a desiderative morph. It occurs on only one verb in the corpus, and that verb is not causative. To understand why *nər* has a strong association with the causative meaning for NDL, we need to examine the other cues that occur in the same word as *nər*. It turns out that all of these cues have strong negative associations with the causative meaning: /á+n/ occurs in the corpus sixteen times, never in a word with a causative meaning; /r+o/ occurs forty-three times, only five times with a causative; and *-o* occurs in the corpus fifty-nine times, only four times paired with a causative. If *nər* had zero association with the causative meaning, the word would strongly inhibit the causative meaning: the activation of causative given the other cues in the word would be 0.93. The RW model adjusts cue weights so that outcome activations stay between zero and one. Whenever the activation of an outcome is below zero

PROOFS

on a trial, weights of connections between the cues present on that trial and the outcome in question are adjusted upward. Similarly, whenever the outcome activation is above one, the weights are adjusted downward. This is the source of the problem in the case of *nər*: in order to keep activation of causative close to zero given the one word that contains *nər*, *nər* must have a strong positive association with the causative meaning. All but one of the strongest cues to the Applicative meaning in NDL are also spurious excitors of that meaning.

NDL is the equilibrium state of the RW model. Thus, RW will also show spurious excitement if it learns fast enough or long enough. However, if learning rate is kept low, spurious excitement can be avoided because activations of outcomes do not (yet) overshoot the limits. Thus, we do not see any spurious excitors in the RW column in Table 5.7: all of the cues listed for a meaning actually do favor it, i.e., the occurrence of the cue predicts the occurrence of the meaning, making the meaning more likely than if the cue did not occur. However, because the learning rate has to be kept low to avoid spurious excitement, the model misses some of the cues to each meaning. This is particularly clear when the cues in the RW column are compared to those in the Logistic column. The Logistic model passes RW activations through the logistic function before comparing them to the limiting values, zero and one, ensuring that the limits are never overshot. Because of this, the Logistic can learn arbitrarily quickly without producing spurious excitement (Kapatsinski 2021a).

The logistic model also turns out to reliably detect more cues to meanings than RW can for the present data. This is because RW shows enormous variability in what cues it considers to be best across runs. Thus, for the applicatives, only the most reliable cue (*é*) is detected to be among the five best in the majority of runs and even it is detected only 62 percent of the time. This great range of variability is also apparent in Figure 5.2: the top cues to Causative, Desiderative, and Applicative for RW are all morphemic half the time, and all cross-morphemic the other half. This variability is much smaller for the logistic model, which is able to identify a larger number of cues across runs, including reliable cues that span morpheme boundaries. A good example is *r-t*, which occurs with the Causative sixty-one out of the sixty-four times it is seen in the corpus. This is because *r-t* is usually the sign of multiple exponence of the Causative, forming the transition between two Causative morphemes. The logistic model is the only one that is able to consistently detect this reliable cue to Causative from the present data.

Although the Logistic appears not to detect *méa*, *niri*, *nari*, and *ri*, compared to RW, in Table 5.7, this is only because the Logistic considers these to be weaker cues. Thus, *méa*, *niri*, and *ri* always fall within the top ten but not within the top five for the model; *nari* occurs within the top ten less frequently, at 20 percent. This appears reasonable because all three cues, and especially *nari*, have lower type frequency than the cues to desiderative that the Logistic considers to be stronger. In contrast, RW does not reliably place any of the cues that only the Logistic detects in Table 5.7 in its top ten, except for *ʔ-n*

PROOFS

(54 percent), *r* (54 percent), and *ú-n* (51 percent). Interestingly, *r-t*, which defines the transition between the two Causative exponents, never appears in the top ten cues to the Causative for RW but always appears in the top five for the Logistic, perhaps, because of stronger cue competition in RW.

Some morphemic cues that the Logistic identifies with meanings in Table 5.7 are not exponents of those meanings. In particular, *koʔ* is a stem meaning 'to eat' rather than a desiderative marker. However, all of the cues it discovers are strongly predictive of the corresponding meaning and have high type frequency. Thus, *koʔ* is a good cue to the Desiderative meaning in this corpus, occurring with that meaning in thirty-eight distinct verb forms, and occurring without it only twice. That is, it predicts that a human learner of this lexicon would expect that the speaker is about to talk about wanting or not wanting to eat (something) when hearing the stem 'eat'. This appears to be an accurate prediction because morphs often absorb meanings that co-occur with them during grammaticalization (Bybee 1988b, Traugott 1988) and develop semantic prosody based on co-occurrence with meanings that are ostensibly 'part of' the surrounding context (Sinclair 1991). For example, *since* has come to mean 'because' rather than merely 'after' because most of the time when the speaker says 'X happened after Y' it can be inferred that X also happened because of Y. This inference became associated with the morpheme *since* over time even though *since* was not originally a causative morpheme (Traugott and Koenig 1991). Similarly *cause* has developed a negative semantic connotation because of occurring in negative contexts, to the point that *cause happiness* sounds odd (Sinclair 1991). A similar development might well happen to the epistemic certainty marker *rá* in CR, which is detected to be a cue to the Future by the Logistic model because it tends to occur in future forms (Table 5.8). Future markers tend to evolve out of epistemic markers because of this correlation, making epistemic > future a common grammaticalization pathway (Bybee and Pagliuca 1987, Bybee et al. 1994). The perceptron is able to account for these types of changes because it learns many-to-many mappings between meanings and forms rather than subscribing to the agglutinative ideal (see also Feldman and Moscoso del Prado Martín this volume).

Many morphemic cues to the Desiderative and Applicative are relatively weak across models, because phonological processes (vowel harmony and vowel reduction) obscure the commonalities between allomorphs. In order to see whether a more abstract encoding would help the models, we replaced all variants of /nVr(V)/ with /nr/. With this change, the RW and Logistic models successfully identified *nr* as the best cue to the desiderative meaning. In contrast, NDL continues to struggle, identifying *kúcha* 'to have children' as the strongest cue to the Desiderative meaning even though it co-occurs with the Desiderative in only one word. However, even with this encoding /nr/ is far from the only cue to the Desiderative for any model, again deviating from the agglutinative ideal.

PROOFS

The great variability in cue weight across runs in RW appears to be due in large part to changes in the token frequency distribution across verbs between runs, rather than to differences in order. The results are quite similar to Table 5.7 if we impose different frequency distributions on different runs, but do not change the order. In contrast, there is much more consistency in cue weight across runs in RW if all words are assumed to have a token frequency of one on each run, but the order of the words in experience varies randomly, although RW is still more variable than NDL or the Logistic across runs in the cues it considers important (Table 5.8). It also still does not discover *r*-$t_{61/64}$ as a cue to Causative or *ʔ*-$n_{38/38}$ as a cue to Desiderative, even though it discovers weaker cues like *á*-$t_{5/19}$ as a cue to Causative. This is likely a result of cue competition because the missing sequences cross boundaries of informative cues. The comparison of Tables 5.7–5.8 indicates that RW is much more sensitive to the token frequency differences between words exemplifying a grammatical pattern than the Logistic is: the Logistic considers largely the same cues to be the best cues to a meaning across token frequency distributions, whereas RW does not. Conversely, the Logistic appears to be more sensitive to type frequency than RW is: the cues considered strongest by the Logistic in Table 5.7 have higher type frequency than those considered strongest by RW. RW cue weights are correlated with type frequency, controlling for reliability, less strongly than Logistic weights are, although for both the correlation is stronger for more reliable cues. Across runs, the mean correlation with type frequency for RW weights is $r = 0.09$ for all cue–outcome associations of reliability >25% (chance), and $r = 0.14$ for reliability >90%. For the Logistic, the correlations are $r = 0.23$ and $r = 0.48$ respectively.[10] We consider the stronger correlations with type frequency to speak in favor of the Logistic as a model of morphological learning, insofar as type frequency is a good predictor of productivity. At the same time, we should note that the Logistic fails to predict that type frequency is more important than token frequency (Perfors et al. 2014, Xu and Tenenbaum 2007). The correlations with token frequency are $r = 0.67$ for associations of reliability >25% and $r = 0.77$ for associations of reliability >90% for the Logistic ($r = 0.20$ and $r = 0.16$ for RW respectively). That is, the Logistic is able to detect the type-frequent cues to the morphological meanings in CR but not because they are type-frequent.

Figures 5.1–5.3 investigate the performance of the models at understanding complete words. Here, the trained models are asked to perform a binary semantic classification of complete forms, e.g., determining whether a given form is Causative or non-Causative. Form–meaning associations are pruned from the models starting with the weakest ones, to determine how well a model can understand words by relying on the cues it considers most reliable.

[10] Logistic weights were transformed back into probability space using logit^{-1}.

PROOFS

If a model successfully identifies the best cues to a meaning, its performance should not suffer dramatically from this procedure, up to a point. The figures show that all models perform very well on the Future Singulars, moderately well on Causatives and rather poorly on Desideratives and Applicatives. The RW and Logistic models are clearly superior to the NDL model on identifying Causatives. RW shows a great deal of variability in what cues it considers to be best across runs.

Figures 5.1–5.3 show that the biggest problem for the models are misses rather than false alarms: 'm' points are consistently higher than 'f' points. That is, the models tend not to hallucinate a meaning when it is absent but often fail to detect a desiderative, causative, or applicative meaning when it is present. This is interesting because these ablated models lack all cues to the absence of a meaning. Ablation of such cues would have produced a lot of false alarms if they were important for the model to extract. However, false alarm rates are quite low. The especially high miss rates in ablated NDL models are attributable to the fact that the strongest cues NDL identifies are often cues of very low type frequency. For example, the strongest NDL cues to the Desiderative meaning in Tables 5.7–5.8 would only allow the model to detect the Desiderative in 17 verbs, compared to 27 for RW, and 175 for the Logistic. The RW and Logistic models perform much better than NDL at not missing the meanings when they do occur, sometimes at the expense of a slightly higher false alarm rate. These results provide additional evidence for the importance of type frequency in learning morphology (Bybee 1985, 1995a) because NDL is not sensitive to type frequency while the other two models are.

6 How Agglutinative? Towards a Metric

Figures 5.1–5.3 show the proportion of cues that are non-morphemic at each estimated cue reliability level. That is, they show whether each model relies on morphemes, transitions between them, or both, to detect a particular meaning. If a model relies primarily on morphemes to detect a meaning, the lines in these figures trend downward. We hoped that the slopes of these lines would provide a metric indicating how agglutinative a particular morphological domain is, with more agglutinative domains having more downward-sloping lines. However, Figures 5.1–5.3 show that the slope direction for each domain is inconsistent across models: the Logistic model always relies on a combination of morphemic and cross-boundary cues, whereas the strongest cues for NDL are often morphemic but associated with the wrong meaning. Furthermore, the slopes vary enormously across runs for RW.

Because NDL often identifies the wrong morphemic cues to a meaning, we decided to filter the morphemic cues to retain only the ones that a human linguist would consider to be exponents of the meaning. We can then ask "how

PROOFS

strong are the morphemic cues to a particular meaning?" or, conversely, "how much can you rely on the morphemic cues alone to detect the meaning?"

Specifically, we found all morphs expressing a particular meaning in the database, and determined their average associations with the meaning they express in each model. We converted these association weights to z-scores by subtracting the mean weight and dividing by the standard deviation. We then weighted each of these associations by how often each morph occurred in the database (with the meaning it expresses). This is necessary because morphs that co-occur with the meaning rarely might justifiably be associated by a model with another meaning. For example, *ri* can express both Causative and Applicative. It accounts for a large proportion of the Causative meaning but a small proportion of the Applicative meaning. We therefore rely on its association weight with the Causative highly in determining the agglutinativity of Causative, but we do not rely on its association weight with the Applicative much to determine the agglutinativity of the Applicative.

Formally, we define A_O, Agglutinativity of an outcome (like Causative), as in (14) where C is a particular morph that expresses that meaning, V is an association weight, $\mu(V)$ and $\sigma(V)$ are the mean and standard deviations of association weights, and N is the type frequency of either the cue–outcome (form–meaning) pair or the outcome (meaning) alone.

$$A_O = \sum_C (V_{C\to O} \times \mu(V) \times N_{C,O} / (N_O \times \sigma(V))) \tag{14}$$

The results of this calculation are shown in Table 5.9. The models agree that Future Singular expression is much more agglutinative than the others, followed by Causative, Applicative, and finally Desiderative. The models disagree somewhat on how agglutinative Applicative is relative to Desiderative, with RW considering it much more agglutinative and the others judging them to be almost equal. Overall, these results align with the conclusions we drew in Section 3.2.5 by examining the distributions of type frequency and reliability across the exponents of each meaning. Future Singular is much easier to detect than the other meanings, with Causative next, followed by Desiderative and Applicative. However, the difficulty of Desiderative compared to Applicative for RW and the relative similarity of Causative to Applicative and Desiderative for the Logistic are somewhat unexpected. Overall, the NDL ranking of the meanings appears most intuitive, even though (or perhaps because!) it is the model most likely to mistakenly rely on rare cues of apparently high reliability when these are available.

The measure defined in (14) is of course only one possible definition of agglutinativity. An alternative way to quantify agglutinativity suggested by Figures 5.1–5.3 is as the slope of the rise in the miss rates. That quantification would not require us to identify a set of morphs that are exponents of a meaning by hand first. As Figures 5.1–5.3 illustrate, the rise in miss

PROOFS

Table 5.9 *Agglutinativity of each construction in CR according to each model.*

	NDL	RW	Logistic
Future Singular	2.53	5.21	0.72
Causative	0.31	0.51	0.04
Desiderative	0.01	–0.04	0.01
Applicative	0.02	0.26	0.02

rates happens with all four meanings, but the slope is shallowest for Future Singular, followed by Causative, followed by Desiderative and Applicative. This aligns with the results in Table 5.9 and with the discussion of distributions of type frequency and reliability across the exponents in Section 3.2. Another possibility to explore is, rather than comparing cue weights, to ask how well the models are able to classify forms as Causative vs. Non-causative or Applicative vs. Non-Applicative using only morphemic cues or using only morphemic cues that are exponents of the meaning based on linguistic analysis.

7 Implications for Learning Mechanisms

In the present chapter, we compared the NDL model, which represents the equilibrium state of the classic RW model of associative learning (Rescorla and Wagner 1972), to an earlier state of the RW model resulting from slow, incremental learning and a version in which activations are passed through a sigmoid logistic function before being compared to the "correct" activation levels (zero and one; Rumelhart et al. 1986). Our approach requires a model to provide numerical association weights for candidate mappings between formal cues and semantic outcomes. However, a variety of models – instantiating different learning mechanisms – could be used to provide such association weights (see Kapatsinski 2018b for a review).

While the NDL model has achieved impressive results in capturing psycholinguistic data from both synthetic and isolating languages (e.g., Baayen et al. 2011, Filipović Đurđević and Milin 2019), the plausibility of the weights on which it relies to accomplish this task has not been examined in detail in previous work. NDL would also be likely to provide promising measures for examining morphological processing in CR: NDL cue weights can be used to derive measures that reflect the transparency of form–meaning mappings involving a morpheme, and transparency has been observed to correlate with various psycholinguistic measures of processing. Yet, the analyses reported in the present chapter reveal significant limitations of NDL as a way to learn reliable form–meaning mappings, some of which are due to the limitations of the

PROOFS

CR corpus, while others can be traced to specific aspects of the RW learning rule that the model instantiates for predicting categorical outcomes.

7.1 Accidentally Exceptionless Generalizations

In particular, the high rates of misses in Figure 5.1 show that the model faces severe difficulties in recognizing three of the four meanings we attempted to recognize if it cannot make use of the full set of cues presented to it. The model predicts that more than ~~a~~ 100 cues, most of them very weak, are needed to reliably recognize applicative and causative meanings. Most of these cues are very rare and do not recur across the lexicon.

For the applicative meaning, this may ring true because the frequent exponents of the applicative meaning are unreliable, since they are homophonous with other morphemes. The applicative morphemes present in the corpus, such as *-é*, *-ni*, *-si*, and *-ki*, have decidedly weak discriminative value for the model. The cue reliability of these morphemes is driven down by the existence of multiple applicative morphemes and applicative stems, which do not bear applicative affixes. The cue validity is driven down by the fact that they are homophonous with non-applicative morphemes. For example, *-ni* occurs in applicatives twenty-three times and in non-applicatives twenty-one times. Thus, given the presence of *-ni* in a word, an applicative meaning is only slightly more likely than a non-applicative meaning. This happens because there is a homophonous desiderative *-ni* morpheme. Similarly, *-si* occurs in applicatives four times and in non-applicatives six times, and *-ki* occurs in applicatives forty-four times and in non-applicatives thirty-seven times. In the absence of strong morphemic cues to the applicative, the model relies on a wide variety of cues, largely associated with the absence of the applicative meaning. This makes the presence of an applicative meaning difficult to detect unless a large variety of cues are extracted.

However, note that the incremental models, RW and Logistic, are able to recover at least some of the traditional applicative morphemes and that this recovery allows the models to achieve lower miss rates, especially for Causatives (Figures 5.1–5.3). The NDL model appears to overweight conditional probability compared to frequency for the present data. This is somewhat unsurprising, because our corpus is a small sample of the overall set of verbs a native speaker of CR encounters in the course of language acquisition. Learners better not train themselves to equilibrium on this relatively small set of words, if they want to avoid overfitting the limited training data.

The overfitting is particularly obvious in the case of the causatives. While the causative is expressed by the suffixes *-ti*, *-t*, *-ri*, *-rə*, and *-r*, the NDL model instead assigns its highest weights to cues that are, in the long run, unreliable. In particular, the model assigns the highest weights to very rare cues that

PROOFS

happen to be always paired with a particular meaning. For example, intuitively, the best cue to the causative is *-ti*, which accounts for 52 percent of all causatives (ninety-seven verbs) and occurs in non-causatives only eight times. However, the NDL model does not consider it a strong cue at all. In contrast, one of the strongest cues to the causative meaning for the model is *simer*, which is not a causative morpheme at all. It occurs in the dataset only once, in a verb that does happen to have a causative meaning, but only because it also contains the causative suffix. The problem is not due to our rather unorthodox use of morphemic cues. When we reran the model using all diphones as cues, the strongest cue to the causative emerged to be /bt/, which also occurs in the dataset only once.

Albright and Hayes (2006) called this "the problem of accidentally exceptionless generalizations." As shown by (4), NDL weights reflect conditional probabilities of outcomes given cues. The number of observations on which the probability is based does not enter into the model's calculations. Thus, if A is followed by X 90 percent of the time, the model does not care whether this probability is based on 10 or 10,000 observations of A. Yet, it is clear that one should be more confident that A is really predictive of X if one has observed AX 9,000 out of 10,000 times rather than 9 out of 10 times one observed A (see also Xu and Tenenbaum 2007). This problem is to some extent due to the use of the equilibrium equations in NDL. The original RW model has a finite learning rate, updating the weights of cues present on a given trial by a certain, usually small amount. Therefore, the weight of a connection would have moved farther from its default (near-zero) value in 10,000 trials than in 10. In contrast, the Danks (2003) equilibrium equations assume that learning has proceeded to completion, converging on the conditional probabilities of outcomes given cues. Figures 5.1–5.3 and Tables 5.7–5.8 show that eliminating this assumption would result in more realistic cue weights. The Logistic model shows stronger correlations with type frequency than RW does, which appears to help it identify reliable cues to morpheme meanings that are expressed by many different exponents such as the CR Applicative.

7.2 Spurious Excitement

Even more puzzling and problematic than cases of cues wiring with meanings based on a single word are cases of cues becoming strongly associated with a meaning with which they have never been paired. As noted earlier, one of the strongest cues to the causative meaning, according to the NDL model, is *-nər*, which is not a causative suffix. It occurs in the dataset only once. Furthermore, the one word that bears *-nər* does not even have a causative meaning. Spurious excitement occurs because the activation that an outcome receives from the cues can exceed the maximum or minimum allowed (1 or 0),

PROOFS

and this is counted as an error. Kruschke (1992: 39) called this property of RW a "strict teacher signal" because exceeding the expected activations is considered an error. Kruschke argued for a "humble teacher" that does not consider it an error to exceed expectations. The Logistic modification also solves this problem but in a different way, by making exceeding expectations impossible. When the sum of weights has passed through the logistic function, it is always between zero and one. For a discussion of the advantages of the logistic solution to spurious excitement, see Kapatsinski (2021a).

Previous studies using NDL did not include morphemes in the set of cues available to the model. Instead, the model is presented with the full set of diphones or triphones. Intuitively, the strict teacher problem could be exacerbated by the inclusion of morphemic cues in the input to the model because the morphemic cues may be more likely to develop into strong inhibitors. To check that the diphone or triphone encoding more typical for NDL does not solve the problem we identified, we reran the models with both encodings. The problem remains. For example, with the diphone encoding, one of the strongest cues to the causative meaning is /wo/, which occurs in the dataset once and not with a causative meaning. Another is /bo/, which occurs in the dataset thirteen times, only twice with a causative sense. As causatives form 22 percent of the dataset, the causative meaning is actually more likely to occur in the absence of /bo/ than in its presence.

Kruschke (1992: 39) and Kapatsinski (2021a) argue that the strict teacher signal of the RW model is inappropriate for cases in which the outcomes are nominal, such as the discrete semantic features or forms of morphosyntax. Kruschke noted that the outcomes predicted by learners in the experiments that motivated the strict teacher assumption of the RW model (see Rescorla and Wagner 1972 for a review) could be conceived of as having a continuous magnitude, such as the strength of an electric shock or the amount of food. This hypothesis is bolstered by recent evidence that the blocking effect that motivated the RW model's assumption of a strict teacher (Kamin 1969) is observed when learners conceptualize the outcome as continuous but not when they conceptualize it as discrete (Lovibond et al. 2003, Packheiser et al. 2020).

Spurious excitement does not arise with the Logistic perceptron model: the forms considered by it to be cues to particular meanings are morphs that are associated with those meanings according to morphological analysis, transitions into and out of these morphemes, or else are common stems that co-occur with those meanings in the data. For example, *ʔ-n* and *i-n* are detected as strong cues to the desiderative, because an initial /n/ is shared by all desiderative morphs. The stem /koʔ/ 'eat,' is also detected to be a strong cue to the desiderative meaning because, in the dataset, 'eat' occurs in thirty-eight Desiderative types and occurs only twice without it (a kind of semantic prosody effect). These cues to the desiderative are stronger in type frequency and reliability than most desiderative allomorphs, showing one way

PROOFS

Table 5.10 *Simplest cue–outcome structure that would produce a spurious excitor.*

Cues	Outcomes
y1	y
x1_y1	x
x2_y1	x
z_x1_x2	x

Note: Given this training set, the RW model eventually learns that z excites y and inhibits x regardless of learning rate. Furthermore, z is the strongest cue to y. The Logistic model avoids this prediction, associating z with ~~z~~ x.

the language departs from the agglutinative ideal, and the logistic model is able to discover this fact about CR.

Although RW can also avoid exceeding zero and one by keeping the learning rate slow, this solution has a significant cost. As shown in Table 5.7, RW fails to consistently discover many of the reliable cues to the meanings of CR that the logistic modification identifies. Furthermore, it is possible to demonstrate that there are cue–outcome structures for which RW will always settle into spurious excitement with enough training. The simplest such structure is shown in Table 5.10. Here, x1 and x2 are exponents of x, the meanings x and y are incompatible (e.g., y is singular and x is plural), and there is a multiple exponence pattern such that the redundant co-occurrence of x1 and x2 is also marked by z. Multiple exponence of this type characterizes Causatives and Applicatives in CR.

When exposed to this training set, the RW model associates y1 with y and learns that x1 and x2 inhibit y and activate x (Kapatsinski 2021a). Because of the strictness of the teacher signal, the model then must learn that z activates y, so that the activation of y not be negative given z_x1_x2. Furthermore, because x1 and x2 overpower y1, z is associated with y more strongly than y1 is. At equilibrium, $w(y1 \rightarrow y) = +1$, $w(x1 \rightarrow y) = w(x2 \rightarrow y) = -1$, and therefore $w(z \rightarrow y) = +2$. That is, a cue that never co-occurs with y is learned to be the strongest cue to y. Therefore, z should be taken to mean y, even though it is part of the multiple exponence pattern for x. Kapatsinski (2021a) exposed human learners to this type of contingency structure instantiated over a simple artificial language and did not observe spurious excitement.

Thus, the spurious excitement problem is not eliminated by slowing the learning rate. Instead, it is inherent to the teacher signal of the RW model and is always exhibited by the model at equilibrium. Although a researcher can set the learning rate to be slow enough not to see spurious excitement in applying the model to a particular corpus, it is only possible to know that a certain

PROOFS

learning rate is too fast by evaluating the associations learned at that rate and determining them to be spurious. It therefore requires the researcher to know what the results "should be," i.e., what the strongest cues to each particular meaning are, which defeats the purpose of using a computational model to discover the best cues to a meaning. The logistic modification ensures that spurious excitement will not arise, no matter the learning rate, which may also help the model discover additional reliable cues to morphological meanings (Tables 5.7–5.8).

7.3 Intermediate Representations

The NDL/RW model is in principle opposed to intermediate levels of representation: it is a "wide" learning model, not a "deep" one (Arnold et al. 2017, Baayen and Hendrix 2016). Like exemplar-based and usage-based approaches to phonological theory (e.g., Bybee 2001, Johnson 1997), NDL tries to commit to a minimum of representational machinery, ideally going directly from acoustics to semantics and semantics to articulation (Arnold et al. 2017). In keeping with this spirit, we first directly mapped surface phonology onto semantics in the present application of NDL. However, this decision is questionable in the sense that regular phonology can make the system appear much less agglutinative under the assumptions made in morphological typology, where only lexically conditioned morpheme alternations are assumed to play a role. Furthermore, as Wilson and Gallagher (2018) point out, the lack of intermediate representations can exacerbate the problem of accidentally exceptionless generalizations. We showed that performance on desideratives is improved if vowel harmony and reduction are "undone." To the extent that comprehenders can actually undo the effects of vowel harmony before using the form to access the meaning, the model may need more than two layers in the form–meaning mapping (see Darcy et al. 2007, Toscano and McMurray 2015 for examples of compensation for context in speech perception).

PROOFS

8 Conclusion

In this chapter, we have operationalized agglutination as the detectability of meanings based exclusively on their exponents, without reference to context. From this perspective, in a perfectly agglutinative language, words would be semantically compositional strings of morphemes, each morpheme phonologically independent of the others. We made a first step towards quantifying the degree to which both individual morphological constructions are agglutinative in this sense by using computational models to learn to predict morpheme meanings from a combination of morphemic and boundary-spanning cues. By applying the methodology to CR, a language that displays both agglutination

and flexion, we have shown that exponents of a meaning develop stronger associations with the meaning when the expression of the meaning is more consistent across contexts. This provides a possible way to quantify the degree to which it is possible to detect the presence of the meaning in a form by relying on its morphemic exponent(s) alone. This intuition can potentially be scaled up to whole corpora and therefore has the potential to allow for whole-language comparisons. Though lack of comparable corpora may make these comparisons difficult, this problem may be ameliorated by the use of bootstrapping and resampling techniques.

Association weights can of course be derived from many alternative models. Here, we applied single-layer error-driven connectionist learners of morphology to learn the form–meaning association weights. We have identified significant issues with estimating the weights of form–meaning mappings using the RW model. First, the model is not very sensitive to type frequency, which results in reliance on cues that are relatively unlikely to occur in words that the model has not yet seen. This problem is particularly severe when a meaning can be expressed in many different ways, as is the case of CR desideratives and applicatives. Second, the model can show spurious excitement, which results in the model learning associations between forms and meanings that never co-occur. Spurious excitement is particularly likely to arise with morphemic cues when the language has multiple exponence, which makes it likely that a rare cue would occur in a word that has more than one synonymous cue. This type of pattern is prevalent in CR, which may have helped us detect spurious excitement in this language. However, spurious excitement can also be observed in the present data even if submorphemic cues are used, suggesting that it is likely to arise in any language learning situation in which the learner needs to predict categorical outcomes. Categorical outcomes appear prevalent in language learning because most forms and meanings vary in probability of occurrence but not in magnitude/intensity across contexts, the definition of a categorical outcome. This makes it important for a model of language learning to avoid spurious excitement.

Spurious excitement arises in RW because activations of outcomes can overshoot the intended maximum and minimum limits on activations and is not observed when the limits are made unachievable by passing activations through a logistic function. Future research should explore the properties and performance of alternative learning rules that avoid spurious excitement in the acquisition of morphology, including Hebbian models that remove the limits (Kapatsinski and Harmon 2017; McMurray et al. 2012, 2013; Yu and Smith 2012), humble teacher signals that tolerate exceeding the limits (Kruschke 1992) and the logistic perceptron, which makes it impossible to exceed the limits (Rumelhart et al. 1986; Kapatsinski 2021a).

PROOFS