Learning non-concatenative morphology

Michelle A. Fullwood
Dept. of Linguistics and Philosophy
Massachusetts Institute of Technology
maf@mit.edu

Timothy J. O'Donnell
Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology
timod@mit.edu

Abstract

Recent work in computational psycholinguistics shows that morpheme lexica can be acquired in an unsupervised manner from a corpus of words by selecting the lexicon that best balances productivity and reuse (e.g. Goldwater et al. (2009) and others). In this paper, we extend such work to the problem of acquiring non-concatenative morphology, proposing a simple model of morphology that can handle both concatenative and non-concatenative morphology and applying Bayesian inference on two datasets of Arabic and English verbs to acquire lexica. We show that our approach successfully extracts the non-contiguous triliteral root from Arabic verb stems.

1 Introduction

What are the basic structure-building operations that enable the creative use of language, and how do children exposed to a language acquire the inventory of primitive units which are used to form new expressions? In the case of word formation, recent work in computational psycholinguistics has shown how an inventory of morphemes can be acquired by selecting a lexicon that best balances the ability of individual sound sequences to combine productively against the reusability of those sequences (e.g., Brent (1999), Goldwater et al. (2009), Feldman et al. (2009), O'Donnell et al. (2011), Lee et al. (2011).) However, this work has focused almost exclusively on one kind of structure-building operation: concatenation. The languages of the world, however, exhibit a variety of other, non-concatenative word-formation processes (Spencer, 1991).

Famously, the predominant mode of Semitic word formation is non-concatenative. For example, the following Arabic words, all related to the concept of writing, share no contiguous sequences of segments (i.e., phones), but they do share a discontinuous subsequence $\sqrt{\text{ktb}}$, which has been traditionally analyzed as an independent morpheme, termed the "root".

kataba	"he wrote"
kutiba	"it was written"
yaktubu	"he writes"
ka:tib	"writer"
kita:b	"book"
kutub	"books"
maktab	"office"

Table 1: List of Arabic words with root $\sqrt{\text{ktb}}$

Many Arabic words appear to be constructed via a process of interleaving segments from different morphemes, as opposed to concatenation.

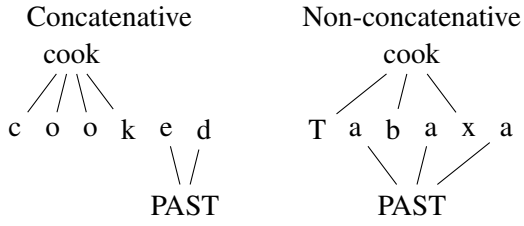

Figure 1: Schematic of concatenative vs non-concatenative morphology

Such non-concatenative morphology is pervasive in the world's languages. Even English, whose morphology is fundamentally concatenative, displays pockets of non-concatenative behavior, for example in the irregular past tenses (see Table 2).

In these words, the stem vowels undergo ablaut changing between tenses. This cannot be handled in a purely concatenative framework unless we consider these words listed exceptions. However, such irregulars do show limited productiv-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 21–27,
Sofia, Bulgaria, August 8, 2013.

bite /bajt/	bit /bɪt/
sing /sɪŋ/	sang /sæŋ/
give /gɪv/	gave /gejv/
feel /fil/	felt /fɛlt/

Table 2: Examples of English irregular verbs

ity (see Albright and Hayes (2003), Prasada and Pinker (1993), Bybee and Slobin (1982), Bybee and Moder (1983), Ambridge (2010)), and in other languages such stem changing processes are fully productive.

In Semitic, it is clear that non-concatenative word formation is productive. Borrowings from other languages are modified to fit the available non-concatenative templates. This has also been tested psycholinguistically: Berman (2003), for instance, shows that Hebrew-speaking preschoolers can productively form novel verbs out of nouns and adjectives, a process that requires the ability to extract roots and apply them to existing verbal templates.

Any model of word formation, therefore, needs to be capable of generalizing to both concatenative and non-concatenative morphological systems. In this paper, we propose a computational model of word formation which is capable of capturing both types of morphology, and explore its ramifications for morphological segmentation.

We apply Bayesian inference on a small corpus of Arabic and English words to learn the morphemes that comprise them, successfully learning the Arabic root with great accuracy, but less successfully English verbal inflectional suffixes. We then examine the shortcomings of the model and propose further directions.

2 Arabic Verbal Morphology

In this paper, we focus on Arabic verbal stem morphology. The Arabic verbal stem is built from the interleaving of a consonantal root and a vocalism that conveys voice (active/passive) and aspect (perfect/imperfect). The stem can then undergo further derivational prefixation or infixation. To this stem inflectional affixes indicating the subject's person, number and gender are then added. In the present work, we focus on stem morphology, leaving inflectional morphology to future extensions of the model.

There are nine common forms of the Arabic verbal stem, also known by the Hebrew grammatical term *binyan*. In Table 3, $\sqrt{\text{fʕl}}$ represents the triconsonantal root. Only the perfect forms are given.

Form	Active	Passive
I	faʕal	fuʕil
II	faʕʕal	fuʕʕil
III	faaʕal	fuuʕil
IV	ʔafʕal	ʔufʕil
V	tafaʕʕal	tufuʕʕil
VI	tafaaʕal	tufuuʕil
VII	ʔinfaʕal	-
VIII	ʔiftaʕal	ʔiftiʕil
X	ʔistafʕal	ʔistufʕil

Table 3: List of common Arabic verbal binyanim

Each of these forms has traditionally been associated with a particular semantics. For example, Form II verbs are generally causatives of Form I verbs, as is *kattab* "to cause to write" (c.f. *katab* "to write"). However, as is commonly the case with derivational morphology, these semantic associations are not completely regular: many forms have been lexicalized with alternative or more specific meanings.

2.1 Theoretical accounts

The traditional Arab grammarians' account of the Arabic verb was as follows: each form was associated with a template with slots labelled C_1, C_2 and C_3, traditionally represented with the consonants $\sqrt{\text{fʕl}}$, as described above. The actual root consonants were slotted into these gaps. Thus the template of the Form VIII active perfect verb stem was taC_1aC_2C_2aC_3. This, combined with the triconsonantal root, made up the verbal stem.

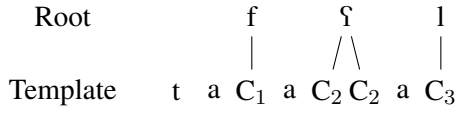

Figure 2: Traditional analysis of Arabic Form V verb

The first generative linguistic treatment of Arabic verbal morphology (McCarthy, 1979; McCarthy, 1981) adopted the notion of the root and template, but split off the derivational prefixes and infixes and vocalism from the template. Borrowing from the technology of autosegmental phonology (Goldsmith, 1976), the template was

now comprised of C(onsonant) and V(owel) slots. Rules governing the spreading of segments ensured that consonants and vowels appeared in the correct positions within a template.

Under McCarthy's model, the analysis for [tafaʔʕal] would be as follows:

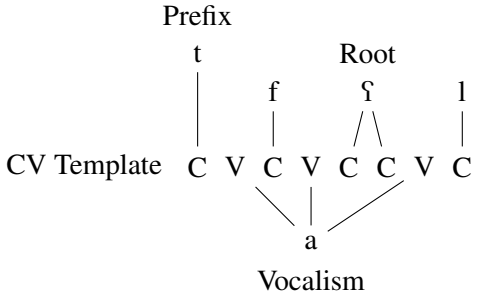

Figure 3: McCarthy analysis of Arabic Form V verb

While increasing the number of morphemes associated with each verb, the McCarthy approach economized on the variety of such units in the lexicon. The inventory of CV templates was limited; there were three vocalisms corresponding to active and passive voice intersecting with perfect and imperfect aspect; and only four derivational prefixes (/ʔ/,/n/,/t/,/st/), one of which became an infix via morphophonological rule in Form VIII.[1]

We adopt a middle ground between the traditional Arab grammarians' description of the verbal stem and McCarthy's analysis as our starting point. We describe this approach in the next section.

3 The Approach

Our initial model of morphology adopts McCarthy's notion of an abstract template, but coalesces the prefixes and infixes with the vocalism into what we term the "residue." Each stem is thus composed of two morphemes: the root and the residue, and their interleaving is dictated by a template with slots for root and residue segments.

For example, ʔiktatab = - - - r - - r - r (template) + ktb (root) + ʔitaa (residue), where r indicates a root segment and - a residue segment.

The residue may be of length 0, effectively making the word consist of a single morpheme. Concatenative morphology may be modelled in this framework by grouping all the root segments together, for example *cooked* [kukt] = r r r - (template) + kuk (root) + t (residue).

The template, root and residue are each drawn from a separate sub-lexicon, modeled using tools from Bayesian non-parametric statistics (see Section 4). These tools put a prior distribution on the lexica that biases them in favour of reusing existing frequent forms and small lexica by promoting maximal sharing of morphemes.

When applied to data, we derive a segmentation for each word into a root and a residue.

4 Model

Following earlier work on Bayesian lexicon learning (e.g. Goldwater et al. (2009), we use a distribution over lexical items known as the Pitman–Yor Process (PYP) (Pitman and Yor, 1995). Let G be a distribution over primitive phonological elements of the lexicon (e.g., words, roots, residues, templates, morphemes, etc.). The behavior of PYP process $\textsc{pyp}(a, b, G)$ with base measure G and parameters a and b can be described as follows. The first time we sample from $\textsc{pyp}(a, b, G)$ a new lexical item will be sampled using G. On subsequent samples from $\textsc{pyp}(a, b, G)$, we either reuse an existing lexical item i with probability $\frac{n_i - a}{N + b}$, where N is the number of lexical items sampled so far, n_i is the number of times that lexical item i has been used in the past, and $0 \leq a \leq 1$ and $b > -a$ are parameters of the model. Alternatively, we sample a new lexical item with probability $\frac{aK + b}{N + b}$, where K is the number of times a *new* lexical item was sampled in the past from the underlying distribution G. Notice that this process induces a rich-get-richer scheme for sampling from the process. The more a particular lexical item has been reused, the more likely it is to be reused in the future. The Pitman–Yor process also produces a bias towards smaller, more compact lexica.

In our model, we maintain three sublexica for templates (L_{Tp}), roots (L_{Rt}), and residues (L_{Rs}) each drawn from a Pitman–Yor process with its own hyperparameters.

$$L_X \sim \textsc{pyp}(a_{\mathrm{X}}, b_X, G_X) \qquad (1)$$

where $X \in \{\mathrm{Tp}, \mathrm{Rt}, \mathrm{Rs}\}$ Words are drawn by first drawing a template, then drawing a root and a residue (of the appropriate length) and inserting the segments from the root and residue in the appropriate positions in the word as indicated by the

[1] Other theories of Arabic morphology that reject the existence of the root are also extant in the literature; see e.g. (Bat-El, 1994) for a stem modification and vowel overwriting approach.

template. Our templates are strings in $\{\mathrm{Rt}, \mathrm{Rs}\}^*$ indicating for each position in a word whether that position is part of the word's root (Rt) or residue (Rs). These templates themselves are drawn from a base measure G_{Tp} which is defined as follows. To add a new template to the template lexicon first draw a length for that template, K, from a Poisson distribution.

$$K \sim \textsc{Poisson}(5) \tag{2}$$

We then sample a template of length K by drawing a Bernoulli random variable t_i for each position $i \in 1..K$ is a root or residue position.

$$t_i \sim \textsc{Bernoulli}(\theta) \tag{3}$$

The base measure over templates, G_{Tp}, is defined as the concatenation of the t_i's.

The base distributions over roots and residues, G_{Rt} and G_{Rs}, are drawn in the following manner. Having drawn a template, T we know the lengths of the root, K_{Rt}, and residue K_{Rt}. For each position in the root or residue r_i where $i \in 1..K_{\mathrm{Rt/Rs}}$, we sample a phone from a uniform distribution over phones.

$$r_i \sim \textsc{uniform}(|\text{alphabet}|) \tag{4}$$

5 Inference

Inference was performed via Metropolis–Hastings sampling. The sampler was initialized by assigning a random template to each word in the training corpus. The algorithm then sampled a new template, root, and residue for each word in the corpus in turn. The proposal distribution over templates for our sampler considered all templates currently in use by another word, as well as a randomly generated template from the prior. Samples from this proposal distribution were corrected into the true distribution using the Metropolis–Hastings criterion.

6 Related work

The approach of this paper builds on previous work on Bayesian lexicon learning starting with Goldwater et al. (2009). However, to our knowledge, this approach has not been applied to non-concatenative morphological segmentation. Where it has been applied to Arabic (e.g. Lee et al. (2011)), it has been applied to unvowelled text, since standard Arabic orthography drops short vowels. However, this has the effect of reducing the problem mostly to one of concatenative morphology.

Non-concatenative morphology has been approached computationally via other research, however. Kataja and Koskenniemi (1988) first showed that Semitic roots and patterns could be described using regular languages. This insight was subsequently computationally implemented using finite state methods by Beesley (1991) and others. Roark and Sproat (2007) present a model of both concatenative and non-concatenative morphology based on the operation of composition that is similar to the one we describe above.

The narrower problem of isolating roots from Semitic words, for instance as a precursor to information retrieval, has also received much attention. Existing approaches appear to be mostly rule-based or dictionary-based (see Al-Shawakfa et al. (2010) for a recent survey).

7 Experiments

We applied the morphological model and inference procedure described in Sections 4 and 5 to two datasets of Arabic and English.

7.1 Data

The Arabic corpus for this experiment consisted of verbal stems taken from the verb concordance of the Quranic Arabic Corpus (Dukes, 2011). All possible active, passive, perfect and imperfect fully-vowelled verbal stems for Forms I–X, excluding the relatively rare Form IX, were generated. We used this corpus rather than a lexicon as our starting point to obtain a list of relatively high frequency verbs.

This list of stems was then filtered in two ways: first, only triconsonantal "strong" roots were considered. The so-called "weak" roots of Arabic either include a vowel or semi-vowel, or a doubled consonant. These undergo segmental changes in various environments, which cannot be handled by our current generative model.

Secondly, the list was filtered through the Buckwalter stem lexicon (Buckwalter, 2002) to obtain only stems that were licit according to the Buckwalter morphological analyzer.

This process yielded 1563 verbal stems, comprising 427 unique roots, 26 residues, and 9 templates. The stems were supplied to the sampler in the Buckwalter transliteration.

The English corpus was constructed along similar lines. All verb forms related to the 299 most frequent lemmas in the Penn Treebank (Marcus et al., 1999) were used, excluding auxiliaries such as *might* or *should*. Each lemma thus had up to five verbal forms associated with it: the bare form (*forget*), the third person singular present (*forgets*), the gerund (*forgetting*), past tense (*forgot*), and past participle (*forgotten*).

This resulted in 1549 verbal forms, comprising 295 unique roots, 108 residues, and 55 templates. CELEX (Baayen et al., 1995) pronunciations for these words were supplied to the sampler in CELEX's DISC transliteration.

Deriving a gold standard analysis for English verbs was less straightforward than in the Arabic case. The following convention was used: The root was any subsequence of segments shared by all the forms related to the same lemma. Thus, for the example lemma of *forget*, the correct template, root and residue were deemed to be:

forget	f@gEt	r r r - r	f@gt	E
forgets	f@gEts	r r r - r -	f@gt	Es
forgot	f@gQt	r r r - r	f@gt	Q
forgetting	f@gEtIN	r r r - r - -	f@gt	EIN
forgotten	f@gQtH	r r r - r -	f@gt	QH

Table 4: Correct analyses under the root/residue model for the lemma *forget*

37 templates were concatenative, and 18 non-concatenative. The latter were necessary to accommodate 46 irregular lemmas associated with 254 forms.

7.2 Results and Discussion

We ran 10 instances of the sampler for 200 sweeps through the data. For the Arabic training set, this number of sweeps typically resulted in the sampler finding a local mode of the posterior, making few further changes to the state during longer runs. An identical experimental set-up was used for English. Evaluation was performed on the final state of each sampler instance.

The correctness of the sampler's output was measured in terms of the accuracy of the templates it predicted for each word. The word-level accuracy indicates the number of words that had their entire template correctly sampled, while the segment-level accuracy metric gives partial credit by considering the average number of correct bits

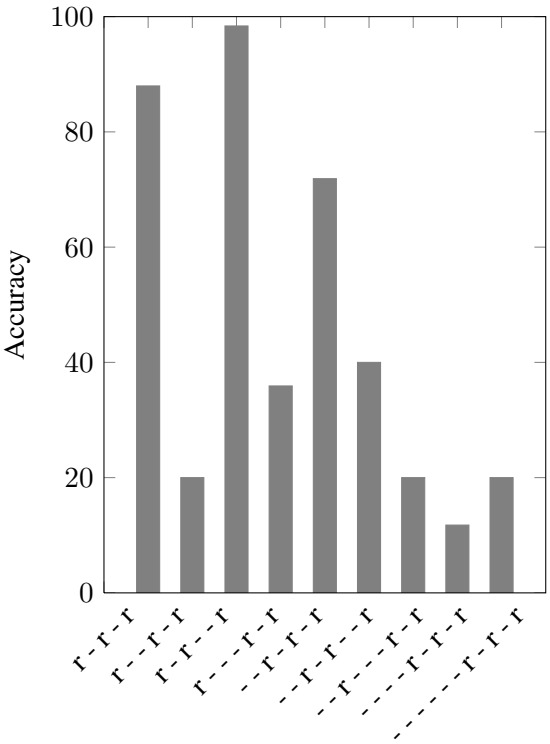

Figure 4: Unweighted accuracy with which each template was sampled

(r versus -) in each sampled template.

Table 5 shows the average accuracy of the 10 samples, weighted by each sample's joint probability.

Accuracy	Word-level	Segment-level
Arabic	92.3%	98.2%
English	43.9%	85.3%

Table 5: Average weighted accuracy of samples

Arabic Analyses Figure 4 shows the average unweighted accuracy with which each of the 9 Arabic templates was sampled.

Figure 4 reveals an effect of both the rarity and the length of each template. For instance, the performance on template r - - r - r (second bar from left) is exceptionally low, but this is the result of there being only one instance of this template in the training set: Euwqib, the passive form of the Form III verb of root Eqb, in the Buckwalter transliteration.[2] In addition, the longer the word,

[2]This is an artifact of Arabic orthography and the Buckwalter transliteration, which puts the active form EAqab with template r - r - r in correspondence with the passive template r - - r - r.

the poorer the performance of the model. This is likely the result of the difficulty of searching over the space of templates for longer forms. Since the number of potential templates increases exponentially with the length of the form, finding the correct template becomes increasingly difficult. This problem can likely be addressed in future models by adopting an analysis similar to McCarthy's whereby the residue is further subdivided into vocalism, prefixes and infixes. Note that even in such long forms, however, the letters belonging to the root were generally isolated in one of the two morphemes.

English Analyses The English experiment yielded poorer results than the Arabic dataset. The statistics of the datasets reveal the cause of the failure of the English model: the English dataset had several times more residues and templates than the Arabic dataset did, thus lacking as much uniform structure. Nevertheless, the relatively high segment-level accuracy shows that the model tended to find templates that were only incorrect in 1 or 2 positions.

The dominant pattern of errors was in the direction of overgeneralization of the concatenative templates to the irregular forms. Out of the 254 words related to a lemma with an irregular past form, 241 received incorrect templates, 232 of which were concatenative, often correctly splitting off the regular suffix where there was one. For example, *sing* and *singing* were parsed as *sing*+∅ and *sing*+*ing*, while *sung* was parsed as a separate root. Note that under an analysis of English irregulars as separate memorized lexical items, the sampler behaved correctly in such cases.

However, out of 1295 words related to perfectly regular lemmas, the sampler determined 628 templates incorrectly. Out of these, 325 were given concatenative templates, but with too much or too little segmental material allocated to the suffix. For example, the word *invert* was analyzed as *inver*+*t*, with its other forms following suit as *inver*+*ted*, *inver*+*ting* and *inver*+*ts*. This is likely due to subregularities in the word corpus: with many words ending with -t, this analysis becomes more attractive.

The remaining 303 regular verbs were given non-concatenative templates. For instance, *identify* was split up into *dfy* and *ienti*. No consistent pattern could be discerned from these cases.

8 Conclusion

We have proposed a model of morpheme-lexicon learning that is capable of handling concatenative and non-concatenative morphology up to the level of two morphemes. We have seen that Bayesian inference on this model with an Arabic dataset of verbal stems successfully learns the non-contiguous root and residue as morphemes.

In future work, we intend to extend our simplified model of morphology to McCarthy's complete model by adding concatenative prefixation and suffixation processes and segment-spreading rules. Besides being capable of handling the inflectional aspects of Arabic morphology, we anticipate that this extension will improve the performance of the model on Arabic verbal stems as well, since the number of non-concatenative templates that have to be learned will decrease. For example, the template for the Form V verb [tafaʕʕal] can be reduced to that for the Form II verb [faʕʕal] plus an additional prefix.

We also anticipate that the performance on English will be vastly improved, since the dominant mode of word formation in English is concatenative, while the small number of irregular past tenses and plurals that undergo ablaut can be handled using the non-concatenative architecture of the model. This would also be more in line with native speakers' intuitions and linguistic analyses of English morphology.

Acknowledgments

Parts of the sampler code were written by Peter Graff. We would also like to thank Adam Albright and audiences at the MIT Phonology Circle and the Northeast Computational Phonology Workshop (NECPhon) for feedback on this project. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374.

References

Emad Al-Shawakfa, Amer Al-Badarneh, Safwan Shatnawi, Khaleel Al-Rabab'ah, and Basel Bani-Ismail. 2010. A comparison study of some Arabic root finding algorithms. *Journal of the American Society for Information Science and Technology*, 61(5):1015–1024.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computa-

tional/experimental study. *Cognition*, 90(2):119–161.

Ben Ambridge. 2010. Children's judgments of regular and irregular novel past–tense forms: New data on the English past–tense debate. *Developmental Psychology*, In Press.

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Outi Bat-El. 1994. Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory*, 12:571–593.

Kenneth R. Beesley. 1991. Computer analysis of Arabic morphology: A two-level approach with detours. In Bernard Comrie and Mushira Eid, editors, *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, pages 155–172. John Benjamins. Read originally at the Third Annual Symposium on Arabic Linguistics, University of Utah, Salt Lake City, Utah, 3-4 March 1989.

Ruth A. Berman. 2003. Children's lexical innovations. In Joseph Shimron, editor, *Language Processing and Acquisition in Languages of Semitic, Root-based, Morphology*, pages 243–292. John Benjamins.

Michael R. Brent. 1999. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, 3(8):294–301, August.

Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0. Technical Report LDC2002L49, Linguistic Data Consortium.

Joan L. Bybee and Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language*, 59(2):251–270, June.

Joan L. Bybee and Daniel I. Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language*, 58(2):265–289.

Kais Dukes. 2011. Quranic Arabic Corpus. http://corpus.quran.com/.

Naomi H. Feldman, Thomas L. Griffiths, and James L. Morgan. 2009. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

John Anton Goldsmith. 1976. *Autosegmental Phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

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

Laura Kataja and Kimmo Koskenniemi. 1988. Finite-state description of Semitic morphology: a case study of Ancient Akkadian. In *Proceedings of the 12th conference on Computational linguistics - Volume 1*, COLING '88, pages 313–315, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2011. Modeling syntactic context improves morphological segmentation. In *Proceedings of the Conference on Natural Language Learning*.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank 3 technical report. Technical report, Linguistic Data Consortium, Philadelphia.

John J. McCarthy. 1979. *Formal Problems in Semitic Phonology and Morphology*. Ph.D. thesis, Massachusetts Institute of Technology.

John J. McCarthy. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12:373–418.

Timothy J. O'Donnell, Jesse Snedeker, Joshua B. Tenenbaum, and Noah D. Goodman. 2011. Productivity and reuse in language. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Jim Pitman and Marc Yor. 1995. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. Technical report, Department of Statistics University of California, Berkeley.

Sandeep Prasada and Steven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8(1):1–56.

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press.

Andrew Spencer. 1991. *Morphological Theory*. Blackwell.