Syllables, Morphemes and Bayesian Computational Models of Acquiring a Word Grammar

Çağrı Çöltekin
Cognitive Science
Middle East Technical University (METU), Ankara 06531 Turkey
cagri@xs4all.nl

Cem Bozşahin
Cognitive Science and Computer Engineering
METU, Ankara 06531 Turkey
bozsahin@metu.edu.tr

Abstract

We report a computational study on the CHILDES database for learning a word grammar of Turkish nouns. The syllable-based model converges to a morpheme-based model in terms of overlaps in the set of lexical hypotheses. Morphology is a hidden variable in all models, and the search problem for hypotheses is narrowed down by a probabilistic conception of universal grammar à la Combinatory Categorial Grammar. The convergence of the syllable model suggests that morphemehood can be an emergent computational property.

Keywords: Morphology, grammar, learning, Bayesian model.

Introduction

How can the meaning and category of words arise in the mind of a child? On one hand, we have the problem of identifying segments of speech as word-like units. On the other, we have the problem of identifying which meanings go with which substrings in speech. The assumption, common to both generative and cognitive linguistics, is that the child has the innate capacity to associate forms with meaning, and it is a question of acquisition to tackle the problem of deciding which forms go with which meanings.

A quick glance over the Turkish fragment of the child-directed speech in the CHILDES database reveals that 44% of the nouns are uninflected; the remaining 56% are inflected by means of affixes and clitics. The question then arises as to how the meaning and category of the inflected words, which constitute the majority, are acquired by the child. A common concept, influential at least since Bloomfield, is that morpheme is the minimal meaning-bearing element in natural languages. Nevertheless, although there are clear phonological and prosodic cues for word boundaries (e.g. Jusczyk, 1999; Thiessen & Saffran, 2003), there are no apparent cues for morpheme boundaries, hence the task of learning morpheme meanings to come up with word meanings is not made easier by labeling some items as morphemes in the competence grammar of adults.[1]

There are indeed phonological and prosodic cues for discerning substrings smaller than words, namely syllables (rhythm), stress and pitch accents. In this work, we report a computational study which starts with the ability to identify syllables, and learns the meaning and category of words and morphemes without the assumption that only words and morphemes have a meaning. The kind of meanings that the system starts with and learns more of is not lexical meanings, such as what it means to be a *dog* or to *sleep* (see Tenenbaum & Xu 2000 for a Bayesian way to tackle that problem), but the combinatory meaning and its syntactic reflex in the form of a category, as a lexical hypothesis, for example how *pisi-ler-e* (kitty-PLU-DAT, Turkish), with the syllables *pi·si·le·re*, can come to be associated with a logical form such as $to'(plu'\,cat')$ and the syntactic type **N** for nouns (and others, such as **VP** modifiers). We show that under a Bayesian scenario of hypothesis revision with the Universal Grammar as the provider of likelihoods and priors, starting with syllables and the assumed ability to associate forms with meanings converges to a lexicalized grammar of words and morphemes, by showing a significant overlap with the lexical hypotheses of a learning model which works with the assumption that only morphemes and words constructed from them have meanings.

Crucially, morphology of words is a hidden variable in our model, and the input to the system are pairs of sequence of syllables (in lieu of phonological form, PF) and a logical form (LF), without any indication as to which syllable contributes to what part of the LF, or which part of the meaning of a morpheme is covered by a syllable. This is unlike the approach of Jack et al. (2006), another syllable-based acquisition model, in which a sequence of syllables is paired not with a possibly ambiguous LF but with a disambiguated representation of world meanings. We do not assume that the child knows *pisi* is kitty, and *ler* is plural; she might (wrongly) hypothesize *pisi* is plural and *ler* means kitty, or the first syllable of *pisi* (*pi*) means kitty, etc. We also differ from Aronoff et al. (2006), whose model detects frequently-occurring sound sequences and hypothesizes that they are morphemes. Our model aims to learn the correct LF of the purported morpheme as well, not just its form.

Universal Grammar

What allows our system to learn with reasonable efficiency is that the search problem for lexical hypotheses is kept manageable by a Universal Grammar (UG) and the current lexi-

[1] Aksu-Koc & Slobin (1985); Peters & Menn (1993) report production data of respectively Turkish and English children of age 2;6 and younger, during which the child produces meaningless filler syllables. Peters & Menn data show this is not idiosyncratic to verbs. Contra the remarks of both work for Turkish without a statistic, morpheme and syllable boundaries do not generally coincide. Only 23% of the syllables in nouns (out of 20,433 syllables) are also morphemes in the CHILDES database. If we only match boundaries (the beginning and end of a morpheme align with a syllable boundary, irrespective of whether the syllable and the morpheme are the same), e.g. *araba-lar* (car-PLU, Turkish) versus the syllables *a·ra·ba·lar* providing two matches out of 4 syllables, the overlap is 57%.

calized grammar, and that the input words are relatively short compared to adult input, so that considering all possibilities of syllable-LF associations is tolerable computationally, as suggested by Steedman & Hockenmaier (2007) for learnability of short utterances involving multiple words.[2]

We shall assume that UG comprises a set of principles and a set of universal combinatory rules, which are completely type-dependent, rather than structure-dependent. Steedman (2000) shows how Combinatory Categorial Grammar (CCG) can fulfill that role, which gives CCG its explanatory edge compared to structure-dependent accounts, to explain the so-called nonstandard constituency in bounded and unbounded dependencies such as in coordination and relativization asymmetries, to integrate intonation structure, information structure and constituent structure as arising from the same derivational system of projecting (PF, LF) pairs from the lexicon to phrases. Steedman & Hockenmaier (2007) show how CCG can bootstrap and facilitate learning a lexicalized grammar of a natural language, with examples involving the use of words from the stage of 2-word syntax onwards.

We will confine ourselves to a few principles related to our work, and to one combinatory rule that seems most relevant to acquiring a word grammar, viz. function application. It is defined as follows:

(1) a. Forward Application:
$$\mathbf{X/Y}\text{: } f \quad \mathbf{Y}\text{: } a \quad \Rightarrow \quad \mathbf{X}\text{: } fa \qquad (>)$$

b. Backward Application:
$$\mathbf{Y}\text{: } a \quad \mathbf{X\backslash Y}\text{: } f \quad \Rightarrow \quad \mathbf{X}\text{: } fa \qquad (<)$$

The elements of a CCG lexical hypothesis are:

$$(2)\quad \overbrace{\textit{like}}^{string}\ \ \underbrace{:=}_{\substack{string\\ type}}\ \ \overbrace{\overbrace{\mathbf{(S\backslash NP)/NP}}^{syn.\,type} :\ \overbrace{\underbrace{\lambda x\lambda y.}_{correspondence}\ \ \underbrace{like'_{\underbrace{(e,(e,t))}_{sem.type}}\,xy}_{logicalform}}^{lambda\ term}}^{category}$$

In a lexicon of morphemes and words, these type assignments and rules engender a derivation of *pisi-ler-e* as follows:

$$(3)\quad \begin{array}{ccc}
\textit{pisi} & \textit{ler} & \textit{e}\\
\text{kitty} & \text{-PLU} & \text{-DAT}\\
\overline{\mathbf{N}\text{: } cat'} & \overline{\mathbf{N\backslash N}\text{: } \lambda x.plu'\,x} & \overline{\mathbf{N}_{\text{dat}}\backslash\mathbf{N}\text{: } \lambda y.to'\,y}\\
\multicolumn{2}{c}{\overline{\qquad\mathbf{N}\text{: } plu'\,cat'\qquad}{<}} & \\
\multicolumn{3}{c}{\overline{\qquad\qquad\mathbf{N}_{\text{dat}}\text{: } to'(plu'\,cat')\qquad\qquad}{<}}\\
\multicolumn{3}{c}{\text{‘to the kitties’}}
\end{array}$$

The following principles of UG narrow down possibilities for universal rules and lexical types, independent of whether the type assignment is to a word, morpheme, affix, clitic, syllable, sign or tone:

(4) *The Principle of Categorial Type Transparency* (PCTT): (Steedman, 2000)

> “For a given language, the semantic type of the interpretation together with a number of language-specific directional parameter settings uniquely determines the syntactic category [syntactic type] of a category.”
>
> In the pair (σ, μ) of a lexical category, the syntactic type σ and the semantic type μ are co-determined: μ is of type $\mathcal{T}\sigma$, and σ is of type $\mathcal{T}^{-1}\mu$, where $\mathcal{T}$ is a relation with inverse. If σ is a syntactic functor $\alpha\backslash\beta$ or α/β, then its semantic type is $\mathcal{T}\sigma = \mathcal{T}\beta \mapsto \mathcal{T}\alpha$.
>
> *The Prin. of Combinatory Type Transparency*:
>
> “All syntactic combinatory rules are type-transparent versions of one of a small number of simple semantic operations over functions.” [They are called **B**, **T** and **S** in Curry’s Combinatory Logic.]
>
> *The Principle of Consistency* (PC):
>
> “All syntactic combinatory rules must be consistent with the directionality of the principal functor.”

For example, (5a–c) are not viable lexical hypotheses (assuming for 5a that the child has not been constantly exposed to **ler-pisi* ‘PLU kitty’ as well by an unduly sarcastic adult). The first one violates PC, and the others violate PCTT: a syntactic functor has to correspond to a predicate, not to a proposition as in (5b); a 2-place syntactic functor cannot originate from a 1-place predicate as in (5c).

(5) a. $\{$*pisiler* := $\mathbf{N}$, *pisi* := $\mathbf{N\backslash N}$, *ler* := $\mathbf{N}\}$ (*)

b. *ler* := $\mathbf{N\backslash N}$: $plu'_{(t)}$ (*)

c. *tut* (catch) := $\mathbf{S\backslash NP\backslash NP}$: $\lambda x.catch'_{(e,t)}x$ (*)

The Models

We have developed three models: 1) A syllable-based model (SBM) in which an LF is associated with a sequence of syllables, 2) a morpheme-based model (MBM) where an LF is associated with a sequence of morphemes, and 3) a random model (RM) in which a randomly-segmented word is associated with an LF.

All models use the same statistical learning mechanism. Each input is a word segmented according to above, depending on the model, with the LF paired with the entire sequence of units in the word. We assume that the number of units per word is always greater than or equal to the number of terms in the LF, so that for example a polysyllabic word can in principle be associated with a one-term LF. We thus distinguish zero-morphemes from root forms. The plural *sheep* would have the LF $sheeps'$, not $plu'sheep'$ if we used English data (we consider relaxing this assumption as future work; Turkish seems to have no zero-morphemes, and composite suffixes such as *-leri* ‘-POSS.3PERS.PLU’ are indeed polysyllabic). The output is a lexicon containing the lexical hypotheses as

[2] The numbers are as follows: *pisilere* := $to'(plu'\,cat')$ example provided earlier requires 4 morpheme-LF pairings to be considered in the morpheme model, and 8 syllable-LF pairings in the syllable model. Average number of pairings are respectively 3.24 and 5.63 in CHILDES. In contrast, an “adult word” such as *kitabındakilerdeki* would require 49 and 343 pairings respectively.

items of a lexicalized grammar, such as (2). A lexical hypothesis is a 4-tuple (PF, σ, LF, w), where σ is the syntactic type (e.g., **N, N$\backslash$N**), and w is the outcome of the system's belief in the hypothesis ($0 \leq w \leq 1$).

All models use Algorithm 1 for learning. Learning is achieved by updating the weights based on new input. The model follows a simple statistical method for updating the weights. The weights in the lexicon are the probability, or system's belief, that the lexical item in question is correct. Each weight update consists of determining the new weight, the probability of the lexical hypothesis h given the new evidence E. The new evidence E is the input word segmented one of three ways, depending on the model. The weight of the lexical item after seeing the input is updated by (6).

(6) $w = w_0 + \alpha w_0 L(1 - w_0)$

where w_0 is the probability (or weight) of the lexical hypothesis before seeing the input E. If the hypothesis is already in the lexicon, w_0 is the weight of the hypothesis in the current lexicon, otherwise an arbitrary initial value is assigned. L in the formula is the likelihood $P(E \mid h)$ in Bayesian terms (7).

(7) $P(h \mid E) = \frac{P(E|h)P(h)}{P(E)} \propto P(E \mid h)P(h)$

L is calculated as the number of parses in which hypothesis h is used, divided by the total number of parses of the word. This determines the contribution of the new input to the posterior probability. The higher the number of parses that the hypothesis supports, the higher the likelihood value will be. If the hypothesis is used by all possible parses of the input, the value is 1. The value gets smaller due to the parses that do not include the hypothesis. The final term in the formula, $1 - w_0$, normalizes the result so that the new weight is in the range (0,1]. α is a constant that is used to control the learning rate. Throughout the experiments it is kept at 0.01 (small perturbations in the neighborhood did not have any effect in preliminary runs).

The final weight, the posterior probability of the hypothesis is increased with a value directly proportional to the prior $P(h)$ and the likelihood $P(E \mid h)$, as shown in (6). Rewriting (6) and (7) together, we get:

(8) $P(h \mid E) = P(h) + \alpha P(h)P(E \mid h)(1 - P(h))$

Our approach is inspired by Bayesian hypothesis revision, but it is not strictly Bayesian. Firstly, the implicit assumption is that there is no negative evidence, as the probabilities do not decrease. One can see no increase in the weight of a hypothesis as less belief in it, compared to its alternatives with higher weight. The problem can be alleviated if we can fit a distribution for $P(E)$, but this is rather difficult if not impossible. We can assume that it is constant for all real word experiences E, therefore it can be ignored in the search for maximum posteriors (cf. Step 3d of Algorithm 1).

Secondly, the system has no grounds to distinguish infrequent but correct hypotheses from incorrect but frequent hypotheses. In the first case, the belief in a hypothesis would not increase much, and in the second case, it will continue to increase, albeit slowly. This is a more serious impediment to approximating the acquisition of grammar by the child in real life, and short of faithfully approximating $P(E)$, the issue remains controversial.

Algorithm 1 Training the three models.

```
1. Inputs: 1) The initial lexicon $L_0$. 2) A pair (PF,LF). PF is the segmented word. LF is the logical form for the entire PF.
2. Output: The final lexicon $L_f$. The procedure stops when no more hypotheses are added to the lexicon.
3. After the $n^{th}$ input, the updated lexicon $L_n$ is determined by the following procedure:
  (a) All possible lexical hypotheses from the input are generated by CCG rules.
  (b) Generated hypotheses are placed in a temporary lexicon, $L_T$. The weights of the items are obtained from the current lexicon $L_{n-1}$. If the lexical hypothesis is not in $L_{n-1}$, an initial weight $w_s$ is assigned for the weight of the item in $L_T$. In the experiments, an initial weight of 0.1 is used.
  (c) All possible parses of PF using $L_T$ are produced. Each parse is assigned a weight proportional to the weights of all the lexical items used in the derivation.
  (d) All the hypotheses used in the derivation of PF with the highest weight are added to $L_n$, with the new weight determined by (6).
```

We use Algorithm 1 to train all models. It is adopted from Zettlemoyer & Collins (2005), who also use CCG as a framework, with a different update mechanism. The crucial point in their algorithm is to allow any contiguous substring of the input to be a lexical item. We use the principles of CCG (4) for eliminating the illicit hypotheses, whereas they can eschew the principles because their inventory of types is specific to a geography database, providing a similarly constrained behavior without UG.

An example

To exemplify the different behavior of the models, starting with an empty lexicon, we go through the process of learning two related words: *oda* (room) and *oda-ya* (room-DAT). We chose short words to save space, one with two syllables (*o·da*) and a single morpheme (*oda*), and the other with three syllables (*o·da·ya*) and two morphemes (*oda-ya*); longer words are attested in CHILDES. For example *adam-lar-a* (man-PLU-DAT) produces 20 hypotheses in MBM, and a staggering number (49) in SBM.[3]

[3] Interestingly, the notorious *-ki* suffix, which causes recursion in morphology to produce indefinitely long words, is nonexistent in recursive form in CHILDES. We counted 30 instances of single use of *-ki*, out of 20,000 morphemes. 17 of them are word-final.

For the first word *oda*, the input to SBM is the pair (*o·da*, $room'$), while MBM gets as PF the whole word as one unit. Step 3a generates the single hypothesis (9) in both MBM and SBM. The input contains only a basic LF, hence no attempt is made to find smaller units in PF.

(9) (*oda*, $\mathbf{N}$, $room'$, 0.1)

With the hypothesis (9) placed in the temporary lexicon in Step 3b, the algorithm generates a single parse of the input. As the only lexical item in the winning parse, the hypothesis is inserted into the lexicon with a weight adjustment according to (6), which increases it to 0.1009.

The second input is segmented as (*oda-ya*, $dat'room'$) for MBM, and (*o·da·ya*, $dat'room'$) for SBM. The morpheme model maintains the hypotheses (10), after Step 3c of Algorithm 1 eliminates potential hypotheses such as *oda* := $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$: $\lambda x.dat'x$ and *ya* := $\mathbf{N}_{\mathrm{dat}}/\mathbf{N}$: $\lambda x.dat'x$, because no universal rule can use them in any derivation of this experience.

(10) (*oda*, $\mathbf{N}$, $room'$, 0.1009) (*ya*, $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$, $\lambda x.dat'x$, 0.1)
(*oda*, $\mathbf{N}_{\mathrm{dat}}/\mathbf{N}$, $\lambda x.dat'x$, 0.1) (*ya*, $\mathbf{N}$, $room'$, 0.1)

Hypotheses such as *oda* := $\mathbf{N}_{\mathrm{plu}}/\mathbf{N}$: $\lambda x.dat'x$ are eliminated by a currently oversimplistic closed-world assumption, by which the child's linguistic world is embodied in the lexicon and the current experience, neither of which includes plurality at this stage.

All hypotheses except (*oda*, $\mathbf{N}$, $room'$, 0.1009) have the weight 0.1, because none of them were in the lexicon. Similarly, the syllable model produces the following set for the input (*o·da·ya*, $dat'room'$) :

(11) (*o*, $\mathbf{N}$, $room'$, 0.1) (*daya*, $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$, $\lambda x.dat'x$, 0.1)
(*o*, $\mathbf{N}_{\mathrm{dat}}/\mathbf{N}$, $\lambda x.dat'x$, 0.1) (*daya*, $\mathbf{N}$, $room'$, 0.1)
(*oda*, $\mathbf{N}$, $room'$, 0.1009) (*ya*, $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$, $\lambda x.dat'x$, 0.1)
(*oda*, $\mathbf{N}_{\mathrm{dat}}/\mathbf{N}$, $\lambda x.dat'x$, 0.1) (*ya*, $\mathbf{N}$, $room'$, 0.1)

After Step 3d, the lexicon contains the items in (12) and (13), with the updated weights respectively for MBM and SBM. Lower weights for SBM are due to likelihood, which is inversely proportional to the total number of parses, which is higher in this example for SBM.

(12) (*oda*, $\mathbf{N}$, $room'$, 0.101354) (*ya*, $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$, $\lambda x.dat'x$, 0.10045)

(13) (*oda*, $\mathbf{N}$, $room'$, 0.101127) (*ya*, $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$, $\lambda x.dat'x$, 0.100225)

The algorithm depends on the occurrences of isolated forms to start up, which are generally the root forms. However, the system makes use of frequently occurring forms to learn other forms without having seen them in isolation. For example, a third input *masaya* (table-DAT), segmented as *masa-ya* for MBM and *ma·sa·ya* for SBM, would cause both systems to add (*masa*, $\mathbf{N}$, $table'$) into their lexicon, as well as increasing the weight of (*ya*, $\mathbf{N}_{\mathrm{dat}}\backslash\mathbf{N}$, $\lambda x.dat'x$).

In closing we note that if initial assumptions that are put in the lexicon are incorrect, they would nevertheless be licensed by UG, to be corrected only by further experience in a Siskindian (1995; 1996) scenario. The role of the current lexicalized grammar as the trigger of negative feedback is crucial in this respect: The child has the current set of hypotheses at her disposal to realize in a new experience that she might have assumed wrongly about which part meant what, as calculated in steps 3b–c of Algorithm 1.

Experiments

We measure the success of the models with usual metrics over the final lexicons (precision, recall, f-score), and with two sets of tests: recognize and generate. The first test measures each model's ability to deal with unseen PFs, and the second, unseen LFs. We use the following items for comparison:

L_r: The reference lexicalized grammar. This is a manually-derived MBM-type adult competence grammar of Turkish nouns in CHILDES. It contains all free and bound morphemes in the data. L_m: The lexicalized grammar learned by MBM. L_s: The lexicalized grammar learned by SBM. L_{rm}: The lexicalized grammar learned by RM.

Data

Our data is the Turkish noun fragment of CHILDES (MacWhinney & Snow, 1990). It contains 51 recording sessions with 33 children. The ages of the children vary between 2;0 to 4;8. The average age of children is 3;4.

We use the child directed speech (CDS) in the corpus. All the nouns in the CDS have been segmented at morpheme boundaries. Each segmented word is tagged with an LF to establish L_r. We left out derivational suffixes as future work. All derivational morphemes are considered part of the nominal root. Due to the nature of CHILDES transcriptions, automated segmentation and tagging was not practical; they were done mostly by hand.

The data for training and testing contains 12,274 nouns out of 33,450 words in the CDS. The total number of morphemes (nominal roots/stems with possible derivations, and inflectional morphemes) is 20,433. The number of syllables is 27,497.

Test Measures

The standard measures translate to the following in our case: precision(p)=$hits/(hits + noise)$, recall(r)=$hits/(hits + misses)$, f-score=$2pr/(p + r)$, where for any lexicon L_x, an entry that is both in L_x and L_r is a hit, an entry that is not in L_x but in L_r is a miss, and an entry that is in L_x but not in L_r is noise. We also compare L_m and L_s.

All models have been trained with equal amount of input. Table 1 shows the value of measures and the size of the lexicons (the number of lexical items with a syntactic type).

The lexicon learned by MBM (L_m) is very similar to L_r. MBM fails to learn 5 of the inflections in the input, and incorrectly learns 4 items which are not in L_r. The misses are due to infrequent salient occurrence of the morphemes. The other errors are due to ambiguous inflections. All the differences in L_m and L_r are due to affixes; the model learned the complete set of root/stem forms in L_r.

Table 1: Test measures over the lexicons.

Lexicon	#of items	precision	recall	f-score
L_r	1041	100.00	100.00	100.00
L_m	1040	99.61	99.51	99.55
L_s	909	81.73	71.37	76.19
L_{rm}	1697	51.73	83.57	63.90

Table 2: Results of the $10\times$ recognition tests.

Lexicon	precision		recall		f-score	
	μ	σ	μ	σ	μ	σ
L_r	87.00	1.63	92.90	1.66	89.83	0.79
L_m	86.70	1.42	92.90	1.66	89.67	0.66
L_s	84.20	4.32	72.80	2.04	78.02	2.06
L_{rm}	57.10	3.35	82.00	2.98	67.21	1.71

As expected, SBM's performance is lower than MBM. However, it is significantly more precise than RM. SBM misses to learn 268 (28%) morphemes while learning an additional 166 (18%) morphemes which were not in L_r. Like MBM, SBM could not learn a number of morphemes that were not frequent enough in the input. However, the majority of the morphemes that SBM fails to learn are the morphemes whose boundaries do not match with syllable boundaries. The additional items that are learned mistakenly also follow a similar pattern. Most of them are due to morpheme-syllable boundary mismatch. For example, for the plural suffix *lar*, L_s contains the lexical item *la* := $\mathbf{N}_{\text{plu}}\backslash\mathbf{N}$: $\lambda x.plu'x$, in addition to the adult reference entry *lar* := $\mathbf{N}_{\text{plu}}\backslash\mathbf{N}$: $\lambda x.plu'x$. Ignoring single-phoneme differences, like in the example above, the number of misses by SBM drops to 165 (15%), and the number of mistakenly learned items drops to 31 (3%).

Recognition and Generation

CCG's lexicalized grammars can act both as recognition mechanisms, deriving an input PF to its possible LFs, and as production mechanisms that output the possible PFs for an LF. To test the success of the models and the differences between them, all models have been run through these tests after equal amount of training. We use the same data set of the lexicon comparison tests.

For both generation and recognition experiments, we employed 10-fold cross-validation. The data set is divided into 10 subsets with equal number of words, and the models have been trained 10 times, each time leaving a different subset as test set and training set. The numbers reported in Table 2 and Table 3 are average and standard deviation of measures for 10 experiments.

The recognition test can produce multiple parses leading to the same or different LFs. Our criteria for hit, miss, and noise in this experiment are as follows: If the lexicon was able to engender the intended semantic form at least once, it is considered a hit. If no parse producing the target LF is found, it is a miss. Any parse that did not lead to expected LF is noise.

Table 3: Results of the $10\times$ generation tests.

Lexicon	precision		recall		f-score	
	μ	σ	μ	σ	μ	σ
L_r	16.20	0.92	92.90	1.66	27.59	1.30
L_m	16.10	0.88	92.90	1.66	27.44	1.24
L_s	15.20	1.14	71.70	1.83	25.08	1.53
L_{rm}	0.90	0.32	91.20	1.52	1.78	0.63

Because of the words that are in the test set but not in the training set due to 10-folding, parsing all the input in the test sets was not always possible, even for L_r. Ambiguous morphemes, on the other hand, caused multiple parses, hence the low precision values for all the models in Table 2 .

Similar to the agreement statistics, L_m performs very close to L_r in recognition. Due to the smaller number of lexical items SBM learns in these experiments, its recall performance is slightly worse than all others. However, it is almost as precise as L_m and L_r.

The generation test produces the possible PFs for the input LF. Due to multiple phonetic alternations, and lack of phonological knowledge in the models, the generation test always overgenerates. For example, the logical form $plu'man'$ generates both *adam-lar* and *adam-ler*, the latter only violating vowel harmony. On average, L_r and L_m generate 5.61 PFs per LF, L_s 4.62 and L_{rm} 80.92.

Except the high rate of noise generated by all models, the results of generation tests are again similar to those of agreement statistics. L_m performs very close to L_r, and L_s does comparably, but slightly worse.

Discussion and Conclusion

With varying degrees of success, all three models we investigated do learn a syntactic type for the (LF, PF) pair, that is, a fragment of grammar, in our case a word grammar.

Our main result is as follows: Although we expected the morpheme model MBM to approximate an adult reference word grammar, the syllable model SBM is unexpectedly not too far behind, at least in the circumstances where word-level ambiguity is kept to a minimum. This is promising for grounding early development of language in perception. SBM's success seems to depend on its ability to consider any contiguous substring of syllables as potential bearer of an LF (or part of it). This is not always possible since morpheme boundaries do not always coincide with syllable boundaries. With a reasonable margin of error (one phoneme), SBM shows comparable performance with MBM in all tests.

The success of the morpheme model is also noteworthy. Even though the input is segmented at morpheme boundaries in MBM, it still needed to match the correct (PF, LF) pairs in the input. It does this successfully, only failing in case of am-

Table 4: Overall comparison of the lexicons. EM (Exact match) is the count of matching items with identical (PF, syn. type, LF). NM (Near match) ignores a single phoneme difference in PF. LFS (LF/syn.type match) ignores the PF completely.

	Roots & Stems	Inflections	Total
# items in L_r	886	155	1041
# items in L_m	886	154	1040
# items in L_s	802	107	909
EM: L_r & L_m	886	150	1036
EM: L_m & L_s	684	59	743
NM: L_m & L_s	774	101	875
LFS: L_r & L_m	886	150	1036
LFS: L_m & L_s	719	83	802

biguous morphemes. Given the lack of disambiguating cues in the models, it seems that word learning in the model is further facilitated by less ambiguity in word structure, compared to syntactic structure.[4]

10-fold cross-validation does not reveal all the similarities between morpheme and syllable models. Taken all together as one lexicalized grammar, the numbers are as in Table 4. It is significant that there is a 71% exact match of lexical hypotheses of the syllable model and the morpheme model. Granted that the exact match of bound morphemes is low (around 40%), we have to keep in mind that the syllable model does not come with root/stem boundaries, therefore the exact match of these forms (77%) is very significant.

In the course of development of the child, there will certainly be more novel free morphemes in the lexicalized grammar than novel bound morphemes. One estimate for Turkish is that children master the nominal paradigm by 24 months or earlier (Aksu-Koc & Slobin, 1985), therefore allomorphy of bound morphemes (all lexical hypotheses for morpheme-like units in our terms, including their correct PF) is already intact by that age in real life. This can only help the syllable model to converge to the morpheme model more, if the same relative success rate can be maintained. We can then surmise that morphemehood need not be a theoretical primitive, since computation might deliver the morphemes without having to start with that assumption.

Both models stand in sharp contrast with the random model. Its performance shows that the success of morpheme and syllable models is not due to chance. Since all models use the same universal grammar to control the explosion of number of hypotheses, the importance of adequately capturing the cues in child-directed speech is clear. As future work we intend to relax the close-world assumption to observe the learning rate of generating more hypotheses early in the course of development. We think that universal grammar will still be indispensable for narrowing down the hypothesis space, but other cues such as stress, intonation and access to possibly ambiguous extra-linguistic world (scenes, objects, events etc.) will have to be incorporated to limit the search to obtain reported early development of word grammar.

References

Aksu-Koc, A. A., & Slobin, D. I. (1985). The acquisition of Turkish. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition, vol.I: The data.* New Jersey: Lawrence Erlbaum.

Aronoff, J. M., Giralt, N., & Mintz, T. H. (2006). Stochastic approaches to morphology acquisition. In *Selected proceedings of the 7th conference on the acquisition of Spanish and Portuguese as first and second languages* (p. 110-121).

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, *29*, 1–36.

Jack, K., Reed, C., & Waller, A. (2006). From syllables to syntax: Investigating staged linguistic development through computational modelling. In *Proceedings of the 28th annual conference of the cognitive science society.*

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Science*, *3*, 323-328.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17*, 457-472.

Peters, A. M., & Menn, L. (1993). False starts and filler syllables: Ways to learn grammatical morphemes. *Language*, *69*(4).

Siskind, J. (1995). Grounding language in perception. *Artificial Intelligence Review*, *8*, 371-391.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39-91.

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: MIT Press.

Steedman, M., & Hockenmaier, J. (2007). *The computational problem of natural language acquisition.* University of Edinburgh. (ms.)

Tenenbaum, J. B., & Xu, F. (2000). Word learning as Bayesian inference. In *Proc. of the 22nd annual conf. of the cognitive science society.* Philadelphia.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706-716.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with Probabilistic Categorial Grammars. In *Proc. of the 21st conf. on Uncertainty in Artificial Intelligence.* Edinburgh.

[4] There is no room for parameters in CCG, because UG is not conceived as the initial state of competence grammar; it is invariant and only the lexicalized grammar is learned. Therefore, a trigger-based scenario of acquisition via parameters such as that of Fodor (1998) where knowledge of structural unambiguity by the learner is assumed is incompatible with CCG. A potential explanation lies toward understanding the work of perceptual cues in narrowing down the hypotheses space allowed by a probabilistic universal grammar.