Language acquisition and cross-modal associations: Computational simulation of the result of infant studies

L. ten Bosch¹, L. Boves¹

¹Department of Linguistics, Radboud University Nijmegen, NL
l.tenbosch@let.ru.nl

Abstract

This paper discusses recent results obtained with a computational model of language acquisition. In previous papers, this model, developed in the ACORNS project, has shown to be able to learn word-like units from stimuli in which utterances are paired with visual information. In this paper we extend the ACORNS experiments to the case where utterances are paired with a ambiguous visual representation, as to obtain a computational correlate of the findings by Smith and Yu in 2008. Smith and Yu stipulate that a young infant is confronted with an uncertainty problem, how to pair a word, embedded in a sentence, and a referent, embedded in a rich visual scene. They show that young infants can resolve the uncertainty problem by evaluating the statistical evidence across many individually ambiguous words and scenes. We investigate to what extent the ACORNS model is able to deal with cross-modal ambiguity. Moreover, we show the positive effect of an 'active' role during learning when confronted with ambiguity, based on internal confidence. Index Terms: Language acquisition, Computational modeling, Ambiguity resolution

1. Introduction

Language acquisition involves the discovery and memorization of units that are potentially semantically relevant. The discovery of words (or word-like units) plays a decisive role in this learning process. In order to learn a novel word-like unit, an infant must perceive and store enough information to reliably distinguish that word from other words that are already known. Several studies have revealed important aspects of this word-discovery process, and how infants generalize previously acquired knowledge to new situations (e.g. [3]). There is substantial evidence that infants start their language acquisition process by storing a large amount of acoustic/prosodic detail, and that abstract representations of sounds and words emerge at a later stage (e.g. [5] [6]). Infants about 8 months old appear to be able to use statistical information for detecting word-like patterns and acoustic onsets of words (e.g. [2]).

The discovery of new words, however, is not solely based on the discovery of recurrent acoustic patterns within the speech modality. Stretches of sound must be associated with a referent in the environment, and this association is usually discovered thanks to a multi-modal context: the relation between sounds and referents is largely guided by cross-modal association (word-referent pairing). And also the mere pairing of a word and a scene is not enough to determine the meaning of the word, as Quine’s well-known ‘gavagai’ example illustrates: Given a certain word and scene, the relation between word and referent may be ambiguous.

To resolve this indeterminacy problem, it is argued that infants apply a mechanism that enables them to accrue statistical evidence across situated experiences to generalize acoustic representations and to strengthen word-referent pairs [7]. In this paper we investigate this indeterminacy problem in language acquisition from a computational perspective. To that end, we employ a computational model of language acquisition (developed in the FP6 FET project ACORNS¹ [9]). This model is able to discover word-like units in continuous speech by taking multi-modal stimuli as input. Each stimulus has an audio part (an utterance in which one or more keywords occur), accompanied by a visual part (which contains the visual representation of the concepts referred to by the keywords). The position of keywords is not known to the model, neither is the phonetic content. The model discovers word-like units on the basis of recurrent patterns in the audio channel in combination with cross-modal associations across the stimuli.

An ecologically and cognitively important issue is the presence of ambiguity in one channel or both channels. In most ACORNS experiments ([10]), ambiguity is restricted to the audio channel (a keyword is embedded in carrier phrases). The visual representation, however, was not ambiguous: The object referred to (and no other) was always available in the visual channel. In this paper we go one step further and investigate the more realistic case when both channels are ambiguous. This will be examined in two ways:

- by varying the number (denoted $N_{obj}$) of objects presented during learning. If $N_{obj} > 1$, randomly drawn distracter objects (foils) are added to the scene ('word ↔ multiple referent ambiguity')
- by assuming "attention noise" ($\theta_{att}$): the fraction of stimuli presented during learning of which the learner perceives an incoherent audio-visual association. This 'attention noise' models the infant looking at an object different from the one referred to in the utterance ('referent ↔ multiple word ambiguity')

The third and final experimental parameter is the active learning threshold ($\theta_{al}$). This determines the degree to which the learner takes an active role during learning [12]. Instead of passively accepting all training inputs, the learner may use its internal confidence for deciding to stick to its own hypothesized association, rather than accepting the association as presented in the stimulus (see section 4.1). The model is tested by using a virtual computational head-turn paradigm (section 4.2). The structure of this paper is as follows. In the following section we will summarize the results obtained by Smith and Yu [7]. In section 3 we will briefly describe the computational model. Sections 4.1, 4.2, 5 and 6 describe the role of active

¹http://www.acorns-project.org
learning, the computational head turn, the experiments and results. The final section contains a discussion and conclusion.

2. Cross-situational learning

Without any constraint, a pairing of a word and scene presents the learner with an infinite number of possible referents, leading to indeterminacy in the speech-referent mapping. [7] shows that infants are able to learn multiple speech-referent pairs by accruing statistical evidence across multiple and individually ambiguous speech-scene pairings. Moreover, this can be achieved for many words and referents. According to [7] this learning process is sufficiently rapid and robust to play a role in early lexical acquisition.

In Smith & Yu’s experiment, 12- and 14-month-old infants were taught six word-referent pairs via a series of individually ambiguous stimuli. On each trial, two word forms and two potential referents were presented with no information about which word went with which referent. The training set consisted of 30 training stimuli, in which each correct word-object pair occurred 10 times. Each object and each word co-occurred with every other word and every other object at least once across the 30 training trials. Word-referent pairings were ambiguous within individual trials, but unambiguous across trials.

In the test infants were presented with a single word and two potential referents, the cross-trial correct referent and a foil (a distracter, drawn from the training set). The infant’s choice was based on the preferential looking method ([1]). The test consisted of 12 trials.

3. The computational model

The ACORNS model aims at the detection of word-like units from multi-modal input data. The representation of the input data, the internal representations, the data processing and the memory structures have been designed and chosen to optimize the cognitive plausibility of the eventual model architecture (e.g. [9]). Prior to learning, the model does not possess any information about the possible words or any other phonetic or linguistic information.

The model employs a structure discovery method based on Non-Negative Matrix Factorization (NMF, [4] [8]). NMF is a member of a family of computational approaches that represent input data in a (large) matrix

\[W\] and decompose this matrix into smaller matrices \(W\) and \(H\) such that \(V \approx WH\). The NMF-variant chosen here minimizes the Kullback-Leibler dissimilarity between \(V\) and its ‘reconstruction’ \(WH\).

In this paper, each column of \(V\) represents a stimulus. It has an audio part and a visual part. The audio part comprises statistics of codebook indices for the audio channel (based on an MFCC representation with \(\Delta\) and \(\Delta^2\) [8]), while the visual part contains a combined representation of all objects in the ‘scene’ ([9]).

After the NMF step, we interpret \(W\) as a set of (unobservable) learner-internal representations, while \(H\) contains the associated internal activations. Similar to \(V\) the columns of \(W\) have an audio and visual part. In combination, the matrices \(W\) and \(H\) approximate the information in \(V\) in a (highly) condensed form. The number of columns in \(W\) (and rows in \(H\)) specifies the maximum number of different internal representations that can be formed (which must be chosen prior to learning).

The other dimension of \(W\) is specified by the dimension of the input. Here, we have used the incremental version of NMF ([11]), which is able to find \(W\) on the basis of one-by-one presentation of \(V\)-vectors.

4. The simulated learning situation

Learning (the word discovery process) takes place in an interaction between the learner and a ‘caregiver’. An interaction runs as follows [9]. The (virtual) caregiver presents one multi-modal stimulus to the (virtual) learner. After the learner receives the input stimulus, NMF is applied to update \(W\) and \(H\). The effect of the learning is that internal representations are hypothesized and strengthened. \(W\) and \(H\) are initialised randomly, and the number of columns of \(W\) is set to 70 (a number well beyond the expected number of keywords in the experiments below).

Based on the audio part of the stimulus and the audio part of \(W\), the learner evaluates \(H\) (the activation of each internal representation). Using these \(H\)-activations and the visual part of \(W\), the learner ‘reconstructs’ by matrix multiplication the estimated visual modality ([8]). This estimation (a vector, comparable to a histogram) provides similarity scores \(sc\) for each object \(i\). The ‘winning’ object is the one with the highest score, and is provided as ‘recognition result’ to the caregiver as the final step of the interaction loop.

Since they are estimated, the scores are not genuine probabilities. We can assign a ‘confidence score’ \(c\) to the winner by using the normalized score difference of the winner \((sc_{1st})\) and the second candidate \((sc_{2nd})\). Imposing \(0 \leq c \leq 1\), \(c\) is defined as

\[c = \frac{sc_{1st} - sc_{2nd}}{\sum_i sc_i}\]

4.1. Active learning

The confidence measure \(c\) enables the learner to play an active role in the learning process [12]. If its confidence is high, the learner could decide to use its own hypothesis about the visual referent, rather than rely on the visual information in the presented stimulus. This behaviour is made explicit as follows:

\[\text{if } c > \theta_{al} \text{ then learner assumes own visual hypothesis is true else the learner accepts the stimulus ‘as is’}.\]

The active learning threshold \(\theta_{al}\) is a model parameter. In theory \(0 \leq \theta_{al} \leq 1\), but the behaviour for values between 0.5 and 1 is very similar. If \(\theta_{al} > 0.5\), the learner is always believing the input stimulus, while if \(\theta_{al} = 0\), the learner is always assuming its own hypothesis.

4.2. The virtual Head Turn

In the text, a computational variant of the Head Turn paradigm is used ([11]). The auditory part of a test stimulus is presented to the learner model in combination with a visual scene in which multiple objects are present, one of which corresponds to the auditory stimulus. The head turn paradigm involves a choice between externally presented objects. As during learning, the model first reconstructs its internal visual hypothesis \((r_{vis})\), based on the audio input. This internal hypothesis is now used by the learner to compare it to the visual representation of each individual object in the scene. Suppose that \(M\) objects are present in the scene, encoded by the learner as visual representations \(\{r_1, \ldots, r_M\}\). (Each \(r\) is comparable to a histogram.) Using the NMF framework, the learner chooses the object with
minimum KL dissimilarity to $r_{hyy}$:

$$\arg \min_{i=1, \ldots, M} (KL(r_i | r_{hyy}))$$  \hspace{1cm} (3)

The underlying hypothesis is that this match corresponds with the (observable) attention time in the head turn paradigm.

5. Experiments

Learning and test sets were selected from the UK database recorded in the ACORNS project [10]. All utterances have a simple syntax, similar to child-directed speech. The total corpus comprises 4000 utterances spoken by two female (F1, F2) and two male (M1, M2) speakers (1000 utt/sp). From this pool, we selected a learning and test set in which each utterance contains a single keyword (Angus, Ewan, bath, book, bottle, car, daddy, mummy, nappy, shoe, telephone). The training set has been used in previous experiments (F1+F2+M1+M2 in [10]), and consists of 520 utterances. The test set consists of 200 independent utterances. The sets were designed such that keywords appear in a fixed and repeating order with a flat distribution.

Performance is measured as the fraction of the correctly recognized items (as determined by the head turn procedure), divided by the total number of presented items. Results are obtained on the test set after learning.

We performed experiments in which the three parameters (the number of objects $N_{obj}$, the active learning threshold $\theta_{alt}$, and the attention noise $\mu_{att}$) were systematically varied during learning.

6. Results

In figure 1 the performance is shown as a function of the ‘attention noise’ in the case of one object per utterance. Five different curves are presented, each corresponding to different values (0 . . . 0.5) of the active learning threshold $\theta_{alt}$. In the upper left corner, the ‘ideal’ learning (attention noise $\mu_{att} = 0$, high active learning threshold $\theta_{alt} = 0.5$) leads to an accuracy of 0.97 ($\sigma = 0.009$). Low values ($\leq 0.2$) of the active learning threshold $\theta_{alt}$ (that is, the learner more and more trusts its own interpretation) lead to decreased performance (McNemar with Yates’ correction, $p < 0.01$), with a collapse to about 10% if the activation threshold equals zero (the lowest curve – this corresponds with a nearly random selection of objects).

If more stimuli are inconsistent ($\mu_{att} > 0.3$), the best learning results are obtained when the learner is more active overriding the information in the stimuli. For example, if $\mu_{att} = 0.5$ (50% of all stimuli inconsistent), a value of the active learning threshold of about 0.1 gives the best performance of about 0.62 ($\sigma = 0.027$), significantly above 0.5 (McNemar with Yates’ correction, $p < 0.01$).

This shows that the model is able to go beyond just ‘correcting’ the inconsistent stimuli. If there are inconsistent stimuli ($\mu_{att} > 0$), a higher performance is obtained by relying on own hypotheses (that is, a low active learning threshold), while if all stimuli are consistent, a high active learning threshold $\theta_{alt} = 0.5$ (i.e., passive learning) is best.

Figure 2 shows the situation in which two objects are presented during training ($N_{obj} = 2$). The prominent difference with the previous figure ($N_{obj} = 1$) is the overall decrease in performance (about 12% relative). Again, an increase of attention noise leads to a gradual decrease in performance. As above, a too low active learning threshold makes the learning ‘collapse’. Again, we see that active learning helps in the case of inconsistent stimuli.
in a qualitative manner. There are two opposing effects. If more ambiguous stimuli are presented during learning, the learner must apply a more selective learning mechanism to be sure that the internal confidence evaluation is reliable, which corresponds to a higher threshold. This, however, may imply that inconsistent stimuli near class boundaries are not corrected (since they do not meet the confidence threshold), which leads to contamination of the internal representations. In the same way, it can be readily understood that the learner model is able to perform significantly beyond 50% accuracy in the case of 50% inconsistent stimuli, but it is more difficult to predict this on the basis of the underlying mathematical model.

Similar to the observations in [7], we think that real-life ambiguities are most likely greater than those modelled in this experiment. Infants can probably rely on more powerful statistically-based approaches to detect word-like units in speech (c.f. [2]). Nevertheless, the computational modelling of language acquisition based on cross-modal ambiguity resolution by accruing statistical information seems a very promising approach.

8. Acknowledgments

ACORNS was an FP6 project, contract number FP6-034362. Currently the first author is related to the FP6 Marie Curie project Sound-to-Sense, and to the Dutch NWO project 360-70-350.

9. References