Stochastic Approaches to Morphology Acquisition

Justin M. Aronoff, Nuria Giralt, and Toben H. Mintz
University of Southern California

1. Introduction

One of the first steps in acquiring a morphology system is discovering which phonetic strings correspond to morphemes. These phonetic strings can then be further analyzed in order to determine their grammatical privileges and contribution to meaning and thus to bootstrap into a functional morphology system. Discovering the relevant phonetic strings is a deceptively easy task. Morpheme discovery presents a number of difficulties that are above and beyond those that occur for the similar task of word discovery and segmentation. Although both require the segmenting of a continuous speech stream, word segmentation can take advantage of the fact that some words are spoken in isolation, and those words can be used to bootstrap into the segmentation of other words. Although this will work for some morphemes (many words are monomorphemic), grammatical morphemes are often bound in many languages, such as English and Spanish, and thus never heard in isolation. Additionally, there is no simple strategy that will universally work for breaking a word into its component morphemes. Although in many languages grammatical morphemes are either at the beginning or the end of a word, simply using an approach whereby the child assumes that the first or last syllable is a morpheme will only work if that assumption aligns with the language environment that the child is exposed to. Since affixing languages of the world can have (multiple) prefixes, suffixes, and infixes, such an approach is likely to fail. Additionally, acquiring grammatical morphemes is much like acquiring function words; unlike nouns, function words have little concrete semantic meaning, likely contributing to the difficulty in learning these types of words (Bird et al. 2001, Caselli et al. 1995, Gentner 1982, Morrison et al. 1997).

The search for morpheme forms does have the advantage that a given morpheme generally occurs within certain syntactic environments (e.g., the morpheme –ing in English generally occurs with verbs). Although it has been noted that morphology can help a child acquire syntax (Morgan et al. 1987), the reverse may also be true. The relationship between morphology and syntax could be beneficial both for discovering bound morphemes and for knowing which words a given bound morpheme can attach to. For instance, -ing might be more readily detected as a suffix when only examining verbs than when examining all words. Additionally, once a child has discovered that –ing can be applied to a particular verb, extending that ending only to other verbs will greatly reduce overgeneralization errors.

There is a long history of research for morphology discovery models (e.g., Brent & Cartwright 1996, Goldsmith 2001, Harris 1955). Many of these systems, such as that by Erjavec and Džeroski (2004) are not designed to model child language acquisition, but rather are designed for computational tasks such as parsing a database. Because we are interested in how children acquire morphological forms, only models of language learning will be discussed here.

In order to model acquisition of morphological forms by children, an automatic morphology discovery system must have the following characteristics. First, since morphemes must be acquired by the child (i.e., they are highly language specific and thus cannot be innate), any morphology discovery system must use a plausible learning mechanism. This entails not only using information available to the language learner, but also using mechanisms that children possess. Second, because morphemes can appear as (multiple) prefixes, suffixes, and infixes in affixing languages, any morpheme discovery system must have flexibility in terms of the position in the word where the morpheme occurs. Third, it must generate a robust list of morphemes which is minimally sufficient to allow the child to bootstrap into the rest of the morphological system. Finally, given that grammatical morphemes generally occur

 Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages, ed. Carol A. Klee and Timothy L. Face, 110-121. Somerville, MA: Cascadilla Proceedings Project.

with a large number of other morphemes (especially root morphemes) across various contexts, grammatical morphemes should play an especially important role in acquiring a morphological system.

A number of approaches have been put forth which meet some, although not all of the above criteria. Although most approaches do generate a substantial list of (grammatical) morphemes, they often either employ implausible learning mechanisms (especially prevalent are supervised learning mechanisms)[1] or lack the ability to detect bound morphemes in variable positions within a word. For example, one approach that puts forth a plausible learning mechanism but has limited morpheme position flexibility is Schone and Jurafsky's (2001). Their multi-step, automatic morphology discoverer first identifies common word beginnings and endings and posits that these are morphemes. It then compares semantic vectors and weights each potential morpheme by affix frequency. Syntactic information is also included as well as other computational techniques to uncover hidden relationships. Although this method is an effective method of discovering morphemes, it makes the assumption that affixes occur at the right or left edge of words, which is only true in a subset of affixing languages, and thus does not fulfill the requirement of morpheme position flexibility.

Other approaches, such as the Whole Word Morphologizer (Neuvel 2002, Neuvel & Fulop 2002), have attempted to handle both variation in morpheme position and learnability requirements, but have failed to generate a system that could be implemented on a large scale as well as ignoring the value of frequency information. The Whole Word Morphologizer searches through a lexicon for words that share either the first or last segments. From these pairs it discovers transformation rules (*ieve → *ief as in believe → belief, relieve → relief) that can then be applied to similar new words. Although this approach does appear to perform well, it is unclear how it would overcome the problems associated with a realistically-sized lexicon. The model was originally tested on a lexicon of 5000 words or less, which is smaller than an adult lexicon by at least an order of magnitude, a factor that will become problematic as the child moves from early word acquisition towards having an adult lexicon. The laborious process of comparing all words in the lexicon would suggest that new words would take an unreasonable amount of time to parse as each new word would need to be compared against the entire lexicon to determine which words it shares patterns with. Additionally, frequency information is largely ignored by this approach although frequency effects are quite prevalent in language processes (Bod et al. 2003).

Some models, such as that by Albright and Hayes (2002) have dealt with the problems associated with a large lexicon, but at the expense of the learnability requirements. Their model requires pairs of morphologically related words in order to learn the underlying relationships. Although this approach does capture the morphological forms of regular and irregular words as well as novel words, its training requirements make it a poor model of morphology acquisition. Given that this approach assumes that the learner already knows a fair amount of the morphological system, it is unrealistic to assume that beginning learners could make use of this system.

Finally, there are models which attempt to have both morpheme position flexibility and learnability, but, because the learnability constraints are based on typological trends rather than strategies demonstrated to be used by language learners, they employ strategies which are more plausible for language researchers than language learners. Baroni (2003) puts forth one such model based on a distribution-driven method for morpheme discovery. Its primary assumption, also adopted here, is that frequent sound sequences tend to be morphemes. There is strong evidence that children and adults attend to frequent patterns both in speech-like stimuli (e.g., Jusczyk et al. 1994, Mattys et al. 1999, Mintz 2002, Saffran et al. 1996) and in other modalities (e.g., Fiser & Aslin 2002). However, Baroni makes two additional assumptions that there is little evidence that naïve learners would actually make. One assumption is that low frequency words are more likely to be morphologically complex. Another is that extracting longer strings is preferred. In the system proposed here, string length is largely ignored. This may be particularly relevant as frequent sequences such as grammatical morphemes tend to be short and thus extracting shorter rather than longer strings will likely yield more accurate morpheme extraction.

[1] This means that the system is presented with parsed morphemes during a training phase and then tested on unparsed morphemes. Although this is an effective way of training programs to detect morphemes, its reliance on a pre-parsed input means that it is an unrealistic model for child language acquisition.

As can be seen, past models of morpheme discovery have problematic mechanisms either in terms of their flexibility for morpheme positions within words, and/or in employing learning mechanisms which are plausible to be used by language learners. The automatic morpheme discovery system put forth here aims to employ both morpheme position flexibility and plausible learning mechanisms.

Because the system does not restrict its search to edge-adjoined sequences (although it does keep track of this positional information), it has a high degree of positional flexibility. An important characteristic is that it takes advantage of the productivity of morphemes (especially grammatical morphemes) and language learners' tendency to search for and respond to frequent patterns in the input (Mintz 2002, Saffran et al. 1996). Given that productive morphemes will likely be highly frequent across a corpus, the morphology discovery system searches for frequently occurring sound sequences and posits that they are morphemes. This is critically different than the strategy put forth by Baroni (2003) which is based in part on the assumption that low frequency words are likely to be morphologically complex.

In order to determine whether the input that children receive has the properties that would be required for the proposed morpheme discovery system, we have devised a series of analyses, carried out across a number of corpora. The first analysis we conducted was aimed at determining if the proposed relationship between frequently occurring sound sequences and morphemes--on which the proposed morpheme discovery technique is predicated--exists. To do this we first analyzed sound sequences across the corpora and determined if the most frequent ones were generally morphemes. In our second analysis we determined whether adding distributionally induced syntactic information increased accuracy of the system. Finally, to examine whether changes in accuracy between the two previous analyses were a result of the added syntactic information *per se*, or resulted from special properties of the distributional induction procedure, we compared the results of that condition with one using experimenter-supplied syntactic categories in a third analysis.

2. Analysis 1: Simple frequency analysis

2.1 Data

Data was obtained from CHILDES transcripts of six children acquiring Spanish as a native language (López Ornat 1994, MacWhinney 2000, Montes 1987, 1992, Romero et al. 1992). Because of the morphological richness of Spanish, as well as its high rate of cliticization (increasing the need for positional flexibility), we have chosen to use this language as a testing ground for our proposed morphology discovery system. These children ranged in age from approximately one to five years of age. Because we are interested in the amount of information that is available in the input to the child, only child-directed speech was analyzed. See Table 1 for information regarding these corpora.

The orthographically encoded transcripts were automatically transcribed into a broad phonetic representation using a series of grapheme to phoneme rules specific for Spanish. Because accents in Spanish represent a change in the acoustical structure of a word that cannot be predicted from the general stress structure of the language, accents were maintained for all words where the accent represents a change in the pronunciation of a phoneme string (i.e., the accent was not included for words such as *él* 'he', where the accent serves to differentiate the word from its homonym *el* 'the' in the orthographic domain, but has no effect on pronunciation).

Corpus	Age range of child	Number of child-directed utterances
Irene	0;11 – 3;2	16,938
Linaza	1;7 – 4;11	3,155
Montes	1;7 - 2;11	4,536
Ornat	1;7 - 3;10	9,419
Romero	2;0 – 2;0	1,030
Vila	0;11 – 4;8	11,453
Total	**0;11 – 4;11**	**46,531**

Table 1: Age range of child and number of child-directed utterances for each corpus

2.2 Procedures

In order to determine which sound sequences frequently occurred within and across the corpora, each word was automatically transcribed phonetically and then broken into all of the possible combinations of contiguous sound sequences. Additionally, for sound sequences that occurred at the left or right edge of a word, word boundary information was included (indicated by the symbol #).[2] Only sequences that contained at least two phonemes or a phoneme and a word boundary were included, since co-occurrence appears to be a significant factor in learning (e.g., Mintz 2002). Figure 1 provides a schematic of this procedure. As each word was broken into its set of contiguous sequences, the sound sequences were added to a list that compared each segment to all of those found in other words. The frequency across all words for each sound segment was then recorded. For all analyses, if two sequences had the same frequency and one was a subset of the other, then only the larger segment was analyzed.

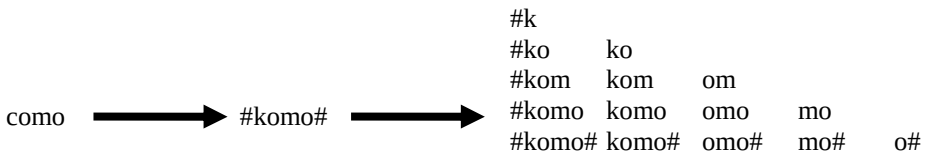

Figure 1: Example of extracting the sound sequences for one word; here the word *como* results in 15 different sequences.

The purpose of the first analysis was to determine whether or not simply assuming that frequent sound sequences are morphemes would result in reasonably accurate morpheme discovery. In order to do this, all sound sequences were ranked by frequency. Given that all words are morphemes or sequences of morphemes, and given that the lowest frequency of any attested sound sequence cannot be less than the lowest frequency of any word, one is confronted with the problem of determining what frequency threshold to use for positing sound sequences as potential morphemes. There are two general types of thresholds that could be used: data-independent and data-dependent.

Data-independent thresholds rely on absolute percentages or rankings, for example, treating the 100 most frequent or top 20% of sequences as morphemes. Although this type of thresholding technique may be sufficient in many applications, ideally one might prefer a system where the threshold is not generically set but is set separately for each data set based on the particular structure of that data. For example, a data-dependent method might make use of jumps in frequency that occur in the data. Figure 2 shows a schematic representation of the frequency distribution of sound sequences in the present data. One can clearly see a division between highly frequent items above a gap in the frequency dimension, and less frequent items below the gap. In cases where such a natural division exists, using a threshold that takes advantage of this information may not only be helpful, but may explain how a child could determine what threshold to use. In this and subsequent analyses we adopt a method for determining this threshold that we call *normalized max derivative*. Using this method we rank the sound sequences by frequency and calculate the frequency difference between each adjacent pair (i.e., we take the derivative of the frequency). This is then normalized based on the frequency of the less frequent member of the pair. We calculate this frequency change for all adjacent pairs where both members have a frequency higher than three. The normalized maximum derivative is indicated by the adjacent pair with the greatest normalized frequency change (i.e., derivative) between the two items. At this point, all sound sequences whose frequency is equal to or greater than that of the more

[2] It has been shown that there is sufficient statistical information in the input to find word divisions (e.g., Christiansen et al. 1998) and that very young children are able to segment words based on statistical properties found in languages (Jusczyk et al. 1994, Mattys et al. 1999, Saffran et al. 1996).

frequent of the two items in the pair with the normalized maximum derivative are considered to be above threshold and thus analyzed.

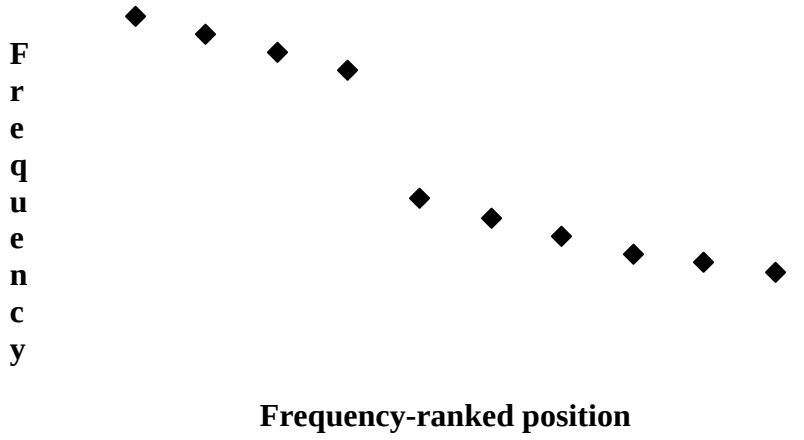

Figure 2: Schematic representation of data. Note the large drop in frequency between the 4th and 5th data points.

For all analyses, a type and token accuracy score was calculated. Each supra-threshold sound sequence was examined for all the words in the data set that it occurred in. For type accuracy, it was determined whether that particular sound sequence ever occurred as a morpheme (regardless of whether it sometimes occurred as a non-morpheme). If it did, the sound sequence was considered a morpheme, if not it was considered a non-morpheme. The type accuracy was then calculated as the number of morphemes over the number of unique sound sequences. For token accuracy, each occurrence of each sound sequence in the data set was analyzed and it was determined if the sound sequence was a morpheme for each word token it occurred in. Thus, if a sound sequence was a morpheme in two of the word tokens that it occurred in but it was not a morpheme in three other word tokens that it occurred in, that sound sequence would receive a score of two morpheme instances and three non-morpheme instances. Token accuracy was then calculated as the number of morpheme instances over the total number of morpheme and non-morpheme instances.

2.3 Results

Using the simple frequency analysis procedure yielded a type accuracy of 88% and an estimated[3] token accuracy of 45%. Additionally, it yielded seven correct morpheme types and one incorrect morpheme type. These are shown in Table 2, transcribed here and in all further examples back into orthographic form when possible for ease of reading.

These results indicate that the proposed morpheme extraction procedure can result in high type accuracy and moderate token accuracy, although it does not result in a large set of morphemes. To determine if this method could be improved by including syntactic information, we conducted a second analysis.

[3] This was estimated based on all words with a frequency of at least 50. Given the thousands of proto-morpheme types generated by our approach, this method was adopted because of the presence of extremely frequent words in the analysis set, which largely determine the overall score. To determine the upper bound on the accuracy score, a second analysis was conducted where the target sound sequence in all uncoded instances was assumed to be a morpheme. In this generous analysis, there was only a moderate increase in accuracy to 54%.

Correct Morpheme Types	Incorrect Morpheme Types
-o#	/k/
-a#	
-e#	
-s#	
#a	
#e	
es	

Table 2: Correct and incorrect morphemes generated by the simple frequency analysis

3. Analysis 2: Including syntactic information derived from distributional information

3.1 Procedures

The data and procedures for Analysis 2 were identical to those for Analysis 1 with the exception that syntactic information was included. As mentioned in the introduction, there is a relationship between morphology and syntax such that morphemes (and generally morphological forms) are limited to one particular syntactic category (e.g., the –ing morpheme is limited to verbs in English). Given this relationship, we were interested in determining whether a procedure that could use grammatical category information—effectively maintaining separate frequency counts and thresholds for each grammatical category—would perform better than the system in Analysis 1. Rather than explicitly providing category information, we adopted a method that uses distributional information in the input to build representations that approximate grammatical categories. The specific approach we adopted is one developed by Mintz (2003), which we herein refer to as the *frames* approach.

The frames approach hypothesizes that the words immediately preceding and following a target word (known as its *frame*) provides useful information regarding the syntactic category of that word. For example, in the sentence, "Who wants some ice-cream?", *who__some* is a frame containing the verb *wants*; the frames approach postulates that frames that occur frequently in a corpus of speech occur with words of the same grammatical category in the position intervening between the frame words. This approach does not depend on any syntactic or semantic knowledge about the frame words themselves, nor does it make any predictions about these. Instead, it merely takes the frame words as a frequent environment and groups words that occur within the same frame, without regard to the syntactic or semantic characteristics of the environment.

To determine whether prior categorization of words into frame-based categories would improve the detection of morphemes, we first categorized words within the 150 most frequent frames. We then selected for further analysis the frame-based categories for which there was a clear data-dependent threshold (normalized max derivative greater than 1) when the morphological analysis was applied to tokens within a frame. We then carried out our morphology extraction procedure separately on each frame-based category.

To verify that using frames did in fact provide syntactic information, we labeled each word in a given frame and determined the amount of syntactic agreement within the frame. This was done by labeling each word with its syntactic category and comparing all possible pairs of words within a given frame. If two words had the same syntactic category the pair was labeled as a "hit," if not, it was labeled as a "false alarm." This was done across all frames and the total number of hits and false alarms across all categories was recorded. Accuracy was calculated as hits / (hits + false alarms) (Mintz 2003). This resulted in an accuracy of 66%.

In order to determine whether the accuracy obtained by using the frames approach was above that which would be obtained by randomly grouping words, chance accuracy was calculated. Each word in the previous analysis was replaced by a place holder. The place holders were then replaced by a word randomly sampled (with replacement) from those that were previously in any of the frames. This resulted in frames with the identical number of tokens, but all tokens were randomly selected.

Additionally, because each newly assigned token was selected from a list of all tokens that previously occurred in any of the frames, tokens that were frequent in the original accuracy analysis were also frequent in this analysis, although the frequency of an item within a frame was not maintained. Accuracy was then calculated following the same procedure described above. This process of reshuffling and accuracy calculation was repeated ten times. The average of all the chance accuracy scores was then calculated and found to be 27% (range: 27% - 28%). Thus, the frames approach was considerably more accurate than chance at deriving proto-syntactic categories.

3.2 Results

In terms of type accuracy, using distributionally derived syntactic categories did not improve the accuracy score over the simple frequency analysis (76% when using frame-based categories vs. 88% for the simple frequency analysis). However, there was a dramatic increase in the token accuracy score up from 45% to 67%, an increase of almost 50%. A qualitative analysis also revealed an impressive increase in the number of morphemes found, including an almost doubling of the number of grammatical morphemes. See Tables 3-6 (compare to Table 2).

	Type Accuracy	Token Accuracy	Morpheme Types discovered	Non-morpheme Types
Analysis 1	88%	45%	6	1
Analysis 2	76%	67%	29	13

Table 3: Comparison of results from Analyses 1 and 2

-a#
-as#
-amos#
-e#
-er#
-o#
-s#

Table 4: Grammatical morphemes discovered when including distributionally derived syntactic information

#cre
#est
#l
#v
ve

Table 5: Root morphemes discovered when including distributionally derived syntactic information

#a#	*#está#*	*#se#*
#casa#	*#has#*	*#va*
#creo#	*#lo#*	*#vamos#*
#en#	*#no#*	*#vas#*
#es#	*#que#*	*#ver#*
#esta#	*#quien#*	

Table 6: Whole words discovered when including distributionally derived syntactic information

The results indicate that including distributionally derived syntactic information does result in improved morpheme extraction both in terms of token accuracy and in terms of the quantity of morphemes that are discovered, especially the critical grammatical morphemes. This raises the question of whether the improvement over using the simple frequency analysis is the result of the added syntactic knowledge or if the frames analysis provides an additional type of information which

affects the morpheme extraction procedure. To test this, we conducted a third analysis using categories supplied by one of the experimenters.

4. Analysis 3: Including experimenter-supplied category information

4.1 Procedures

The data and procedures for Analysis 3 were identical to those for Analysis 2 with the exception of how syntactic category information was determined. In order to test whether the effects of using distributionally derived syntactic information extend beyond that of simply using syntactic information, the highest frequency words were categorized by a native speaker as being either a noun or a verb (words from other syntactic categories such as adjectives and adverbs were not included in this analysis because of their relative low frequency of occurrence). Using this procedure, the 200 most frequent noun forms and the 200 most frequent verb forms, collapsed across all the corpora, were placed into two separate groups based on their syntactic category (note that 11 word forms, which can be both nouns and verbs, were placed in both categories). Selection was based solely on frequency and syntactic category membership, and as a result, nouns or verbs could occur multiple times in different forms (e.g., singular and plural). The morpheme extraction procedure was then applied to the words in each category separately.

4.2 Results

Type accuracy for the noun category was 100%, a considerably better type accuracy than for either previous analysis, and 50% for the verb category. Token accuracy also varied dramatically across the two categories. Token accuracy for nouns was 87%, considerably higher than for either of the previous analyses. In contrast, token accuracy for verbs was only 36%, below that found for both previous analyses. Qualitatively, there is a striking difference in terms of the actual morphemes that were discovered using this analysis and the previous analyses. Both the noun and the verb categories generated only two potential morphemes each (*-o#* and *–a#* for nouns, and *es* and the incorrect *–s#*[4] for verbs). See Table 7 for a summary of all three analyses.

Given that the threshold in this analysis (as well as the first analysis) was based on the largest normalized max derivative, which was numerically smaller than the threshold for the frames analysis, it is possible that the threshold is simply too high and that a lower threshold would demonstrate a high degree of accuracy along with a large number of correct morphemes. To test this, the threshold for verbs was lowered to allow the same number of correct morphemes as was found using the syntactic categories based on frames (29). This resulted in a small increase in token accuracy from 36% to 41%. One possible objection to this measure is that not all of the morphemes found using the frames approach are from verbs. Limiting the reanalysis to include only as many correct morphemes as verb morphemes found with the frames approach (19) yields a similar token accuracy, 40%. Thus, the low performance of the verb category is unlikely to be a result of the threshold being too high. With nouns, it is even less likely that the limited number of morpheme types is a result of the threshold being set too high given that, within the most frequent 50 sound segment types, there are only 6 actual morpheme types.

Combining the results across syntactic category indicates that the type accuracy, token accuracy, and, most strikingly, the number of morpheme types discovered using experimenter-supplied syntactic information are all lower than when using distributionally derived syntactic information (see Table 7). These results suggest that using the frames approach provides a benefit above and beyond that which would result simply from having additional syntactic information. This also presents an interesting question. In what way do noun and verb categories differ such that performance is so much better for nouns in terms of accuracy? It also invites the question of whether what caused the improved

[4] Under some analyses *-s#* could be considered a verb morpheme, but categorizing it as such would assume that the root for a word such as hablar is habla- rather than habl- which is a fairly liberal assumption, not made in the previous analyses. If it were included here, the token accuracy would be 63%, still lower than when using syntactic categories derived from distributional information.

performance on the noun category also caused the improved performance resulting from using the frames approach.

	Type Accuracy	Token Accuracy	Morpheme Types Discovered	Non-morpheme Types
Analysis 1	88%	45%	6	1
Analysis 2	76%	67%	29	13
Analysis 3 (combined)	75%	62%	3	1
Analysis 3 (nouns)	100%	87%	2	0
Analysis 3 (verbs)	50%	36%	1	1

Table 7: Summary of results from all three analyses

5. Homogenous environments

An analysis of the noun and verb categories reveals a striking difference in terms of the number of available grammatical morphemes. Although there are a number of exceptions, nouns generally have at least one of three frequent grammatical morphemes; verbs, on the other hand, have a much larger set (see Table 8). The result of the small set of grammatical morphemes for nouns is that there is less competition among the morphemes resulting in a higher likelihood of a small number of morphemes occurring with a much higher frequency than anything else, rather than a cascade of frequencies running across a large set of actual morphemes and intermixing with the background noise. This explanation is further supported by the results when the threshold was lowered for verbs: although additional morphemes were found, nearly an equal amount of non-morphemes were also posited to be morphemes.

Nouns	Verbs	
-o	AR	ER/IR
-a	-o	-o
-s	-as	-es
	-a	-e
	-amos	-emos
	-ais	-eis
	-an	-en

Table 8: Grammatical morphemes for nouns and present tense verbs

Examining the results of using the frames approach suggests that a similar mechanism explains its improvement over Analysis 1. A number of the frames, such as *el*_x_*de* 'the (masc., sing.) x from' and *a_x_a* 'to x to' generally only allow words containing a small set of grammatical morphemes. Thus, much like with the noun category in Analysis 3, there is a high likelihood that those morphemes will be far more frequent than any other sound sequence. Unlike with categories provided by the experimenter, the number of potentially homogenous environments is considerably larger when using distributionally derived syntactic information. For instance, the frames *el* x *de*, *la* x *de*, and *los* x *de* 'the (masc. sing. / fem. sing. / masc. plur.) x from / of ' are all sufficient for creating homogenous environments. More importantly, each of these frames creates a homogenous environment that is advantageous for a different morpheme (i.e., *–o*, *-a*, and *–s*, respectively). Because there are a number of frames which provide limited but differently biased environments (e.g., one biased for singular masculine nouns and another for plural masculine nouns), using the frames approach can discover more grammatical morphemes than using experimenter-provided categories. The proliferation of

environments when using frames also explains why Analysis 2 resulted in so many falsely-posited morpheme types while performing so well in terms of token accuracy.

6. Conclusions and discussion

In summary, the first analysis, using a simple frequency approach, demonstrated that searching for frequent sound sequences is a plausible mechanism to start acquiring morphological forms. Using this method, the automatic morpheme discoverer was able to find a number of morphemes with moderate accuracy. The second analysis, using distributionally derived syntactic categories, indicated that syntactic information can improve the frequent sound sequence approach to morpheme extraction by providing more morpheme types, including a larger number of grammatical morphemes, along with much greater token accuracy. The third analysis, using experimenter-supplied syntactic categories, suggested that, although syntactic information is helpful for extracting morphological forms, creating multiple local syntactic environments improves performance considerably over using simple noun/verb level categories.

This last point is especially notable because of the conceptual similarities between the current approach and that used by others, particularly the Minimal Generalization (Albright & Hayes 2002) and Whole Word Morphologizer (Neuvel 2002, Neuvel & Fulop 2002) approaches. Both of these approaches create local environments in which to search for morphological forms. The Minimal Generalization approach restricts its morpheme search to pairs of morphologically-related words and the Whole Word Morphologizer restricts its morpheme search to pairs of left- or right-aligned words. Although the overarching concept of creating local environments is shared by all three approaches, there are critical differences between them. These are discussed below.

Both the Minimal Generalization and the Whole Word Morphologizer rely on mechanisms to create the local environments which are problematic from an acquisition perspective. The Minimal Generalization approach would require the child to know the morphology system in order to learn the morphology system, which may be useful for later morpheme acquisition, but provides no manner to start acquiring morphology. The Whole Word Morphologizer requires a search across the entire lexicon, which would suggest that children would take longer to segregate the morphemes in a new word later in acquisition than earlier in acquisition. Additionally, it bases its analysis on finding pairs of words which contain overlapping strings. This means that it is primarily focused on finding transformations from one form of a word to another (e.g., converting from present to past tense). Because it uses the similarities within the pairs it finds to be morphologically related, the generalizability of morphemes that are found is limited to familiar phonological environments, which seems to be in contrast with children's tendency to overgeneralize regular morphemes (e.g., Cazden 1968, Marchman et al. 1997, Marcus 1995). Additionally it fails to take into account token or type frequency, thus ignoring the effect of frequency on learning. In contrast, the current approach starts with the assumption that the child does not have any knowledge of the morphological system they are learning. It then takes into account the important role of frequency and allows for the application of new morphemes within a restricted syntactic category (i.e., within a frame).

It should be noted, however, that the method presented here does produce a number of errors, which would need to be avoided by the child in production, as well as missing a number of forms. Although the missing forms may be found with a larger corpus, the incorrectly identified morphemes pose a greater problem and future research will be needed to resolve this, likely by examining the stability of proto-morphemes across environments.

In conclusion, the frequent sound sequence approach, when combined with the frames approach, provides a highly plausible model of the beginning stages of morpheme acquisition. Given the important role of homogenous environments, one can speculate that this method would be effective cross-linguistically, and thus could represent a general approach used by children acquiring a wide variety of affixing languages.

Acknowledgements

This research was supported in part by a grant from the National Institutes of Health (HD040368) to Toben Mintz.

References

Albright, Adam, and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. Proceedings of the sixth meeting of the ACL special interest group in computational phonology, ed. by Michael Maxwell, 58-69. Philadelphia: ACL.

Baroni, Marco. 2003. Distribution-driven morpheme discovery: a computational/experimental study. Yearbook of morphology, ed. by Geert Booij and Jaap van Marle, 213-48. Dordrecht: Springer.

Bird, Helen; Sue Franklin; and David Howard. 2001. Age of acquisition and imageability ratings for a large set of words, including verbs and function words. Behavior Research Methods, Instruments, & Computers 33.73-9.

Bod, Rens; Jennifer Hay; and Stefanie Jannedy (eds.) 2003. Probabilistic linguistics. Cambridge, MA: MIT Press.

Brent, Michael, and Timothy Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. Cognition 61.93-125.

Caselli, Maria Cristina; Elizabeth Bates; Paola Casadio; Judi Fenson; Larry Fenson; Lisa Sanderl; and Judy Weir. 1995. A cross-linguistic study of early lexical development. Cognitive Development 10.159-99.

Cazden, Courtney B. 1968. The acquisition of noun and verb inflections. Child Development 39.433-48.

Christiansen, Morten H.; Joseph Allen; and Mark S. 1998. Learning to segment speech using multiple cues: a connectionist model. Language and Cognitive Processes 13.221-68.

Erjavec, Tomaž, and Sašo Džeroski. 2004. Machine learning of morphosyntactic structure: lemmatizing unknown Slovene words. Applied Artificial Intelligence 19.17-41.

Fiser, József, and Richard N. Aslin. 2002. Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences 99.15822-6.

Gentner, Dedre. 1982. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. Language development: language, thought and culture, ed. by Stan I. Kuczaj, 301-34. Hillsdale, NJ: Erlbaum.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. Computational Linguistics 27.153-98.

Harris, Zellig. 1955. From phoneme to morpheme. Language 31.190-222.

Jusczyk, Peter W.; Paul A. Luce; and Jan Charles-Luce. 1994. Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language 33.630-45.

López Ornat, Susana. 1994. La adquisición de la lengua española. Madrid: Siglo XXI.

MacWhinney, Brian. 2000. The CHILDES project: tools for analyzing talk. Vol. 2: the database, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Marchman, Virginia A.; Kim Plunkett; and Judith Goodman. 1997. Overregularization in English plural and past tense inflectional morphology: a response to Marcus (1995). Journal of Child Language 24.767-79.

Marcus, Gary F. 1995. Children's overregularization of English plurals: a quantitative analysis. Journal of Child Language 22.447-59.

Mattys, Sven; Peter W. Jusczyk; Paul A. Luce; and James L. Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. Cognitive Psychology 38.465-94.

Mintz, Toben H. 2002. Category induction from distributional cues in an artificial language. Memory and Cognition 30.678-86.

Mintz, Toben H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. Cognition 90.91-117.

Montes, Rosa. 1987. Secuencias de clarificación en conversaciones con niños. Morphe 3-4.167-84.

Montes, Rosa G. 1992. Achieving understanding: repair mechanisms in mother–child conversations. Washington, D.C.: Georgetown University dissertation.

Morgan, James L.; Richard P. Meier; and Elissa L. Newport. 1987. Structural packaging in the input to language learning: contributions of prosodic and morphological marking of phrases to the acquisition of language. Cognitive Psychology 19.498-550.

Morrison, Catriona M.; Tameron D. Chappell; and Andrew W. Ellis. 1997. Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. Quarterly Journal of Experimental Psychology 50A.528-59.

Neuvel, Sylvain. 2002. Whole Word Morphologizer. Expanding the word-based lexicon: a nonstochastic computational approach. Brain and Language 81.454-63.

Neuvel, Sylvain, and Sean A. Fulop. 2002. Unsupervised learning of morphology without morphemes. Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology, ed. by Michael Maxwell, 31-40. Philadelphia: ACL.

Romero, S.; A. Santos; and D. Pellicer. 1992. The construction of communicative competence in Mexican Spanish speaking children (6 months to 7 years). Mexico City: University of the Americas.

Saffran, Jenny R.; Richard N. Aslin; and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. Science 274.1926-8.

Schone, Patrick, and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. 2nd meeting of the North American Chapter of the Association for Computational Linguistics: proceedings of the conference, June 2-7, 2001, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 183-91. San Francisco: ACL.

Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages

edited by Carol A. Klee and Timothy L. Face

Cascadilla Proceedings Project Somerville, MA 2006

ISBN 1-57473-409-1 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Aronoff, Justin M., Nuria Giralt, and Toben H. Mintz. 2006. Stochastic Approaches to Morphology Acquisition. In *Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, ed. Carol A. Klee and Timothy L. Face, 110-121. Somerville, MA: Cascadilla Proceedings Project.

or:

Aronoff, Justin M., Nuria Giralt, and Toben H. Mintz. 2006. Stochastic Approaches to Morphology Acquisition. In *Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, ed. Carol A. Klee and Timothy L. Face, 110-121. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1279.