Segmentation and UR Acquisition with UR Constraints*

Max Nelson
University of Massachusetts Amherst
manelson@umass.edu

Abstract

This paper presents a model that treats segmentation and underlying representation acquisition as parallel, interacting processes. A probability distribution over mappings from underlying to surface representations is defined using a Maximum Entropy grammar which weights a set of underlying representation constraints (URCs) (Apoussidou, 2007; Pater et al., 2012). URCs are induced from observed surface strings and used to generate candidates. Structural ambiguity arising from the comparison of segmented outputs to unsegmented surface strings is handled with Expectation Maximization (Dempster et al., 1977; Jarosz, 2013). The model successfully learns a simple voicing assimilation rule and segmentation via correspondences between surface phones and input meanings. The trained grammar is also able to segment novel forms affixed with familiar morphemes.

1 Introduction

Segmentation is the task by which continuous speech is broken up into discrete words. This task is complicated by the fact that there are no universal cues to word boundary location. Language-specific morphological, phonotactic, and prosodic cues to word boundaries do exist, but these cues are unavailable in early acquisition because their cooccurrence with word boundaries has not yet been observed (Perruchet and Vinter, 1998).

The lexicon and accompanying phonological knowledge provide a rich source of potential information about boundary location. If any substring from an utterance can be mapped onto a lexical item, then boundaries can be inferred by identifying the correspondences between the phones in the surface string and those in the underlying form. However, using this knowledge requires that some of the lexicon is known to the learner and that segmentation has already been used successfully to identify surface forms.

The fact that segmentation is a prerequisite to build the lexicon only precludes lexical information from being used in segmentation if the two processes take place in serial, with learners developing the ability to segment speech before storing any lexical information. Previous models of segmentation either ignore the acquisition of the lexicon (Saffran et al., 1996a; Saffran et al., 1996b; Perruchet and Vinter, 1998) or do not fully utilize the richness of lexical knowledge (Johnson et al., 2015; Goldwater et al., 2009). This paper presents a model of segmentation in which the lexicon, represented by phonological underlying forms which correspond to meanings, is being acquired in parallel with segmentation, and the two processes are mutually informing. This type of joint inference has been explored elsewhere, particularly with regards to the interaction of segmentation with phonetic categorization and lexical acquisition (Elsner et al., 2013; Elsner et al., 2016), but little work has been done on the interaction of other processes with the acquisition of phonological alternations.

*Thank you to Katherine Blake, Gaja Jarosz, Andrew Lamont, Joe Pater, Brandon Prickett, UMass Phonology Reading Group, and three anonymous reviewers for comments. Remaining errors are my own.

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 60-68.
New York City, New York, January 3-6, 2019

2 Background

2.1 Segmentation

Early work on segmentation excluded the use of phonological knowledge by design. Saffran et al. (1996a; 1996b) conducted a series of experiments in which both infants and adults were tasked with segmenting continuous speech that had no prosodic cues to word boundaries, finding in all cases that participants were able to segment the data into the composite words. This led to the hypothesis that learners are able to identify word boundaries solely by tracking transitional probability minima in the input.

However, the storage and update of transitional probabilities is computationally costly and statistical models have been shown to be successful without relying on their direct computation. One such model is Perruchet and Vinter's PARSER (1998). The PARSER model takes advantage of the fact that any randomly selected set of syllables is more likely to reoccur if the syllables are a word than if they are not, storing a set of weights on encountered substrings rather than explicitly storing and computing transitional probabilities.

Both of these approaches model segmentation in isolation. Johnson and Jusczyk (2001) suggested that when phonological cues to word boundaries are available, they supercede statistics in word boundary identification. Infants in their study were more likely to learn word boundaries cued by prosodic/phonological information than competing boundaries cued by statistical information. Furthermore, segmentation is a necessary step toward the identification of phonological surface forms, which are the necessary precursor to the learning of phonotactics, phonological grammars and underlying representations. Phonological acquisition both feeds and is fed by segmentation; therefore a model of segmentation that does not incorporate phonological processes and underlying forms is incomplete.

Similarly, a model of segmentation should not only model acquisition, but should also model adult-like behavior. A simple Wug task (Berko Gleason, 1958) involves the use of lexical and phonological knowledge to identify correspondences between phonological content in the surface form and known morphemes. This results in a segmentation of the novel word, but this kind of segmentation task is largely absent from the literature. The use of lexical knowledge to predict segmentations requires that the learner entertain multiple possible lexical entries for a given meaning. The model presented below uses underlying representation constraints (URCs) within a standard constraint-based grammar (Prince and Smolensky, 19932004; Pater et al., 2012; Smith, 2015) to allow the learner to entertain multiple possible URs. The likelihood of a segmentation is affected by the likelihood of the corresponding URs and phonological alternations.

2.2 Underlying Representation Constraints

Underlying representation constraints (URCs), also referred to as lexical constraints, specify the underlying form for a meaning and are violated when an alternative underlying form is chosen (Apoussidou, 2007; Kager, 2008; Eisenstat, 2009). URCs allow the selection of underlying forms to happen in parallel with phonological optimization, allowing the grammar to choose between multiple URs with an eye to the phonological consequences of the decision (Pater et al., 2012). A sample UR constraint is defined below, using language from Smith (2015). This constraint specifies the underlying form /ə/ for the indefinite article.

> {IND}=/ə/ : Assign one violation for every input set of morphosyntactic features corresponding to IND (indefinite determiner) that is not realized by /ə/

URCs represent non-discrete lexical entries. The phonological representation of a lexical item is distributed over the set of relevant URCs. When there are multiple candidate URs and corresponding surface allomorphs, the choice between URs is made in the phonology in parallel with other phonological operations (Pater et al., 2012; Smith, 2015) . Inputs to the phonology are sets of meanings without any inherent phonological material; following Smith (2015), these are formalized as sets of morphosyntactic features. Candidates evaluated by the grammar are mappings from underlying to surface forms.

To illustrate how UR constraints interact with the rest of the phonology, consider the *a*∼*an* alternation in English. Simplifying slightly by ignoring vowel reduction, the indefinite determiner surfaces as [ə]

before a consonant and [ən] before a vowel.

If the UR of the indefinite determiner were always /ə/, describing this process would require /n/ insertion and the analyst would be tasked with accounting for the fact that [n]-epenthesis occurs in only this specific environment. Likewise, if the UR were assumed to be /ən/, this process would require preconsonantal /n/-deletion and the analyst would have to account for the lack of /n/-deletion elsewhere. With UR constraints however there is a third possibility: UR selection. The tableaux in (1) and (2) illustrate how UR selection can result in a non-default form surfacing due to pressure from the standard markedness constraint HIATUS, which penalizes adjacent vowels.

{IND} + DOG	HIATUS	IND=ə	IND=ən
☞ a. ə+dɔg → ədɔg			∗
b. ən+dɔg → əndɔg		∗W	L

Tableau 1: The default UR, /ə/, is chosen when there is no interaction with markedness constraints

In Tableau (1) there is no possible HIATUS violation so /ə/, the default UR, is chosen. The default status of /ə/ is captured by the ranking IND=ə≫IND=ən. Tableau (2) illustrates how a potential HIATUS violation can result in the selection of a non-default form, creating a surface alternation. When a markedness constraint outranks the constraint specifying the default UR then a non-default UR can be chosen to repair the markedness violation.

{IND} + ANT	HIATUS	IND=ə	IND=ən
a. ə+ænt → æənt	∗W	L	∗W
☞ b. ən+ænt → ənænt		∗	

Tableau 2: The non-default UR, /ən/ is rendered optimal by a high ranked markedness constraint

The tableaux in (1) and (2) do not consider candidates in which the UR→SR mapping is unfaithful. In the URC model, faithfulness constraints evaluate faithfulness between the selected UR and corresponding surface form. To illustrate the role of faithfulness in UR selection Tableaux (1) and (2) are repeated in Tableaux (3) and (4) with MAX and DEP added to the constraint set and the relevant unfaithful candidates considered.

{IND} + DOG	DEP	MAX	HIATUS	IND=ə	IND=ən
☞ a. ə+dɔg → ədɔg					∗
b. ən+dɔg → əndɔg				∗W	L
c. ə+dɔg → əndɔg	∗W				∗
d. ən+dɔg → ədɔg		∗W		∗W	L

Tableau 3: The default UR is chosen and sufaces faithfully when there are no interacting markedness constraints

Candidate (d) in (3) illustrates why an /n/-deletion account if the *a∼an* alternation does not work, it is harmonically bounded by (b) due to the lack of a markedness constraint motivating deletion and (a) due to the lack of a markedness constraint motivating non-default UR selection.

{IND} + ANT	DEP	MAX	HIATUS	IND=ə	IND=ən
a. ə+ænt → æənt			∗W	L	∗W
☞ b. ən+ænt → ənænt				∗	
c. ə+ænt → ənænt	∗W			L	∗W
d. ən+ænt → æənt		∗W	∗W	∗	

Tableau 4: High ranked faithfulness prevent an unfaithful mapping from the default UR from being optimal

Candidate (c) in (4) illustrates why an /n/-epenthesis analysis of the *a∼an* alternation does not work, it is ruled out by high ranked DEP which is necessary to account for the with the lack of /n/-epenthesis elsewhere in English in response to HIATUS violations.

UR selection is a viable alternative to faithfulness-violating phonological alternations when the alternation is either unmotivated or highly restricted. However, UR selection as described thus far remains a possibility even in cases in which a standard phonological explanation is preferred. The URC model provides no convincing reason that UR selection should not be used in, for example, the English plural alternation. Smith (2015) holds that the use of UR selection is limited by the fact that not all inputs have multiple UR constraints. UR selection is limited to suppletive forms because only those forms have multiple URCs. This claim creates problems for the learnability of URCs. A learner cannot restrict the creation of URCs to suppletive forms without first knowing that those forms are suppletive. The model presented below shows that this stipulation is unnecessary. Removing these restrictions makes the URC induction task tractable and does not result in rampant use of UR selection when a simple phonological solution is available.

3 Model

3.1 The grammar and learning algorithm

The present model uses URCs along with standard phonological constraints in a Maximum Entropy (MaxEnt) grammer (Goldwater and Johnson, 2003) to learn a probability distribution over segmented phonological surface forms for any input set of morphosyntactic objects. The training data consists of mappings from morphosyntactic objects to surface forms, which have no surface-apparent segmentation. At no point are segmentations provided to the learner: segmentations of inputs emerge as a result of the acquisition of URs, through the induction and weighting of URCs, and the acquisition of phonological alternations.

In a MaxEnt grammar, constraints are weighted and candidates' violations of constraints are represented by negative integers. The weighted sum of constraint violations is referred to as the *harmony* of a candidate. The closer to 0 the harmony is, the more likely that candidate is to surface. The probability distribution over the set of candidates is calculated by applying the softmax function to the set of harmonies. In this case a single candidate x is a mapping from underlying to surface form and an input M is a set of morphosyntactic features. This is shown explicitly in the formula in (1), where c_i represents the number of violations the mapping of M to x incurs on constraint i, w_i represents the current weight of i, and Ω_M represents the set of all candidates in the tableau for input M.

$$p(x \mid M) = \frac{e^{-(\sum_i w_i c_i(M,x))}}{\sum_{x' \in \Omega_M} e^{-(\sum_i w_i c_i(M,x'))}} \quad (1)$$

The learner's goal is to find the set of weights that maximize the likelihood of the training data T or, in other words, minimize the negative log likelihood:

$$\mathcal{L} = -\sum_{x \in T} \log p(x) \quad (2)$$

This is used as the current model's objective function with no regularization. Learning is error-driven and trained via stochastic gradient descent. In standard MaxEnt the calculation of the gradient is relatively simple. For a single training datum y, which in this case is a mapping from a set of morphosyntactic features to a surface string, the gradient of the loss function with respect to a given weight can be calculated as follows:

$$\frac{\partial \mathcal{L}}{\partial w_i} = c_i(M, y) - \sum_{x \in \Omega_M} c_i(M, x) p(x) \quad (3)$$

The update to a constraint's weight given a training datum is the learning rate times the difference between the observed number of violations of that constraint, $c_i(M, y)$, and the expected number of violations based on the current state of the model, $\sum_{x \in \Omega_M} c_i(M, x) p(x)$.

In this model, however, things are complicated by the fact that there can be multiple possible segmented outputs that, when segmentation is removed, produce the observed surface string. As framed here, the segmentation problem is therefore a problem of learning structural ambiguity - a topic of much recent work in the phonological learning literature (see Jarosz (2019) for a recent review). This creates two challenges for standard stochastic gradient descent in MaxEnt.

First, the definition of an error must be revised. In standard error-driven learning it is straighforward to compare the predicted output and the observed form. However, in this case the predicted output has more structure than the observed. Tesar and Smolensky's (1998) Robust Interpretive Parsing algorithm overcomes this issue by using the current grammar to assign structure to the observed form before making a prediction, allowing for the observed and predicted forms to both be fully structured. Jarosz's hidden structure learning algorithm, Expected Interpretive Parsing (2013), is the basis for the algorithm used here, and the definition of 'error' adopted follows her account: an error occurs when the predicted form, stripped of structure, does not match the observed form. The learner is therefore agnostic about segmentation with regard to errors. Both ðə$_1$#dɔg$_2$ and ðəd$_1$#ɔg$_2$ are acceptable segmented outputs for the input {DEF}$_1$+{DOG}$_2$, where # represents a word boundary.

Second, in the update rule above, $c_i(M, y)$ refers to the number of violations of a constraint incurred by the observed form. However, because the observed form has no structure, the corresponding structured candidate in the tableau is unknown and the violations cannot be counted. A solution to this

problem relies on the use of expectation maximization (Dempster et al., 1977; Jarosz, 2013; Jarosz, 2015). An estimate of the observed violations of a constraint can be made given the grammar's current belief about the likelihood of the different segmentations of the unsegmented input. Given a training datum y, an estimate of the observed violations for a constraint i can be calculated as in (4), where Z_y is the set of all outputs that are possible segmentations of the observed string.

$$\hat{c}_i(M, y) = \sum_{z \in Z_y} c_i(M, z) \frac{p(z)}{\sum_{z \in Z_y} p(z)} \quad (4)$$

This is equivalent to defining a probability distribution over the set of segmented candidates that overtly produce the unsegmented observed form, and then assigning a probabilistic segmentation to the observed form that is the average of all possible segmentations weighted by their probabilities.

3.2 URC induction

The training data take the form of observed surface strings and their underlying sets of morphosyntactic objects. Upon encountering a novel datum, the learner first constructs the complete set of UR constraints for all present morphosyntactic objects given that datum and adds them to the current grammar. These constraints are then immediately used in the generation of the candidate set and evaluation of the grammar.

Given a string and a set of n corresponding morphosyntactic objects, URC induction begins by computing every possible partition of the string into n non-empty substrings. A URC is then added to the grammar specifying that every substring is the UR for every morphosyntactic object in the input. This process is illustrated below for a sample training datum: the observed surface form [abc] for the morphosyntactic objects {M1}+{M2}.

Segmentation	New UR constraints	
a#bc	{M1}=a,	{M1}=bc,
	{M2}=a,	{M2}=bc
ab#c	{M1}=ab,	{M1}=c,
	{M2}=ab,	{M2}=c

Table 1: UR constraints generated from the two possible segmentations of [abc] into two non-vacuous substrings

This method of constraint induction implicitly assumes that all morphosyntactic objects will have some phonological exponent. It also provides no mechanism for URCs to specify strings that do not occur at any point in the training data. In other words, every underlying form must surface faithfully at least once in order to be considered a possible UR. This assumption is shared by other models of UR acquisition, such as Albright (2002), and of segmentation and UR acquisition (Johnson et al., 2015).

3.3 Candidate generation

For each tableau, candidates are generated from the input and constraint set. Each URC that makes reference to a morphosyntactic object in the input defines a possible UR for that object. Candidates are generated by combining every possible UR for each morphosyntactic object in the input. Tableau (5) illustrates the set of candidates that would be generated for the {M1}+{M2} input given the constraints that had been induced from the [abc] surface form in Table (5). For the sake of brevity the M1 preceding M2 order is assumed, cutting the number of constraints and candidates in half by eliminating all candidates that place the exponents of {M2} before that of {M1}. The actual model assumes no knowledge of the relative orderings of morphosyntactic objects, and the candidates with opposite correspondence relations would also be generated. Candidates shown in bold are consistent with the observed surface form [abc] and would not produce an error in training.

$\{M1\}_1$+$\{M2\}_2$	{M1}=a	{M1}=ab	{M2}=bc	{M2}=c
a. **a_1#bc_2**		-1		-1
b. **ab_1#c_2**	-1		-1	
c. a_1#c_2		-1	-1	
d. ab_1#bc_2	-1			-1

Tableau 5: Candidates and violations generated from the constraints in (5) - the ordering {M1} precedes {M2} is imposed for brevity

4 Test case: English plural

Voicing assimiliation of the English plural morpheme, by which the plural morpheme surfaces as [s] after a voiceless consonant and [z] after a voiced consonant or vowel, was used as a test case for the model. The model was tasked with segmenting utterences that contained either the definite or indef-

English Phrase	Input String	Input Morphemes
a dog	ədɔg	IND, DOG
a cat	əkæt	IND, CAT
the dog	ðədɔg	DEF, DOG
the cat	ðəkæt	DEF, CAT
the dogs	ðədɔgz	DEF, DOG, PLURAL
the cats	ðəkæts	DEF, CAT, PLURAL

Table 2: Training strings and corresponding sets of morphosyntactic objects for the English plural alternation

inite determiner (DEF/IND) and a singular or plural noun that ended with either a voiced or voiceless consonant. The complete set of inputs to the learner is listed in Table (2); sets of morphemes are unordered. The task of the learner then, is to learn the segmentation of the input strings and underlying representations for the definite and indefinite determiners, the roots DOG and CAT, and the plural. To make possible the learning of voicing assimilation, the constraints AGREE(VOICE), which assigns violations to adjacent consonants that do not share the same voicing specification, and IDENT(VOICE), which assigns violations to corresponding consonants in the UR and surface form that have different voicing specifications, are added to the constraint set. The candidate generation algorithm is expanded to have the ability to generate all IDENT(VOICE) violating candidates.

It is worth addressing the small size of this test language, especially in comparison to the corpora often used to train and test models of segmentation alone. While small for a model of segmentation, toy languages of similar size are often used to test models of phonological alternations (Tesar, 2006; Pater et al., 2012; Jarosz, 2016) and are justified by the complexity of the task. The constraint set increases linearly with the number of unique utterances in the language, but the number of candidates for any input increases exponentially with the size of the constraint set. The small test case was chosen to minimize the computational cost of evaluating an exponentially increasing number of candidates in each tableau and to ensure an interpretable output.

In all simulations the learner was able segment with near perfect accuracy. Table (3) shows the total probability assigned to correct segmentations for all six inputs after 1000 epochs with a learning rate of 0.1 and an initialization of 1.0 for all weights.

A segmentation was considered correct if a morpheme's phonological exponent was correctly identified as corresponding with that morpheme, even if the resulting phonological surface form was incorrect. For example, the probability that the grammar maps the input $\{\sqrt{\text{DOG}}\}_1$+$\{\text{PLURAL}\}_2$ to the phonological mapping /dɔg$_1$+s$_2$/ $\rightarrow$[dɔg$_1$#s$_2$] would be included in the total probability for a correct parse of *dogs* even though the phonological surface form is incorrect. However, the probabilities assigned to correct segmentations with incorrect surface forms were very small in all simulations and should make minimal difference to the total probability of correct segmentations.

Input String	Segmentation	Probability
ədɔg	ə#dɔg	0.962
əkæt	ə#kæt	0.959
ðədɔg	ðə#dɔg	0.947
ðəkæt	ðə#kæt	0.948
ðədɔgz	ðə#dɔg#z	0.954
ðəkæts	ðə#kæt#s	0.933

Table 3: Probability assigned to the correct segmentation of all phrases after training

The UR learning problem as given to the model has three solutions. There are two standard solutions in which there is a fixed underlying representation for the plural, either /s/ or /z/, and it either voices or devoices, violating IDENT, in order to satisfy AGREE. Given the data in Table (2) there is no reason to believe that /s/ or /z/ is a more likely UR for the plural, so the learner should reach these two solutions with equal likelihood. The third solution is UR selection, which is specific to the use of URCs, and involves choosing between the URs /s/ and /z/ to satisfy AGREE without violating IDENT. The data in (2) do not suggest that any one solution is preferable over another, so any solution is considered correct as long as it results in the desired outputs and segmentations.

In 100 simulations with all weights initialized at 1.0, the learner converged on a single voicing assimilation solution to the critical data points, *the cats* and *the dogs*, 51 times. A dominant solution is defined here as a solution in which there is a single candidate in both relevant tableau with a probability greater than 0.70. In 24 of these 51 solutions the plural was underlyingly voiceless and mapped unfaith-

	4.00	3.75	3.70	0.00		
{DOG}+{ PLURAL}	{PLURAL}=/z/	AGREE	{PLURAL}=/s/	IDENT	$\mathscr{H}$	p
/dɔg+z/→[dɔg#z]	0	0	-1	0	-3.70	0.57
/dɔg+s/→[dɔg#z]	-1	0	0	-1	-4.00	0.42
/dɔg+s/→[dɔg#s]	-1	-1	0	0	-7.75	0.01

Tableau 6: A final grammar with free variation between voicing assimilation and UR selection

fully to [+VOICE] after /dɔg/, in the remaining 27 the plural was underlyingly [+VOICE] and mapped unfaithully to [-VOICE] after /kæt/.

In the other 49 runs, UR selection was used to an extent, but there was no clear dominant solution. In these cases there was free variation between UR selection and voicing assimilation candidates which yielded the same phonological surface form. Because an error is defined as a mismatch between an observed surface form and a structureless version of the predicted surface form, the learner has no reason to select between two candidates with equivalent surface forms. An example of this type of solution is shown in Tableau (6). The data as presented in Table (2) do not favor voicing assimilation or UR selection, so it is expected that the learner converge on these kinds of ambiguous solutions frequently.

The effect of a data point that forced one specific solution to be preferred was tested by adding the vowel final word *eye* to the training data in the singular and plural. The plural form, *eyes*, surfaces as [aiz], taking a [+VOICE] plural morpheme with no possible AGREE violation. To the analyst this suggests that /-z/ is the underlying form of the plural morpheme and that voicing assimilation is responsible for the [s] that surfaces after voiceless consonants. In 100 more simulations identical to those described above but with the *eye(s)* data points added to the language, the learner now converged on voicing assimilation with a [+VOICE] UR 96 times. The remaining four final grammars represented ambiguous solutions similar to that shown in Tableau (6) and segmentation accuracy remained near ceiling.

Finally, to test the ability of the model to perform adult like parsing of novel words the trained grammar from one of the previous 100 simulations was used to make predictions about the segmentations of the previously unencountered surface forms [wuks] and [wugz] from the morphosyntactic elements {WUK} and {WUG} plus {PLURAL}. The probabilities of key candidate segmentations are shown in Table (4). Generated candidates which are not possible segmentations of the input string, such as [$\text{wug}_1\#\text{ugz}_2$] are not included in (4), but are all assigned near zero probabilities.

Input	Segmentation	Probability
$\{\text{WUG}\}_1, \{\text{PLURAL}\}_2$	$\text{w}_1\#\text{ugz}_2$	0.024
	$\text{wu}_1\#\text{gz}_2$	0.009
	$\text{wug}_1\#\text{z}_2$	0.967
$\{\text{WUK}\}_1, \{\text{PLURAL}\}_2$	$\text{w}_1\#\text{uks}_2$	0.061
	$\text{wu}_1\#\text{ks}_2$	0.061
	$\text{wuk}_1\#\text{s}_2$	0.878

Table 4: Probability of key segmentations of novel words suffixed with the plural morpheme

In these cases the model was able to correctly segment the novel words based solely on a high ranked constraint that the underlying form for the plural morpheme is /z/. In the [wuks] case, the probability of the correct segmentation is slightly hurt by the lack of a surface [z] but [s] here is a possible and likely phonological exponent of underlying /z/, making the correct segmentation drastically more likely than its competitors.

5 Discussion

When the *eye(s)* data points were included the training data were no longer agnostic towards the solution and the learner converged on the expected assimilation solution nearly all of the time. Recall that Smith (2015) stipulates that only suppletive forms can have multiple URCs in order to prevent the rampant use of UR selection rather than unfaithful phonological mappings. In this case, there were a large number of URCs for every word in the lexicon but the UR selection solution was reached only 4 out of 100 times. Consequently the restriction placed on URCs by Smith seems unnecessary. While there exists a solution to the dataset in which UR selection is responsible for every alternation, that solution ap-

pears strongly disfavored by the learner.

Assimilation represents a large portion of the space of possible weights compared to UR selection, making it easier for the learner to find. Setting aside extraneous UR constraints, the Hasse diagram in Figure (1) shows the necessary rankings for assimilation and UR selection. A direct line between two constraints means that the weight of the higher constraint must be greater than that of the lower one.

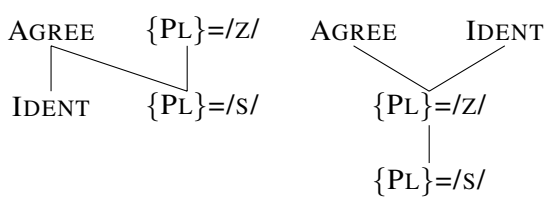

Figure 1: Ranking arguments for assimilation and UR selection

Randomly sampling one million sets of weights from the uniform distribution between 0 and 5, the range of the final weights of most simulations run above, the ranking arguments for assimilation are satisfied 14.68% of the time, and for UR selection only 3.86%. Assimilation occupies roughly 80% of the solution space. The model implemented here used no regularization term, but regularization will further decrease the likelihood of UR selection as the assimilation solution requires that two constraints have weights greater than 0 (AGREE and IDENT) where the UR selection solution requires three (AGREE, IDENT, and {PL}=/z/).

6 Conclusions

The acquisition of segmentation, underlying representations, and phonological alternations are treated here as parallel and interacting processes. The result is a model that succeeds in learning phonological alternations while also learning segmentation with near perfect accuracy, albeit on a very simple test case.

This model succeeds at segmentation for the same reason that the transitional probability and PARSER models work. A UR constraint that refers to a correct UR, such as {$\surd$DOG}=/dɔg/, will be reinforced by every observed output, regardless of the word's context. A UR constraint that refers to an 'incorrect' UR, such as {$\surd$DOG}=/dɔgz/, will be reinforced only by surface forms that result from one particular concatenation of morphemes. Because of transitional probability minima, the correct UR constraints will end up highly ranked. Like PARSER, this approach effectively tracks statistical trends in the data without the need to explicitly store them. Unlike PARSER, this model does so using a pre-existing phonological framework which allows for the incorporation of segmentation into a larger model of phonological learning.

This model relies on the strong assumption that the meaning of the utterance is known to the learner as a set of morphosyntactic objects. Consequently, this model cannot account for Saffran et al.'s (1996) result, in which participants were able to segment a language consisting only of nonce words. However, the Saffran et al. tasks are far removed from naturalistic language acquisition. Segmentation is not learned in isolation before the rest of acquisition. Information regarding segmentation, phonological processes, and underlying representations are made available to the learner simultaneously.

The assumption that the set of meanings are known to the learner greatly reduces the complexity of the segmentation task by providing the learner with the number of boundaries to be drawn, however this does not necessarily reduce the validity of the model. A slightly relaxed assumption, that infants have at least partial knowledge about the meaning of an utterance and are actively trying to identify correspondences between the phonological material and this partial meaning, does not seem empirically unsound. It is likely that infants are making use of contextual cues to make hypotheses about the semantic content of sentences from an early stage of learning, as evidenced by research showing that lexical representations are present as early as 6 months (Bergelson and Aslin, 2017). There is no reason that the infant needs to directly discover how many boundaries are in an utterance, they need only look for as many substrings as there are hypothesized meanings.

Beyond acquisition, this model captures the ability of adult speakers to segment novel words after a single exposure. Statistical models assume the minimum amount of linguistic knowledge of the learner, often relying only on representations of phonemes or syllables. This may be a sound assumption to make about infants in the earliest stages of acquisition, but

it fails to allow a mechanism for higher level linguistic information to be incorporated as it is acquired. The end state of the presented model represents a speaker that is able to make simultaneous use of lexical and phonological knowledge to segment novel forms.

References

Adam Albright. 2002. *The identification of bases in morphological paradigms*. Ph.D. thesis, University of California, Los Angeles.

Diana Apoussidou. 2007. *The learnability of metrical phonology*. Ph.D. thesis, University of Amsterdam.

Elika Bergelson and Richard N. Aslin. 2017. Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, 114(49).

Jean Berko Gleason. 1958. The child's learning of English morphology. *Word*, 14, 08.

Arthur Dempster, Natalie Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38.

Sarah Eisenstat. 2009. Learning underlying forms with MaxEnt. Master's thesis, Brown University.

Micha Elsner, Sharon Goldwater, Naomi H. Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 42–54.

Micha Elsner, Stephanie Antetomaso, and Naomi H. Feldman. 2016. Joint word segmentation and phonetic category induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 59–65.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation in Optimality Theory*, pages 111–120.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

Gaja Jarosz. 2013. Learning with hidden structure in optimality theory and Harmonic Grammar: beyond robust interpretive parsing. *Phonology*, 30(1):27–71.

Gaja Jarosz. 2015. Expectation driven learning of phonology. *Unpublished Manuscript*.

Gaja Jarosz. 2016. Learning opaque and transparent interactions in Harmonic Serialism. In *Proceedings of the 2015 Annual Meetings and Phonology, Vancouver BC*.

Gaja Jarosz. 2019. Computational modeling of phonological learning. *Annual Review of Linguistics*, 5:To appear.

Elizabeth K. Johnson and Peter W. Jusczyk. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567.

Mark Johnson, Joe Pater, Robert Staubs, and Emmanuel Dupoux. 2015. Sign constraints on feature weights improve a joint model of word segmentation and phonology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313. Association for Computational Linguistics.

René Kager. 2008. Lexical irregularity and the typology of contrast. In Kristin Hanson and Sharon Inkelas, editors, *The Nature of the Word: Studies in Honor of Paul Kiparsky*, pages 397–432. MIT Press.

Joe Pater, Robert Staubs, Karen Jesney, and Brian Smith. 2012. Learning probabilities over underlying representations. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*.

Pierre Perruchet and Annie Vinter. 1998. PARSER: A model for word segmentation. *Journal of Memory and Language*, 39:246–263.

Alan Prince and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing, Malden, MA.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996a. Statistical learning by 8-month-old infants. *Science*, 274.

Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996b. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35:606–621.

Brian Smith. 2015. *Phonologically conditioned allomorphy and UR constraints*. Ph.D. thesis, University of Massachusetts Amherst.

Bruce Tesar and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry*, 29:229–268.

Bruce Tesar. 2006. Faithful contrastive features in learning. *Cognitive Science*, 30:863–903.