Morpheme Segmentation from Distributional Information

Sara Finley and Elissa L. Newport[1]
University of Rochester

1. Introduction

Morphology is the study of how meaningful components of form are combined to make complex words. Understanding how such complex words can be 'broken apart' into their morphological constituents is the problem of morpheme segmentation. While words that have similar meanings tend to share similar forms (e.g., *run* and *running*), many morphemes do not have transparently shared meanings. For example, *canning* and *running* are only abstractly related in meaning through the progressive morpheme *–ing*. Further, sharing phonological material is not sufficient for morphological relatedness. Many words contain overlapping phonetic material without being morphologically related (e.g., words of the same cohort, such as *canning* and *canopy*). The question, then, is how learners find morphologically related words in the linguistic input, even when these words may not share common meanings (or before we know what the words mean).

The hypothesis that we address in this paper is that learners use the distributional information from the forms they hear in order to infer morphological regularities. For example, many words end in *–ing*, while many fewer begin with *can*. It is possible that learners are able to extract this regularity across many words in the language to identify the patterns that characterize the structure of words. In the present study, we expose adult learners to words that fit a stem-affix pattern (stem+suffix). We then test whether learners are able to extract the regularities of the stem-affix pattern through generalization to novel stem-affix combinations. If learners are able to do this without the use of semantic cues, it suggests that learners can form morphological parses from distributional cues.

Studies in another area of language learning, focused on word categories and sub-categories, have argued that phonological cues (Brooks, Braine, Catalano, & Brody, 1993; Frigo & McDonald, 1998; Gerken, Wilson, & Lewis, 2005) and semantic cues (Braine, et al., 1990) are highly important aids in learning (MacWhinney, Leinbach, Taranan, & McDonald, 1989; Maratsos & Chalkley, 1980). Most of these studies have suggested that learners have great difficulty in learning categories without phonological or semantic cues, and some have argued that learning categories is impossible without these cues to category structure (Gomez & Gerken, 2000). However, recent evidence shows that learners can use distributional information alone to acquire categories, as long as the distributional regularities are rich enough to support the induction of grammatical classes (Reeder, Newport, & Aslin, 2009).

While there is thus evidence that learners can use distributional information to learn categories, previous research on affix learning has focused on learning the meaning of the affix rather than the distribution of forms (Braine, et al., 1990; MacWhinney, 1983). For example, Braine et al. (1990) taught children inflectional locative affixes (e.g., *to*, *from*, *at*) in an artificial grammar learning setting. In this case, learning the form of the affix was dependent on the semantic context associated with the form. However, in natural languages the systematic pattern of affixation is not necessarily dependent of the meaning associated with the affix.

There are two reasons why it is important to study how morpheme segmentation can be done without semantics. First, a morpheme is more than just semantics; it involves both form and meaning. One must explain why words that are morphologically related are likely to be phonologically related as well. Understanding how learners can infer morphological relatedness through form relatedness is important for understanding the structure of the lexicon as well as the restrictions on morphological

[1] We are grateful for the following people who have provided helpful comments, advice and time: Kelly Johnston, Patricia Reeder, Neil Bardhan, Richard Aslin, Paul Smolensky, Colin Wilson, Neal Snider, and members of the University of Rochester ANLab. Funding was provided by NIH Grant DC00167.

systems in languages of the world. For example, if learners hypothesize that words with similar phonological forms might be morphologically related, they must have a way of differentiating these from the many morphologically unrelated words that also have similar phonological forms (e.g., cohort members, lexical neighbors). An important question is thus how these two kinds of similarity are distinguished. Second, children do not have access to the semantic information of their language for several months, but they do have access to phonological information of the language. If we assume that the learning process begins as soon as the child is exposed to language, much learning will take place when the child only has access to non-semantic distributional cues to morpheme segmentation. Thus distributional cues should be a primary source of evidence in early morpheme segmentation.

Further evidence that learners use distributional information for morpheme segmentation comes from computational solutions to the problem. In computational linguistics, the problem of morpheme segmentation involves finding the best algorithm to parse the morphemes of a particular corpus from a given language. These corpora typically contain either the orthographic or phonological representations for morphologically simple and morphologically complex words. The algorithms have no access to semantic or prosodic information, but will nonetheless output a morphological parse of all words in the corpus. While there are several computational techniques for performing the morpheme segmentation, the most common is the minimum description length (MDL) approach (Baroni, 2003; Baroni, Matiasek, & Trost, 2002; Brent, Murthy, & Lundberg, 1995; Clark, 2001; Creutz & Lagus, 2007; Gaussier; Goldsmith, 2006; Harris, 1955; Jacquemin, 1997; Johnson & Martin, 2003; Kazakov & Manandhar, 2001; Kontrovich, Ron, & Singer, 2003; Manning, 1998; Wicentowski, 2004). In this approach, the algorithm searches for the smallest possible description of the corpus as a whole. The best way to make a 'minimal' description of the corpus is to build each word from separate parts or morphemes. For example, rather than listing '*run*', '*running'*, '*can*' and '*canning*' as separate words, the algorithm will list '*run'*, *'can'* and '*-ing,'* thereby decreasing the number of bits that the description of the corpus will take. Thus there is an advantage to finding the morphemes of the language. There is also an advantage for morphemes to be as long as possible. That is, the algorithm does not simply list 'a' through 'z' and segment words by using orthographic elements (letters) as morphemes. The algorithm works best when it finds '*ing*' as the common denominator between '*running*' and '*canning*' rather than '*i'* and *'ng'*.

While the MDL approach to morpheme segmentation is relatively successful at parsing corpora into stems and affixes, it is an open question as to how human learners use distributional information to segment morphological material. There are several differences between the MDL algorithm and a child learning morphology. First, the MDL has access to an entire corpus at once and is not restricted by memory or processing limitations. Second, the MDL algorithm has access to type frequency but not token frequency, while the child learner has access to both. It may be that token frequency influences the particular parse of a given word. Third, the MDL algorithm has access to stable orthographic representations, while the child has access to highly graded and variable acoustic, phonetic and phonological representations of the word. Fourth, the MDL is able to use language-specific heuristics for parsing morphemes. For example, in English, stems that end in the same segment as the first segment of a suffix will generally have an [e] added between the stem and the suffix (e.g., *bus*, *buses*). The MDL algorithm can be programmed to search for specific orthographic changes as result of affixation, but the child must simultaneously learn both the phonological changes that result from affixation and morphological affixation itself. Finally, it is not clear whether the errors that the MDL algorithms make in parsing are analogous to the errors that children make in learning morphological structure.

Despite these differences, the relative success of the MDL algorithm at morpheme segmentation without any prosodic or semantic cues suggests that learners might be capable of morpheme segmentation using only the distributional cues of the lexicon. The present experiments test this ability in adult learners.

2. The Experiment

The goal of the present experiment is to explore the ability of adult learners to parse words into constituent parts (stems and suffixes) using only distributional information.

2.1 Participants

Eight participants were recruited from the University of Rochester community. All participants were adult monolingual native English speakers and were paid $10 for their participation. Participants were randomly assigned to one of two languages (A and B), with 4 participants in each condition.

2.2 Design

The experiment was designed to test the ability of adult learners to parse morphologically complex words. This was done by creating a language with all of the phonological properties of a suffixing language, but without associated meaning. Stems were all of the shape CVCV and suffixes were all of the shape CV, creating tri-syllabic words of the form CVCVCV. All consonants and vowels were drawn from the inventory [p, t, k, b, d, g, m, n, s, v, z, f] for consonants and [a, e, i, o, u] for vowels. Each consonant and vowel appeared equally often in each position, and none of the words were actual English words.

To ensure that the results were not due to any unnoticed peculiarity of the stimuli, we created two languages (Language A and Language B) with the same properties, but consisting of different sets of stems and suffixes. Each language had 24 stems and 4 suffixes. All suffixes were unique in that no stem contained the same syllable as a suffix.

Each stem was paired with 2 suffixes in the training phase (that is, each suffix was paired with 12 stems). This created a total of 48 affixed forms in training. Pairing of stem with affixes was balanced such that there was no pattern regarding which set of stems went with which affixes or which affixes went with particular stems (i.e., there were no categories or sub-categories within the affixed forms). Participants heard the 48 tri-syllabic items eight times during exposure, in randomized blocks (participants heard all 48 items before hearing the same set again in a different order). Examples of training stimuli are presented in (1) below.

(1) Examples of Training Stimuli

Language A			
befabu	basoke	basodo	dipimi
demebu	befake	demedo	fegemi
fegebu	dipike	fibado	fibami
tisebu	pumuke	tenodo	vopimi
zikubu	tiseke	zovado	zovami
Language B			
numopa	bikagu	bikase	dopono
sofepa	gisigu	fenuse	fenuno
finapa	vipogu	vipose	gisino
kidipa	zefogu	zefose	kobuno
tegipa	tegigu	numose	vemano

Following exposure, participants were given a two-alternative forced-choice test. This test was used to determine whether learners had parsed the words into stems and affixes. There were three different types of test items. The first set of test items tested familiarity of the words presented during exposure, comparing a familiar word (ABX) with a scrambled counterpart (AXB). The second set of test items tested the ability to generalize the stem + suffix pattern by presenting a stem with a suffix that had not been heard with it previously. We compared a new stem-affix combination (ABY) with a scrambled familiar item (AXB). If learners have extracted the general form of Stem+Affix and have learned the affixes, they should choose both the ABX and ABY items significantly above chance. The third set of test items tests whether learners choose a familiar word (ABX) equally with, or more frequently than, a stem-suffix combination that is grammatical but never heard before (ABY). There were 12 items in each test condition. Examples of test items are presented in the table in (2).

(2) Examples of Test Items

	Language A		**Language B**	
ABX-AXB	befabu	befado	bovepa	bopave
	demeke	dememi	dopogu	dogupo
	tisebu	tisedo	sovise	sosevi
ABY-AXB	tenoke	tekeno	tegipa	tegise
	vopibu	vobupi	zefogu	zefopa
	zikubu	zibuku	degapa	degano
ABX-ABY	basobu	bakeso	finano	fipana
	demebu	dedome	gisipa	ginosi
	fegebu	femige	muvuse	mupavu

Stimuli were recorded in a sound-attenuated booth by an adult female native English speaker. While the speaker was aware that the stimuli were to be used for an artificial grammar learning study, she was unaware of the hypothesis of the study. Tokens were individually recorded, with main stress on the initial syllable and with no reduction to schwa for the vowels in unstressed syllables. Each token was spoken 4 times in list format. A single token was chosen from the second or third position of the set in order to keep the prosody as uniform as possible (thus avoiding the marked prosodic characteristics of the production at the beginning or end of the list).

2.3 Procedure

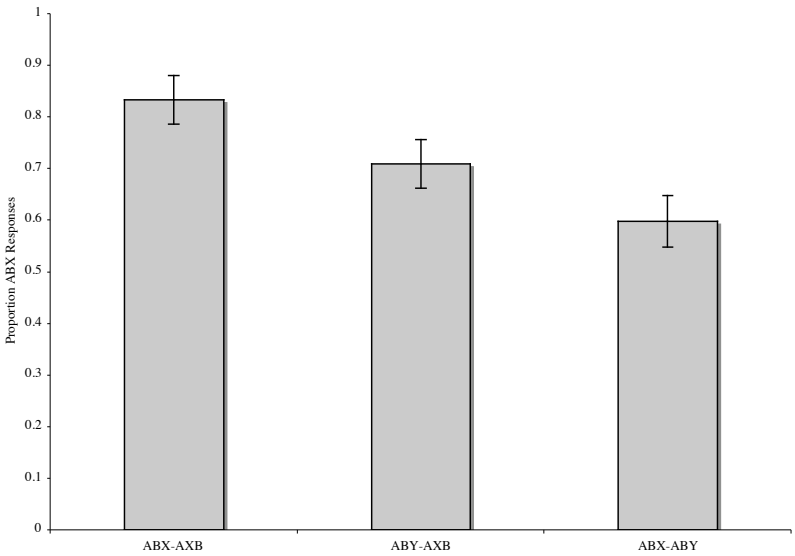
Participants were told that they would be listening to words from a language they had never heard before, and that their task was to listen to the way the novel language sounded but that they need not try to memorize the forms. The experiment took approximately 20 minutes to complete.

2.4 Results

Proportion of ABX (or ABY) responses was recorded for each participant, as shown in the table in (3). To ensure that there were no differences between Language A and Language B, we compared response patterns for Language A and Language via a 2 (Language) x 3 (Test Condition) ANOVA. Because there was no difference between the two languages (F <1) and no interaction (F <1), we combined the data for Language A and Language B. However, there was a significant effect of Test Items ($F(2, 12) = 8.86$, $p < 0.01$). This reflects the fact that the choice of one consistent response in the ABX vs. ABY test items (both grammatical) was significantly lower than in the ABX vs. AXB and the ABY vs. AXB test items ($F(1, 6) = 11.25$, $p < 0.05$). This is what we would expect if participants learned an abstract suffixing pattern: they should be more likely to select grammatical forms over ungrammatical forms than to select a familiar grammatical form over an unfamiliar grammatical form.

We separately compared each of the means for each type of test items to chance (50%) via one sample t-tests. The familiar test items (ABX vs. AXB) were significantly above chance (mean = 0.80, $t(7)= 6.41$, $p < 0.001$), showing that learners had acquired familiarity with the training items when compared with a scrambled item. The new grammatical items (ABY vs. ABX) were also chosen significantly above chance (mean = 0.78, $t(7)=5.97$, $p = 0.001$), showing that they had learned a productive suffixing pattern. Importantly, the selection of familiar vs. novel grammatical items (ABX vs. ABY) was not significantly above chance (mean = 0.55, $t(7)=1.05$, $p = 0.329$), suggesting that learners had no preference for familiar grammatical items over novel grammatical items.

(3)



Together, these results provide evidence that participants can learn a productive suffixing pattern using distributional information. In the language we presented to participants, there were no semantic cues to the morphology; the only reliable cue to morpheme segmentation was the distributional properties of the words themselves. Participants showed generalization to novel stem-suffix combinations in that they were able to select novel stem-affix combinations. In addition, participants were less likely to respond selectively to familiar items if the alternative was a grammatical item. This suggests that participants learned an abstract, productive suffixing pattern.

3. General Discussion

In the present experiment, adult learners used the distributional information of the phonological form of the words in an artificial language to reliably segment three-syllable words into stems and suffixes. Specifically, learners were able to infer a productive morphological pattern and did not simply memorize forms heard in training. Participants were substantially more likely to reject an ungrammatical form than a grammatical but unfamiliar form.

While the results of the present study demonstrate that suffixes can be learned through distributional information, more work is needed to understand the precise role of this information in learning morphological systems. In additional experiments, we have shown that learners are able to parse stems from prefixes as well as suffixes, thus acquiring a prefix pattern as well as a suffix pattern. We are currently investigating how learners parse more complex morphological patterns, such as infixation and non-concatenative morphology. Our preliminary results indicate that these more complex patterns require different distributional cues for learning than the prefix and suffix systems. For infixation and non-concatenative morphology, segmenting stems versus affixes cannot be accomplished simply through using distributional cues from adjacent syllables. In non-concatentative morphology, the relevant distributional contrasts are found at the segment level (e.g., consonant and vowel tiers) rather than in terms of syllables or strings of adjacent segments, as with prefixes and suffixes. Our future work will investigate how different types of statistical cues can be used to acquire these more complex morphological systems. We also plan to investigate the role of distributional information in learning systems utilizing multiple types of morphemes (e.g., both prefixes and suffixes that can appear together or in alternation) and involving variation across syntactic classes (e.g., verb vs. noun paradigms).

Future work will also investigate developmental changes in learning morphological patterns through distributional cues. In preliminary work we have tested older children (ages 7-10), and have found that, like adults, children are able to use distributional information to learn morphological patterns. However, some differences that appear to be emerging are differences in memory for individual items. Children are more likely than adults to accept a novel grammatical item as familiar. These differences in memory and recognition may lead to greater generalization of patterns. Our future work will examine how similarities and differences between children and adults affect the types of morphological patterns that are learned.

4. Conclusion

Understanding the role of distribution in morphological learning is important for developing a theory of the mechanisms that underlie language development. As we begin to understand the biases that learners have in using various statistical cues, we can uncover the ways in which these learning biases shape the patterns of languages throughout the world.

References:

Baroni, M. (2003). Distribution-driven morpheme discovery: A computational/experimental study. *Yearbook of Morphology*, 213-248.

Baroni, M., Matiasek, J., & Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002* (pp. 48-57).

Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., et al. (1990). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language, 29*, 591-610.

Brent, M. R., Murthy, S., & Lundberg, A. (1995). Discovering morphemic suffixes: A case study in minimum description length induction. *Fifth International Workshop on Artificial Intelligence and Statistics*. Fort Lauderdale, Florida.

Brooks, P. J., Braine, M. D. S., Catalano, L., & Brody, R. E. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language, 32*, 76-95.

Clark, A. (2001). Partially supervised learning of morphology with stochastic transducers *Proceedings of Natural language processing Pacifc Rim symposium, NLPRS 2001* (pp. 341-348). Tokyo, Japan.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing, 4*(1), 1-34.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasss. *Journal of Memory and Language, 39*, 218-245.

Gaussier, E. Unsupervised learning of derivational morphology from inflectional lexicons. 24-30.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*(249-268), 249-268.

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. *Computational Linguistics, 12*(4), 353-371.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Science, 4*(5), 178-186.

Harris, Z. S. (1955). From phoneme to morpheme. *Language, 31*(2), 190-222.

Jacquemin, C. (1997). Guessing morphology from, terms and corpora. *International ACM SIGR Conference on Research and Developmental Retrieval (SIGR '97)*. Philadelphia, PA.

Johnson, H., & Martin, J. (2003). Unsupervised learning of morphology for English and Inuktitut *HLT-NAACL 2003, Human language technology Conference of the North American Chapter of the Association for Computational Lingustics*. Edmonton, Candada.

Kazakov, D., & Manandhar, S. (2001). Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning, 43*, 1210162.

Kontrovich, L., Ron, D., & Singer, Y. (2003). A Markov model for the acquisition of morphological structure.

MacWhinney, B. (1983). Miniature Linguistic Systems as Tests of the Use of Universal Operating Principles in Second-Language Learning by Children and Adults. *Journal of Psycholinguistic Research, 12*(5), 467-478.

MacWhinney, B., Leinbach, J., Taranan, R., & McDonald, J. L. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28*, 255-277.

Manning, C. D. (1998). The segmentation problem in morphology learning. In D. M. W. Powers (Ed.), *NeMlaP3/CoNLL98 Workshop on paradigms and grounding in language learning* (pp. 299-305): ACL.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language* (Vol. 2, pp. 127-189). NY: Gardner Press.

Reeder, P., Newport, E., & Aslin, R. (2009). The role of distributional information in linguistic category formation. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Wicentowski, R. (2004). Multilingual noise-robust supervised morphological analysis using the word-frame model. *Proceedings of the ACL Special Interest Group on Computational Phonology (SIGPHON)* (pp. 70-77).