Learnability shapes typology: the case of the midpoint pathology[*]

Juliet Stanton, *MIT*

ABSTRACT

The *midpoint pathology* (in the sense of Kager 2012) characterizes a type of unattested stress system in which the stressable window contracts to a single word-internal syllable in some words, but not others. Kager (2012) shows that the pathology is a prediction of analyses employing contextual lapse constraints (e.g. *EXTLAPSER; no 000 strings at the right edge), and argues that the only way to avoid it is to eliminate these constraints from CON. This paper explores an alternative: that systems exhibiting the midpoint pathology are unattested not because the constraints that would generate them are absent from CON, but because they are difficult to learn. This study belongs to a growing body of work exploring the idea that phonological typology is shaped by considerations of learnability (e.g. Boersma 2003, Blevins 2004, Alderete 2008, Heinz 2009).

KEYWORDS: phonology, typology, stress, learnability

1 Introduction

One of the goals of linguistic research is to construct theories that make the right typological predictions: theories that predict the existence of all and only those patterns attested in the world's languages. In constraint-based theories of phonology, such as Optimality Theory (OT; Prince & Smolensky 2004), the typological predictions of a constraint set can be evaluated by exploring its factorial typology. The notion of factorial typology is grounded in the classical assumption that the set of constraints (CON) is universal, but that constraint rankings are language-specific. If all constraints are freely rankable, then the set of systems predicted by a given constraint set, its *factorial typology*, is equivalent to the set of systems generated by each possible ranking of constraints.

When evaluating a factorial typology, there are at least two important questions that the analyst must ask. First, does the predicted typology *undergenerate:* does it fail to predict certain attested patterns? Second, does the predicted typology *overgenerate:* does it fail to predict *only* attested patterns? Undergeneration is typically viewed as a serious problem, as we want our theories to be able to account for the full range of linguistic variation. Thus the usual response to undergeneration is to modify the contents or the structure of CON, with the goal of including all attested patterns in the predicted typology. The response to overgeneration, however, is much more nuanced. Because there are multiple reasons why a proposal might overgenerate, there are multiple possible responses.

[*] My thanks to Adam Albright, Donca Steriade, Edward Flemming, René Kager, and participants at MIT's Ling Lunch, AMP 2014, NECPhon 8, and the GLOW 38 phonology workshop, for helpful comments and feedback. I am also grateful to Francesca Cicileo, for help with counting words. Comments from two anonymous *Language* reviewers and the associate editor have also led to many improvements in the current version of this paper.

One common response to overgeneration is to take a closer look at CON, and propose modifications that exclude the predicted but unattested patterns from the factorial typology. These strategies can be roughly divided into two groups. Some researchers dispute the idea that all constraints are freely rankable and propose that, in order to model typological generalizations, certain constraint rankings must be universal and therefore immutable (see e.g. Prince & Smolensky 2004 on fixed rankings for peak and margin hierarchies, Steriade 2001 on fixed rankings of correspondence constraints). Others focus on the contents of CON, arguing that certain unwanted predictions can be avoided if we exclude certain (classes of) constraints, and/or include certain others (see e.g. McCarthy 2003 on gradient ALIGN, Alber 2005 on ALL-FEET-RIGHT). What all of these proposals have in common is that they modify the contents or structure of CON in order to constrain its factorial typology. To exclude a pattern in this way is to claim that it cannot be represented by a learner: a pattern that cannot be generated by CON is not part of the learner's hypothesis space.

In addition to investigating the contents and structure of CON, much recent work in phonological theory has begun to investigate the hypothesis that other, additional factors play a role in shaping phonological typology. This paper is part of a growing body of research exploring the idea that one such factor is *learnability*. Work in this area has helped us understand why only certain types of phonotactic patterns are attested (Heinz 2010); why (classes of) gaps in the typology of stress systems exist (Boersma 2003, Heinz 2009, Staubs 2014a,b); why gang effects, predicted by weighted constraint models, are not attested in the larger typology (Hughto et al. 2015); and why some types of phonological patterns are restricted to certain morphological domains (Alderete 2008). While details of implementation vary, the basic insight across these works is the same. Whether a pattern can be generated by a particular constraint set or not, we should only expect for it to be attested if it can be learned, given the input available to an average human learner. In other words, the idea is that learnability shapes typology: the range of attested patterns that we see is shaped by limitations on the kinds of patterns that can be accurately and reliably learned.

In Section 2, I introduce Kager's (2012) midpoint pathology as a type of unattested system that is predicted by a popular set of metrical constraints, anti-lapse constraints (Elenbaas & Kager 1999, Gordon 2002), and discuss how Kager proposes to modify CON in order to exclude midpoint systems from the predicted typology. In Section 3, I introduce an alternative: midpoint systems are unattested not because the constraints necessary to generate them are absent from CON, but because they are difficult to learn. Sections 4 and 5 present results from machine learning simulations as supporting evidence for this proposal, and show that predictions of the analysis are borne out. Section 6 provides general discussion and conclusions. It is important to note at the outset that this paper is not intended as a criticism of Kager's (2012) proposal, nor does it attempt to argue that the proposed alternative is superior. Rather, the goal of the paper is only to develop the alternative – that considerations of learnability can suffice to shape this particular aspect of stress typology – and to explore some of its predictions.

2 The midpoint pathology

In many languages, stress is required to fall within a certain fixed distance from a word edge. Kager (2012) refers to these kinds of systems as *metrical window systems*, and identifies four types: right-edge with a window of two syllables, right-edge with a window of three syllables, left-edge with a window of two syllables, and left-edge with a window of three syllables. Schematic representations, the number of languages in Kager's survey, and one example of each type, are provided in (1).

(1) Summary of accentual windows (see Kager 2012: 1464; { } = accentual domain)

Edge	**Window Length**	
	Two syllables	*Three syllables*
Right	$\sigma\sigma\sigma\{\sigma\sigma\}$ (82 lgs.) Kobon (Trans-New Guinea, Davies 1980)	$\sigma\sigma\{\sigma\sigma\sigma\}$ (38 lgs.) Latin (Indo-European, Mester 1994)
Left	$\{\sigma\sigma\}\sigma\sigma\sigma$ (39 lgs.) Hopi (Uto-Aztecan, Jeanne 1982)	$\{\sigma\sigma\sigma\}\sigma\sigma$ (1 lg.) Choguita Rarámuri (Uto-Aztecan, Caballero 2011)

One possible analysis of the typology of window systems employs contextual anti-lapse constraints (Elenbaas & Kager 1999, Gordon 2002), which forbid lapses from occurring within certain specified domains of the word. For example, a system enforcing a right-edge trisyllabic window can be modeled with the constraint *ExtendedLapseRight (or *ExtLapseR, defined in (2)), which forbids a sequence of three stressless syllables from occupying the word's right edge. When *ExtLapseR is active, stress must fall on one of the final three syllables (3); other constraints will determine its exact placement within this window.

(2) *ExtLapseR: one * if the final three syllables of the word are stressless.

(3)

	/$\sigma\sigma\sigma\sigma\sigma$/	*ExtLapseR
☞	a. $\sigma\sigma\{\sigma\sigma\boldsymbol{\acute{\sigma}}\}$	
☞	b. $\sigma\sigma\{\sigma\boldsymbol{\acute{\sigma}}\sigma\}$	
☞	c. $\sigma\sigma\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}$	
	d. $\sigma\boldsymbol{\acute{\sigma}}\{\sigma\sigma\sigma\}$	*!

Other anti-lapse constraints can be employed to analyze the three remaining types of system in (1). *ExtendedLapseL (or *ExtLapseL) enforces a left-edge trisyllabic window, by forcing stress to fall within that domain; *LapseRight (*LapseR) and *LapseLeft (*LapseL) enforce right- and left-edge disyllabic windows, respectively (4-6).

(4) *ExtLapseL: one * if the initial three syllables of the word are stressless.

(5) *LapseR: one * if the final two syllables of the word are stressless.

(6) *LapseL: one * if the initial two syllables of the word are stressless.

Kager (2012) shows that a grammar including anti-lapse constraints can predict all attested window systems; however, it also overgenerates. In particular, anti-lapse constraints give rise to the *midpoint pathology*. The midpoint pathology is a term used to describe a type of system in which the stressable window contracts to a single word-internal syllable in some words, but not others.[1] Generally speaking, midpoint systems arise when two anti-lapse constraints dominate all others. For example, when the two-top ranked constraints are *ExtLapseL and *ExtLapseR (in that order), the system in (7) results. As before, accentual domains where *ExtLapseL and *ExtLapseR can be satisfied are bracketed and subscripted (or superscripted) with L and R, respectively; stressable syllables given the constraint ranking in (7) are bolded.

[1] The discussion in this paper focuses exclusively on Kager's 2012 referent of the term *midpoint pathology*. Earlier uses of the term, by Eisner (1997) and Hyde (2008), describe a different kind of pattern, in which stress gravitates towards the middle of all words, regardless of their length. These kinds of patterns are predicted by the Generalized Alignment approach to alignment constraints (Prince & Smolensky 2004); see Hyde (2015) for discussion.

(7) *ExtLapseL >> *ExtLapseR

a. $_{\text{L}}\{^{\text{R}}\{\boldsymbol{\sigma\sigma}\}_{\text{L}}\}^{\text{R}}$

b. $_{\text{L}}\{^{\text{R}}\{\boldsymbol{\sigma\sigma\sigma}\}_{\text{L}}\}^{\text{R}}$

c. $_{\text{L}}\{\sigma^{\text{R}}\{\boldsymbol{\sigma\sigma}\}_{\text{L}}\sigma\}^{\text{R}}$

d. $_{\text{L}}\{\sigma\sigma^{\text{R}}\{\boldsymbol{\sigma}\}_{\text{L}}\sigma\sigma\}^{\text{R}}$

e. $_{\text{L}}\{\boldsymbol{\sigma\sigma\sigma}\}_{\text{L}}{}^{\text{R}}\{\sigma\sigma\sigma\}^{\text{R}}$

f. $_{\text{L}}\{\boldsymbol{\sigma\sigma\sigma}\}_{\text{L}}\sigma^{\text{R}}\{\sigma\sigma\sigma\}^{\text{R}}$

In (7a) and (7b), stress may fall on any syllable in the word, as all options satisfy *ExtLapseL and *ExtLapseR. In other words, the accentual domains of *ExtLapseL and *ExtLapseR overlap entirely. In (7c) and (7d), however, the accentual domains of *ExtLapseL and *ExtLapseR only partially overlap. To satisfy both window constraints in a four-syllable word (7c), it is necessary to stress either the second or the third syllable; in a five-syllable word (7d), the accentual domain is restricted to the word's middle syllable (its *midpoint*). In words of six syllables or longer (7e-f), the domains of the two window constraints no longer overlap. Because *ExtLapseL dominates *ExtLapseR, one of the initial three syllables must be stressed.

Many other midpoint systems can be created through different combinations of context-sensitive varieties of *Lapse and *ExtLapse. For example, a system where *LapseR >> *LapseL is schematized in (8), and a system where *ExtLapseL >> *LapseR is in (9). While the specifics of the patterns in (8-9) are slightly different from those of the pattern in (7), the overall situation is the same: two contextual anti-lapse constraints compete, and in words of a certain length, the stressable domain is restricted to a single syllable in the middle of the word.

(8) *LapseR >> *LapseL

a. $_{\text{L}}\{^{\text{R}}\{\boldsymbol{\sigma\sigma}\}_{\text{L}}\}^{\text{R}}$

b. $_{\text{L}}\{\sigma^{\text{R}}\{\boldsymbol{\sigma}\}_{\text{L}}\sigma\}^{\text{R}}$

c. $_{\text{L}}\{\sigma\sigma\}_{\text{L}}{}^{\text{R}}\{\boldsymbol{\sigma\sigma}\}^{\text{R}}$

d. $_{\text{L}}\{\sigma\sigma\}_{\text{L}}\sigma^{\text{R}}\{\boldsymbol{\sigma\sigma}\}^{\text{R}}$

e. $_{\text{L}}\{\sigma\sigma\}_{\text{L}}\sigma\sigma^{\text{R}}\{\boldsymbol{\sigma\sigma}\}^{\text{R}}$

f. $_{\text{L}}\{\sigma\sigma\}_{\text{L}}\sigma\sigma\sigma^{\text{R}}\{\boldsymbol{\sigma\sigma}\}^{\text{R}}$

(9) *ExtLapseL >> *LapseR

a. $_{\text{L}}\{^{\text{R}}\{\boldsymbol{\sigma\sigma}\}_{\text{L}}\}^{\text{R}}$

b. $_{\text{L}}\{\sigma^{\text{R}}\{\boldsymbol{\sigma\sigma}\}_{\text{L}}\}^{\text{R}}$

c. $_{\text{L}}\{\sigma\sigma^{\text{R}}\{\boldsymbol{\sigma}\}_{\text{L}}\sigma\}^{\text{R}}$

d. $_{\text{L}}\{\boldsymbol{\sigma\sigma\sigma}\}_{\text{L}}{}^{\text{R}}\{\sigma\sigma\}^{\text{R}}$

e. $_{\text{L}}\{\boldsymbol{\sigma\sigma\sigma}\}_{\text{L}}\sigma^{\text{R}}\{\sigma\sigma\}^{\text{R}}$

f. $_{\text{L}}\{\boldsymbol{\sigma\sigma\sigma}\}_{\text{L}}\sigma\sigma^{\text{R}}\{\sigma\sigma\}^{\text{R}}$

While systems in which the size of the stressable window is dependent on word length do exist (see section 5.2), systems like (7-9), where the stressable window narrows and then widens again, are unattested. Thus we have a situation where a particular constraint set overgenerates: while including contextual anti-lapse constraints (e.g. *ExtLapseL, *ExtLapseR) in Con results in a theory that generates all attested window systems, it also generates some unattested systems: midpoint systems, like those in (7-9).

2.1 Expected frequency of midpoint systems

When dealing with cases of overgeneration, there is always the possibility that a predicted but unattested system is an accidental gap: that is, it exists, but hasn't been discovered yet. While it is impossible to rule this possibility out, we can evaluate whether or not it is a realistic one by determining how frequent we might expect the predicted pattern to be.

One way to evaluate the expected frequency of a pattern is to determine the number of constraint rankings that are compatible with it. If each possible permutation of rankings within a given constraint set is equally probable[2], then we might expect that the more rankings are consistent with a single surface pattern, the more frequent that pattern should be. For example, if pattern A can be generated by either of two rankings (10a), but pattern B by only one (10b), then all else equal we might expect for pattern A to be twice as frequent as pattern B: twice as many rankings generate Pattern A as do Pattern B.

(10) a. Pattern A:
CONST1 >> CONST2 and CONST3 >> CONST4

b. Pattern B:
CONST5 >> CONST6

Work by Anttila (1997) and Bane & Riggle (2008) has confirmed this expectation: in the domains that have been investigated (i.e. the typology of quantity insensitive stress systems), the patterns that are most frequent are also the patterns that can be generated by the largest number of rankings.

To know how frequent we expect midpoint systems to be, we first have to determine how many constraint rankings generate them. Kager (2012) claims that midpoint systems are generated when two opposite-edge anti-lapse constraints sit at the top of the hierarchy; assuming a single stress per word, 322,560 (or 8.89%) rankings of Kager's anti-lapse constraint set (p. 1479) fit this description.[3] But this precondition is not specific enough: in order to generate a midpoint system, several other ranking conditions must hold. For example, in quantity insensitive midpoint systems, stress must be aligned to the outer edge of the window for the 'overlapping domains' effect to be visible. This is illustrated in (13) and (14), both of which have *EXTLAPSEL and *EXTLAPSER ranked at the top of the hierarchy (as in (7)). In (13), stress is pulled towards the outer edge of the window by ALIGNL (11); in (14), it is pulled towards the inner edge of the window by ALIGNR (12).

(11) ALIGNL: one * for each syllable separating stress from the left edge of the word.

(12) ALIGNR: one * for each syllable separating stress from the right edge of the word.

(13) *EXTLAPSEL >> *EXTLAPSER
>> ALIGNL

a. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\}^{\mathrm{R}}$

b. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\mathrm{L}}\}^{\mathrm{R}}$

c. ${}_{\mathrm{L}}\{\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma\}^{\mathrm{R}}$

d. ${}_{\mathrm{L}}\{\sigma\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}}\}_{\mathrm{L}}\sigma\sigma\}^{\mathrm{R}}$

e. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\mathrm{L}}{}^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$

f. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\mathrm{L}}\sigma^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$

(14) *EXTLAPSEL >> *EXTLAPSER
>> ALIGNR

a. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\sigma\acute{\sigma}}\}_{\mathrm{L}}\}^{\mathrm{R}}$

b. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\sigma\sigma\acute{\sigma}}\}_{\mathrm{L}}\}^{\mathrm{R}}$

c. ${}_{\mathrm{L}}\{\sigma^{\mathrm{R}}\{\boldsymbol{\sigma\acute{\sigma}}\}_{\mathrm{L}}\sigma\}^{\mathrm{R}}$

d. ${}_{\mathrm{L}}\{\sigma\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}}\}_{\mathrm{L}}\sigma\sigma\}^{\mathrm{R}}$

e. ${}_{\mathrm{L}}\{\boldsymbol{\sigma\sigma\acute{\sigma}}\}_{\mathrm{L}}{}^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$

f. ${}_{\mathrm{L}}\{\boldsymbol{\sigma\sigma\acute{\sigma}}\}_{\mathrm{L}}\sigma^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$

[2] This is a simplifying assumption: it isn't always the case that each possible permutation of constraints in a given constraint set will be equally probable. If the ranking of two or more constraints is fixed, permuting them will be impossible. I abstract away from this complication here, as the issue of fixed rankings does not arise in this paper.

[3] The expected frequencies in this section were calculated by hand. To keep the size of the typology manageable, the candidates considered were limited to forms with a single stress. This decision was made because midpoint systems are single-stress systems, so the typology of single-stress systems is a logical comparison class. Calculations are available online for the reader to verify at `http://web.mit.edu/juliets/www/expected-midpoint.xlsx`.

In (13), the overlapping domains effect is visible: when the size of the window shrinks in (13c-d), stress is pulled from its default initial position towards the middle of the word, only to return to its default initial position once the domains no longer overlap. But (14), where stress is right-aligned towards the inner edge of the stress window, is indistinguishable from a system with post-peninitial stress, as in Choguita Rarámuri (Caballero 2011) or Ho-Chunk (Winnebago; Miner 1989). In other words, while the ranking in (13) generates a midpoint system, the ranking in (14) does not.

Excluding cases like (14) and others, only 166,480 (or 4.58%) rankings of Kager's constraint set give rise to midpoint systems. Translating this into expected frequency of attestation, we expect that midpoint systems should make up 4.58% of all languages with one stress per word. However, no midpoint systems are attested in either Kager's (2012) survey of accentual window systems or Gordon's (2002) survey of quantity-insensitive stress systems (a summary of the latter is in (15)).[4]

(15) Summary of Gordon's (2002) single-stress survey (Gordon 2002:5)

	# of lgs.	**%**	**Example**
Initial	61	30.8%	Tinrin (Austronesian, Osumi 1995)
Penultimate	55	27.8%	Mohawk (Iroquoian, Michelson 1988)
Final	63	31.8%	Mazatec (Oto-Manguean, Jamieson 1977)
Antepenultimate	7	3.5%	Wappo (Yuki, Radin 1929)
Peninitial	12	6.1%	Basque (isolate, Hualde 1991)
Total	**198**	**100%**	

If Gordon's (2002) survey is representative, and 75.75% of all languages have one stress per word, then 3.47% of all languages should be midpoint systems. As of August 2014, 510 languages were included in StressTyp (Goedemans & van der Hulst 2009). 3.47% of these, or 18, should be midpoint systems. 18 is not a huge number, but the difference between 18/510 (expected) and 0/510 (attested) is significant (binomial test, p $<$ .001). Thus appealing to the expected frequency of midpoint systems is not sufficient to explain their absence: given that these systems are expected to be reasonably frequent, we must continue to look for reasons why they are unattested.

2.2 One solution: Kager (2012) and weakly layered feet

Kager (2012) proposes to exclude midpoint systems from the predicted typology by removing contextual anti-lapse constraints from CON, and by introducing weakly layered feet.[5] Weakly layered feet are composed of two constituents: a maximally binary head (so ✓([$\sigma\sigma$]) or ✓([σ]), but not *([$\sigma\sigma\sigma$])), and an optional monosyllabic adjunct (✓(σ[$\sigma\sigma$]) or ✓([σ]σ), for example). In Kager's theory, feet are maximally ternary, and this is assumed to be hard-wired into GEN: feet with more than one adjunct (e.g. *(σ[$\sigma\sigma$]σ)), or more than one head (e.g. *σ[$\sigma\sigma$][$\sigma\sigma$])), are not admitted as possible candidates. Thus the foot inventory includes the following (see Kager 2012: 1482):

[4] It should be noted, however, that Gordon's (2002) proposal undergenerates to some degree, and that the survey is not completely comprehensive. For example, the proposal precludes the possibility of languages with post-peninitial stress (as in Ho-Chunk, e.g Miner 1989), and the typology does not include any examples of systems where stress placement depends on word length (e.g. Içuã Tupi, Abrahamson 1968; on these systems see section 5.2).

[5] Earlier works on weak layering cited by Kager are Hewitt 1992, Itô & Mester 1992, Rice 1992, Kager 1994, Rifkin 2003, Blevins & Harrison 1999, Zoll 2004, Caballero 2011. Kager also notes that these OT models continue a tradition of exploring weak layering in the prosodic hierarchy; see e.g. Prince 1980, Selkirk 1980, Dresher & Lahiri 1991. See also Martínez-Paricio (2013), Martínez-Paricio & Kager (2014), and Kager & Martínez-Paricio (2014).

(16) Foot inventory (Kager 2012: 1482; ( ) = foot, [ ] = foot head)

	head + adjunct	adjunct + head	no adjunct
binary head, trochee	$([\acute{\sigma}\sigma]\sigma)$	$(\sigma[\acute{\sigma}\sigma])$	$([\acute{\sigma}\sigma])$
binary head, iamb	$([\sigma\acute{\sigma}]\sigma)$	$(\sigma[\sigma\acute{\sigma}])$	$([\sigma\acute{\sigma}])$
unary head	$([\acute{\sigma}]\sigma)$	$(\sigma[\acute{\sigma}])$	$([\acute{\sigma}])$

The size and composition of the foot is determined by a number of constraints that regulate foot form. In addition to standard foot-based constraints, i.e. IAMB and TROCHEE (see Kager 2012: 1482 for the full list and definitions), two new constraints, ALIGNHDL (= *"heads are left-aligned with feet"*) and ALIGNHEADR (= *"heads are right-aligned with feet"*), regulate the linear ordering of the foot's head and its adjunct. When ALIGNHEADL >> ALIGNHEADR, the adjunct appears on the right, as in $([\acute{\sigma}\sigma]\sigma)$; when ALIGNHEADR >> ALIGNHEADL, the adjunct appears on the left, as in $(\sigma[\acute{\sigma}\sigma])$. Binary feet arise when the foot head must be both right- and left-aligned with the foot boundary: in other words, when an adjunct is not allowed to intervene on either side.

The crucial property of the weakly layered model that allows it to avoid the midpoint pathology is that the constraints governing foot form are independent of the constraints that specify the foot's location within the word (see Kager 2012:1484). Put more precisely, foot form constraints that determine the *size* and *shape* of the constituent do not interact with alignment constraints that determine its *location*. This independence of foot form and alignment constraints makes it impossible to trap stress in the middle of some words, but not others: there is no way to derive the 'overlapping domains' effect that midpoint systems exhibit. Kager shows that this property of the weakly layered model does indeed prevent it from predicting midpoint systems: a factorial typology explored with the weakly layered constraint set excludes them entirely (see Kager 2012: 1485ff).

Our focus here is not on the details of the weakly layered model; interested readers are referred to Kager (2012) for more information. What is important to take away from this short discussion is only the nature of the proposal, and its implications. Kager identifies the midpoint pathology as an unattested prediction of contextual anti-lapse constraints. By removing these constraints from CON, and modifying the structure of GEN, he constructs a theory whose predicted typology closely matches the attested typology. And these modifications to CON, if they are the correct response to the problem posed by the midpoint pathology, have important theoretical consequences. As mentioned earlier, foot-free theories of stress (e.g. Gordon 2002) *depend* on contextual anti-lapse constraints to model the typology of stress windows. If the midpoint pathology is indicative of a fundamental problem with contextual anti-lapse constraints, it is indicative of a fundamental problem with foot-free theories of stress. And if there is no alternative to Kager's (2012) explanation for the absence of midpoint systems – that anti-lapse constraints are not part of CON – then the midpoint pathology is a strong argument for the necessity of weakly layered feet in metrical theory.

3 An alternative: midpoint systems are hard to learn

The remainder of this paper explores an alternative hypothesis for the absence of midpoint systems: namely, that they *are* part of the learner's hypothesis space, but they are unattested because they are difficult to learn. I will show that there are two distinct learnability problems that midpoint systems pose to a machine learner (hereafter just 'the learner'). The first problem arises because the forms necessary to learn certain kinds of midpoint systems are only rarely presented to the learner; this is the *long-word problem* (section 4). The second problem arises because the learner, when attempting to acquire a midpoint system, receives inconsistent information about the placement of stress relative to the word edge; this is the *credit problem* (section 5). The current section provides some necessary background information on the choice of learner to be used to explore this alternative hypothesis, as well as the specifics of how the learner functions.

3.1 Selecting a learner

When trying to determine whether some system *x* would be difficult for a child to acquire, the most straightforward way to test such a hypothesis would be to observe how first-language acquisition of *x* proceeds. As midpoint systems are unattested, however, this option is unavailable. As proxy for a human learner, in this paper we will focus on the performance of a machine learner as it attempts to learn midpoint systems. The learner used in this paper is Magri's (2012) convergent implementation of the Gradual Learning Algorithm (the GLA; see also Boersma 1997, Boersma & Hayes 2001).

The main motivation for selecting the GLA is that it is frequently cited as a plausible model of human phonological acquisition. Studies taking into account natural language data have shown that the GLA is capable of realistically modeling generalizations regarding order of acquisition and learning curves: for example, Boersma & Levelt (2000) show that a GLA learner accurately predicts the order of acquisition of syllable types in Dutch. The GLA is also able to predict that children's repair strategies in response to marked structures can change over time (McLeod et al. 2001), or differ from child to child (Pater & Barlow 2003); see Magri 2012:23 for discussion. In addition, recent work has suggested that, with regards to certain kinds of phonotactic learning, the GLA converges on more restrictive grammars than competing alternatives (Magri 2014).

Although the GLA is a plausible model of acquisition, its apparent failure in some cases to make the correct empirical predictions has led some researchers to develop and endorse other learning models (e.g. Pater 2009, Tessier 2009). As can be expected, these different models make different predictions regarding the kinds of systems that are easiest, or most difficult, for a learner to acquire. This, in turn, means that the results of the present investigation are to some extent dependent on the choice of learner. Below, I outline those properties of the GLA that are necessary to derive the results discussed in sections 4 and 5. First, the learner assumes *strict domination* (17); second, it assumes that learning is *error-driven* (18); and third, it assumes that learning is *gradual* (19).

(17) **Constraints are ranked**
All constraints stand in relations of strict domination: two lower-ranked constraints cannot gang up to overcome a higher-ranked constraint. This can be contrasted with ranking algorithms in which constraints are weighted, and ganging is possible (e.g. Goldwater & Johnson 2003, Jäger 2007, Jesney & Tessier 2011, Boersma & Pater 2016).

(18) **Learning is error-driven**
The learner only adjusts its grammar when it guesses the incorrect output form for the current piece of data it is considering. This can be contrasted with learners that adjust their grammar in response to all forms, even those on which it guesses correctly (Jarosz 2013).

(19) **Learning is gradual**
The learner's grammar is adjusted in response to individual pieces of data. The learner cannot access data it has seen previously, nor can it determine whether the adjustment precipitated by an individual form is consistent with the forms it has previously seen. This can be compared to a batch or ERC learner, which can make decisions about how to adjust its

grammar based on generalizations extrapolated from the entirety of data presented to it (e.g. Hayes 2004, Prince & Tesar 2004, Tessier 2009, Brasoveanu & Prince 2011).

Later on, where it becomes more relevant, I will flag those portions of the modeling results that would likely look much different given a learning model that differs according to one or more of the above assumptions. What I aim to show in the remainder of this paper, then, is that midpoint systems as a class are difficult to learn for a GLA learner, and more broadly the class of machine learners with the properties in (17-19). Whether or not human learners also exhibit these properties is an open question. If it can be shown that they do not, then the viability of a learnability-based explanation for the absence of midpoint systems will have to be reconsidered.

3.2 How the learner works

The GLA learner used in the simulations is provided with three kinds of information: (i) a constraint set, (ii) a set of input and candidate output forms, and (iii) advance knowledge of which forms are consistent with the system it is trying to learn. The learner's task is to discover a constraint ranking that is guaranteed to generate all of the forms present in its input. The constraint set that will be used in the midpoint simulations, adapted from Kager 2012:1479, is provided in (20). It includes *general* anti-lapse constraints (those that penalize sequences of stressless syllables), *contextual* anti-lapse constraints (those that penalize sequences of stressless syllables in certain locations), *alignment* constraints (those that prefer for stress to be at some edge), and NONFINALITY, a markedness constraint penalizing words with final stress.

(20) Adaptation of Kager's (2012) anti-lapse constraint set (based on Gordon 2002)

a. *General anti-lapse constraints:*
   - i. *LAPSE: assign one * for each sequence of two stressless syllables.
   - ii. *EXTLAPSE: assign one * for each sequence of three stressless syllables.

b. *Contextual anti-lapse constraints:*
   - i. *LAPSEL: assign one * if neither of the initial two syllables is stressed.
   - ii. *LAPSER: assign one * if neither of the final two syllables is stressed.
   - iii. *EXTLAPSEL: assign one * if none of the initial three syllables is stressed.
   - iv. *EXTLAPSER: assign one * if none of the final three syllables is stressed.

c. *Alignment constraints:*
   - i. ALIGNL: assign one * for each syllable separating stress from the left edge.
   - ii. ALIGNR: assign one * for each syllable separating stress from the right edge.

d. NONFINALITY: assign one * if the final syllable is stressed.

I assume that the learner is exposed to forms of one through seven syllables. The candidate set I assume makes a couple of expository simplifications, none of which are crucial here. First, the learner is only exposed to words containing all light syllables: our focus will be on systems where quantity is not at issue. Second, I assume that each word has one and only one stress. The set of inputs and outputs provided to the learner, then, is fairly small: see (21) for a full list.[6]

[6] While enriching the candidate set to include inputs with heavy syllables and multiple stresses generally drives up the total number of trials required for convergence, it does not appear to affect how quickly a given system converges *relative to other systems*, which is our focus in the remainder of the discussion.

(21) Inputs and outputs considered

Input	σ	$\sigma\sigma$	$\sigma\sigma\sigma$	$\sigma\sigma\sigma\sigma$	$\sigma\sigma\sigma\sigma\sigma$	$\sigma\sigma\sigma\sigma\sigma\sigma$	$\sigma\sigma\sigma\sigma\sigma\sigma\sigma$
Candidate(s)	$\boldsymbol{\acute{\sigma}}$	$\boldsymbol{\acute{\sigma}}\sigma$	$\boldsymbol{\acute{\sigma}}\sigma\sigma$	$\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$	$\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma$	$\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma\sigma$	$\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma\sigma\sigma$
		$\sigma\boldsymbol{\acute{\sigma}}$	$\sigma\boldsymbol{\acute{\sigma}}\sigma$	$\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$	$\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$	$\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma$	$\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma\sigma$
			$\sigma\sigma\boldsymbol{\acute{\sigma}}$	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma$
				$\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}$	$\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$	$\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$	$\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$
					$\sigma\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}$	$\sigma\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$	$\sigma\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$
						$\sigma\sigma\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}$	$\sigma\sigma\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$
							$\sigma\sigma\sigma\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}$

The learner is provided with information about which of the output forms in (21) is optimal given the system it is learning, as well as the frequency at which that particular form is attested in its input. Frequency information and its effects on learning will be further discussed in section 4.

To illustrate how learning proceeds, consider the following simplified demonstration, in which a learner is taught a system with penultimate stress. I will assume that the learner is equipped with the constraint set in (22), a simplified version of the constraint set in (20). Furthermore, I will assume here and throughout that all of the constraints in (20) are unranked with respect to one another at the beginning of learning (the *initial state*), as they are all markedness constraints (see Tessier 2009:13 for an explicit statement of this common assumption). Arbitrarily, as shown in (22), all of the constraints will begin with a ranking value of 100. Values can be directly translated into rankings: if constraint A has a value of 100 and constraint B has a value of 99, then A >> B; if the two have identical values, then there is no crucial ranking between them.

(22) Sample simulation: initial ranking values

Constraint	**Ranking value**
*LAPSEL	100
*LAPSER	100
ALIGNL	100
ALIGNR	100

Let us assume that the first input form is disyllabic. As the learner's initial state is one in which all constraints are unranked with respect to one another, the learner will randomly permute all of the constraints to form a fully stratified hierarchy (see Tesar & Smolensky 2000:47-50). Let us further assume that the resulting grammar causes the learner makes the wrong guess, $\sigma\boldsymbol{\acute{\sigma}}$. The learner, informed of its error, must update its grammar. Each update consists of two parts: promotion of the *winner-preferring* constraints (those penalizing the incorrect guess more than the correct one; here ALIGNL), and demotion of the *loser-preferring* constraints (those penalizing the correct guess more than the incorrect ones; here ALIGNR). Constraints that prefer neither the winner nor the loser (*LAPSE, *LAPSEL, and *LAPSER) remain at their current values. The learner's updated grammar is in (23); the update rule (i.e. the relative amounts of promotion and demotion) is Magri's (2012).

(23) Sample simulation: first update

Constraint	**Ranking value**
ALIGNL	100.5
*LAPSEL	100
*LAPSER	100
ALIGNR	99

This process – the presentation of a form, the learner's guess, and the update in response to an error (if there is one) is referred to as a *trial*, and the entire learning procedure, composed of a number of trials, is referred to as a *run*. On Trial 2 of this run, let's again assume that the learner encounters a disyllabic form. This time, $\boldsymbol{\acute{\sigma}}\sigma$ is the optimal choice: incorrect $\sigma\boldsymbol{\acute{\sigma}}$ is penalized by high-ranked ALIGNL. The learner thus correctly guesses that the output form is $\boldsymbol{\acute{\sigma}}\sigma$, and no update is necessary.

At Trial 3, let us assume that the learner encounters a four-syllable form, $\sigma\sigma\sigma\sigma$. As the learner's grammar is now one in which initial stress is preferred, it will guess that the form should have initial stress ($\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$) – but this is incorrect, as the language it is learning is one with penultimate stress (so $\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$). In response to this error, the learner will promote the winner-preferrers (*LAPSER and ALIGNR) and demote the loser-preferrers (ALIGNL and *LAPSEL), resulting in (24).

(24) Sample simulation: second update

Constraint	Ranking value
*LAPSER	100.66
ALIGNR	99.66
ALIGNL	99.5
*LAPSEL	99

At Trial 4, the learner encounters another four-syllable form. Notice, in (24), that the update in Trial 3 has caused the relative ranking of ALIGNR and ALIGNL to switch. The learner will therefore guess that four-syllable $\sigma\sigma\sigma\sigma$ should have final stress ($\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}$), when in fact it should have penultimate stress ($\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$). In response to this error, the learner promotes ALIGNL (the winner-preferrer) and demotes ALIGNR (the loser-preferrer), resulting in the grammar in (25).

(25) Sample simulation: third update

Constraint	Ranking value
*LAPSER	100.66
ALIGNL	100
*LAPSEL	99
ALIGNR	98.66

The learner has now converged at Trial 4: it will cease to make errors, as the constraint ranking it has reached is consistent with the data it receives (in which each word has penultimate stress).

Throughout, I will treat the number of trials required for the learner to converge on a ranking that generates some system *x* as a rough indication of the difficulty of acquiring that system. Although we do not yet know what the human equivalent is of a single machine learning trial – could a human learner infer, based on only four forms, that it is learning a system with penultimate stress? – it seems reasonable to believe that there is a positive correlation between the number of trials required for the learner to converge and the difficulty of the system that it is attempting to learn. For example, if a learner takes 4 trials to converge on Grammar A and 400 trials to converge on Grammar B, I will assume that Grammar B is more difficult than Grammar A for the learner to acquire.

In the following two sections, I show that a GLA learner equipped with the constraint set in (20) and the input-output set in (21) takes longer on average to converge on rankings that generate midpoint systems than it does on rankings that generate superficially similar, but attested, systems. I show that the learner's difficulty in acquiring midpoint systems stems from fundamental properties of gradual error-driven learning: rankings that generate midpoint systems are difficult for the learner to discover. The hypothesis that midpoint systems are unattested because they are difficult to learn also makes broader predictions about stress typology; I show that these predictions are borne out.

4 The long-word problem

In this section, I argue that many classes of midpoint system suffer from the *long-word problem*: certain crucial rankings needed to derive a midpoint system are only available in long (5+ syllable) words. Results from a cross-linguistic study on word length distribution show that long words are rare in most languages, and modeling results presented in section 4.1 show that the rarity of long words makes learning midpoint systems difficult. In section 4.2, I explore consequences of the long-word problem and demonstrate that the minority of attested stress systems suffering from the long-word problem also happen to be attested in languages with many long words.

4.1 Modeling results

Consider the trio of midpoint systems in (26-28). While all display the behavior characteristic of midpoint systems (the stressable window shrinks, then expands), they differ in one crucial respect: the minimum word length in which the relative ranking of the two anti-lapse constraints can be determined. In (26), it is possible to infer from all words of four or more syllables that *LapseL >> *LapseR; I refer to (26) as a *limited* midpoint system, as the conflicting anti-lapse constraints are both varieties of *Lapse. In (27), words of five or more syllables are required to infer that *LapseL >> *ExtLapseR; this is a *mixed* midpoint system, as one anti-lapse constraint is a variety of *Lapse, and the other is a variety of *ExtLapse. And finally, in (28), words of six or more syllables are necessary to determine that *ExtLapseL >> *ExtLapseR; systems like (28), where both anti-lapse constraints are varieties of *ExtLapse, are *extended* midpoint systems.

(26) Limited Midpoint
*LapseL >> *LapseR
>> AlignL
a. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\}^{\mathrm{R}}$
b. ${}_{\mathrm{L}}\{\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}}\}_{\mathrm{L}}\sigma\}^{\mathrm{R}}$
c. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}{}^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$
d. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$
e. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma\sigma^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$
f. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma\sigma\sigma^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$

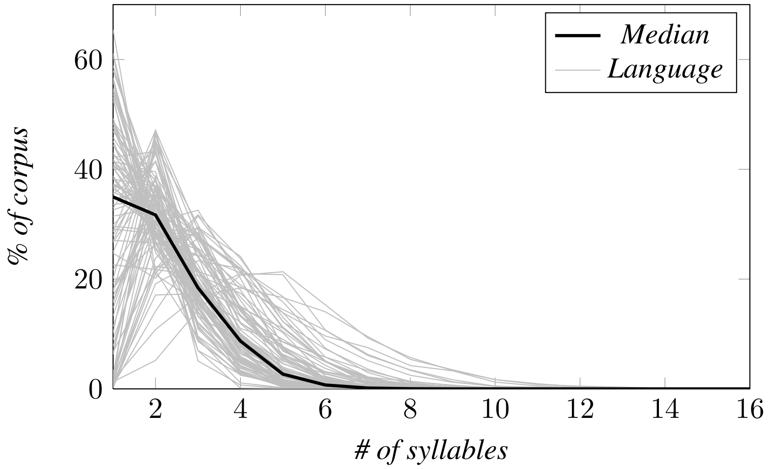
(27) Mixed Midpoint
*LapseL >> *ExtLapseR
>> AlignL
a. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\}^{\mathrm{R}}$
b. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma\}^{\mathrm{R}}$
c. ${}_{\mathrm{L}}\{\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}}\}_{\mathrm{L}}\sigma\sigma\}^{\mathrm{R}}$
d. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}{}^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$
e. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$
f. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma\sigma^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$

(28) Extended Midpoint
*ExtLapseL >> *ExtLapseR
>> AlignL
a. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\}^{\mathrm{R}}$
b. ${}_{\mathrm{L}}\{^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\mathrm{L}}\}^{\mathrm{R}}$
c. ${}_{\mathrm{L}}\{\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\mathrm{L}}\sigma\}^{\mathrm{R}}$
d. ${}_{\mathrm{L}}\{\sigma\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}}\}_{\mathrm{L}}\sigma\sigma\}^{\mathrm{R}}$
e. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\mathrm{L}}{}^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$
f. ${}_{\mathrm{L}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\mathrm{L}}\sigma^{\mathrm{R}}\{\sigma\sigma\sigma\}^{\mathrm{R}}$

In the case of the mixed and extended midpoint systems, the fact that certain crucial rankings are only visible in longer (5+ syllable) words has implications for acquisition. For example, in order for a learner to successfully acquire all rankings associated with (28), she would have to be exposed to words that are six syllables or longer. A survey of text corpora from 102 languages reveals that this situation is, on average, unrealistic: long words are infrequent (on the distribution of word lengths, see also Hatzigeorgiu et al. 2001, Sigurd et al. 2004, Piantadosi et al. 2011, Kalimeri et al. 2015). The results of the survey are presented in Figure 1: each thin gray line represents the frequency distribution of an individual language, while the thicker black line represents the median values.

More details about how the survey was conducted, as well as more information on the surveyed languages (including frequencies by language, genetic classification information, and sources of the data) are in the Appendix.

Figure 1: Results of the word counting study (see the Appendix for more details)



#	Median
1	35%
2	32%
3	18%
4	9%
5	3%
6	1%
7	0%

The important point to take away from Figure 1 is that, assuming that the median values represent approximately what the average learner would be exposed to, words of five or more syllables make up only 4% of the learner's input, and words of six or more syllables make up only 1%. What this means, then, is that for a learner attempting to learn a midpoint system like (27) or (28), evidence as to the relative ranking of the anti-lapse constraints comes from a small minority of forms present in the input. As there is reason to believe that long words are even less frequent in child-directed speech (see e.g. Vihman et al. 1994:656 for properties of child-directed speech in English, French, and Swedish, where 1-2 syllable words predominate), patterns where crucial rankings are only available in these longer words might therefore be difficult for a child to acquire.

The rest of this subsection focuses on the following question: if a learner samples long words at the rate they are attested cross-linguistically, does it have a difficult time learning midpoint systems? To address this question, we will focus on the learner's behavior as we steadily decrease the number of long words that it encounters. To model this decrease in long words, I selected five word length distributions from the word counting study (29). Here, Portuguese represents the "average" language, as its distribution is closest to the median. Inuktitut represents the upper bound, as it has more long words than any other language in the study; Haitian represents the lower bound, as it has very few. English and Ganda represent intermediate points along the continuum.

(29) Word length distributions used in modeling

Distribution	**1σ**	**2σ**	**3σ**	**4σ**	**5σ**	**6σ**	**7+σ**
Inuktitut	1.3%	5.2%	14.4%	20.7%	21.3%	15.3%	21.8%
Ganda	22.6%	21.7%	20.7%	17.4%	10.4%	5.0%	2.1%
Portuguese	32.6%	35.4%	18.2%	10.0%	3.0%	0.7%	0.1%
English	56.6%	28.0%	11.5%	3.0%	0.6%	0.3%	<0.1%
Creole	58.0%	36.1%	5.1%	0.7%	0.1%	0.0%	0.0%

To probe the effects of word length distribution on learning different systems, I taught each learner five different systems: the three midpoint systems in (26-28), a system with initial stress (Initial, (30)) and one with antepenultimate stress (AP, (31)). Notice that, for AP, words of four syllables or longer are required to establish that *EXTLAPSER >> ALIGNL, as only in words of this length is it clear that there is a right-edge window actively prohibiting ALIGNL from being fully satisfied. In this sense, AP is exactly like the limited midpoint system in (26) in that four-syllable words are required to establish all crucial rankings.

(30) Initial (ALIGNL >> all)

a. $\boldsymbol{\acute{\sigma}}\sigma$

b. $\boldsymbol{\acute{\sigma}}\sigma\sigma$

c. $\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$

d. $\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma$

e. $\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma\sigma$

f. $\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma\sigma\sigma\sigma$

(31) AP (*EXTLAPSER >> ALIGNL)

a. $^{R}\{\boldsymbol{\acute{\sigma}}\sigma\}^{R}$

b. $^{R}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{R}$

c. $\sigma^{R}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{R}$

d. $\sigma\sigma^{R}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{R}$

e. $\sigma\sigma\sigma^{R}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{R}$

f. $\sigma\sigma\sigma\sigma^{R}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{R}$

Each system was presented to each learner ten times, for a maximum of 2,000 trials each. The results are in (32). Across word length distributions, the two attested systems (Initial and AP) were learned very quickly. For Initial, word length distribution has little effect on the number of trials required for convergence. For AP, word length distribution has some effect: compare the Haitian learner's average of 287 trials to the Portuguese learner's 27. This is not surprising, as four-syllable words (necessary to infer all crucial rankings for AP) are rarely presented to the Haitian learner. The limited midpoint system (LM, in (26)) is also learned relatively quickly by all learners; the fact that the learner takes slightly longer on average to converge on LM than it does on AP is reflective of the fact that LM poses an additional problem to the learner (discussed in section 5).

For the two remaining systems – the mixed midpoint system in (27) (MM) and the extended midpoint system in (28) (EM) – the rate at which the learner is exposed to long words has a marked effect on the number of trials required for convergence. Although MM and EM are learned relatively quickly by the Inuktitut and the Ganda learners, the Portuguese and English learners take longer to converge on EM and MM than they do on the other three systems. The difficulty that these systems pose is only made clearer by the performance of the Haitian learner, which failed to converge on the correct ranking within 2,000 trials for three of the MM runs and all ten of the EM runs.

(32) Modeling results

Distribution	**Initial**	**AP**	**LM**	**MM**	**EM**
Inuktitut	2	10	25	29	17
Ganda	3	14	23	35	48
Portuguese	4	27	39	98	199
English	7	58	109	289	689
Haitian	18	287	317	1,593+	2,000+

The results in (32) support the hypothesis that the long-word problem plays a significant role in the absence of some types of midpoint system. For MM and EM, the number of long words presented to the learner is inversely correlated with the number of trials necessary to converge on a ranking that generates a midpoint system. But there is still an additional question: given that the data presented to the learner are *consistent* with a ranking that generates a midpoint system, why is this not the preferred analysis? In other words: in the absence of overt evidence that the learner is attempting to acquire a midpoint system, why is it systematically biased against this hypothesis?

To explore this question, we will focus on the Haitian learner's failed attempts to learn EM. When the Haitian learner attempts to learn EM, it is never exposed to (33d-e), as words of six or more syllables are entirely absent from its input. Without these forms, (33) is identical to what the Haitian learner sees when learning AP (34). When long words are absent, the data are ambiguous.

(33) 2-5 syllable forms for EM

a. $_{\text{L}}\{^{\text{R}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\text{L}}\}^{\text{R}}$

b. $_{\text{L}}\{\sigma^{\text{R}}\{\boldsymbol{\acute{\sigma}\sigma}\}_{\text{L}}\sigma\}^{\text{R}}$

c. $_{\text{L}}\{\sigma\sigma^{\text{R}}\{\boldsymbol{\acute{\sigma}}\}_{\text{L}}\sigma\sigma\}^{\text{R}}$

d. ~~$_{\text{L}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\text{L}}{}^{\text{R}}\{\sigma\sigma\sigma\}^{\text{R}}$~~

e. ~~$_{\text{L}}\{\boldsymbol{\acute{\sigma}\sigma\sigma}\}_{\text{L}}\sigma^{\text{R}}\{\sigma\sigma\sigma\}^{\text{R}}$~~

(34) 2-5 syllable forms for AP

a. $^{\text{R}}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{\text{R}}$

b. $\sigma^{\text{R}}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{\text{R}}$

c. $\sigma\sigma^{\text{R}}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{\text{R}}$

d. ~~$\sigma\sigma\sigma^{\text{R}}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{\text{R}}$~~

e. ~~$\sigma\sigma\sigma\sigma^{\text{R}}\{\boldsymbol{\acute{\sigma}}\sigma\sigma\}^{\text{R}}$~~

Given the data from EM in (33), both AP and EM are possible hypotheses[7] – but AP is the *preferred* one. Every time the Haitian learner is exposed to (33) or (34), it converges on a grammar that generates AP. To see where this preference comes from, consider the schematic learning trajectory presented in Figure 3 (based on Run 1 of the Haitian learner's EM trials). Forms that the learner encounters (1-5 syllables) are in black; forms that the learner does not (6+ syllables) are in gray.

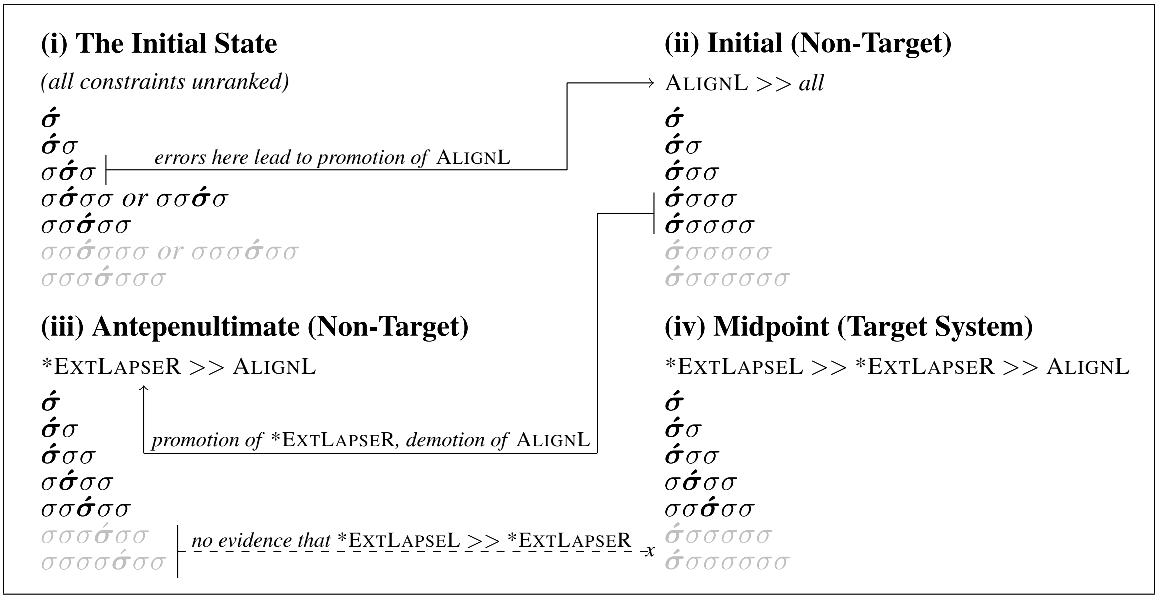
At the beginning of the learning procedure, the learner's initial state (State i) is one in which all markedness constraints are unranked with respect to one another. When the learner is presented with monosyllabic and disyllabic forms, all of the constraints in its grammar are randomly permuted to form a fully stratified ranking. As for the stress of monosyllabic forms, the learner will always make the correct guess because there is no other option (I assume that *all* forms must bear a stress). For the disyllabic forms, more of the possible fully stratified rankings prefer $\boldsymbol{\acute{\sigma}}\sigma$ (with initial stress; target) to $\sigma\boldsymbol{\acute{\sigma}}$ (with final stress; non-target), so the learner often makes the correct guess.

When presented with a trisyllabic form, the guess compatible with the most fully stratified rankings is second-syllable stress ($\sigma\boldsymbol{\acute{\sigma}}\sigma$), but this is the wrong guess: in the target system (State iv), trisyllabic forms have initial stress. In response to its error, the learner promotes ALIGNL (and demotes several other constraints[8]), resulting in a grammar in which all forms have initial stress (State ii). When the learner encounters a four– or five-syllable form, it makes another error: the learner guesses that the form should have initial stress (e.g. $\boldsymbol{\acute{\sigma}}\sigma\sigma\sigma$), when in fact stress should be antepenultimate (e.g. $\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$). In response to this error, the learner promotes *EXTLAPSER (preferring the correct form with antepenultimate stress) and demotes ALIGNL (preferring the incorrect forms with initial stress). As learning is gradual, it can take the learner more than one error of this type to eventually converge on a grammar that generates AP (State iii).

[7] There are other hypotheses given the data in (33a-d): for example, the system could be one in which post-peninitial stress is preferred, but stressing either of the final two syllables is impossible. For simplicity, I do not discuss these.

[8] The full story: for this trial, the learner promotes the winner-preferrer ALIGNL, and demotes the loser-preferrers *LAPSE, *LAPSER, and ALIGNR. For simplicity, here and in what follows I focus on only a subset of the updates.

Figure 2: Learning trajectory for the Haitian Midpoint learner


(i) The Initial State
(all constraints unranked)
σ́
σ́σ
σσ́σ
σσ́σσ or σσσ́σ
σσσ́σσ
σσσ́σσσ or σσσσ́σσ
σσσσ́σσσ
errors here lead to promotion of ALIGNL
(ii) Initial (Non-Target)
ALIGNL >> all
σ́
σ́σ
σ́σσ
σ́σσσ
σ́σσσσ
σ́σσσσσ
σ́σσσσσσ
(iii) Antepenultimate (Non-Target)
*EXTLAPSER >> ALIGNL
σ́
σ́σ
σ́σσ
σσ́σσ
σσσ́σσ
σσσσ́σσ
σσσσσ́σσ
promotion of *EXTLAPSER, demotion of ALIGNL
no evidence that *EXTLAPSEL >> *EXTLAPSER
x
(iv) Midpoint (Target System)
*EXTLAPSEL >> *EXTLAPSER >> ALIGNL
σ́
σ́σ
σ́σσ
σσ́σσ
σσσ́σσ
σ́σσσσσ
σ́σσσσσσ

Although the learner has not yet reached the target grammar (State iv), learning ends when it reaches a grammar that generates antepenultimate stress (State iii). This is because some of the crucial rankings necessary to reach the target grammar (State iv) are not motivated by data that the learner encounters. To reach the target grammar, it is necessary for the learner to infer that *EXTLAPSEL >> *EXTLAPSER, but the learner never sees any evidence that *EXTLAPSEL needs to be promoted. This is because the errors that would cause the learner to promote *EXTLAPSEL are incompatible with other aspects of the system that it is learning. For example, if the learner were to encounter four-syllable *σσσσ* and incorrectly guess that it should be stressed as *σσσ**σ́***, *EXTLAPSEL would be promoted, as the target form *σ**σ́**σσ* shows the learner that stress must fall within an initial trisyllabic window. But the learner never makes this error or others like it, because it learns very early on that ALIGNL is high-ranked: it has no reason to ever guess that a word should have final stress. In short, AP is the preferred hypothesis given the data in (33a-d) because the rankings needed to derive it are supported by data that the learner encounters.

As I am assuming that all markedness constraints are equally ranked in the initial state, another way of viewing the preference for AP is as a preference for the *simplest possible hypothesis*, meaning here the hypothesis that involves the fewest deviations from the initial state. Comparing the grammars necessary to generate AP (35) and EM (36) reveals a fundamental difference between them: the grammar for EM involves more strata. If all markedness constraints are unranked with respect to one another in the initial state, then (36) represents a more significant departure. From the learner's perspective, there is no reason to assume that the additional ranking differentiating (35) from (36) is necessary.[9]

[9] Stated this way, the preference for AP stress over Midpoint is one that relies heavily on the GLA's assumption that constraints stand in relations of strict domination: the learner is biased to acquire systems in which there are very few constraint strata. There is no reason to believe that this same metric of simplicity should apply, however, if we assume that constraints are weighted, and that several low-weighted constraints can gang up on a higher-weighted one. In fact, when attempting to teach the ambiguous data presented to the Haitian learner to a Noisy Harmonic Grammar (NHG)

(35) Grammar for AP
*ExtLapseR
↓
AlignL

(36) Grammar for EM
*ExtLapseL
↓
*ExtLapseR
↓
AlignL

This result continues to hold even if we adopt other proposals in the literature arguing for more refined initial rankings. Within a constraint class, i.e. markedness or faithfulness, proposals about biases in the initial state have typically appealed to the difference between specific and general constraints (though cf. Tesar & Smolensky 2000:68-70 for other proposals regarding the initial ranking of foot form and quantity-sensitive constraints). For example, Hayes (2004) proposes that if both a specific and a general faithfulness constraint can be used to rule out a single losing candidate, the specific constraint should be selected. Favoring specific faithfulness constraints allows the learner to maintain a more restrictive grammar, which helps avoid overgeneration (Hayes 2004:22). With this logic, if learners should favor restrictive grammars, we might expect that general markedness constraints should be favored over specific ones, as general markedness constraints penalize a wider variety of forms (see Albright & Do 2013). But preferring general over specific markedness constraints does not affect the results discussed above, as neither *ExtLapseL nor *ExtLapseR is more specific than the other. Even if we assume the opposite – that specific markedness constraints should be favored over general ones (Do 2013:123) – the result still holds. In fact, it is quite difficult to envision a reason why a learner would be biased to prefer one of *ExtLapseL and *ExtLapseR over the other, as the two constraints are completely symmetrical.[10]

In sum, this subsection has suggested that the long-word problem can help us understand why certain types of midpoint system are unattested. As demonstrated above, a GLA learner trying to acquire an extended midpoint system has to be exposed to words of six syllables or longer in order to reach the target grammar. In many cases, this is an unrealistic situation: in the word count study described in the Appendix, words of six or more syllables make up a negligible portion of the corpus (0.4% or lower) in 39 of the 102 surveyed languages. The situation is similar, though much less dire, for a learner trying to acquire a mixed midpoint system (as in (27)): five-syllable words make up on average 3% of the entire corpus, with 9 of the 102 surveyed languages having very few words of this length (0.4% or lower). As exposure to long words is in many cases necessary for a learner to reliably acquire a midpoint system, the cross-linguistic rarity of long words poses a general problem for the acquisition of these systems.

learner (these simulations done in OTSoft, Hayes et al. 2013), using the same constraint set and input-output pairs, the NHG learner was biased to acquire a *midpoint* system. This result, however, only arises because AlignL and AlignR are assessed gradiently. If we replace these constraints with categorical ones, however, then the weighted and ranked constraint learners behave identically: both are biased to prefer antepenultimate stress.

[10] It is true that, typologically speaking, more languages exhibit a right-edge window (where *LapseR or *ExtLapseR is ranked high) than a left-edge window (where *LapseL or *ExtLapseL is ranked high); see Kager (2012) for details. There might then be a plausible reason to prefer right-edge contextual lapse constraints over left-edge ones. But whatever the source of this preference, favoring right-edge over left-edge constraints will just make the midpoint system under discussion even more difficult to learn.

4.2 Consequences of the long-word problem

Above, I showed that appealing to the cross-linguistic rarity of long words can help us make progress in understanding why certain types of midpoint system are unattested. This claim, if correct, has broader typological consequences: stress systems in which some crucial rankings are only visible in long words should only arise in the small minority of languages in which long words are frequent. In this subsection, I explore these consequences by investigating several types of stress system in which long (6+ syllable) words appear to be necessary for all crucial rankings to be established. The results of this investigation suggest that stress systems can be divided into two classes: (i) those in which the stress in long words is predictable given the stress in shorter words (section 4.2.1), and (ii) those in which the stress in long words is *not* predictable given the stress in shorter words (section 4.2.2). As expected, the (ii)-type systems appear to only be attested in languages that have far more long words than average.

4.2.1 Behavior of long words is predictable: binary plus clash systems

To illustrate how the stress of long (6+ syllable) words can be predicted given the stress of shorter words, we will focus on the typology of binary plus clash systems (name due to Gordon 2002). In these systems, stress generally alternates in a binary fashion, but clashes arise in words of certain lengths. For example, in Passamaquoddy (LeSourd 1988, LeSourd 1993), odd-parity words license a clash (underlined) at their left edge (37).

(37) Stress in Passamaquoddy (LeSourd 1988:140-143)

a.	wá.sis	‘child’	$\boldsymbol{\acute{\sigma}}\sigma$
b.	wà.sí.sək	‘dirt, soil’	$\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\acute{\sigma}}}\sigma$
c.	wì.coh.ké.mal	‘he helps the other’	$\boldsymbol{\grave{\sigma}}\sigma\boldsymbol{\acute{\sigma}}\sigma$
d.	wì.còh.ke.ké.mo	‘he helps out’	$\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma\boldsymbol{\acute{\sigma}}\sigma$
e.	wì.coh.kè.ta.há.mal	‘he thinks of helping the other’	$\boldsymbol{\grave{\sigma}}\sigma\boldsymbol{\grave{\sigma}}\sigma\boldsymbol{\acute{\sigma}}\sigma$
f.	tèh.sàh.kwa.pà.sol.tí.ne	‘let’s walk around on top’	$\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma\boldsymbol{\grave{\sigma}}\sigma\boldsymbol{\acute{\sigma}}\sigma$

As another example, consider the pattern attested in Southern Paiute (Sapir 1930; also Harms 1966, Wheeler 1979), where even-parity words license a clash (underlined) between the penult and the antepenult. In (38), vowel devoicing is indicated by capitalization, following Sapir.

(38) Stress in Southern Paiute (Sapir 1930:28-40; see also Sapir p. 39 and van Urk 2013:11)

a.	ú.mA	‘with it’	$\boldsymbol{\acute{\sigma}}\sigma$
b.	tï.qá.qːA	‘several eat’	$\sigma\boldsymbol{\acute{\sigma}}\sigma$
c.	qa.ní.à.ŋA	‘his house’	$\sigma\underline{\boldsymbol{\acute{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma$
d.	pU.cá.ɣa.ì.pːì.ɣa	‘looked for’	$\sigma\boldsymbol{\acute{\sigma}}\sigma\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma$
e.	nam.pú.cːa.ɣà.I.pì.ɣa	‘looked for trail’	$\sigma\boldsymbol{\acute{\sigma}}\sigma\boldsymbol{\grave{\sigma}}\sigma\boldsymbol{\grave{\sigma}}\sigma$
f.	tï.v^{w}á.qːaŋ.wà.i.yù.càm.pA	‘though not killing game’	$\sigma\boldsymbol{\acute{\sigma}}\sigma\boldsymbol{\grave{\sigma}}\sigma\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma$

In both of these systems, if we focus only on words of five syllables or fewer, the preferred location of the clash is ambiguous. In Passamaquoddy, five-syllable <wì.còh.ke.ké.mo> ($\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma\boldsymbol{\acute{\sigma}}\sigma$) is consistent with two seven-syllable forms: the attested <tèh.sàh.kwa.pà.sol.tí.ne> ($\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma\boldsymbol{\grave{\sigma}}\sigma\boldsymbol{\acute{\sigma}}\sigma$), with a clash at the edge, and the unattested *<tèh.sah.kwà.pà.sol.tí.ne> ($^*\boldsymbol{\grave{\sigma}}\sigma\underline{\boldsymbol{\grave{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma\boldsymbol{\acute{\sigma}}\sigma$), with a word-internal clash. In Southern Paiute, the stress pattern of four-syllable [qa.ní.à.ŋA] ($\sigma\underline{\boldsymbol{\acute{\sigma}}\boldsymbol{\grave{\sigma}}}\sigma$) is consistent with two possible six-syllable forms: the attested [pU.cá.ɣa.ì.pːì.ɣa] (σ**σ́**σ**σ̀σ̀**σ), with a clash between two secondary stresses, or the unattested *[pU.cá.ɣà.i.pːì.ɣa] (*σ**σ́σ̀**σ**σ̀**σ), with a clash between a primary and a secondary. At face value, then, it appears that a learner would have to be exposed to long words in order to successfully acquire the systems in (37) and (38). Assuming that long words are relatively infrequent in Southern Paiute and Passamaquoddy (as is the case for most of the world's languages), this situation is perhaps unrealistic.

But as Kager (2001) and van Urk (2013) note, there are typological generalizations regarding the typology of binary plus clash systems that render the stress patterns in these long words entirely predictable. The first generalization is that stress clash is typically realized *away from the primary stress*: this is the case for both Passamaquoddy (37) and Southern Paiute (38).[11] The second generalization, also evident in both languages discussed above, is that stress clash is typically realized *at or close to the edge of a word*. More precisely, in quantity-insensitive systems, stress clashes that are separated from both word edges by another stress (e.g. **σ́**σ**σ̀σ̀**σ**σ̀**) are unattested. To encode these asymmetries, two constraints have been proposed: CLASH-AT-EDGE, penalizing all word-internal clashes (definition adapted from van Urk 2013:21; see also Kager 2001:11), and *CLASH-AT-PEAK, penalizing all primary-adjacent clashes (definition adapted from Kager 2001:10).

(39) CLASH-AT-EDGE: assign one * for each sequence of two stressed syllables that is both preceded and followed by another stressed syllable.

(40) *CLASH-AT-PEAK: assign one * if the syllable bearing primary stress is immediately adjacent to one or more syllables bearing secondary stress.

Simply admitting CLASH-AT-EDGE and *CLASH-AT-PEAK into CON is sufficient to render the seven-syllable forms of Passamaquoddy, and the six-syllable forms of Southern Paiute, predictable. In Passamaquoddy, five-syllable <wì.còh.ke.ké.mo> (**σ̀σ̀**σ**σ́**σ) and all shorter forms show us that the initial and penult must receive stress, and that *LAPSE is inviolable. Given this, it is predictable that, in seven-syllable <tèh.sàh.kwa.pà.sol.tí.ne> (**σ̀σ̀**σ**σ̀**σ**σ́**σ), the clash should be realized at the edge opposite the primary stress. As shown in (41), the alternatives are harmonically bounded.

(41) Passamaquoddy [tèh.sàh.kwa.pà.sol.tí.ne] (**σ̀σ̀**σ**σ̀**σ**σ́**σ) is predicted

		/σσσσσσσ/	CLASH-AT-EDGE	*CLASH-AT-PEAK
☞	a.	**σ̀σ̀**σ**σ̀**σ**σ́**σ		
	b.	**σ̀**σ**σ̀σ̀**σ**σ́**σ	*!	
	c.	**σ̀**σ**σ̀**σ**σ̀σ́**σ		*!

Similar considerations apply for Southern Paiute. A learner can infer from four-syllable [qa.ní.àn.gA] (σ**σ́σ̀**σ) and other shorter forms that stressing peripheral syllables is dispreferred (the initial is only stressed in disyllabic forms, e.g. (38a), to avoid stressing the final), and that *LAPSE is inviolable. Once we take CLASH-AT-EDGE (39) and *CLASH-AT-PEAK (40) into account, it is predictable that six-syllable [pU.cá.xa.ì.pì.xa] (σ**σ́**σ**σ̀σ̀**σ) and longer even-parity words should license their clashes at the side of the word not adjacent to the primary stress.

[11] The only exception to this generalization discussed by van Urk (2013) is South Conchucos Quechua (SCQ, Hintz 2006). In SCQ, we find clash occurring between a primary and a secondary ([tú.shù.ku.nà.qḁ], **σ́σ̀**σ**σ̀**σ). All this shows, though, is that the preference to avoid clashes between primaries and secondaries can be overruled by the preference to place clashes at the edge of the word. In other words, in SCQ, CLASH-AT-EDGE (39) >> *CLASH-AT-PEAK (40).

In short, binary plus clash systems are not systems in which very long (6+ syllable) words are necessary to establish any crucial rankings: the stress of these words is predictable given the stress of shorter (5- syllable) words. We might expect, then, that a learner would not face any difficulty in learning binary plus clash systems, as exposure to long words is not necessary to reach the target grammar. This expectation is borne out: a learner equipped with CLASH-AT-EDGE and *CLASH-AT-PEAK takes 68 trials on average to learn Passamaquoddy, and 69 to learn Southern Paiute.[12]

The discussion in this subsection has focused entirely on binary plus clash systems, but there are other classes of systems in which the stress patterns of long (e.g. 7 syllable) words are predictable given the stress of shorter words. Another class of examples comes from the typology of binary plus lapse systems (see Kager 2001, Gordon 2002), where several typological generalizations also render the stress of long words predictable. Kager (2001) shows that when lapses are licensed in quantity-insensitive systems, they are realized either (i) adjacent to the peak, or (ii) at the right edge of the word. Given five-syllable $\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma$ (as in Garawa; see Furby 1974), the learner can infer that the lapse must occur adjacent to the primary stress: the seven-syllable form must then be the attested $\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}\sigma$, and not the unattested $*\acute{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma$. Thus, in languages exhibiting binary plus lapse patterns, are also languages in which the stress of long (7+ syllable) words is predictable given the stress of shorter words.

4.2.2 Long-word phenomena in languages with many long words: ternary stress

I turn now to the class of systems in which the stress of long words is *not* predictable from the stress of shorter words; our case study will be languages exhibiting ternary stress patterns. In these systems, each stress is preferably separated by two stressless syllables from another stress (e.g. $\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma\sigma$). Some ternary systems pose a potential challenge for the long-word hypothesis because they are systems in which a learner must be exposed to long (6+ syllable) words in order to infer all crucial rankings. To illustrate, we will focus on two such systems: Chugach Alutiiq Yupik (42) (hereafter 'Chugach'; Leer 1985a,b, Hewitt 1992) and Cayuvava (43) (Key 1961, 1967).

The stress pattern of Chugach, as described by Kager (1993:412-413), is as follows: every syllable in position 3*n*-2 is stressed, and in words with 3*n*+1 syllables the final is stressed, too. In (42), I abstract away from the effects of quantity-sensitivity and consider only the patterns found in words with all light syllables (but see Kager 1993:412 for an overview of the rest of the data).

(42) Stress in Chugach (Leer 1985a for a, c, d; Leer 1985b for b, e)[13]

a.	pa.lá.yaq	'rectangular skiff'	$\sigma\acute{\boldsymbol{\sigma}}\sigma$
b.	a.kú.ta.mèk	'*akutaq* (a food), abl.sg.'	$\sigma\acute{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}$
c.	ta.qú.ma.lu.nì	'apparently getting done'	$\sigma\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}$
d.	a.kú.tar.tu.nìr.tuq	'he stopped eating akutaq'	$\sigma\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma$
e.	ma.ngár.su.qu.tà.qu.nì	'if he (refl.) is going to hunt porpoise'	$\sigma\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}$
f.	*inferred*		$\sigma\acute{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}\sigma\sigma\grave{\boldsymbol{\sigma}}$

Note that there are no eight-syllable words containing all light syllables provided, but the pattern in (42f) can be inferred through consideration of equivalent words with heavy syllables. Eight-

[12] The learner used in this subsection differs in non-crucial ways from the learner introduced in section 3.2. For example, it has to consider and evaluate candidates with more than one stress. For comparison, this learner takes 580 trials on average to learn EM (28), which is many more trials than required to learn either of the binary plus clash systems.

[13] Leer reports all stresses as equal. I assume here that the leftmost is primary, as it is more common in quantity-insensitive systems for main stress to remain at a consistent distance from some edge (on this see section 5.2).

syllable <tán.er.lir.sú.qu.ta.qú.ni> ‘if he is going to hunt’ (Leer 1985a:113; $\boldsymbol{\acute{\sigma}}\sigma\sigma\boldsymbol{\grave{\sigma}}\sigma\sigma\boldsymbol{\grave{\sigma}}\sigma$), with a heavy initial syllable, shows that Chugach stress is optimally ternary: each stressed syllable prefers to be followed by two stressless ones. From this, I infer that an eight-syllable word with all light syllables would be stressed as $\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma\boldsymbol{\grave{\sigma}}\sigma\sigma\boldsymbol{\grave{\sigma}}$; this is also what Kager’s (1993) description implies.

In Cayuvava, primary stress falls on the antepenultimate syllable, and secondary stresses occur on every third syllable counting back from the primary stress (43).[14]

(43) Stress in Cayuvava (all data from Key 1961:143-150)

a.	dá.pa	‘canoe’	$\boldsymbol{\acute{\sigma}}\sigma$
b.	tó.mo.ho	‘small water container’	$\boldsymbol{\acute{\sigma}}\sigma\sigma$
c.	a.rí.po.ro	‘he already turned around’	$\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$
d.	a.ri.pí.ri.to	‘already planted’	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$
e.	à.ri.hi.hí.be.e	‘I have already put the top on’	$\boldsymbol{\grave{\sigma}}\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$
f.	ma.rà.ha.ha.é.i.ki	‘their blankets’	$\sigma\boldsymbol{\grave{\sigma}}\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$
g.	i.ki.tà.pa.re.ré.pe.ha	‘the water is clean’	$\sigma\sigma\boldsymbol{\grave{\sigma}}\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma\sigma$

In both Chugach and Cayuvava, a learner would have to be exposed to long words to infer all crucial rankings. In Chugach, it is only clear in words of eight syllables or longer (42f) that ternary alternation is completely general in this language, not just licensed at the peak. In Cayuvava, it is only clear in words of six syllables or longer (43e-g) that there is more than one stress per word; in all shorter words, the system could just as well be one with a single antepenultimate stress.

The fact that the Chugach and Cayuvava patterns require long words to become clear is completely consistent with the fact that, in both languages, it is probable that long words are frequent. Although neither Chugach nor any other variety of Yupik has an accessible text collection, in Inuktitut and Inupiatun (both are related Eskimo-Aleut languages), long words are extremely frequent. In Inuktitut, words of eight or more syllables make up 12.50% of the word count study; in Inupiatun, they make up 12.11% (compare this to the average of 0.4%; see the Appendix for the full frequency distributions). While we do not know for sure that the word length distribution of Chugach is similar, another dialect of Yupik, Central Alaskan, certainly has long words. I assume that the form in (44), from Miyaoka 2012:132, is a single prosodic word; see Miyaoka pp. 70-71 for information on CAY prosody, where it appears that morphologically simple and complex forms are treated alike.

(44) angya-cuara-li-yu-kapigte-llru-nric-aaq-sugnarq-llru-yugnarz-aanga
boat-small-make-DES-ITS-PST-NEG-CTR-INF-PAST-INF-IND-3sg.1sg.
‘I’m in doubt that he actually didn’t really want to make me a small both (but he did)’

If we make the assumption that the word length distribution of Chugach roughly resembles that of Inuktitut, we can ask if a learner is able to easily acquire the pattern in (42) by sampling different word lengths at the rate that they are attested in the Inuktitut word count study. As we might expect, the pattern is easy to acquire: the learner takes around 40 trials to converge.

As Cayuvava also has no accessible text collection (nor does it have any known relatives), it is hard to know what the word length distribution in this language was. A short text in Key (1967), however, gives us an idea: as shown in (45), words of six syllables or longer make up 26.4% of this

[14] The associate editor raises a concern that evidence for Cayuvava’s stress pattern is insufficient, and that the reported facts cannot be confirmed. While we should indeed be cautious of accepting impressionistic descriptions of stress patterns (see de Lacy 2014), the Cayuvava data can still be used as illustration of a broader point, as long as they have not been publicly disputed. As de Lacy notes, much of the data that stress typologies are based on is open to question.

76-word text. This is a significant percentage, especially compared to the average of 3.3% (see the Appendix). As expected, a learner that samples words at the high rate that they are attested in (45) easily learns the Cayuvava pattern: it requires 45 trials on average to converge.

(45) Word length distribution of Key's (1967) text

Syllables	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
# of syllables	5	0	19	21	8	10	5	4
% of total	6.9%	0%	26.4%	29.2%	11.1%	13.9%	6.9%	5.6%

Although it is currently unclear exactly how many long words a learner would have to encounter for a pattern dependent on long words (like those found in Chugach and Cayuvava) to be easily learned, we have seen in this section that the long-word hypothesis makes correct predictions about stress typology. Crucially, systems where some crucial rankings are only discernible in long words appear to only be attested in languages that independently have many long words, far more than is cross-linguistically average. We can therefore safely point to the general cross-linguistic rarity of long words as a contributing factor to the unattested status of certain types of midpoint systems.

5 The credit problem

In section 4, we saw that it is difficult for the learner to acquire midpoint systems when it is deprived of the forms necessary to infer all crucial rankings. But this cannot be the only factor leading to the absence of midpoint systems from the attested typology, for two reasons. First, not all learners are deprived of long words: although languages like Inuktitut (where we would expect learners to be exposed to many long words) are rare, they do exist: in 6 of the 102 languages included in the word length study, words of six or more syllables make up 10% or more of the corpus. Second, not all kinds of midpoint system require a learner to be exposed to long words for expedient acquisition: in the limited midpoint systems discussed in section 4.1, learners only need to be exposed to words of four or more syllables, which are cross-linguistically frequent.

This section discusses an additional factor that makes midpoint systems difficult for a learner to acquire. In section 5.1, I show that the inconsistent placement of stress with respect to a word edge poses a *credit problem* for the learner: in short, updates in response to words of different lengths are mutually antagonistic, causing the learner to make many errors that cancel each other out before it finally converges, more or less by chance, on the correct grammar. Unlike the long-word problem, the credit problem is fully general: it applies to *all* kinds of midpoint systems. In section 5.2, I show that this dispreference for midpoint systems is part of a much larger dispreference for systems in which the placement of stress depends on word length, as is predicted by the results in section 5.1.

5.1 Modeling results

To examine the credit problem in more detail, we will focus on the Portuguese learner's attempts at learning midpoint systems. The results for the Portuguese learner are recapitulated from section 4.1 in (46). While the Portuguese learner always converges on a grammar that generate the midpoint systems, it takes longer on average to learn these systems than it does to learn either Initial or AP.

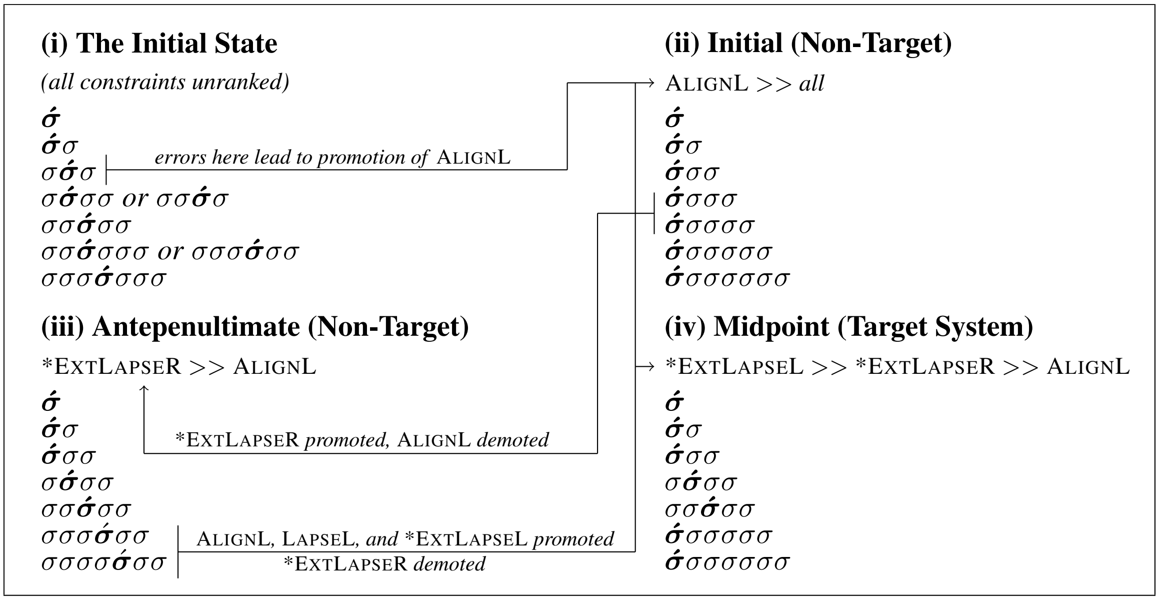
(46) Results from section 4.1, Portuguese learner

Distribution	**Initial**	**AP**	**LM**	**MM**	**EM**
Portuguese	4	27	39	98	199

To understand why it takes the learner a relatively long time to converge on the correct ranking for EM, consider the schematic learning trajectory presented in Figure 3 (based on a simplified Run 6 of the Portuguese learner's EM trials). As was the case with the simulations discussed in section 4.1, at the beginning of the learning procedure, the learner's initial state (State i) is one in which all markedness constraints are unranked with respect to one another; when the learner is presented with monosyllabic and disyllabic forms, it correctly guesses that they should have initial stress. When presented with a trisyllabic form, the learner makes an error: it guesses that the form should have second-syllable stress (*σ**σ́**σ*), when it should in fact have initial stress (***σ́**σσ*). In response to this error, the learner promotes ALIGNL, leading to a grammar that predicts all forms should have initial stress (State ii). But this is incorrect: when the learner encounters a four– or five-syllable word, it predicts that the form should have initial stress (***σ́**σσσ*), when it should have second-syllable stress (*σ**σ́**σσ*). This error causes the learner to promote *EXTLAPSER and demote ALIGNL, eventually reaching a grammar that generates forms with antepenultimate stress (State iii).

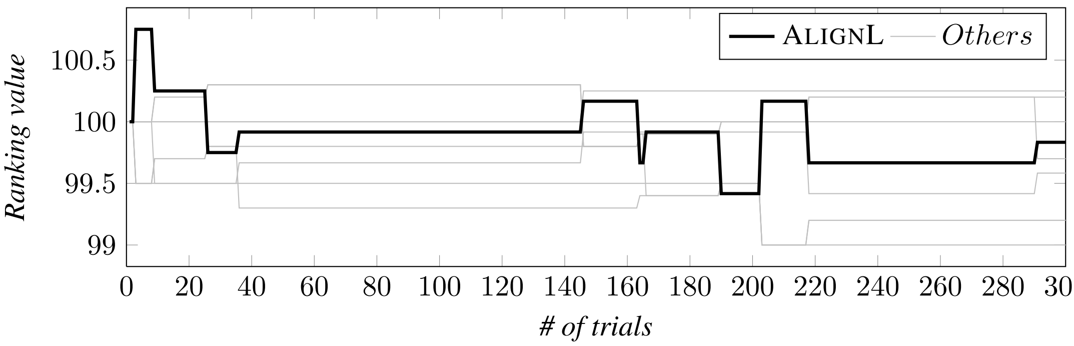
So far, the learning trajectory is identical to the Haitian learner's trajectory (see section 4.1). The difference is that the Portuguese learner is exposed to six– and seven-syllable words. While in State iii, a Portuguese learner seeing a six-(or seven-)syllable word will make an incorrect guess: its grammar tells it that stress should be antepenultimate (*σσσ**σ́**σσ*), when in fact it should be initial (***σ́**σσσσσ*). When comparing the winning form (***σ́**σσσσσ*) and the losing form (*σσσ**σ́**σσ*, notice that the winning form satisfies a number of constraints demanding that stress fall near the left edge: ALIGNL, *EXTLAPSEL, and *LAPSEL. The learner, not knowing which of these constraints is responsible for choosing the attested form, promotes *all of them* – in most cases, this update takes the learner back to State ii, causing it to again believe that all words should have initial stress. EM thus poses a *credit problem*: not knowing which of several markedness constraints is responsible for the attested form, the learner must promote all of them, causing it to revert to an earlier hypothesis.

Figure 3: Learning trajectory for the Portuguese Midpoint learner


(i) The Initial State
(all constraints unranked)
σ́
σ́σ
σσ́σ
σσ́σσ or σσσ́σ
σσσ́σσ
σσσ́σσσ or σσσσ́σσ
σσσσ́σσσ
errors here lead to promotion of ALIGNL
(ii) Initial (Non-Target)
ALIGNL >> all
σ́
σ́σ
σ́σσ
σ́σσσ
σ́σσσσ
σ́σσσσσ
σ́σσσσσσ
(iii) Antepenultimate (Non-Target)
*EXTLAPSER >> ALIGNL
σ́
σ́σ
σ́σσ
σσ́σσ
σσσ́σσ
σσσσ́σσ
σσσσσ́σσ
*EXTLAPSER promoted, ALIGNL demoted
ALIGNL, LAPSEL, and *EXTLAPSEL promoted
*EXTLAPSER demoted
(iv) Midpoint (Target System)
*EXTLAPSEL >> *EXTLAPSER >> ALIGNL
σ́
σ́σ
σ́σσ
σσ́σσ
σσσ́σσ
σ́σσσσσ
σ́σσσσσσ

Learning continues on in this manner for a while longer, with the learner bouncing back and forth between a grammar that generates initial stress (State ii), antepenultimate stress (State iii), and other incorrect hypotheses before converging, more or less by chance, on the target grammar that generates EM (State iv). What is immediately noticeable about the learning trajectory is that the value of ALIGNL is in constant flux: as the position of stress is inconsistent across words of different lengths, the learner receives inconsistence evidence about the relative importance of satisfying ALIGNL. The ranking trace in Figure 4, from the Portuguese learner's Run 6, illustrates: with each update, the ranking value of ALIGNL changes.

Figure 4: Ranking value of ALIGNL over time (EM)



The erratic behavior of ALIGNL demonstrates visually why midpoint systems are difficult for the learner. Recall that when the learner (at State ii) incorrectly guesses that a six-syllable form should have antepenultimate stress ($\sigma\sigma\sigma\acute{\boldsymbol{\sigma}}\sigma\sigma$), there are three constraints that could be responsible for the attested $\acute{\boldsymbol{\sigma}}\sigma\sigma\sigma\sigma\sigma$: *EXTLAPSEL, *LAPSEL, and ALIGNL. The learner is agnostic as to which constraint is responsible for the attested form, so it promotes all three. We could imagine a different response to this error: the learner could evaluate its current grammar, see that previous updates has caused it to rank *EXTLAPSER above ALIGNL, and refuse to promote ALIGNL. This reference to previously established rankings is a property of batch and ERC learners (see citations in section 3.1): the updates a learner performs in response to errors are informed by the crucial rankings that it has already learned. Such a learner would not encounter the credit problem described above, as it would require the learner to retain the ranking *EXTLAPSER >> ALIGNL, once it had been established.[15] This is not how the GLA works, though, and the GLA's lack of reference to previously established rankings is in fact a desirable property of the algorithm. If for example the first word a child hears is a speech error – an adult intends to produce the form $\acute{\boldsymbol{\sigma}}\sigma$, but instead produces $\sigma\acute{\boldsymbol{\sigma}}$ – we do not want the child to learn that ALIGNR >> ALIGNL is irreversible. In other words, a learner must be able to unlearn incorrect crucial rankings established in response to misproduced or misperceived forms. This ability of the learner to unlearn rankings established in response to previous errors is, in turn, exactly what makes EM so difficult to acquire.

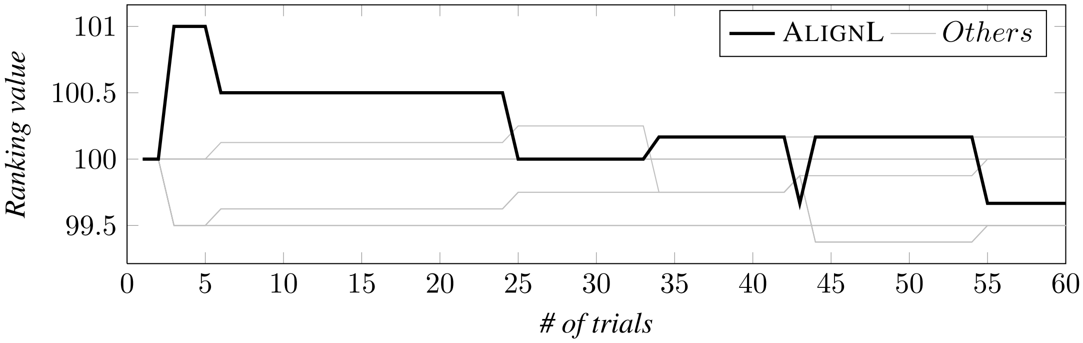
[15] The situation for ERC-based learners might not be so optimistic if ERCs are stored in cache (see Tessier 2007). In the case of EM, as long words are only infrequently encountered by the learner, they are not statistically strong patterns that the ranking algorithm has to deal with. Thanks to an anonymous reviewer for pointing this out.

The important point of this section is that the credit problem posed by the midpoint system in (28), as illustrated above, is completely general: *all* midpoint systems, whether they involve context-sensitive varieties of *ExtLapse or *Lapse, are difficult for the learner in this respect; *all* midpoint systems involve the inconsistent placement of stress across words of different lengths. For example: a learner acquiring a limited midpoint system, like the system discussed in section 4.1 (LM in (26), in an abbreviated format as (47)), do not need to be exposed to long words.

(47) LM (*LapseL >> *LapseR >> AlignL)
a. $_{\text{L}}\{^{\text{R}}\{\acute{\boldsymbol{\sigma}}\sigma\}_{\text{L}}\}^{\text{R}}$
b. $_{\text{L}}\{\sigma^{\text{R}}\{\acute{\boldsymbol{\sigma}}\}_{\text{L}}\sigma\}^{\text{R}}$
c. $_{\text{L}}\{\acute{\boldsymbol{\sigma}}\sigma\}_{\text{L}}{}^{\text{R}}\{\sigma\sigma\}^{\text{R}}$

While the relative ranking of *LapseL and *LapseR can be inferred from all forms of four syllables or longer, acquiring the ranking that generates (47) is still difficult for the learner, as the updates performed in response to the trisyllabic form with second-syllable stress (47b) are not consistent with those performed in response to the other forms. In other words, the learner runs into the same credit problem as it does when attempting to learn EM. The ranking trace in Figure 5 illustrates: the value of AlignL oscillates with each update that the learner performs.

Figure 5: Ranking value of AlignL over time (LM, Run 8)



In sum, the characteristic of midpoint systems responsible for the credit problem is the variable positioning of stress with respect to a word edge: stress is located at a word edge in words of some lengths, but not others. This inconsistency causes the learner to overgeneralize. When the learner sees a word with initial stress, it will often update its grammar to one that prefers initial stress, regardless of what has come before. The variable positioning of stress with respect to a word edge is a signature characteristic of all midpoint systems; thus, the credit problem can help us understand why these systems, as an entire class, are dispreferred.

5.2 Extensions: a dispreference for inconsistent stress placement

The main observation in section 5.1 is that midpoint systems are hard to learn because the data presented to the learner are not self-consistent: updates performed in response to words of different lengths, in effect, cancel one another out. In this subsection, I suggest that this observation can help us understand why several other classes of systems are underattested, relative to what we might expect. The general picture that emerges is that the absence of midpoint systems is just one symptom of a more general dispreference for systems in which the placement of main stress depends on word length (see also the Stress-Harmony constraint; Bailey 1995:204-205). The fact that it is *possible* for stress placement to depend on word length means that we cannot exclude these systems from the predicted typology; the fact that they are *rare* is consistent with a learnability-based explanation.

The first class of systems we will discuss here, which I refer to as *shrinking window* systems, are systems in which the size of the accentual window shrinks as the word lengthens. For example, in North Kyungsang Korean (NKK; Kenstowicz & Sohn 2001), pitch accent can occur on either the penultimate or final syllable in words of up to three syllables (so $\checkmark\sigma\boldsymbol{\acute{\sigma}}\sigma$ and $\checkmark\sigma\sigma\boldsymbol{\acute{\sigma}}$), but is fixed on the penult in words of four syllables or longer (so $\checkmark\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$, but $*\sigma\sigma\sigma\boldsymbol{\acute{\sigma}}$). Thus, in NKK, the right-edge disyllabic window found in shorter words (1-3 syllables) *shrinks* in longer words (4+ syllables). A similar pattern arises in Kimatuumbi (Odden 1996:179), where a process shifting final high tone one mora to the left applies only in words of four (and presumably more) moras (48).

(48) Leftward tone shift in Kimatuumbi nouns (Odden 1996:179)

a.	ngalibá	'female circumciser'	$\sigma\sigma\boldsymbol{\acute{\sigma}}$
	ma-ngalíba	'female circumcisers'	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$
b.	ngalawá	'canoe'	$\sigma\sigma\boldsymbol{\acute{\sigma}}$
	ka-ngaláwa	'little canoe'	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$
c.	mbutuká	'gazelle'	$\sigma\sigma\boldsymbol{\acute{\sigma}}$
	ma-putúka	'gazelles'	$\sigma\sigma\boldsymbol{\acute{\sigma}}\sigma$

The effects of this process mirror the more general fact that almost all tetrasyllabic noun stems in Kimatuumbi carry high tone on the penultimate mora (e.g. *changaláwe* 'gravel', p. 179), whereas shorter stems (1-3 syllables) can carry high tone on either the penult (e.g. *ndogólo*, p. 178) or the final (e.g. *ngalawá* 'canoe', p. 179). A related pattern can also found in Içuã Tupi (Abrahamson 1968:17-18), where accent occurs predominantly on the penult in words of up to four syllables (e.g. [í.tĩŋ] 'it is white', [pa.ti.u̯á.pɛ] 'bark pan'), but on the antepenult in words of five or more ([a.bi.dá.bi.dab$^{\mathrm{m}}$]). Note however that the Içuã Tupi pattern is subject to some variation: Abrahamson (pp. 17-18) notes that words of up to four syllables can have antepenultimate stress ([ta.tá.pũ.ĩ] 'ashes'), and six-syllable words can have third-syllable stress ([a.hɛ.á.bɛ.bui] 'his lung'.

Shrinking window systems can be modeled as resulting from an interaction between contextual anti-lapse constraints (like *LAPSER) and general anti-lapse constraints (like *LAPSE). In both Kimatuumbi and NKK, in words of all lengths, the position of accent is restricted by a right-edge disyllabic window. As before, we will model this window with the constraint *LAPSER; candidates where stress does not fall on one of the final two syllables receive a fatal violation (as in (49a)). Within the window, accent is free to fall on either the penult (49b) or the final (49c): I will assume that this freedom is due to a variable ranking between ALIGNR and NONFINALITY.

(49) Freedom of accent in shorter (1-3 syllable) words

		$\sigma\sigma\sigma$	*LAPSER	ALIGNR	NONFINALITY
	a.	$\boldsymbol{\acute{\sigma}}^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$	*!	**	
☞	b.	$\sigma^{\mathrm{R}}\{\boldsymbol{\acute{\sigma}}\sigma\}^{\mathrm{R}}$		*	
☞	c.	$\sigma^{\mathrm{R}}\{\sigma\boldsymbol{\acute{\sigma}}\}^{\mathrm{R}}$			*

In longer (4+ syllable) words, the desire to keep accent at the left edge of the stressable window (i.e. on the penult) can be attributed to *EXTLAPSE, a context-free anti-lapse constraint. As demonstrated by losing candidate (50d), stressing the final syllable in a four-syllable word results in a sequence of three stressless syllables, and a fatal violation of *EXTLAPSE. Candidate (50c), with penultimate stress, is selected as the winner, as it is the only other candidate that satisfies *LAPSER.

(50) Restriction of accent in longer (4+ syllable) words

	$\sigma\sigma\sigma\sigma$	*LAPSER	*EXTLAPSE	ALIGNR	NONFINALITY
	a. $\acute{\boldsymbol{\sigma}}\sigma^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$	*!		***	
	b. $\sigma\acute{\boldsymbol{\sigma}}^{\mathrm{R}}\{\sigma\sigma\}^{\mathrm{R}}$	*!		**	
	c. $\sigma\sigma^{\mathrm{R}}\{\acute{\boldsymbol{\sigma}}\sigma\}^{\mathrm{R}}$			*	
☞	d. $\sigma\sigma^{\mathrm{R}}\{\sigma\acute{\boldsymbol{\sigma}}\}^{\mathrm{R}}$		*!		*

For the Içuã Tupi pattern described above, the analysis is similar: retraction of stress to the antepenult in 5+ syllable words (as in [a.bi.dá.bi.dab$^{\mathrm{m}}$]) can be analyzed as a general tendency to avoid *EXTLAPSE violations, subject to the constraints of a right-edge trisyllabic window.

The three systems just discussed are the only examples of shrinking window systems that I am aware of. Although these systems are only marginally attested, their expected rate of attestation is quite high: at least 20% of the rankings of Kager's (2012) anti-lapse constraint set, assuming a single stress per word, generate shrinking window systems.[16] Assuming that 75.75% of all systems have only one stress per word, as is the case in Gordon's (2002) survey, the joint probability is that shrinking window systems should make up at least 15.2% of all languages, or at least 77 of the 510 languages in StressTyp (Goedemans & van der Hulst 2009). The fact that they are severely underattested relative to what we might expect is consistent with the discussion in section 5.1: shrinking window systems, like midpoint systems, are difficult to learn because they pose a credit problem. The fact that shrinking window systems are attested, and midpoint systems are not, just reflects the fact that shrinking window systems are expected to be more frequent in the first place.[17]

I turn now to a stark asymmetry in the typology of binary stress systems. In the majority of iterative binary stress systems (143/158 in StressTyp, see Staubs 2014b:2), the placement of primary stress is correlated with the direction of iterative stress (see also Gordon 2002:31). Systems where the primary stress is rightmost generally exhibit right-to-left iteration; systems where primary stress is leftmost generally exhibit left-to-right iteration. The result is a system in which the location of primary stress is consistent across words of all lengths, as in the Maranungku examples below (51).

(51) Iterative binary stress in Maranungku (Tryon 1970:10 for a-b; 9 for c-d[18])

a.	tíralk	'saliva'	$\acute{\boldsymbol{\sigma}}\sigma$
b.	mǽræpæ̀t	'beard'	$\acute{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}$
c.	jáŋarmàta	'the Pleiades'	$\acute{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}\sigma$
d.	ŋáltirìtirì	'tongue'	$\acute{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}$

[16] For a partial calculation, see: `http://web.mit.edu/juliets/www/expected-shrinking.xlsx`.

[17] It is worth noting, however, that given the data from the shrinking window systems discussed above, there are alternative analyses available. Kimatuumbi and NKK could just as well be systems with post-peninitial stress (like Ho-Chunk, Miner 1989); the possibility of penultimate stress in shorter stems could reflect a dispreference for final stress. Içua Tupí could also be a system with predominantly post-peninitial stress, in which NONFINALITY is inviolable. In the text, I have followed the authors' characterizations of the patterns, but the data necessary to determine which analysis is correct are not available. If these are all post-peninitial stress systems, then the claim of this part of the section could be strengthened: shrinking window systems, like midpoint systems, are unattested because they are difficult to learn.

[18] Stresses for the forms in (51c-d) are inferred from Tryon's (1970) description of the stress pattern on p. 10.

In a smaller number of languages (15/158 in StressTyp, Staubs 2014b:2), the placement of primary stress *opposes* the direction of iterative parsing: these are systems with right-to-left parsing where the primary stress is leftmost, or left-to-right parsing where the primary stress is rightmost. The result is a 'count' system (also van der Hulst 1996, McGarrity 2003), where the position of primary stress varies as a function of word parity. Data from Nyawagi (Dixon 1983) illustrate in (52).

(52) Iterative binary stress in Nyawagi (Dixon 1983:443)

a.	ɟíŋa	'man'	$\acute{\boldsymbol{\sigma}}\sigma$
b.	bulbíri	'quail'	$\sigma\acute{\boldsymbol{\sigma}}\sigma$
c.	bíyaɟàla	'water snake'	$\acute{\boldsymbol{\sigma}}\sigma\grave{\boldsymbol{\sigma}}\sigma$

Staubs (2014b) shows that a MAXENT learner faces a greater difficulty in acquiring systems like (52), where the position of main stress varies according to word parity, than it does in acquiring systems like (51), where the position of main stress is fixed with respect to some edge. This clear asymmetry in the typology of binary stress systems further illustrates the dispreference for systems in which the placement of main stress varies as a function of syllable count.[19]

As with the comparison between the shrinking window and the midpoint systems, perhaps the reason why count systems are attested, and midpoint systems are not, is because count systems are expected to be far more frequent in the first place. Assuming that the relative placement of primary stress is a parametric choice (leftmost or rightmost: Prince 1983:25, Gordon 2002:20), and that both directionalities of parsing are equally probable, we would expect that in fully half of all binary systems, the placement of main stress should oppose the direction of parsing. Assuming that Gordon's (2002) survey is representative, and that binary stress systems make up 23.28% of the total typology, we would expect 11.64% of all systems to exhibit the kind of binary alternation displayed in (52). We can now compare this expected rate of attestation to that of the midpoint systems, which are expected to make up only 3.47% of all systems (section 2.2). Given this large difference in expected frequency, it is unsurprising that we also find a difference in attested frequency.

6 Discussion and conclusions

At this point, we can enumerate a number of factors that potentially contribute to the absence of midpoint systems from the attested typology. First, as discussed in section 2.2, their expected rate of attestation is fairly low: we expect midpoint systems to comprise roughly 3.47% of the total typology. Second, we have seen that a learner attempting to acquire a midpoint system is faced with several difficulties. The credit problem incurred by all midpoint systems, together with the long-word problem incurred by a subset of them, causes midpoint systems as a class to be difficult for a learner to acquire. It is important to note that no one of these factors is independently responsible for the absence of midpoint systems from the attested typology: no one of them is sufficient to explain the absence of the entire class. Rather, the hypothesis is that it is *all* of these factors, working together, that drive the attested frequency of midpoint systems down to zero.

[19] There are, of course, other classes of attested systems in which the placement of main stress relative to an edge is inconsistent. Two examples are: quantity-sensitive accentual window systems, where stress can fall anywhere within a certain domain (e.g. English); and languages with qualitatively-driven stress (e.g. Nanti, Crowhurst & Michael 2005). More work is required to determine whether or not these systems are also underattested relative to what we would expect, and if not, what differentiates them from the classes of systems discussed above.

This multi-part story, then, is the alternative to Kager's (2012) proposal that midpoint systems should be eliminated from the learner's hypothesis space. And if this learnability-based alternative is successful, there are important theoretical consequences. Recall that Kager's proposed modifications to CON involve the elimination of contextual anti-lapse constraints. As foot-free theories of stress depend on contextual anti-lapse constraints to model the typology of stress windows, the elimination of these constraints poses a serious problem for foot-free theories of stress. If, however, it is possible to show that the absence of midpoint systems can be explained in another way, then there is no need to exclude contextual anti-lapse constraints from CON. This means that the midpoint pathology no longer poses a problem for foot-free theories of stress, nor does it serve as an argument for the necessity of weakly layered feet in metrical theory.

It is important to keep in mind, however, that what *has* been shown in this exploration is only a small part of what *must* be shown for this learnability-based alternative to be truly viable. What the above discussion has established is that midpoint systems are difficult to learn for a specific type of machine learner. Our interest, however, is ultimately in the behavior of *human* learners: would they find midpoint systems difficult to learn, just like the machine learner does? In order for this alternative to be a valid one, it would be necessary to show that human learners behave like the machine learner, in that midpoint systems are difficult to acquire. While the behavior of human learners could potentially be assessed through artificial grammar learning experiments (see e.g. Carpenter 2010, Greenwood 2014 for experiments involving stress systems), for the time being I leave these questions open. This paper has shown that the learnability problem posed by midpoint systems is a possible explanation for their absence from the attested typology. Further work is necessary to determine whether or not this explanation is the correct one.

References

Abrahamson, Arne. 1968. Contrastive distribution of phoneme classes in Içuã Tupi. Anthropological Linguistics 10(6). 11–21.

Alber, Birgit. 2005. Clash, lapse and directionality. Natural Language and Linguistic Theory 23. 485–542.

Albright, Adam & Young Ah Do. 2013. Biased learning of phonological alternations. Paper presented at the 21st Manchester Phonology Meeting.

Alderete, John. 2008. Using learnability as a filter on factorial typology: A new approach to Anderson and Browne's generalization. Lingua 118. 1177–1220.

Anttila, Arto. 1997. Deriving variation from grammar. In Amsterdam studies in the theory and history of linguistic science series, vol. 4, chap. 35-68.

Bailey, Todd Mark. 1995. Nonmetrical constraints on stress: University of Minnesota dissertation.

Bane, Max & Jason Riggle. 2008. Three Correlates of the Typological Frequency of Quantity-Insensitive Stress Systems. In Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, 29–38. Columbus, Ohio.

Blevins, Juliette. 2004. Evolutionary phonology: The emergence of sound patterns. Cambridge: Cambridge University Press.

Blevins, Juliette & Sheldon P. Harrison. 1999. Trimoraic feet in Gilbertese. Oceanic Linguistics 38. 203–230.

Boersma, Paul. 1997. Functional phonology: formalizing the interactions between articulatory and perceptual drives: University of Amsterdam dissertation.

Boersma, Paul. 2003. Bruce Tesar and Paul Smolensky (2000). Learnability in Optimality Theory. Cambridge, Mass.: MIT Press. Pp. vii + 140. Phonology 20. 436–446.

Boersma, Paul & Bruce Hayes. 2001. Empirical Tests of the Gradual Learning Algorithm. Linguistic Inquiry 32. 45–86.

Boersma, Paul & Clara Levelt. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. In Proceedings of child language research forum, vol. 30, 229–237.

Boersma, Paul & Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy & Joe Pater (eds.), Harmonic Grammar and Harmonic Serialism, London: Equinox Press.

Brasoveanu, Adrian & Alan Prince. 2011. Ranking and necessity: the Fusional Reduction Algorithm. Natural Language and Linguistic Theory 29. 3–70.

Caballero, Gabriela. 2011. Morphologically conditioned stress assignment in Choguita Rarámuri. Linguistics 49. 749–790.

Carpenter, Angela C. 2010. A nautralness bias in learning stress. Phonology 27. 345–392.

Crowhurst, Megan & Lev D. Michael. 2005. Iterative Footing and Prominence-Driven Stress in Nanti (Kampa). Language 81. 47–95.

Davies, H. John. 1980. Kobon phonology Pacific Linguistics Series B, No. 87. Canberra: Australian National University.

Dixon, R. M. W. 1983. Nyawaygi. In R. M. W. Dixon & Barry J. Blake (eds.), Handbook of Australian Languages, vol. 3, 431–525. The Australian National University Press.

Do, Young Ah. 2013. Biased learning of phonological alternations: MIT dissertation.

Dresher, Elan & Aditi Lahiri. 1991. The Germanic foot: metrical coherence in Old English. Linguistic Inquiry 22. 251–286.

Eisner, Jason. 1997. What constraints should OT allow? Paper presented at the Annual Meeting of the Linguistic Society of America, Chicago. ROA-204.

Elenbaas, Nine & René Kager. 1999. Ternary rhythm and the lapse constraint. Phonology 16. 273–329.

Furby, Christine. 1974. Garawa phonology Pacific Linguistics, series A. Canberra: Australian National University.

Goedemans, Rob & Harry van der Hulst. 2009. StressTyp: a database for word accentual patterns in the world's languages. In The use of databases in cross-linguistics research, 235–282. New York: Mouton de Gruyter.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Proceedings of the Workshop on Variation within Optimality Theory, 111–120.

Gordon, Matt. 2002. A factorial typology of quantity insensitive stress. Natural Language and Linguistic Theory 20. 491–552.

Greenwood, Anna. 2014. Unpacking the effects of naturalness and simplicity biases in stress pattern learning. Talk presented at the 45th Meeting of the North East Linguistics Society, Massachusetts Institute of Technology.

Harms, Robert. 1966. Stress, voice, and length in Southern Paiute. International Journal of American Linguistics 32. 228–235.

Hatzigeorgiu, Nick, George Mikros & George Carayannis. 2001. Word length, word frequencies and Zipf's law in the Greek language. Journal of Quantitative Linguistics 8(3).

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In René Kager, Joe Pater & Wim Zonneveld (eds.), Fixing priorities: Constraints in phonological acquisition, Cambridge: Cambridge University Press.

Hayes, Bruce, Bruce Tesar & Kie Zuraw. 2013. OTSoft 2.3.2, software package. http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Heinz, Jeffrey. 2009. On the role of locality in learning stress patterns. Phonology 26. 303–351.

Heinz, Jeffrey. 2010. Learning Long-Distance Phonotactics. Linguistic Inquiry 41. 623–661.

Hewitt, Mark S. 1992. Vertical maximization and metrical theory: Brandeis University dissertation.

Hintz, Diane. 2006. Stress in South Conchucos Quechua: A phonetic and phonological study. International Journal of American Linguistics 72. 477–521.

Hualde, José Ignacio. 1991. Basque phonology. New York: Routledge.

Hughto, Coral, Joe Pater & Robert Staubs. 2015. Grammatical agent-based modeling of typology. Paper presented at the GLOW 2015 Phonology Workshop, Paris, France.

van der Hulst, Harry. 1996. Separating primary and secondary accent. In Rob Goedemans, Harry van der Hulst & Ellis Visch (eds.), Stress Patterns of the World, 1–25. The Hague: Holland Academic Graphics.

Hyde, Brett. 2008. Alignment continued: distance-sensitivity, order-sensitivity, and the midpoint pathology. Washington University, ms. ROA-998.

Hyde, Brett. 2015. The Midpoint Pathology: What it is and what it isn't. ROA-1231.

Itô, Junko & Armin Mester. 1992. Weak layering and word binarity. Ms., University of California, Santa Cruz.

Jäger, Gerhard. 2007. Maximum Entropy Models and Stochastic Optimality Theory. In A. Zaenen, J. Simpson, T. H. King, J. Grimshaw, J. Maling & C. Manning (eds.), Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan, CSLI Publications, Stanford.

Jamieson, Allan. 1977. Chiquihuitlan Mazatec phonology. In William Merrifield (ed.), Studies in Otomanguean Phonology, 93–106. Arlington, TX: Summer Institute of Linguistics.

Jarosz, Gaja. 2013. Naive Parameter Learning for Optimality Theory - the Hidden Structure Problem. In Seda Kan, Claire Moore-Cantwell & Robert Staubs (eds.), NELS 40: Proceedings of the 40th Annual Meeting of the North East Linguistic Society, Amherst, MA: University of Massachusetts Graduate Student Association.

Jeanne, LaVerne Masayesva. 1982. Some phonological rules of Hopi. International Journal of American Linguistics 48. 245–270.

Jesney, Karen & Anne-Michelle Tessier. 2011. Biases in Harmonic Grammar: the road to restrictive learning. Natural Language and Linguistic Theory 29. 251–290.

Kager, René. 1993. Alternatives to the Iambic-Trochaic Law. Natural Language and Linguistic Theory 11. 381–432.

Kager, René. 1994. Ternary rhythm in alginment theory. Ms., Utrecht University.

Kager, René. 2001. Rhythmic directionality by positional licensing. Paper presented at the Fifth HIL Phonology Conference (HILP 5).

Kager, René. 2012. Stress in windows: Language typology and factorial typology. Lingua 122. 1454–1493.

Kager, René & Violeta Martínez-Paricio. 2014. Minimally recursive foot structure. Paper presented at the Annual Meeting on Phonology 2014, MIT, Cambridge, MA.

Kalimeri, Maria, Vassilios Constantoudis, Constantinos Papadimitriou, Konstantinos Karamanos, Fotis K. Diakonos & Harris Papageorgiu. 2015. Word-length Entropies and Correlations of Natural Language Written Texts. Journal of Quantitative Linguistics 22. 101–118.

Kenstowicz, Michael & Hyang-Sook Sohn. 2001. Accentual adaptations in North Kyungsang Korean. In Michael Kenstowicz (ed.), Ken Hale: a Life in Language, 239–270. Cambridge: MIT Press.

Key, Harold H. 1961. Phonotactics of Cayuvava. International Journal of American Linguistics 27. 143–150.

Key, Harold H. 1967. Morphology of Cayuvava. The Hague: Mouton.

de Lacy, Paul. 2014. Evaluating evidence for stress systems. In Harry van der Hulst (ed.), Word stress: Theoretical and typological issues, 149–193. Cambridge: Cambridge University Press.

Leer, Jeff. 1985a. Prosody in Alutiiq. In Michael Krauss (ed.), Yupik Eskimo prosodic systems: descriptive and comparative studies, 77–134. Fairbanks, AK: Alaska Native Language Center, University of Alaska.

Leer, Jeff. 1985b. Towards a Metrical Interpretation of Yupik Prosody. In Michael Krauss (ed.), Yupik Eskimo prosodic systems: descriptive and comparative studies, chap. 159-172. Fairbanks, AK: Alaska Native Language Center.

LeSourd, Phil. 1988. Accent and syllable structure in Passamaquoddy: MIT dissertation.

LeSourd, Phil. 1993. Accent and syllable structure in Passamaquoddy. New York: Garland.

Magri, Giorgio. 2012. Convergence of error-driven ranking algorithms. Phonology 29. 213–269.

Magri, Giorgio. 2014. Error-driven versus batch models of the acquisition of phonotactics: David defeats Goliath. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.), Supplemental proceedings of the 2013 Meeting on Phonology, Linguistic Society of America.

Martínez-Paricio, Violeta. 2013. An exploration of minimal and maximal metrical feet: University of Tromsø dissertation.

Martínez-Paricio, Violeta & René Kager. 2014. Non-intervention constraints and the binary-to-ternary rhythmic continuum. Ms.

McCarthy, John J. 2003. OT constraints are categorical. Phonology 20. 75–138.

McGarrity, Laura. 2003. Constraints on patterns of primary and secondary stress. Bloomington, IN: Indiana University dissertation.

McLeod, S., J. van Doorn & V. Reed. 2001. Normal acquisition of consonant clusters. American Journal of Speech-Language Pathology 10. 99–110.

Mester, R. Armin. 1994. The quantitative trochee in Latin. Natural Language and Linguistic Theory 12. 1–61.

Michelson, Karin. 1988. A comparative study of Lake-Iroquoian accent. Dordrecht: Kluwer.

Miner, Kenneth L. 1989. Winnebago accent: The rest of the data. Anthropological Linguistics 31. 148–172.

Miyaoka, Osahito. 2012. A grammar of Central Alaskan Yupik (CAY). New York: Mouton.

Odden, David. 1985. An accentual approach to tone in Kimatuumbi. In Didier L. Goyvaerts (ed.), African linguistics: Essays in memory of m.w.k. semikenke, 345–420. Amsterdam: John Benjamins.

Odden, David. 1996. Phonology and morphology of Kimatuumbi. Oxford: Oxford University Press.

Osumi, Midori. 1995. Tinrin grammar. Honolulu: University of Hawai'i Press.

Pater, Joe. 2009. Weighted Constraints in Generative Linguistics. Cognitive Science 33. 999–1035.

Pater, Joe & Jessica A. Barlow. 2003. Constraint conflict in cluster reduction. Journal of Child Language 30. 487–526.

Piantadosi, Steven T., Harry Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. In Paul Kay (ed.), Proceedings of the National Academy of Sciences of the United States of America, vol. 108 9, 3526–3529.

Prince, Alan. 1980. A metrical theory for Estonian quantity. Linguistic Inquiry 11. 511–562.

Prince, Alan & Paul Smolensky. 2004. Optimality Theory: Constraint interaction in generative grammar. Oxford: Blackwell.

Prince, Alan & Bruce Tesar. 2004. Learning phonotactic distributions. In R. Kager, W. Zonneveld & J. Pater (eds.), Fixing priorities: Constraints in phonological acquisition, Cambridge: Cambridge University Press.

Prince, Alan S. 1983. Relating to the grid. Linguistic Inquiry 14(1). 19–100.

Pulleyblank, Douglas. 1983. Accent in Kimatuumbi. In Jonathan Kaye, Hilda Koopman, Dominique Sportiche & André Dugas (eds.), Current approaches to african lingustics, vol. 2, 195–216. Dordrecht: Foris Publications.

Radin, Paul. 1929. A grammar of the Wappo language. Berkeley: University of California Press.

Rice, Curt. 1992. Binarity and ternarity in metrical theory: parametric extensions: University of Texas at Austin dissertation.

Rifkin, Jay I. 2003. Ternarity is Prosodic Word binarity. In Jeroen van de Weijer, Vincent van Heueven & Harry van der Hulst (eds.), The Phonological Spectrum, vol. ii: Suprasegmental Structure, 127–150. Amsterdam: John Benjamins.

Sapir, Edward. 1930. Southern Paiute, a Shoshonean language. In Proceedings of the academy of arts and sciences, vol. 65, .

Selkirk, Elisabeth O. 1980. The role of prosodic categories in English word stress. Linguistic Inquiry 11. 563–605.

Sietsema, Brian Mark. 1989. Metrical dependencies in tone assignment: MIT dissertation.

Sigurd, Bengt, Mats Eeg-Olofsson & Joost van de Weijer. 2004. Word length, sentence length and frequency – Zipf revisited. Studia Linguistica 58. 37–52.

Staubs, Robert. 2014a. Computational modeling of learning biases in stress typology: UMass Amherst dissertation.

Staubs, Robert. 2014b. Learning and the position of primary stress. In Robert E. Santana-LaBarge (ed.), Proceedings of the 31st West Coast Conference on Formal Linguistics, 428–437.

Steriade, Donca. 2001. The phonology of perceptibility effects: The P-map and its consequences for constraint organization. Ms., UCLA.

Tesar, Bruce & Paul Smolensky. 2000. Learmability in optimality theory. Cambridge: MIT Press.

Tessier, Anne-Michelle. 2007. Biases and stages in phonological acquisition: UMass Amherst dissertation.

Tessier, Anne-Michelle. 2009. Frequency of violation and constraint-based phonological learning. Lingua 119. 6–38.

Tryon, Darrell. 1970. An Introduction to Maranungku. Canberra: Australian National University.

van Urk, Coppe. 2013. A typology of clash-tolerating languages. Ms., MIT.

Vihman, Marilyn May, Edwin Kay, Bénédicte de Boysson-Bardies, Catherine Durand & Ulla Sundberg. 1994. External Sources of Individual Differences? A Cross-Linguistic Analysis of the Phonetics of Mothers' Speech to 1-Year-Old Children. Developmental Psychology 30. 651–662.

Wheeler, Deirdre. 1979. A metrical analysis of stress and related processes in Southern Paiute and Tübatulabal. In University of massachusetts working papers in linguistics, vol. 5, 145–175.

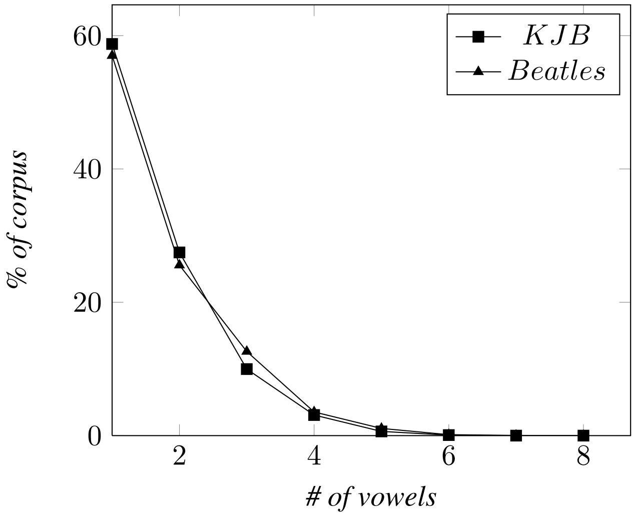
Zoll, Cheryl. 2004. Ternarity vs. final exclusion: a synthesis. Ms., Massachusetts Institute of Technology.

Appendix for "Learnability shapes typology: the case of the midpoint pathology"
Presentation and discussion of the word count data[1]

Although it has been shown for a number of lexica that short words outnumber long ones, there has been less work investigating the distributional properties of natural language corpora, i.e. an approximation of what a learner would encounter (references in 4.1). In order to make a meaningful cross-linguistic comparison, it is necessary to find a standardized text corpus that is readily available for a number of languages. To satisfy this criterion, the Bible was chosen, as it is a single text that has been translated into hundreds of languages. Counts were obtained from the book of Mark; in the rare case that Mark was unavailable, other books were substituted. Analysis was automated with a script that counted the number of words per number of orthographic vowels in a given corpus. Scripts and all other word-counting resources are available from the author upon request.

The Bible has the advantage of being a text that is freely available in hundreds of languages, but using it introduces several methodological issues. First, it is unclear exactly how closely the word length distribution of the Bible mirrors the word length distribution of every day speech. To explore this, I compared the word length distribution of the King James Bible (790,028 words; representing Biblical language) to a selection of interviews with various former members of the Beatles, from the Beatles Interviews Database[2] (90,713 words; representing everyday conversational speech). As shown below, despite the difference in both corpus size and corpus type, their word length distributions are nearly identical.

Figure 1: The Beatles Interviews Database (Beatles) vs. the King James Bible (KJB)



At least in English, then, the word length distributions in a somewhat archaic Bible translation and more contemporary everyday speech do not appear to differ too greatly. I have not yet investigated to what extent this result holds across other corpora, or in other languages.

Another methodological question surrounds the question of what counts as a vowel. In many languages, phoneme-to-orthography conversion is not one-to-one, and some graphemes can function as either a vowel or a consonant. English <y>, for example, is pronounced as the diphthong

[1] I am grateful to Francesca Cicileo for her help with this part of the project.

[2] `http://www.beatlesinterviews.org`. All interviews dated 1970 or later were included.

[aɪ] in the word *by*, but as the glide [j] in the word *you*. To avoid under-counting, English <y>, and phonemic chameleons in other languages, were always counted as vowels. More generally, if a given grapheme could function as a vowel in any context, it was counted as a vowel in all contexts; no attempt was made to account for language-specific, context-sensitive processes like glide formation. The set of vowels counted for a given language was generally determined by consulting online resources, i.e. Wikitravel's phrasebooks.[3] When this information was unavailable, the most likely set of vowels was determined by examining the distributional properties of suspect graphemes.

As each orthographic vowel was counted individually, this means that sequences of orthographic vowels were treated as sequences of monophthongs, rather than diphthongs (or triphthongs, etc.). It should be noted that counting each individual vowel, rather than each syllable, can sometimes lead to artificial inflation of the counts in languages where the phoneme-to-orthography conversion is not one-to-one. In English, for example, the word *beat* has two vowel symbols, though it has only one syllable (/bit/); Romanian *pierd-ea-i* ('you used to lose') has five vowel symbols, but only two syllables (/pi̯erde̯ai̯/).[4] No attempt was made to determine which vowel sequences constitute a syllable in a given language, or more generally, to control for differing orthographic conventions.

One final question regards the notion of wordhood. Languages often differ in which combinations of morphemes can be represented together as a single word. Even within a language, orthographic conventions are often inconsistent. Consider, for example, the pronominal clitics of French. In the imperative *Mange-les* ('eat them'), *les* is appended to the verb. In the indicative *Il les mange* ('he is eating them'), however, the clitic is written as a separate word. This orthographic difference does not correspond to a difference in phonology; in neither case is the clitic prosodically independent. Although differing conventions regarding orthographic wordhood introduce a confound, I did not attempt to address it. Here, a space constitutes a word boundary.

At the time of writing, the sample consisted of data from 102 languages, selected on the basis of the availability of online resources. The surveyed languages hail from 26 major language families: Afro-Asiatic (3), Algic (1), Austro-Asiatic (1), Austronesian (10), Eskimo-Aleut (2), Creole (English-based: 3; French-based: 1), Gunwingguan (3), Daly (1), Indo-European (33), Iroquoian (1), Japonic (1), Jivaroan (1), Koreanic (1), Mayan (6), Niger-Congo (8), Nilo-Saharan (2), Pama-Nyungan (10), Quechuan (1), Sepik (1), Sino-Tibetan (2), Trans-New Guinea (2), Turkic (2), Uto-Aztecan (2), Uralic (3), and one isolate (Basque). The number of words counted ranged from 6,415 (Inuktitut; Eskimo-Aleut) to 38,266 (Anindilyakwa; Pama-Nyungan), with an average of 14,402.

The investigation yielded two main results. First, the distribution of word lengths is extremely variable across languages. This is immediately visible in Figure 2 (on the next page), where each individual gray line represents the word length distribution of a single language. Despite this variability, however, in the large majority of languages, long words are extremely rare. The darker line in Figure 2 represents the median value for each word length; these values are also provided in (1).

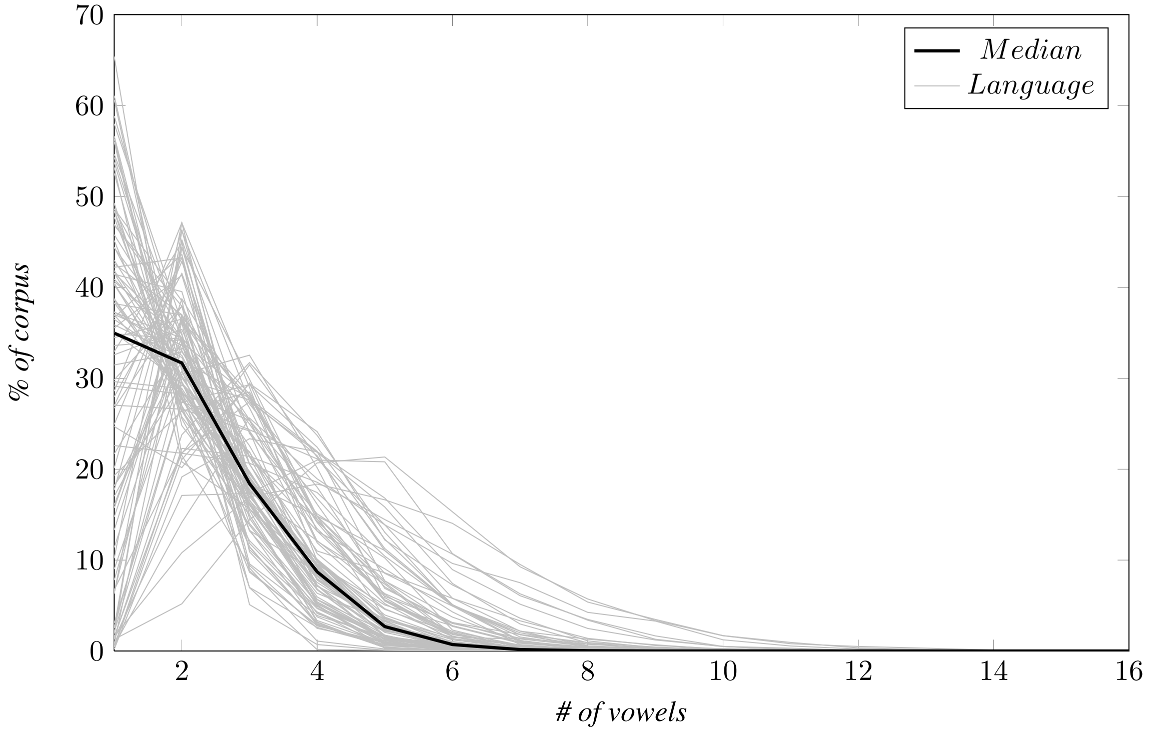
(1) Median values for the word count data

# of vowels	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Median	35%	32%	18%	9%	3%	1%	0%

[3] http://wikitravel.org/en/List_of_phrasebooks

[4] This example due to Donca Steriade (p.c.).

Figure 2: Results of word length study



In the pages that follow, I have included a breakdown of the word length distributions for each of the languages in the survey, together with language family information (from Ethnologue, Lewis et al. 2015) and the online data source. An editable Excel spreadsheet containing this information, as well as some additional material (raw numbers, information on which graphemes were counted as vowels, etc.) is available from the author upon request. The key for the sources is in (2).

(2)

Key	Source Name	Website
AB	eBaibul	`http://aboriginalbibles.org.au`
AKT	Alkitab TOBA	`http://alkitabtoba.wordpress.com`
B	YouVersion	`http://bible.com`
BB	Baibala Hemolele	`http://baibala.org`
BG	BibleGateway	`http://www.biblegateway.com`
BIT	SABDA-Web	`http://bit.net.id/SABDA-Web`
BL	Biblica	`http://www.biblica.com`
G	Project Gutenberg	`http://www.gutenberg.org`
JA	Jesus Army	`http://www.jesus-army.com`
PB	Da Hawai'i Pidgin Bible	`http://www.pidginbible.org`
U	The Unbound Bible	`http://unbound.biola.edu`
WB	WorldBibles.org	`http://worldbibles.org`
WP	WordProject	`http://www.wordproject.org/bibles`

Language	Family	Book(s) counted	Source	1	2	3	4	5	6	7	8	9	10+	Total
Abau	Sepik	Mark	B	37%	34%	18%	7%	2%	1%	0%	0%	0%	0%	18,678
Afrikaans	Indo-European	Mark	JA	47%	38%	10%	4%	2%	0%	0%	0%	0%	0%	15,680
Aguaruna	Jivaroan	Mark	B	3%	30%	28%	21%	11%	5%	2%	1%	0%	0%	11,309
Albanian	Indo-European	Mark	JA	49%	35%	12%	4%	0%	0%	0%	0%	0%	0%	14,337
Alyawarr	Pama-Nyungan	Mark	AB	11%	34%	28%	15%	8%	3%	1%	0%	0%	0%	12,405
Ama	Nilo-Saharan	Mark	B	25%	20%	29%	13%	6%	3%	2%	1%	0%	0%	18,805
Anindilyakwa	Gunwingguan	Luke	AB	1%	14%	26%	19%	14%	10%	8%	4%	3%	2%	38,266
Anmatyerre	Pama-Nyungan	Mark	AB	27%	38%	21%	10%	4%	1%	0%	0%	0%	0%	14,969
Arabic (Chadic, Romanized)	Afro-Asiatic	Mark	AB	29%	28%	25%	11%	5%	2%	0%	0%	0%	0%	13,792
Armenian (Western)	Indo-European	Mark	U	34%	37%	17%	8%	3%	1%	0%	0%	0%	0%	10,602
Arrarnta (Western)	Pama-Nyungan	Mark	AB	1%	36%	28%	22%	8%	4%	1%	0%	0%	0%	13,100
Aukan	Creole (English)	Mark	B	54%	30%	13%	3%	0%	0%	0%	0%	0%	0%	25,241
Azerbaijani (North)	Turkic	Mark	JA	16%	36%	25%	14%	6%	2%	1%	0%	0%	0%	10,069
Bengali	Indo-European	Mark	B	6%	44%	30%	15%	4%	1%	0%	0%	0%	0%	13,206
Bargam	Trans-New Guinea	Mark	B	36%	41%	17%	5%	1%	0%	0%	0%	0%	0%	20,417
Basque	(Isolate)	Mark	U	6%	37%	27%	18%	7%	3%	1%	0%	0%	0%	11,143
Batak Toba	Austronesian	Mark	AKT	35%	34%	20%	8%	3%	1%	0%	0%	0%	0%	14,382
Breton	Indo-European	Mark	WB	53%	26%	14%	5%	2%	1%	0%	0%	0%	0%	15,047
Bulgarian	Indo-European	Mark	BG	39%	28%	22%	8%	3%	0%	0%	0%	0%	0%	12,455
Burrara	Gunwingguan	Mark	AB	3%	45%	24%	15%	10%	2%	0%	0%	0%	0%	17,756
Cebuano	Austronesian	Mark	BG	37%	34%	19%	9%	1%	0%	0%	0%	0%	0%	16,386
Chamorro	Austronesian	Mark	U	22%	34%	21%	15%	6%	2%	1%	0%	0%	0%	12,767
Cherokee	Iroquoian	Mark	BG	1%	22%	21%	19%	14%	11%	6%	3%	2%	1%	8,974
Chinese (Mandarin)	Sino-Tibetan	Mark	WP	42%	43%	11%	3%	0%	0%	0%	0%	0%	0%	13,267
Croatian	Indo-European	Mark	JA	40%	30%	19%	8%	2%	0%	0%	0%	0%	0%	11,080
Czech	Indo-European	Mark	JA	42%	34%	16%	6%	2%	0%	0%	0%	0%	0%	10,997
Danish	Indo-European	Mark	BG	61%	27%	9%	3%	1%	0%	0%	0%	0%	0%	15,333
Dholuo	Nilo-Saharan	Mark	BL	30%	38%	24%	7%	1%	0%	0%	0%	0%	0%	13,078
Djambarrpuygnu	Pama-Nyungan	Mark	G	18%	44%	19%	10%	5%	3%	1%	0%	0%	0%	21,250
Dutch	Indo-European	Mark	JA	47%	36%	11%	4%	1%	0%	0%	0%	0%	0%	15,174
English	Indo-European	Mark	BL	57%	28%	11%	3%	1%	0%	0%	0%	0%	0%	14,337
Estonian	Uralic	Mark	JA	25%	43%	21%	8%	2%	1%	0%	0%	0%	0%	11,838
Éwé	Niger-Congo	Mark	BL	44%	34%	14%	5%	1%	0%	0%	0%	0%	0%	17,836
Faiwol	Trans-New Guinea	Mark	B	29%	47%	17%	5%	1%	0%	0%	0%	0%	0%	19,572
Finnish	Uralic	Mark	WB	19%	26%	28%	15%	7%	3%	1%	0%	0%	0%	11,063

Language	Family	Book(s) counted	Source	1	2	3	4	5	6	7	8	9	10+	Total
French	Indo-European	Mark	JA	42%	31%	16%	8%	3%	0%	0%	0%	0%	0%	13,873
Ganda	Niger-Congo	Mark	WP	23%	22%	21%	17%	10%	5%	2%	0%	0%	0%	9,463
German	Indo-European	Mark	BG	43%	37%	15%	4%	1%	0%	0%	0%	0%	0%	14,138
Gumatj	Pama-Nyungan	Mark	AB	8%	33%	25%	17%	9%	5%	2%	1%	0%	0%	22,906
Haitian	Creole (French)	Mark	BG	58%	36%	5%	1%	0%	0%	0%	0%	0%	0%	17,011
Hawaiian	Austronesian	Mark	BB	38%	37%	16%	6%	2%	1%	0%	0%	0%	0%	20,373
Hawai'i Pidgin	Creole (English)	Mark	PB	55%	35%	9%	1%	0%	0%	0%	0%	0%	0%	19,364
Hiligaynon	Austronesian	Mark	BG	41%	34%	13%	9%	2%	1%	0%	0%	0%	0%	16,981
Hindi (Romanized)	Indo-European	Mark	WP	41%	33%	19%	5%	1%	0%	0%	0%	0%	0%	15,094
Hmar	Sino-Tibetan	Mark	JA	65%	22%	9%	2%	1%	0%	0%	0%	0%	0%	15,925
Hungarian	Uralic	Mark	WB	30%	29%	18%	12%	6%	3%	1%	1%	0%	0%	11,466
Icelandic	Indo-European	Mark	BG	49%	39%	9%	3%	1%	0%	0%	0%	0%	0%	12,439
Indonesian	Austronesian	Mark	BIT	10%	47%	29%	10%	2%	1%	0%	0%	0%	0%	13,718
Inuktitut	Eskimo-Aleut	Mark	B	1%	5%	14%	21%	21%	15%	9%	6%	3%	3%	6,415
Inupiatun (NW Alaska)	Eskimo-Aleut	Mark	B	2%	11%	17%	18%	17%	14%	10%	5%	3%	3%	7,588
Irish	Indo-European	Mark	WB	49%	26%	16%	6%	2%	1%	0%	0%	0%	0%	14,999
Italian	Indo-European	Mark	WB	35%	29%	22%	10%	3%	1%	0%	0%	0%	0%	12,227
Jakalteko	Mayan	Mark	BG	27%	43%	18%	8%	2%	1%	0%	0%	0%	0%	17,911
Japanese	Japonic	Mark	WP	48%	30%	16%	5%	1%	0%	0%	0%	0%	0%	19,239
Kabyle	Afro-Asiatic	Mark	U	28%	42%	21%	7%	2%	0%	0%	0%	0%	0%	10,464
Kaqchikel	Mayan	Mark	BG	48%	26%	16%	7%	2%	0%	0%	0%	0%	0%	25,548
K'iche'	Mayan	Mark	BG	56%	26%	12%	5%	1%	0%	0%	0%	0%	0%	21,155
Korean	Koreanic	Mark	WP	9%	30%	33%	20%	6%	2%	0%	0%	0%	0%	8,700
Kriol	Creole (English)	Mark	AB	37%	44%	15%	4%	1%	0%	0%	0%	0%	0%	20,490
Kuku-Yalanji	Pama-Nyungan	Mark	AB	2%	46%	27%	16%	6%	2%	0%	0%	0%	0%	11,998
Latvian	Indo-European	Mark	JA	36%	31%	19%	9%	3%	1%	0%	0%	0%	0%	11,283
Lithuanian	Indo-European	Mark	JA	27%	27%	24%	13%	6%	2%	1%	0%	0%	0%	9,786
Macedonian	Indo-European	Mark	BG	40%	28%	20%	8%	3%	0%	0%	0%	0%	0%	13,258
Malagasy	Austronesian	Mark	JA	18%	32%	28%	12%	7%	2%	1%	0%	0%	0%	13,392
Mam (Central)	Mayan	Mark	BG	54%	31%	11%	3%	0%	0%	0%	0%	0%	0%	16,539
Manx	Indo-European	Mark	U	53%	21%	16%	6%	3%	0%	0%	0%	0%	0%	15,744
Maori	Austronesian	Mark	BG	44%	30%	18%	5%	2%	1%	0%	0%	0%	0%	18,861
Murrinh-Patha	Daly	Various[5]	AB	20%	46%	19%	7%	4%	2%	1%	1%	0%	0%	12,063
Nahuatl	Uto-Aztecan	Mark	BG	18%	29%	21%	11%	9%	6%	3%	1%	1%	0%	13,568
Ndebele	Niger-Congo	Mark	BL	1%	24%	32%	24%	12%	7%	4%	1%	0%	0%	8,466

Language	Family	Book(s) counted	Source	1	2	3	4	5	6	7	8	9	10+	Total
Ngaanyatjarra	Pama-Nyungan	Mark	AB	0%	21%	29%	24%	12%	7%	4%	1%	0%	0%	10,987
Norwegian	Indo-European	Mark	BG	61%	28%	7%	3%	1%	0%	0%	0%	0%	0%	14,786
Nunggubuyu	Gunwingguan	Mark	AB	1%	17%	17%	21%	21%	11%	6%	3%	1%	0%	12,801
Pipil	Uto-Aztecan	Mark	BG	37%	29%	18%	10%	5%	2%	0%	0%	0%	0%	11,475
Polish	Indo-European	Mark	JA	36%	32%	19%	8%	3%	1%	0%	0%	0%	0%	11,268
Portuguese	Indo-European	Mark	JA	33%	35%	18%	10%	3%	1%	0%	0%	0%	0%	12,494
Potawatomi	Algic	Mark	U	17%	31%	18%	15%	11%	5%	2%	1%	0%	0%	15,233
Q'eqchi'	Mayan	Mark	BG	45%	27%	16%	9%	3%	1%	0%	0%	0%	0%	18,264
Quechua	Quechuan	Mark	BG	2%	22%	27%	22%	16%	7%	3%	1%	0%	0%	10,163
Romani	Indo-European	Mark	JA	33%	45%	17%	5%	0%	0%	0%	0%	0%	0%	13,735
Romanian	Indo-European	Mark	JA	42%	30%	17%	8%	3%	1%	0%	0%	0%	0%	13,980
Russian	Indo-European	Mark	BG	34%	34%	19%	9%	2%	1%	0%	0%	0%	0%	11,292
Scottish Gaelic	Indo-European	Mark	U	49%	26%	16%	6%	2%	1%	0%	0%	0%	0%	15,024
Serbian	Indo-European	Mark	JA	43%	31%	17%	7%	1%	0%	0%	0%	0%	0%	11,119
Slovak	Indo-European	Mark	BG	37%	33%	18%	9%	2%	0%	0%	0%	0%	0%	11,818
Somali	Afro-Asiatic	Mark	WP	16%	35%	25%	14%	8%	2%	0%	0%	0%	0%	12,839
Spanish	Indo-European	Mark	WB	40%	28%	20%	10%	2%	0%	0%	0%	0%	0%	13,009
Swahili	Niger-Congo	Mark	WB	15%	37%	21%	13%	9%	4%	1%	0%	0%	0%	10,528
Swedish	Indo-European	Mark	WB	46%	34%	13%	5%	2%	1%	0%	0%	0%	0%	14,735
Tagalog	Austronesian	Mark	WB	38%	28%	18%	10%	4%	1%	0%	0%	0%	0%	14,833
Turkish	Turkic	Mark	JA	13%	37%	28%	14%	5%	2%	0%	0%	0%	0%	9,325
Twi	Niger-Congo	Mark	BG	48%	28%	16%	7%	1%	0%	0%	0%	0%	0%	14,622
Ukrainian	Indo-European	Mark	BG	31%	33%	23%	9%	3%	1%	0%	0%	0%	0%	11,925
Uma	Austronesian	Mark	U	16%	39%	27%	12%	5%	1%	0%	0%	0%	0%	14,301
Uspanteko	Mayan	Mark	BG	41%	40%	15%	3%	0%	0%	0%	0%	0%	0%	17,743
Vietnamese	Austro-Asiatic	Mark	JA	59%	34%	7%	0%	0%	0%	0%	0%	0%	0%	15,613
Walmajarri	Pama-Nyungan	Various[6]	AB	3%	27%	24%	14%	6%	2%	1%	1%	0%	0%	10,939
Warlpiri	Pama-Nyungan	Mark	AB	0%	19%	23%	22%	17%	9%	5%	2%	0%	0%	16,151
Wik-Mungkan	Pama-Nyungan	Mark	AB	39%	28%	18%	10%	4%	1%	0%	0%	0%	0%	20,784
Wolof	Niger-Congo	Mark	WB	56%	25%	15%	3%	1%	0%	0%	0%	0%	0%	13,418
Xhosa	Niger-Congo	Mark	JA	7%	22%	31%	21%	11%	5%	2%	1%	0%	0%	9,078
Zulu	Niger-Congo	Mark	WP	2%	27%	29%	22%	12%	5%	1%	0%	0%	0%	7,685

[6] Genesis 1-9,11 + Jonah 1-4 + 1 Thessalonians 1-5 + 2 Thessalonians 1-3 (gospels not available)

[6] Mark (1,2,4,5-9,11,14-16) + John 8, 11, 20, 21 + Matthew 1,2,6,11,27. These do not all appear to be full chapters.

References

2015. Alkitab TOBA. Online at http://alkitabtoba.wordpress.com.

2015. Baibala Hemolele: The Hawaiian Bible. Ulukau: Hawaiian Electronic Library. Online at http://baibala.org.

2015. The Beatles Interview Database. Online at http://beatlesinterviews.org.

2015. BibleGateway. Online at http://biblegateway.com.

2015. Biblica: Transforming lives through God's Word. Online at http://www.biblica.com.

2015. Da Hawai'i Pidgin Bible. Online at http://www.pidginbible.org.

2015. eBaibul: Bible text in Australian Aboriginal languages. Online at http://aboriginalbibles.org.au.

2015. Jesus Army. Online at http://jesus-army.com.

2015. List of phrasebooks. Online at http://wikitravel.org/en/List_of_phrasebooks.

2015. Project Gutenberg. Online at http://www.gutenberg.org.

2015. SABDA-Web. Online at http://http://www.bit.net.id/SABDA-Web/.

2015. The Unbound Bible. Online at http://unbound.biola.edu.

2015. WordProject. Online at http://www.wordproject.org/bibles.

2015. WorldBibles.org: Helping you to find God's Word in over 4,000 languages. Online at http://worldbibles.org.

2015. YouVersion. Online at http://bible.com.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig. 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.