Distribution-driven morpheme discovery: A computational/experimental study

MARCO BARONI
SSLMIT (University of Bologna)
Corso della Repubblica 136
47100 Forlì (FC), Italy

baroni@sslmit.unibo.it

October 2002

Abstract

In an early stage of morphological acquisition, children must discover which strings correspond to affixes of their language, and which of the words containing those strings are actually affixed. For example, a child acquiring English must be able to discover that the word-initial string *re-* is a prefix, but also that the word *remake* is prefixed, whereas the word *retail*, probably, is not, even though it begins with *re-*. In this study, I present a computational model of how the task of morpheme (in particular, prefix) discovery could be performed on the basis of distributional cues (cues based on the co-occurrence patterns, frequency and length of words and their substrings in the input). The results of a simulation using this model with an input corpus of English words show that distributional evidence could in principle be very helpful to learners having to perform the task of morpheme discovery. Moreover, I show that the morphological parses assigned by the distribution-driven model to a set of potentially prefixed but semantically opaque words are correlated with morphological complexity ratings assigned to the same words by native English speakers. I argue that this convergence between the model and the speakers, in a domain in which speakers cannot rely on semantic cues, constitutes evidence that humans do rely on distributional cues similar to the ones exploited by my model, when assigning morphological structure to words.

1. INTRODUCTION[1]

During the process of language acquisition, learners must discover which strings constitute the affixes of their language and which words of the language can be decomposed into affixes and other components. These are prerequisites to morphological acquisition.

Ultimately, a learner acquiring a language must discover the syntactic and semantic properties associated with each affix of the language, in order to be able to produce and understand new words. For example, a learner acquiring English must discover that *re-* is a prefix that attaches to verbs to create other verbs with an iterative meaning.

However, in order to learn the morphological properties of an affix, learners must first of all notice the existence of that affix. Moreover, in order to discover the linguistic

properties associated with the affix, the learner must inspect the semantic, syntactic and morphological characteristics of a set of words containing that affix.

For example, in order to discover the properties of the prefix *re-*, English learners must first of all, of course, notice that the string *re-* is a prefix. Moreover, the learners must collect and analyze a number of words containing the prefix *re-* (*redo, rename, remake…*),[2] in order to extract the correct generalizations about this prefix.

However, not all the words containing a string identical to an affix actually contain that affix. In order to discover the correct generalizations about the properties of the affix, the learners must have a preliminary idea of which of the words containing a string identical to the affix are actually affixed. If an English learner tried to decide what is the meaning and function of *re-* on the basis of, say, *redo, retail* and *really*, the learner would probably come up with the wrong generalizations about the prefix or, more likely, she would not notice any generalization at all and she would conclude that *re-* is not a prefix.

Of course, if the string corresponding to an affix mostly occurs in words which do indeed contain the affix, the learner is probably going to extract the correct generalizations even if there are a few pseudo-affixed words (i.e., words containing the string corresponding to the affix without actually being synchronically analyzable as affixed). However, this is not always the case. For example, Schreuder and Baayen (1994) have shown that, for several common English and Dutch prefixes, the number of pseudo-prefixed words is higher than the number of truly prefixed words (at least in terms of token frequency).

Thus, it would not be safe, for a learner, to assume *a priori* that any word containing a string identical to an affix does indeed contain the affix from a morphological point of view. Consequently, the learner must decide which of the words containing a potential affix are truly morphologically complex, and which are pseudo-affixed, i.e., the learner must assign preliminary morphological parses to the words she hears.

While I presented the task of discovering that a certain string is an affix (or more generally a morpheme) and the task of assigning parses to words as separate aspects of morpheme discovery, the two tasks are obviously closely related. A learner is not likely to hear affixes in isolation. Thus, the task of discovering the affixes will typically involve assigning morphological parses to words. A string is an affix of the language if at least one of the words containing the string in the language is parsed as morphologically complex, and the string constitutes one of the morphological components in the parse.

Morpheme discovery is a difficult task. Not only does the learner have to consider many possible segmentations of each potentially complex word she hears, but she does not *a priori* know which meanings and/or syntactic functions are expressed by morphemes in her language, and consequently she cannot *a priori* know whether a word is morphologically complex or not. Furthermore, the learner does not know which types of morphemes (prefixes, suffixes, circumfixes, infixes, autosegments, templates...) are present in the language. Thus, even if the learner had some reason to expect a certain word to be morphologically complex, she still would have to determine whether the word should be divided into a prefix and a stem, or into a stem and a suffix, or into consonantal and vocalic templates, or into other morpheme combinations.

It is probable that learners follow a number of different morpheme discovery strategies, looking for phonological, syntactic and semantic cues. Moreover, distributional evidence, i.e., evidence based on the frequency and co-occurrence patterns of words and their substrings, provides potentially useful cues that learners can exploit. While each of these approaches can help the learner in the morpheme discovery task, none of them is likely to be sufficient by itself.

The primary goal of this study is to contribute to a better understanding of how language learners perform morpheme discovery. In particular, the study provides evidence in favor of the hypothesis that distributional cues play a significant role in this process.

Some recent studies have provided new support for the idea that distributional information plays an important role in language learning (see Redington and Chater 1998

for a review of both the classic objections to distributional approaches and recent distribution-driven learning models). Thus, another goal of the present study is to provide further support for the general claim that language learners make crucial use of distributional cues.

As I will shortly discuss in the conclusion, another goal of this study is to provide a (partial) explanation for an interesting datum emerging from experimental studies of morphological processing and representation, i.e., that speakers can represent words as morphologically complex even if they lack semantic compositionality (i.e., the meaning of the whole word is not the product of the meanings of the component morphemes). One can argue that this phenomenon (complex representation/treatment of semantically opaque words) is at least in part a by-product of distribution-driven morpheme discovery, and the empirical evidence presented here provides some support for this hypothesis.

In order to assess the potential role of distributional cues in morpheme discovery, I designed an automated learner, which performs a simplified version of this task on the sole basis of the distributional evidence it can extract from a corpus of untagged words. The most obvious simplification in the task performed by this computational model is that it only looks for prefixes and stems, and not also for other kinds of morphemes.

The strategy followed by the automated learner in its search for prefixes and stems is based on a simple fact about the distributional nature of morphemes: morphemes are independent linguistic units, and as such, they occur in a number of different words where they combine with other morphemes. The nonrandom distribution of morphemes makes them detectible, in many cases, by statistical methods.

Given an input corpus of English words, the automated learner, equipped with a small number of simple distributional heuristics, is able to discover a large set of actual English prefixes, finding very few "false positives" (strings which are not English prefixes but are treated by the learner as such). Moreover, the morphological parses (prefix + stem vs. monomorphemic) assigned by the learner to the words in the input corpus are correlated with intuitions of native English speakers about the morphological structure of the same words.

Thus, the computational simulation presented here demonstrates first of all that a limited number of simple distributional heuristics can help a morpheme discoverer a great deal, i.e., that there is in principle a large amount of evidence about morphological constituency that children could extract from simple distributional cues.

Moreover, I show that the morphological parses assigned by the distribution-driven model to a set of potentially prefixed but semantically opaque words are correlated with morphological complexity ratings assigned to the same words by native English speakers. I argue that this convergence between the model and the speakers, in a domain in which speakers could not have relied on semantic cues, constitutes evidence that humans do indeed rely on distributional cues similar to the ones exploited by my model, when assigning morphological structure to words (see section 5 below for a full discussion of this argument).

Notice that in the current project I am modeling morpheme discovery as a purely distribution-driven task because I am interested in trying to determine how much and what kind of information a learner could in principle extract from distributional evidence alone. I am *not* trying to argue that this is the only kind of evidence used by human learners. It is plausible that learners would use distributional cues at the earliest stages of morpheme discovery, since distributional information can be straightforwardly extracted from the data, and it can be exploited prior to any linguistic analysis. More sophisticated linguistic information can later be used to refine the coarse guesses on morphological structure made on the basis of distributional cues.

The remainder of this study is organized as follows: In section 2, I shortly review some related work. In section 3, I present and discuss the computational model I am proposing. In section 4, I present the results of a simulation in which this model was tested, and I compare the morphological parses assigned by the model to morphological

complexity ratings assigned by humans. In section 5, I compare the performance of the model to that of native speakers in parsing semantically opaque (but potentially complex) words. Finally, in the conclusion I briefly discuss future directions that this project could take.

2. RELATED WORK

After the pioneering work of Harris in the fifties (see, e.g., Harris 1955) and until very recently, modeling morpheme discovery has been a relatively unpopular research domain. However, in the last few years there has been a resurgence of interest in the topic, and several supervised and unsupervised algorithms performing morpheme discovery or related tasks have been recently proposed: See, for example, most of the papers collected in Maxwell 2002, and the references quoted there.

A recent approach that is closely related to the one I propose below is the one of Goldsmith 2001.[3] Goldsmith's model, like mine, is based on the idea that morphological segmentation can be re-phrased as a data compression problem. However, Goldsmith's model differs from the one presented here in several respects.

The most obvious (and probably least interesting) difference is that Goldsmith's model looks for suffixation patterns, whereas my model focuses on prefixation.

More importantly, Goldsmith's model is based on an information-theoretically rigorous model of data compression constructed using the Minimum Description Length (MDL) principle of Rissanen 1978. The model proposed below, on the other hand, is only loosely inspired by the MDL idea, and the data compression scheme it assumes is not valid from an information-theoretic point of view (see Baroni 2000b:3.3.7 on why I decided to abandon the more rigorous MDL-based approach I adopted in Baroni 2000a).

Furthermore, Goldsmith's model generates a single analysis using alternative heuristic strategies, and then uses the MDL criterion to refine such analysis. On the other hand, the model presented here uses heuristics to generate a set of alternative analyses, and then applies the maximal data compression criterion to choose the best of these alternatives.

From a strictly linguistic point of view, Goldsmith's model has two desirable properties that are missing in the current model, i.e. it can fully decompose words containing multiple affixes, and it groups stems and affixes into primitive forms of paradigms called signatures.

Last but not least, Goldsmith's main interest seems to lie in the possibility of extracting morphological analysis tools from unlabeled corpora using an automated procedure, whereas I developed the model I describe below because I am interested in testing some hypotheses about the role of distributional learning during human morphological acquisition. Thus, the focus here is less on the technical aspects of the model, and more on how the outputs it produces compare to human intuitions about morphological structure.

Another model that is closely related to mine is the utterance segmentation method proposed in Brent and Cartwright (1996). Indeed, my algorithm can be seen as an adaptation of Brent and Cartwright's lexicon selection and generation methods to the morpheme discovery problem. Thus, my algorithm takes words, rather than unsegmented utterances as input, and it returns maximally binary segmentations.

Moreover, my model is biased so that it favors parses in which one element has affix-like distributional properties, and the other element has stem-like properties.

Finally, Brent and Cartwright's model, like the one proposed by Goldsmith, is based on an information-theoretically sound data compression scheme, whereas the model I propose below is only justified by the fact that it captures intuitively plausible morpheme-segmentation heuristics.

3. DDPL: AN AUTOMATED DISTRIBUTION-DRIVEN PREFIX LEARNER

In order to assess the effectiveness of distributional heuristics in morpheme discovery, I designed and implemented a learning model which performs a particular aspect of this task --prefix discovery-- on the sole basis of distributional evidence.

The algorithm presented here takes a corpus of untagged orthographically or phonetically transcribed words as its input and outputs a lexicon composed of a list of prefixes and stems. Moreover, the algorithm assigns morphological parses (prefix + stem or monomorphemic parses) to all the word types in the input corpus.[4] The algorithm relies entirely on the distributional information that can be extracted from the input. From here on, I will refer to the algorithm presented here with the acronym *DDPL*, which stands for *Distribution-Driven Prefix Learner*.

DDPL is based on a "generation and selection" strategy: a large number of lexica compatible with the input data are generated, a certain measure is computed for each lexicon, and the lexicon with the lowest value of this measure is selected. The formula used to compute this measure constitutes the conceptual core of the algorithm, and it is based on the idea that the best morphological analysis of the input is also the one allowing maximal data compression of the input, given certain assumptions about how the data compression process should work.

As we will see, the data compression scheme from which the DDPL lexicon selection formula is derived favors lexical analyses respecting the following three principles, which in turn can be interpreted as plausible morpheme discovering strategies:

- Substrings occurring in a high number of different words are likely to be morphemes;
- Substrings which tend to co-occur with other potential morphemes are more likely to be morphemes;
- All else being equal, low frequency words are more likely to be morphologically complex than high frequency words.

The first two principles should be fairly intuitive: Morphemes -- especially affixes -- tend to occur in a number of different words. Thus, they will tend to have a high type frequency. Moreover, real morphemes are not simply substrings that occur in a high number of random words, but rather substrings that can co-occur with other morphemes to form complex words. This explains the second principle.

The third principle is perhaps less intuitive. Consider, however, the following. At one extreme, if a morphologically complex word is very frequent, the word is likely to have its own lexical entry, distinct from the entries of its component parts (at the very least, for reasons of ease of lexical access). However, once a word has an independent lexical entry, the word can acquire its own semantic features and thus it is likely to lose, over the course of time, its connection with its component parts. In other words, high frequency words are less likely to be morphologically complex because, even if they *were* complex from an etymological point of view, they will tend to acquire a lexicalized meaning due to heavy usage.

At the other extreme, productively formed complex words must be *hapax legomena* (words with a token frequency of 1), or in any event have a very low token frequency. Indeed, Baayen has shown in several studies (see for example Baayen 1994, Baayen and Lieber 1991) that the number of *hapax legomena* containing a certain morpheme is a good indicator of the productivity of the morpheme. If a morpheme is productive, then the morpheme is often used to create new forms, and new forms, being new, are likely to have a very low frequency.

Thus, all else being equal, it would make sense for a learner to be more willing to guess that a word is complex if the word has a low token frequency than if the word has a high token frequency.

While in the following sections I will concentrate on how morpheme discovery can be rephrased as a data compression problem, the reader should keep in mind that I am by no means assuming that morpheme discovery *is* indeed a data compression problem (indeed, as I discuss in Baroni (2000b), the data compression method proposed here, when seen as an actual data compression algorithm, is far from optimal, and it is not truly implementable). Rather, the interest of the data compression approach lies in the fact that it allows us to derive an explicit formula that, as I will show, favors the same type of lexical analysis that would be favored by distribution-based morpheme discovery strategies such as the ones I just discussed.

3.1 Data compression and morphological analysis: the shortest lexicon + encoding criterion

The criterion used by DDPL to select the best lexicon is based on the idea that the lexicon generated by the most plausible morphological analysis is also the best lexicon for purposes of data compression, given certain restrictions on how the compression procedure must work. The rationale behind this intuition is the following: Since morphemes are syntagmatically independent units, which occur in different words and combine with each other, a lexicon containing morphemes is going to be "shorter" (in the literal sense that it can be represented using a small number of characters) than a lexicon containing random substrings, or a lexicon in which no word is decomposed. The advantage of reducing the problem of morpheme discovery to a matter of (constrained) data compression is the following: There are no straightforward ways to decide which one, among a set of possible lexica, is the best one from the point of view of morphology, but it is relatively simple to estimate which lexicon allows maximal data compression.

Given that the connection between data compression and morphological analysis is not very intuitive, I will illustrate it with a set of simple examples.

Let us suppose that we are given a list of words, and our goal is to find a compact format to store information from which the very same list can be reconstructed. In particular, we take a "lexicon and encoding" approach to this task. We construct a compact *lexicon* from which all the input words (plus, possibly, others) can be reconstructed. We associate an index to each lexical unit, and then we rewrite the corpus as a sequence of these indices. I will refer to the rewriting of the corpus as a sequence of indices with the term *encoding.*

As long as some lexical units occur very frequently in the input corpus (and/or many units occur relatively frequently), and the lexical indices are, on average, shorter than the units they represent, the lexicon + encoding strategy will allow us to represent the corpus in a shorter format than the original one. In order to make sure that the second requirement is satisfied (i.e., lexical indices are on average shorter than the input words they represent), I assume here that all indices are exactly one character long (I will revise this in 3.3 below). This is one of the main aspects in which the method presented here is not a realistic data compression scheme.

The following example, which has nothing to do with morphological decomposition, is presented to give a first, general idea of how and why the lexicon and encoding strategy works. Suppose that we are given the following input corpus:

(1) dog cat dog dog cat cat

In order to write this list, we need 18 characters. Following the compression method described above, we can instead write the words *dog* and *cat* only once, assigning a one-character index to each of them (this is the lexicon component of the compressed data), and then rewrite the words in the input as a sequence of indices (the encoding):

(2) *Lexicon*
dog 1
cat 2

Length of lexicon: 8

Encoding of (*1*)
1 (= dog)
2 (= cat)
1 (= dog)
1 (= dog)
2 (= cat)
2 (= cat)

Length of encoding: 6

Total length (lexicon + encoding): 14

To store the word types *dog* and *cat* in a lexicon, and then rewrite the word tokens in the input as a sequence of indices is more economical than writing down the list of input word tokens as it is (in the sense that it requires a smaller number of characters: 14 vs. 18). Notice that from the lexicon and the list of indices we can reconstruct the original input. Thus, we can store a corpus in the more economical lexicon and encoding format without any loss of information.

Now, suppose that we are allowed to decompose input words into two constituents, in order to further compress the data. We assume the following encoding scheme: if an input word is identical to a lexical entry, then the input word is encoded using the index associated with that lexical entry (as in the example above); however, if a word does not have a corresponding entry, and must be reconstructed by concatenating two lexical units, then the word is encoded as the sequence of the index associated with the first component, a one-character concatenation operator (represented here by the symbol ˚) and the index associated with the second component.

For example, suppose that a corpus contains the word *redo*. If this word is listed in the lexicon, for example associated with the index *1*, then the word is represented by a *1* in the encoded input. However, if *redo* is not listed in the lexicon, and it has to be reconstructed from the entries *re*, associated with the index *1*, and *do*, associated with the index *2*, then the word will be represented by the sequence *1˚2* in the encoded corpus.

While it can be convenient to store frequent substrings in the lexicon, in order to make the lexicon shorter, there is a counterbalancing factor, namely the length of the encoding. On the one hand, treating substrings occurring in a number of words as independent lexical entries will make the lexicon shorter. On the other hand, since it takes three characters (two indices plus the concatenation operator) instead of one to encode an input word not listed in the lexicon, any decomposition which makes the lexicon shorter will also make the encoding longer. Thus, only decompositions allowing a decrease in lexical length which more than compensates for the corresponding increase in encoding length are worth performing. From the point of view of morpheme discovery, this trade-off ensures that the only decompositions that will be performed will be those motivated by strong distributional evidence.

For example, compare the shortest lexicon + encoding representations of the lists in (3) vs. (6). Consider first the list in (3):

(3) redo do remake undo make unmake

The shortest lexicon + encoding representation of this list is the following:

(4) *lexicon*

re	1	do	3
un	2	make	4

length of lexicon: 14

encoding of (3)

1˚3 (= re˚do)	1˚4 (= re˚make)	4 (= make)
3 (= do)	2˚3 (= un˚do)	2˚4 (= un˚make)

length of encoding: 14

total length (lexicon + encoding): 28

In particular, this is a shorter representation than the one in which no decomposition is attempted:

(5) *lexicon*

redo	1	remake	3	make	5
do	2	undo	4	unmake	6

length of lexicon: 32

encoding of (3)

1 (= redo)	3 (= remake)	5 (= make)
2 (= do)	4 (= undo)	6 (= unmake)

length of encoding: 6

total length (lexicon + encoding): 38

The representation in (4), which is based on a plausible morphological decomposition of the input, is ten characters shorter than the representation in (5), where no decomposition is attempted. The reason for this is that the analysis of the input upon which (4) is based provides a very compact lexical component, since the units *re, un, do* and *make* are all morphemes which occur in at least two input words. Thus, even if the encoding in (4) is longer than the encoding in (5), the lexicon in (4) is so much shorter than the one of (5) that, overall, (4) is the analysis to be selected on the basis of the shortest lexicon + encoding criterion.

However, consider now the case of the input in (6):

(6) dog tag mug

This time, we must choose the lexical analysis of (7), in which no word is decomposed:

(7) *lexicon*

dog	1	tag	2	mug	3

length of lexicon: 12

encoding of (6)

1 (= dog)	2 (= tag)	3 (= mug)

length of encoding: 3

total length (lexicon + encoding): 15

The total length of lexicon and encoding in (7) (where the input words are not decomposed) is shorter than the one of (8), where *g* is treated as an independent lexical unit:

(8) *lexicon*

do	1	mu	3
ta	2	g	4

length of lexicon: 11

encoding of (6)

1˚4 (= do˚g) 2˚4 (= ta˚g) 3˚4 (= mu˚g)

length of encoding: 9

total length (lexicon + encoding): 20

While the lexicon in (8) is shorter than the one in (7), the small decrease in lexical length does not justify the increase in encoding length due to the fact that one needs three characters to represent each word not listed as an independent unit in the lexicon.

As the previous examples show, the shortest lexicon + encoding criterion favors the representation of substrings as lexical entries only when this approach leads to considerable savings in lexical length. True morphemes, unlike arbitrary substrings, are more likely to lead to such savings, and, thus, to be treated as independent lexical entries.

3.2 Capturing the morpheme discovery heuristics

In this section, I present a series of related examples which have the function of illustrating how the shortest lexicon + encoding approach principle favors lexical analyses that are also optimal from the point of view of the distributional morpheme discovery heuristics that I listed in section 3.[5]

3.2.1 The high frequency heuristic

The first example I present shows how the first heuristic ("frequent substrings are more likely to be morphemes") is captured by the shortest lexicon + encoding criterion. Consider the data sample in (9):

(9) disarray disdain disintegrate disadvantage disaster

The analysis in which the word-initial substring *dis* is treated as an independent lexical entry (10.a) allows a more compact lexicon + encoding representation than the analysis in which the words are not decomposed (10.b)

(10) a. *lexicon*

dis	1	integrate	4
array	2	advantage	5
dain	3	aster	6

length of lexicon: 40

encoding of (9)

1˚2 (= dis˚array)	1˚5 (= dis˚advantage)
1˚3 (= dis˚dain)	1˚6 (=dis˚aster)
1˚4 (= dis˚integrate)	

length of encoding: 15

total length (lexicon + encoding): 55

b. *lexicon*

disarray	1	disadvantage	4
disdain	2	disaster	5
disintegrate	3		

length of lexicon: 52

encoding of (9)

1 (= disarray)	4 (= disadvantage)
2 (= disdain)	5 (= disaster)
3 (= disintegrate)	

length of encoding: 5

total length (lexicon + encoding): 57

The analysis in (10.a) is shorter than the analysis in (10.b) simply in virtue of the fact that *dis* is a frequent word-initial string (of course, what counts as "frequent" depends in general on the size of the input corpus -- in this and the following sections we analyze very small corpora, and, as a consequence, strings occurring in a small number of words will behave as frequent). The lexicon which would be favored by the "frequent strings are morphemes" heuristic is also the lexicon favored by the shortest lexicon + encoding criterion.

However, notice also that the difference in length between the analyses is rather small (55 vs. 57). Indeed, it is sufficient to remove one word from the list in (9)...

(11) disarray disdain disintegrate disadvantage

... and the analysis in which *dis* is treated as an independent entry is no longer shorter than the analysis in which the input words are not decomposed:

(12) a. *lexicon*

dis	1	integrate	4
array	2	advantage	5
dain	3		

length of lexicon: 35

encoding of (11)

1˚2 (= dis˚array)	1˚4 (= dis˚integrate)
1˚3 (= dis˚dain)	1˚5 (= dis˚advantage)

length of encoding: 12

total length (lexicon + encoding): 47

b. *lexicon*

disarray	1	disintegrate	3
disdain	2	disadvantage	4

length of lexicon: 43

encoding of (11)

1 (= disarray)	3 (= disintegrate)
2 (= disdain)	4 (= disadvantage)

length of encoding: 4

total length (lexicon + encoding): 47

As this example shows, the "store frequent strings as independent units" principle plays a secondary role in our data compression scheme. This is also reasonable when we look at our approach as a way to select the best lexicon from a morphological point of view: Morphemes are not simply frequent substrings, but, as the second heuristic stated above states, substrings that frequently co-occur with other syntagmatically independent substrings (i.e., other morphemes) to form complex words. As the following section shows, lexical analyses respecting this heuristic lead to shorter representations than the ones obtained by simply treating frequent substrings as independent units.[6]

3.2.2 Co-occurrence with other potential morphemes

As I just stated, a sensible morpheme discovery strategy should not simply be based on absolute frequency, but on the number of times a string tends to co-occur with other "potential morphemes", i.e., strings also occurring elsewhere in the corpus.

In the lexicon + encoding model, independent lexical entries for strings which tend to combine with other independently occurring strings lead to larger savings than simply treating frequent strings as lexical entries. Consider first the sample in (13):

(13) redo go remake replug dogs

Even if the string *re* occurs three times in this corpus, the representation in which this string is not treated as an independent lexical unit (14.b) is shorter than the one in which the words beginning with *re* are decomposed (14.a):

(14) a. *lexicon*

re	1	go	3	plug	5
do	2	make	4	dogs	6

length of lexicon: 24

encoding of (13)

1˚2 (= re˚do)	1˚5 (= re˚plug)
3 (= go)	6 (dog)
1˚4 (= re˚make)	

length of encoding: 11

total length (lexicon + encoding): 35

b. *lexicon*

redo	1	remake	3	dogs	5
go	2	replug	4		

length of lexicon: 27

encoding of (13)

1 (= redo)	4 (= replug)
2 (= go)	5 (= dogs)
3 (= remake)	

length of encoding: 5

total length (lexicon + encoding): 32

Compare now the input in (13) with the input in (15):

(15) redo do remake sprint make

The two inputs are similar in that they are both composed of six words of length 4, 2, 6, 6 and 4 respectively. Moreover, the two words containing *re* in (15) are also present in (13). However, on the one hand the string *re* only occurs two times in (15) (vs. three times in (13)); on the other hand, in (15) the string *re* occurs before strings which also occur elsewhere in the list (both *do* and *make* also occur as independent words). In this case, the analysis in which *re* is treated as an independent lexical entry (16.a) is much shorter than the analysis in which the words beginning with *re* are not decomposed (16.b):

(16) a. *lexicon*

re	1	make	3
do	2	sprint	4

length of lexicon: 18

encoding of (15)

1˚2 (= re˚do)	4 (= sprint)
2 (= do)	3 (= make)
1˚3 (= re˚make)	

length of encoding: 9

total length (lexicon + encoding): 27

b. *lexicon*

redo	1	remake	3	make	5
do	2	sprint	4		

length of lexicon: 27

encoding of (15)

1 (= redo)	4 (= sprint)
2 (= do)	5 (= make)
3 (= remake)	

length of encoding: 5

total length (lexicon + encoding): 32

The examples from (13) to (16) show how, in our model, strings which co-occur with other strings that also independently occur in the corpus are more likely to be treated as independent units than strings which simply frequently occur in the corpus, even if the latter are more frequent, in absolute terms, than the former.[7] The reason for this preference is that, if both elements forming a complex word are already stored in the lexicon, then there is no need to store any extra item in order to be able to reconstruct the word. On the other hand, if only one of the constituents of a word is already stored in the lexicon, we still need to create a lexical entry, complete with an index, for the remainder of the form.

3.2.3 Word frequency and morphological complexity

We observed that, all else being equal, learners should be more willing to treat words as morphologically complex if they rarely occur in the corpus than if they are frequent. Again, there is a parallel with the shortest lexicon + encoding approach to data compression, as the following examples show. Consider first the input in (17):

(17) disarray array disobey obey

Of course, the shortest analysis of this input is the one in which both *disarray* and *disobey* are decomposed (18.a) (both analyses presented here are shorter than the one in which *disobey* is not decomposed):

(18) a. *lexicon*

dis	1	obey	3
array	2		

length of lexicon: 15

encoding of (17)

1˚2 (= dis˚array)	1˚3 (= dis˚obey)
2 (= array)	3 (= obey)

length of encoding: 8

total length (lexicon + encoding): 23

b. *lexicon*

disarray	1	dis	3
array	2	obey	4

length of lexicon: 24

encoding of (17)

1 (= disarray)	3˚4 (= dis˚obey)
2 (= array)	4 (= obey)

length of encoding: 6

total length (lexicon + encoding): 30

However, consider now the input in (19), which is identical to (17), except that I added four more tokens of the word *disarray*.

(19) disarray array disobey obey disarray disarray
disarray disarray

Now, the best analysis becomes the one in which *disarray* is *not* decomposed into *dis* and *array* (20.b) (both analyses are shorter than the one in which neither *disarray* nor *disobey* are decomposed):

(20) a. *lexicon*

dis	1	obey	3
array	2		

length of lexicon: 15

encoding of (19)

1˚2 (= dis˚array)	1˚2 (= dis˚array)
2 (= array)	1˚2 (= dis˚array)
1˚3 (= dis˚obey)	1˚2 (= dis˚array)
3 (= obey)	1˚2 (= dis˚array)

length of encoding: 20

total length (lexicon + encoding): 35

b. *lexicon*

disarray	1	dis	3
array	2	obey	4

length of lexicon: 24

encoding of (19)

1	(= disarray)	1	(= disarray)
2	(= array)	1	(= disarray)
3˚4	(= dis˚obey)	1	(= disarray)
4	(= obey)	1	(= disarray)

length of encoding: 10

total length (lexicon + encoding): 34

These examples show how, all else being equal, in the lexicon + encoding model more frequent words are more likely to be stored in the lexicon in non-decomposed format than less frequent words. The only difference between the distribution of *disarray* in (17) and (19) is that in (19) this word occurs five times, whereas in (17) it occurs only once. Because of this difference in frequency, *disarray* get its own lexical entry in the shortest analysis of (19), whereas in the shortest analysis of (17) it must be reconstructed from the components *dis* and *array.*

The reason why this model favors independent storage of frequent words, even when both of their components are also specified in the lexicon, is the following: Given that each occurrence of a decomposed word in the encoded corpus requires three indices instead of one, if a word occurs frequently in the corpus, it is more convenient to use some characters to build a lexical entry for it, in order to have an economical way to encode it. On the other hand, if a word is rare, it is more convenient to save the characters required to represent the word in the lexicon, and encode the word in a costly way the few times in which it occurs in the corpus.

As observed by a reviewer, the model also predicts that longer words will be more resistant to independent storage, since, being longer, more characters are needed to store them in the lexicon. Future research should try to assess to what extent a heuristic along these lines (longer words are more likely to be morphologically complex) is empirically founded.

The example (19)/(20.b) also illustrates an interesting general property of the DDPL model: Not only is this model flexible enough to allow words to be treated as morphologically simple (i.e., represented in the lexicon as independent units), even if they begin with a substring which is specified as a prefix in the lexicon, but words can be represented as units in the lexicon even if *both* their component parts are also lexical entries: in the case at hand, both the word *disarray* and its constituents *dis* and *array* are listed in the lexicon. As it is a common feature of many models of lexical-morphological processing (see Schreuder and Baayen 1995 and the other models reviewed there) to assume that words can have an independent lexical representation even if they could be entirely derived from morphemic constituents also stored in the lexicon, I believe that it is a desirable property of our learning model that it can select lexica in which this situation arises.

3.3 Prefix-stem asymmetries in the DDPL model

The actual formula used to estimate lexicon + encoding length in the DDPL model assumes a representation strategy slightly different from the one presented in the examples above, in order to account for the fact that prefixes and stems have different distributional properties.

In particular, affixes tend to be more frequent units than stems. Given a prefixed input word, it is very likely that its prefix also occurs in a number of other input words, whereas the stem probably only occurs in very few other words. For example, it is plausible that in a corpus of English containing the form *reconsider*, the prefix *re* also occurs in hundreds of other words, whereas the stem *consider* only occurs in this prefixed form, as an independent word and in no more than five or six other derived forms.

In order to take the prefix-stem asymmetry into account, a bias against prefixes has been introduced in their lexical representation: prefixes are associated with indices that are slightly longer than the ones assigned to stems. Specifically, prefixes are associated with indices that are 1.25 characters long, whereas stems are associated with indices that are one character long. The value 1.25 was determined empirically, by fitting against the data described in section 4 below. The length of the index need not be an integer, since it simply represents an arbitrary penalty, and, for our purposes, it does not have to correspond to a plausible encoding scheme.

The effect of this representational bias is that prefixes have to be more frequent (and/or more frequently co-occur with potential morphemes) than stems in order to be represented as independent lexical units in the shortest analysis of the input.

An important consequence of the bias against infrequent prefixes is that it disfavors analyses in which stems of suffixed forms are mistakenly treated as prefixes. Consider for example the word *lovely*. Given that, in a reasonably sized English corpus, both *love* and *ly* probably also occur in other forms, there is the risk that DDPL could treat *love* as a prefix and *ly* as a bound stem. However, compared to a real prefix, *love-* is likely to occur in a very limited number of forms: for example, in the PHLEX database (see section 4 below) the word-initial substring *love* only occurs in a total of 8 forms, whereas even a rare prefix such as *para-* occurs in 22 words. Thus, given the anti-infrequent-prefix bias, it is unlikely that in the shortest analysis of an input corpus a string such as *love* will be actually treated as a prefix.

Of course, this bias only makes sense in a morpheme-discovery model, such as the one I am assuming, in which prefix- and suffix-discovery are conducted in separate stages (see Baroni 2000b:3.2.1.2 for discussion).

3.4 Finding the best solution

The discussion above has laid out the criteria by which the DDPL defines the optimum analysis of the data submitted to it. In addition, it is necessary to find an efficient procedure whereby this solution can be located, through generation and selection. For a detailed description of the algorithm used to generate alternative lexical analyses of the input corpus, see Baroni (2000b:3.3.9).

To summarize briefly, the algorithm is based on a greedy strategy similar to the one proposed by Brent and Cartwright (1996) for sentence segmentation. A lexicon constructed with *n* morphological splits is only evaluated if, of those *n* splits, *n-1* are identical to the ones that were used to generate the shortest lexicon constructed with *n-1* splits. In other words, in the exploration of possible, increasingly complex morphological analyses, morpheme breaks can only be assigned, never removed.

Moreover, the DDPL lexicon generation strategy uses a series of heuristics to further constrain the set of lexical analyses to be evaluated. In particular, the number of lexica to be evaluated is strongly reduced by the requirement that only word-initial strings that frequently occur in the input before relatively long word-final strings can be treated as prefixes in a DDPL lexicon.

The alternative lexical analyses generated in this way are compared, and the one allowing the shortest lexicon + encoding is selected as the best analysis. The length of a lexicon and the corresponding encoding are estimated using the following formula:

(21) *Formula for estimating lexicon + encoding length*

$$dl = \sum_{ent \ entries} length(ent) + 2|stem\_entries| + 2.25|prefix\_entries| +$$

$$+ |stem\_occurrences| + 2.25|prefix\_occurrences|$$

dl:	description length;
ent entries:	any entry, prefix or stem, in the DDPL lexicon;
length(ent):	length in characters of an entry;
|*stem_entries*|:	total number of entries in the stem list;
|*prefix_entries*|:	total number of entries in the prefix list;
|*stem_occurrences*|:	total number of occurrences of all stem entries in the input corpus;
|*prefix_occurrences*|:	total number of occurrences of all prefix entries in the input corpus.

See Baroni (2000b:3.3.8) for an explicit derivation of (21). In brief, the need to minimize the first three terms of (21) favors distributionally motivated decompositions (as such decompositions will reduce the number of entries in the prefix and stem lists, and they will reduce the lengths of lexical entries), whereas the last two terms will disfavor decompositions (after each decomposition, the total number of occurrences of entries in the encoding increases). The extra weights added to the two prefix-specific factors will insure that, all else being equal, more distributional evidence is needed to postulate a prefix than to postulate a stem.

Given that the formula in (21) computes the length of a certain analysis as if it were represented in the lexicon + encoding format, the very same analyses that are selected by the shortest lexicon + encoding criterion (and, thus, by the morpheme discovery heuristics presented in section 3) are also selected when using (21) to select the best lexicon (to be

precise, the encoding scheme from which (21) is derived differs from the one described above in that it assigns an extra-character to each entry in a lexical analysis, in order to mark which entries are prefixes and which entries are stems).

4. DISCOVERING ENGLISH PREFIXES WITH THE DDPL MODEL

The performance of DDPL was tested with a corpus of untagged orthographically transcribed English words from the PHLEX database (Seitz, Bernstein, Auer and MacEachern 1998) as its input. In this section, I will present the results of the simulation.

The PHLEX database contains, among other word lists, a list of the 20,000 most common word types in the Brown corpus (Kucera and Francis 1967), together with their frequency of occurrence in the Brown corpus. I removed from this list all word-types containing non-alphabetic symbols (digits and diacritics such as -). This trimming yielded a set composed of 18,460 word types. The input corpus for the DDPL simulation was generated by multiplying each of the types in this set by its frequency in the Brown corpus. For example, the word *kindergarten* has a frequency of 3 in the Brown corpus, and thus it occurred three times in the DDPL input used for the simulation presented here. In total, the corpus generated in this way contained 959,655 orthographically transcribed word tokens.

4.1 Prefixes discovered by the DDPL

Given the input described in the previous section, DDPL generated an output lexicon containing the following 29 prefixes:

(22) *prefixes postulated by DDPL*

ad-	auto-	co-	com-	con-	cor-	de-	dis-	ex-
extra-	juris-	in-	inter-	man-	mis-	non-	over-	para-
pre-	psycho-	radio-	re-	sub-	sup-	super-	sur-	tele-
un-	under-							

A first inspection of this list shows that DDPL was quite successful and accurate in finding (almost) only actual English prefixes.

Notice that *com-*, *con-* and *cor-* are actually allomorphs of the same prefix, and so are *sub-* and *sup-* (cf. *suppress*). The purpose of the DDPL model is simply to find the list of strings that correspond to prefixes (and stems) of a language. The model does not attempt to group strings that are allomorphs of the same morpheme into the same entry. I do not think that allomorph grouping is a task which should be performed on the sole basis of distributional cues, as syntactic, semantic and phonological cues would, obviously, be of great help.

It is interesting to observe that the DDPL model was able to find prefixes that constitute substrings of other prefixes: *co-* is a substring of *com-/con-/cor-*; *ex-* is a substring of *extra-*; *in-* is a substring of *inter-*; *un-* is a substring of *under-*.

The list contains only one obvious false positive: the string *man-*, which is a full noun, not a prefix. As *man-* often occurs as the first member of compounds, and DDPL does not have access to (morpho-)syntactic information, the model mistakenly treated forms such as *manservant* and *manslaughter* as prefixed.

Besides *man-*, there are three ambiguous cases: The strings *juris-*, *radio-* and *psycho-* are not classified as prefixes in standard references such as Marchand (1969) or Quirk, Greenbaum and Svartvik (1985) (the Merriam-Webster's dictionary classifies *radio-* and *psycho-* as "combining forms"). Thus, we should perhaps count them as false positives. However, as these strings correspond to bound word-initial units associated with

specific semantic features, even if they might not be prefixes under some definition of what a prefix is, they are "prefix-like" enough that I am reluctant to classify them as "real" false positives.

Interestingly, the DDPL did not classify any frequent but linguistically insignificant word-initial string (strings such as *pa-*, *pr-*...) as a prefix.

It seems legitimate, I think, to claim that the DDPL was very accurate in avoiding false positives.

On the other hand, the following 19 prefixes listed in Quirk *et al.* (1985) were not discovered by DDPL (notice that *il-*, *im-* and *ir-* are allomorphs of *in-*, a prefix that *was* found by DDPL):

(23) *prefixes missed by DDPL*

a-	an-	anti-	arch-	contra-	counter-	fore-
hyper-	il-	im-	ir-	mal-	mini-	out-
post-	pro-	pseudo-	trans-	ultra-		

Neo-classical prefixes (such as *hemi-* and *paleo-*) and conversion prefixes (such as *en-* and *be-*) listed in Quirk *et al.* (1985) but missed by DDPL are not reported in (23). I believe that misses from these classes are not problematic, as they concern prefixes that are very cultivated and/or not very productive.

Of the set in (23), the following 6 misses are due to the nature of the input corpus, which did not contain enough forms to motivate their treatment as prefixes:

(24) a- an- arch- hyper- mini- pseudo-

The string *hyper-* never occurs in the corpus, whereas the other five strings in (24) only occur in two or fewer words in which they function as prefixes (even counting completely semantically opaque, highly lexicalized prefixed forms).

Of the remaining 13 misses, the following 9 are due to the heuristic constraint on lexicon generation mentioned in 3.4 above, according to which only word-initial strings occurring a certain number of times before long word-final strings can be evaluated as candidate prefixes:

(25) counter- fore- il- im- ir- mal- out- post- ultra-

None of these prefixes occurs frequently enough before long word-final strings in the input to avoid being filtered out by this constraint.

This leaves us with 4 unexplained misses:

(26) anti- contra- pro- trans-

The input corpus contains several truly prefixed forms displaying these prefixes in combination with independently occurring stems (although the prefixes *contra-* and *pro-* only occur in lexicalized formations with bound stems, such as *contraception* and *proceed*). Thus, these misses cannot be attributed to the nature of the input. I plan to explore the issue of why DDPL failed to discover these prefixes in future research.

To conclude, the list of prefixes found by DDPL is more accurate than exhaustive, i.e., we can say that the model is better in terms of precision than in terms of recall. On the one hand, the false positives in the list are few and linguistically motivated. On the other hand, even if the prefixes in (24) are excluded from the count, DDPL still missed 13 productive English prefixes.

In particular, most of these misses are due to the constraint on lexicon generation requiring word-initial strings to occur a certain number of times before "long" word-final

strings. This suggests that the first step in future revisions of DDPL should concentrate on the lexicon generation component of the model.

4.2 Assessing the model against native speaker intuition

Besides finding a list of prefixes, DDPL assigns morphological parses (prefixed vs. monomorphemic) to all the words in the input corpus. Of course, only the parses assigned by DDPL to potentially prefixed words, i.e., words beginning with a string identical to one of the prefixes found by the algorithm, are of interest, as all other input words are treated as monomorphemic.

Assessing the plausibility of the parses assigned by DDPL to potentially prefixed words is not a trivial task, as the morphological status of many potentially prefixed words is not clear. While probably everybody would agree that the word *redo* is prefixed and the word *red* is not, there are many intermediate cases (such as *resume, recitation, remove*) about whose status morphologists would probably disagree.

Thus, rather than trying to decide on my own which of the parses assigned by DDPL were "right" and which ones were "wrong", I conducted a survey in which I collected morphological complexity ratings from native English speakers, and then I tested whether the parses assigned by DDPL to the same words could predict the distribution of such ratings (see Smith 1988, Wurm 1997 and Hay 2000 for other studies using morphological complexity ratings). The idea was to see if speakers' intuitions on the prefixed status of such words would agree with the parses assigned by the algorithm.

4.2.1 Word list construction, methodology and data collection

All the words in the survey corpus begin with a string corresponding to one of the prefixes postulated by DDPL, but half of the words were selected from the set of forms that were treated as complex by the model, the other half from the set of forms which, although they begin with a string identical to a prefix, were *not* treated as complex by the model.

In particular, the survey corpus contained 300 forms that were randomly selected from the words in the DDPL output that began with one of the prefixes postulated by the algorithm (excluding *man-*, *juris-* and *radio-*). 150 of these forms were randomly selected from the set of words that DDPL treated as prefixed. The other 150 forms were randomly selected from the set of words that DDPL treated as monomorphemic (non-prefixed).

The complex-for-DDPL set contained 22 distinct prefixes, the simple-for-DDPL set contained 21 distinct prefixes. The two sets shared 17 prefixes. The average length of the words in the complex-for-DDPL set was 9.9 characters, the average length of their potential stems was 7.1 characters. The average length of the words in the simple-for-DDPL set was 9 characters, the average length of their potential stems was 6.5 characters. The average frequency of the words in the complex-for-DDPL set was 3.6, the average frequency of their potential stems was 192.9. The average frequency of words in the simple set was 25.7, the average frequency of their potential stems was 189.2. All words in the complex set had potential stems occurring as independent strings in the corpus; 101 words in the simple set had potential stems that did not occur as independent strings.

The word lists used in this and the following survey, together with DDPL parses, average speaker ratings for each word and other statistics are available from http://sslmit.unibo.it/~baroni.

A group of eight native English speakers were asked to rate the set of potentially prefixed words ("potentially prefixed" in the sense that they begin with a word-initial string identical to a prefix) on a scale from 1 to 5, assigning 1 to words that they definitely felt to be non-prefixed, 5 to words that they definitely felt to be prefixed.

In the instructions, the participants were presented with an example of a word that should receive a 1-rating (the word *cocoon*) and an example of a word that should receive a 5-rating (the word *coexist*).

The 300 words in the survey corpus were presented in a different random order to each participant. Words were presented in list format. To avoid possible ambiguities due to the fact that some of the prefixes under investigation are substrings of other prefixes (e.g., *in-* is a substring of *inter-*), each word in the list was followed by the potential prefix that participants were to consider.

Participants were given unlimited time to complete the task, but they were asked to write down their ratings as quickly as possible, and to avoid revising a rating once they had written it down.

Clearly, in order to take part in the survey, participants had to be familiar with basic notions of morphology (at least, the notions of prefix and prefixed form). Thus, I selected the participants among undergraduate and graduate linguistics students (the participants in the surveys, however, were not aware of the goals of the study).

4.2.2 Results and discussion

The rating patterns of all eight participants were highly correlated (the Pearson correlation coefficients computed pairwise for each pair of raters were always higher than .55; the Spearman correlation coefficients computed pairwise for each pair of raters were always higher than .6.). Thus, I computed the per-word average rating across all participants, and I conducted a one-way ANOVA in which these ratings were grouped on the basis of the corresponding DDPL parses (simple vs. complex). The difference between ratings assigned by native speakers to words treated as simple by DDPL and ratings assigned to words treated as complex by DDPL is highly significant ($F(1,298) = 209.7$, $p < .0000$).

Thus, we can conclude that, besides being able to find a number of actual English prefixes, DDPL also assigned plausible morphological parses to potentially prefixed words.

Of course, while the relation between DDPL and the speakers' ratings is significant, it is by no means "perfect", as there are some discrepancies between the speakers' ratings and the DDPL parses.

Several reasons can explain these discrepancies. First, as there are individual differences in morphological intuitions among speakers (indeed, the ratings of some of the speakers appear to be less correlated with each other than DDPL and the speakers' average), even if DDPL were a perfect model of human morpheme discovery, we should not expect a 100% correlation between its parses and an average of human intuitions.

More importantly, the DDPL is intended to model a hypothesized early stage of language acquisition, in which morpheme discovery is performed. However, the speakers who participated in the survey are adults who successfully completed the task of morphological acquisition, and are aware of the semantic and syntactic properties associated with prefixes and stems.

From this perspective, it is actually surprising that the purely distributionally-driven parses assigned by DDPL are as well correlated with adult speakers' ratings as the results indicate.

Interestingly, the discrepancies between DDPL and English speakers appear to be attributable to the fact that DDPL is too "conservative", i.e., DDPL was more likely to treat obviously prefixed forms as simple than obviously simple forms as complex. While it is easy to find obvious misses among the forms treated as simple by DDPL (*unconsciously, distrust, subgroups, unavoidable*…), only two of the forms treated as complex by DDPL are obviously non-prefixed (*comin* -- probably a spelling of the colloquial form of *coming* -- and *constable*).

Indeed, the average mean rating across all forms that were treated as complex by DDPL is a rather high 4.05 (recall that speakers had to rate forms on a scale from 1 to 5,

assigning 5 to clearly prefixed forms). This indicates that in general speakers largely agree with DDPL on the status of forms that the algorithm treated as complex. On the other hand, the average mean rating across all forms that were treated as simple by DDPL is 2.11. This is still lower than chance level, but does suggest that there was less overlap between DDPL parses and speakers' intuitions in the domain of forms that are treated as simple by the computational model.

As with the list of prefixes found by DDPL, what emerges here is that the analysis generated by the model is quite accurate (very few "false positives") but not exhaustive (many "misses"). Precision is high, recall is low.

I observed in the introduction that the purpose of assigning parses to potentially complex words in morpheme discovery is to have a set of forms to analyze in order to discover the semantic and grammatical properties of affixes. In this perspective, it seems that morpheme discovery should indeed favor precision over recall: a relatively small set of words containing a certain prefix is probably more helpful, in identifying the properties of that prefix, than a larger set that also includes many pseudo-prefixed forms (see Snover and Brent 2001 for a similar line of reasoning).

5. TREATMENT OF SEMANTICALLY OPAQUE WORDS

The analysis of the results of the DDPL simulation shows that distribution-driven principles such as the ones implemented by this model can be quite helpful in morpheme discovery, both in terms of finding the prefixes of a language and in terms of assigning morphological parses to words.

The success of this computational simulation constitutes evidence against the claim that children cannot *in principle* learn something about morphology from distributional evidence, the claim being that distributional evidence would not provide enough useful cues. Clearly, even the relatively simple distributional cues used by DDPL might be of great help to language learners.

However, this does not *per se* constitute evidence that humans *do* rely on such cues, since humans, unlike the automated learner, could have used different types of evidence -- most plausibly, semantic evidence -- in order to discover the same structures found by the automated learner. For example, the learner, exploiting distributional cues only, came to the conclusion that *renamed* is a prefixed word, composed of the prefix *re-* and the stem *named*. Although all the native speakers surveyed shared the intuition that *renamed* is indeed a prefixed form composed of *re-* and *named*, this convergence between the automated learner and humans does not prove that humans exploited distributional cues in morpheme discovery, since humans could have decided that *renamed* is prefixed simply on the basis of its meaning.

However, the comparison of the output parses assigned by the automated prefix learner to English words with morphological intuitions of native speakers can potentially provide a form of more direct empirical evidence supporting the hypothesis that learners resort to distributional cues in morpheme discovery. This evidence would emerge from the analysis of semantically opaque but potentially morphologically complex words such as *recitation* or *remain*. Words of this kind are potentially prefixed, at least in the sense that they begin with a string identical to a prefix (*re-*, in this case). However, the meaning of *recitation* is not synchronically related to the meaning of the prefix *re-* nor to the meaning of the stem *citation* (or *cite*). In the case of *remain*, not only is the meaning of the word not related to the meaning of the components, but it is not even clear that the potential stem, the bound verbal form *-main*, is associated with any semantic content.

However, several experimental studies have shown that speakers do treat some of these semantically opaque forms as morphologically complex. For example, the following studies have presented evidence from a variety of experimental tasks that speakers are aware of the morphological structure of words that are (partially or completely)

semantically opaque: Emmorey (1989), Bentin and Feldman (1990), Feldman and Stotko (unpublished -- quoted in Stolz and Feldman 1995), Roelofs and Baayen (1997), Baroni (2001) and Baayen, Schreuder and Burani (2001).

Now, if it turned out that there is a convergence between the parses assigned by the distribution-driven learner and the speakers' intuitions about semantically opaque forms, then this would constitute a stronger form of evidence in favor of the hypothesis that speakers used distributional cues to assign morphological structure to words.

For example, if it turned out that both the automated learner and the speakers treated *recitation* as morphologically complex (*re+citation*), but *remain* as monomorphemic, then it would be reasonable to conclude that speakers are sensitive to distributional cues similar to the ones implemented in the automated learner, since they could not have assigned a morphological structure to *recitation* on the basis of its meaning (and, also, it is unlikely that they could have used syntactic or phonological cues to distinguish *recitation* from *remain*).

Indeed, the results of the second survey I conducted show that, even when only semantically opaque words are considered, there is a significant correlation between the parses assigned by the learner and speakers' intuitions. Thus, this study provides strong support for the claim that humans use distributional cues in morpheme discovery.

Notice that this type of evidence in favor of distribution-driven learning is not available in other domains. For example, even if it has been shown that distributional cues can be very effective for segmenting utterances into words (Brent and Cartwright 1996), there is no clear equivalent to semantically opaque morphemes in the domain of syntactic segmentation.

In particular, idiomatic phrases such as *kick the bucket* are not the equivalent of semantically opaque morphologically complex forms such as *permit*. First, idiomatic phrases also have a literal, semantically transparent meaning, and it is unlikely that speakers are not aware of this meaning. Second, words occurring in idioms also occur in non-idiomatic sentences. This is not the case of a bound stem like *-mit*, which occurs only in opaque forms.

5.1 Constructing the semantically opaque word list

Following a standard practice in morphological processing studies (see, for example, Marslen-Wilson, Tyler, Waksler and Older 1994), I first conducted a survey in which three judges were asked to rate a set of forms from the DDPL output for semantic transparency, and then I selected forms that received a low average semantic transparency rating to construct the survey corpus.

The DDPL output contains a total of 3,651 forms beginning with strings corresponding to one of the prefixes postulated by the model. Of these, 382 are actually treated by the model as prefixed. Clearly, it was not feasible to ask the semantic transparency judges to assign a rating to all 3,651 forms. Thus, the corpus presented to the judges was constructed in the following way.

First, I made a preliminary division of the 382 words treated as prefixed by DDPL into two categories: words that I judged to be obviously prefixed (productively formed, semantically transparent), and words that may or may not be prefixed (this preliminary list included a wide range of types, from obviously non-prefixed words such as *adage* to only slightly lexicalized forms such as *inhumane*). The first list was composed of 101 words, the second list of 181 words. I randomly selected 10 words from the first list, and I kept all the 181 words from the second list.

From the list of the remaining 3,269 words treated as simple by DDPL, I then randomly selected 10 more words that were obviously prefixed and completely transparent, and 200 words that may or may not be prefixed.

The corpus presented to the three judges was composed of the 20 completely transparent words and 381 "ambiguous" words selected in this way. The 20 completely transparent words served both as a control and, more importantly, they were added in order to try to minimize the risk that judges would assign high ratings to some semantically opaque forms merely in order to make use of the whole range of the rating scale.

The judges were two graduate students and one postdoctoral fellow in the UCLA Linguistics Department, and were selected because of their strong background in morphology and morphological processing. I selected expert judges because I wanted to make sure that they would understand the task, and in particular that they would understand the distinction between rating forms on the basis of semantic transparency vs. morphological complexity.

Judges were asked to rate the words in the corpus on a scale from 1 to 5, assigning 1 to completely opaque words and 5 to completely transparent words.

A series of correlation analyses showed that the judges' ratings were highly correlated (both Pearson and Spearman correlation coefficients in all pairwise comparisons were higher than .7). Thus, I computed the average cross-judge rating for each word in the corpus.

As expected, the 20 transparent words received very high ratings (the mean rating for this set of words was 4.89). Of the remaining forms, 97 out of the 181 words treated as prefixed by DDPL received an average rating lower than 2.5; 183 out of the 200 words treated as simple by DDPL received an average rating lower than 2.5. Notice the asymmetry between the two sets: just a little more than half of the complex-for-DDPL words that were pre-selected as potentially opaque are indeed semantically opaque, whereas 90% of the simple-for-DDPL words that were pre-selected as potentially opaque are indeed semantically opaque. This suggests that, although DDPL did not have access to semantic information, the model did show a preference for treating semantically opaque words as simple. This is good from the point of view of a general assessment of the DDPL performance, but it made it harder to design the survey presented here.

The corpus for the second survey was thus composed of the 97 complex-for-DDPL forms that had a semantic rating lower than 2.5, and 97 randomly selected words from the 183 simple-for-DDPL words with a semantic rating lower than 2.5. I decided not to add a control set of semantically transparent forms, as I wanted to maximize the participants' sensitivity to differences in morphological status among opaque words. If some semantically transparent words had been inserted, speakers would have probably reserved the high values of the rating scale for such forms, "squeezing" the ratings of semantically opaque words within a narrow range at the bottom of the scale.

The average semantic rating across the complex-for-DDPL forms in this list was 1.54; the average rating across the simple-for-DDPL forms in this list was 1.21. One of the judges was also asked to rate the 194 forms in the corpus by assigning ratings on a 5 point scale on the sole basis of the degree of semantic transparency of the potential *prefix* of each form. The average prefix transparency rating across forms treated as complex by DDPL was 1.86; the average prefix transparency rating across forms treated as simple by DDPL was 1.46. Thus, while there is a noticeable and slightly worrisome difference in the degree of prefix transparency between the two sets, it seems safe to state that not only the forms in both sets are semantically opaque when considered as wholes, but also that the potential prefixes occurring in them tend to be opaque.

The complex-for-DDPL set contained 17 distinct prefixes, the simple-for-DDPL set contained 16 distinct prefixes. The two sets shared 14 prefixes. The average length of the words in the complex-for-DDPL set was 9 characters, the average length of their potential stems was 6.4 characters. The average length of the words in the simple-for-DDPL set was 8.6 characters, the average length of their potential stems was 6.3 characters. The average frequency of the words in the complex-for-DDPL set was 3.6, the average frequency of their potential stems was 263.4. The average frequency of words in the simple set was 21.1, the average frequency of their potential stems was 102.9. One word in the complex

set had a potential stem that did not occur as an independent string in the corpus; 77 words in the simple set had potential stems that did not occur as independent strings.

5.2 Methodology and data collection

The same methodology and data collection procedure described in section 4.2.1 above was followed in the second survey.

A group of eight English native speakers, all graduate or undergraduate students or post-doctoral fellows in linguistics, took part in the survey. None of them had participated in the previous survey.

5.3 Results and discussion

Pairwise Pearson and Spearman correlation coefficients were computed for the ratings of all pairs of participants. The patterns of three participants were poorly correlated with those of the other participants and with each other (for each of these three participants, the correlation coefficient between her/his ratings and those of a majority of other speakers was lower than .4). Thus, their data were discarded.

As the ratings of the remaining participants were highly correlated (all pairwise Pearson and Spearman coefficients were higher than .5), the per-word average rating value across them was computed, and the resulting variable was compared to the parses assigned by DDPL to the same words in a one-way ANOVA in which the average ratings were grouped on the basis of the DDPL parses (simple vs. complex). The results of the ANOVA indicates that, in this case as well, the difference between ratings assigned by native speakers to words treated as simple vs. complex by DDPL is highly significant ($F(1,192) = 49.2$, $p < .0000$).

If the participants in the survey had mostly relied on semantic cues when assigning ratings to the words in the list, they should have assigned uniformly low ratings to all words. However, this was not the case: as shown by the correlation between the average ratings and DDPL parses, in general speakers assigned higher ratings to words that DDPL treated as complex, lower ratings to words that DDPL treated as simple. The average mean rating across words that were complex for DDPL was 3.78; the average mean rating across words that were simple for DDPL was 2.81.

The most plausible explanation for this asymmetry is that the way in which speakers represent potentially complex words is affected by distributional factors such as the ones implemented in DDPL.[8] In turn, a plausible hypothesis about why such distributional factors have an effect on speakers' morphological intuitions is that speakers relied on distributional cues during morpheme discovery.

On the other hand, adult speakers are obviously also sensitive to semantic cues, when rating words for morphological complexity. As all the words in the survey corpus were semantically opaque, it is not surprising that the results of this second survey are less clear-cut than those of the previous survey (as shown by the fact that this time there is less of a difference between the average mean ratings assigned to DDPL simple and complex words).

I suspect that semantics influenced the results both directly and indirectly. First, the morphological representations of adult speakers are almost certainly affected by the semantic structure of words. Thus, while speakers seem to distinguish words that are complex on purely distributional grounds from simple words, it is likely that such words occupy a middle ground, in terms of morphological complexity, between simple words and semantically transparent words (see Seidenberg and Gonnerman 2000 for similar considerations). Indeed, if no correlation between DDPL and the speakers had emerged, we could not have been sure that the negative result was due to the fact that speakers do not

rely on distributional cues such as the ones employed by DDPL during morpheme discovery. The negative result could have instead been due to the fact that, once speakers acquire sufficient evidence about the semantic properties associated with morphemes, they revise their morphological representation of forms, and they change (from complex to simple) the representation of those forms that were originally treated as complex on distributional grounds, but whose complex representation is not supported by semantic evidence.

Moreover, as a consequence of the fact that the distinction between semantically opaque but complex forms and simple forms is probably not as clear-cut as the distinction between complex and transparent words and simple words, the participants in the second survey had to provide ratings based on more subtle judgments, requiring more sophisticated metalinguistic introspection skills. Thus, as this was a harder task, it is likely that the participants in the second survey had more difficulty with it than the participants in the first survey, and that the less marked difference between sets is in part due to "noise" in the ratings.

However, beyond these considerations, what is truly important from our point of view is that there *is* a high correlation between DDPL parses and the speakers' ratings of semantically opaque words. Thus, the survey results provide support for the hypothesis that humans are sensitive to distributional cues to morphological constituency such as the ones used by DDPL.

6. CONCLUSION

The results of the simulation reported above provide support for the general hypothesis that distributional information of the kind encoded in the DDPL model can in principle be helpful in morpheme discovery. Moreover, the convergence between the DDPL parses and speakers' ratings of a set of semantically opaque words provides some preliminary support for the hypothesis that humans rely on distributional cues such as the ones employed by the automated learner when assigning morphological parses to words. A plausible explanation of this finding is that speakers are sensitive to such cues because they employed them in order to assign morphological parses during morpheme discovery.

Moreover, these results are also potentially relevant to the theory of morphological processing, in that they could provide the basis for a (partial) explanation of the fact that, as various psycholinguistic studies have shown, speakers treat some semantically opaque words as morphologically complex: They do so because, during morpheme discovery, they used distributional schemes to search for the morphemes of their language, and these schemes lead them to analyze some words as morphologically complex even in the lack of semantic cues supporting the complex analysis.

Clearly, while I believe that the results presented here are encouraging, many questions are still open, and much more research has to be done before we can reach safe conclusions about the nature and role of distributional evidence in morpheme discovery.

The DDPL model could be improved and extended in various ways. Obviously, the model should be extended to suffixation and other types of affixation. Furthermore, algorithms in which the distributional information used by DDPL is integrated with other types of information (such as syntactic category information) could be developed. Also, alternative lexicon generation algorithms, exploring a larger (or, better, more morphologically sensible) area of the hypothesis space, should be investigated.

The reviewers pointed out recent work by Jennifer Hay (see, e.g., Hay 2000) suggesting that what matters in morphological processing is not the *absolute* frequency of derived forms, but the *relative* frequency of derived forms and their bases. In short, if a potentially complex form is more frequent than its potential base, the form is more likely to be parsed as a whole, whereas, if the base is more frequent than the complex form, then the complex form is more likely to be decomposed. In this setting, the absolute frequency

heuristic used by DDPL can be seen as an approximation of a more realistic relative-frequency-based heuristic. In future research, it will be extremely interesting to test a revised version of the model that takes relative frequency effects into account.[9]

Finally, a reviewer also suggested that it would interesting to develop a version of DDPL that returns values on a continuous complexity scale, rather than binary complex vs. simple parses. This would allow a more direct comparison with human ratings, and it would correspond, perhaps, to a more realistic model of human morphological processing (see, e.g., Seidenberg and Gonnerman 2000 and Baroni 2001 for arguments in favor of gradient morphological representations).

From the point of view of testing the model, we should first of all test DDPL in simulations with other English corpora, both in orthographic and phonetic transcriptions. Furthermore, DDPL should be tested using input corpora from other languages.

In terms of collecting empirical evidence, we should first of all collect data from more speakers, possibly re-designing the survey task in order to make it feasible for speakers with no linguistics background. Furthermore, it would be interesting to collect data using other methods (for example, using a morphological priming paradigm), to make sure that the results we obtained are task-independent. Obviously, it would also be important to collect developmental data from children, to have a more concrete idea of when and how human learners perform morpheme discovery.

Last but not least, a more sophisticated analysis of the empirical results obtained should try to assess whether all the cues exploited by DDPL are relevant in predicting the response patterns of the speakers, and/or what is their relative importance as predictors.

While all these lines of research should be pursued in the near future, and I am sure that readers will raise other important issues that were not dealt with here, I believe that the current results are already shedding some (weak) light on the role of distributional cues in the domain of morpheme discovery.

NOTES

[1] I would like to thank Donca Steriade, Adam Albright, Lynne Bernstein, Harald Baayen, Amy Schafer, Kie Zuraw, the reviewers for the Yearbook of Morphology and, especially, Ed Stabler, Carson Schütze and Bruce Hayes for help and advice. Of course, none of them is responsible for any of the claims I make. A more detailed discussion of several of the issues discussed here can be found in Baroni (2000b), which is downloadable from http://sslmit.unibo.it/~baroni.

[2] In this study, I illustrate my points through examples presented in orthographic transcription. The same points could have been also illustrated by the same examples (or similar ones) presented in phonetic transcription. A preliminary experiment with a corpus of phonetically transcribed words suggests that, because of the different distributional properties of specific morphemes in spoken and written language, the morpheme-discovery algorithm presented here performs in a similar but slightly different way when presented with orthographic vs. phonetically transcribed input. See Baroni (2000b:4.6) for discussion.

[3] See also Brent (1993) and Brent, Murthy and Lundberg (1995). The model of Brent and colleagues represents, as far as I know, the first attempt to apply the Minimum Description Length principle to the problem of morpheme discovery

[4] In Baroni (2000b), I motivate and defend the assumptions about morpheme discovery that went into the design of the algorithm described here, i.e., that it makes sense to model morpheme discovery as a separate task from utterance segmentation, that it makes sense to model prefix discovery as a separate subtask within morpheme discovery, and that it makes sense to consider an approach to the task in which only binary (prefix+stem) parses of words are evaluated.

[5] Baroni (2000b) discusses further how the lexicon + encoding criterion reflects morphological heuristics (including, among other things, a discussion of how the heuristics interact and of how such interaction insures that the "frequent substrings are likely to be morphemes" heuristic is interpreted in terms of type rather than token frequency).

[6] All else being equal, in the data compression scheme proposed here longer substrings are more likely to constitute independent lexical entries than shorter substrings. For example, at the same frequency of occurrence in the input corpus, a substring like *dis-* is more likely to be treated as an independent entry than a substring like *a-*. Again, I would like to claim that this also makes sense from the point of view of morpheme discovery.

[7] To keep things simple, I presented here an example in which stems occur elsewhere in the corpus as independent words -- i.e., they are free stems. However, the same pattern takes place even if the relevant stems never occur in independent words, but are the product of the parse of other prefixed forms -- i.e., they are bound stems.

[8] See Baroni (2000b:4.5.3) for a *post-hoc* analysis that seems to rule out the possibility that the asymmetry can be explained by phonological cues.

[9] Interestingly, the earlier model presented in Baroni 2000a did take relative frequency effects into account, at least to a certain extent, by assigning shorter indices to more frequent lexical entries, thus making the likelihood that a form will be parsed as complex dependent not only on the frequency of the form itself, but also on the frequency of its potential base.

REFERENCES

Baayen, Harald. 1994. "Productivity in language production". *Language and Cognitive Processes* 9, 447-469.

Baayen, Harald; and Lieber, Rochelle 1991. "Productivity and English derivation: A corpus-based study". *Linguistics* 29, 801-843.

Baayen, Harald; Schreuder, Robert; and Burani, Cristina. 1997. "Parsing and semantic opacity". in *Morphology and the mental lexicon*, E. Assink and D. Sandra (eds), Dordrecht: Kluwer. In press.

Baroni, Marco. 2000a. "Using distributional information to discover morphemes: A distribution-driven prefix learner". Paper presented at the LSA Meeting, Chicago.

Baroni, Marco. 2000b. *Distributional cues in morpheme discovery: A computational model and empirical evidence*. UCLA dissertation.

Baroni, Marco. 2001. "The representation of prefixed forms in the Italian lexicon: Evidence from the distribution of intervocalic [s] and [z] in northern Italian". *Yearbook of Morphology 1999*, 121-152.

Bentin, Shlomo; and Feldman, Laurie. 1990. "The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew". *Quarterly Journal of Experimental Psychology* 42A, 693-711.

Brent, Michael. 1993. "Minimal generative explanations: A middle ground between neurons and triggers". *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, 28-36.

Brent, Michael; and Cartwright, Timothy. 1996. "Distributional regularity and phonotactic constraints are useful for segmentation". *Cognition* 61, 93-125.

Brent, Michael; Murthy, Sreerama; and Lundberg, Andrew. 1995. "Discovering morphemic suffixes: A case study in minimum description length induction". Presented at the *Fifth International Workshop on AI and Statistics*.

Emmorey, Karen. 1989. "Auditory morphological priming in the lexicon". *Language and Cognitive Processes* 4, 73-92.

Feldman, Laurie (ed). 1995. *Morphological aspects of language processing*. Hillsdale: LEA.

Goldsmith, John. 2001. "Unsupervised learning of the morphology of a natural language". *Computational Linguistics* 27, 153-198.

Harris, Zellig. 1955. "From phoneme to morpheme". *Language* 31, 190-222.

Hay, Jennifer. 2000. *Causes and consequences of word structure.* Northwestern University dissertation, 2000.

Kucera, Henry; and Francis, Nelson. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

Marchand, Hans. 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. Munich: Beck.

Marslen-Wilson, William; Tyler, Lorraine; Waksler, Rachelle; and Older, Lianne. 1994. "Morphology and meaning in the English mental lexicon". *Psychological Review* 101, 3-33.

Maxwell, Michael (ed). 2002. *Morphological and phonological learning: Proceedings of the sixth ACL-SIGPHON meeting.* Philadelphia: ACL.

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffry; and Svartvik, Jan. 1985. *A comprehensive grammar of the English language*. London: Longman.

Redington, Martin; and Chater, Nick. 1998. "Connectionist and statistical approaches to language acquisition: A distributional perspective". *Language and Cognitive Processes* 13, 129-191.

Rissanen, Jorma. 1978. "Modeling by shortest data description". *Automatica* 14, 456-471.

Roelofs, Ardi; and Baayen, Harald. 2001. "Morphology by itself in planning the production of spoken words". *Psychonomic bulletin and review*. In press.

Schreuder, Robert; and Baayen, Harald. 1994. "Prefix stripping re-revisited". *Journal of Memory and Language* 33, 357-375.

Schreuder, Robert; and Baayen, Harald. 1995. "Modeling morphological processing". In Feldman (1995), 131-154.

Seidenberg, Mark; and Gonnerman, Laura. 2000. "Explaining derivational morphology as the convergence of codes". *Trends in Cognitive Sciences* 4, 353-361.

Seitz, Philip; Bernstein, Lynne; Auer, Edward; and MacEachern, Margaret. 1998. *The PHLEX Database*. Los Angeles: House Ear Institute.

Smith, Philip. 1988. "How to conduct experiments with morphologically complex words". *Linguistics* 26, 699-714.

Snover, Matthew; and Brent, Michael. 2001. "A Bayesian model for morpheme and paradigm identification". *Proceedings of ACL 39*, 482-490.

Stolz, Jennifer; and Feldman, Laurie. 1995. "The role of orthographic and semantic transparency of the base morpheme in morphological processing". in Feldman (1995), 109-129.

Wurm, Lee. 1997. "Auditory processing of prefixed English words is both continuous and decompositional". *Journal of Memory and Language* 37, 438-461.