Cross-situational cues are relevant for early word segmentation

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, Aalto, Finland

Heikki Rasilo (heikki.rasilo@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, Aalto, Finland

Abstract

Existing models of infant word learning have mainly assumed that the learner is capable of segmenting words from speech before grounding them to their referential meaning, while segmentation itself has been treated relatively independently of meaning acquisition. In this paper, we argue that situated cues such as visually perceived concrete objects or actions are not just important for word-to-meaning mapping, but that they are useful in pre-linguistic word segmentation, thereby helping the learner to bootstrap the language learning process. We present a model where joint acquisition of proto-lexical segments and their meanings maximizes the referential quality of the lexicon, and where learning can occur without any a priori knowledge of the language or its linguistically relevant units. We investigate the behavior of the model using a computational implementation of statistical learning, showing successful word segmentation under varying degrees of referential uncertainty.

Keywords: word learning; segmentation; meaning acquisition; computational modeling; synergies in word learning; language acquisition

Introduction

One of the largest challenges faced by language learning infants is the problem of word learning. From a linguistic point of view, the problem is often posed as the question of 1) how to segment the incoming speech input into words and 2) how to associate the segmented words to their correct referents in the surrounding environment in order to acquire meaning of the words. In this paper, we describe how these two tasks can be approached as a single learning problem.

Several behavioral and computational studies have addressed the segmentation problem and it is now known that infants may utilize different cues, such as statistical regularities (Saffran, Aslin & Newport, 1996), prosody (Cutler & Norris, 1988; Thiessen & Saffran, 2003), or other properties of infant directed speech (Thiessen, Hill & Saffran, 2005), in order to find word-like units from speech. Similarly, computational modeling studies show that segmentation into recurring word-like units is possible at varying levels of language representation (e.g., Brent, 1999; Frank, Goldwater, Griffiths & Tenenbaum, 2010; Pearl, Goldwater & Steyvers, 2010; Räsänen, 2011). However, the main problem with these models is that they either require strong constraints or heuristics to drive the segmentation, or they assume representations of language such as phonetic transcriptions that are not available for an infant trying to bootstrap the language acquisition process.

Likewise, the problem of associating segmented words to their referents has been widely addressed in earlier research. One of the prominent mechanisms in this area is the so-called cross-situational learning (XSL; Pinker, 1989; Gleitman, 1990). According to the XSL hypothesis, infants learn meanings of words by accumulating statistical information on the co-occurrences of spoken words and the possible word referents (e.g., objects and events) across multiple communicative contexts. While each individual communicative situation may be referentially ambiguous, the ambiguity is gradually resolved as the learner integrates co-occurrence statistics over multiple such scenarios. A large body of evidence shows that infants and adults are sensitive to cross-situational statistics between auditory words and visual referents (see, e.g., Yu & Smith, 2012 for a recent overview) and that these statistics are accumulated and used incrementally across subsequent exposures to the word-referent co-occurrences (e.g., Yu, Zhong & Fricker, 2012).

In this work, we argue that, instead of looking at the early word segmentation and meaning acquisition separately, the two problems should be approached as one. Then the learning problem can be formulated as *how to segment speech into meaningful units?* When defined this way, there is no longer the implication that successful segmentation precedes meaning acquisition; rather, segment meaningfulness as such should be the main criterion for speech segmentation (for similar ideas, see ten Bosch et al., 2009; Johnson, Demuth, Frank & Jones, 2010 and Fourtassi & Dupoux, 2014). Since word meanings are acquired through grounding of word forms to their referents, it would be natural to utilize the statistical regularities in the referential domain also in the acquisition of word forms themselves.

We base our argument on the idea that the role of language is to describe the external world as accurately as possible, making all speech potentially referential. In this context, an effective learner is the one that finds maximally informative mapping from the initially ambiguous acoustic speech stream to the word referents that consistently co-occur with the speech contents. Solving this mapping problem simultaneously solves the acquisition of word meanings (speech-to-referents associations) and word segmentation (mutually exclusive segments of speech). This

provides the basis for a functional proto-lexicon (Nazzi & Bertoncini, 2003) that has functional significance to a language learner that does not yet master the phonological structure of the language, and upon which more sophisticated language processing and parsing can build.

Learning word segments through cross-situational learning

The idea of learning word segments and their referential meanings simultaneously is not new. Several computational models have made use of joint inference at both levels with (Roy & Pentland, 2002; Yu & Ballard; 2004) and without (ten Bosch et al. 2009; Aimetti, 2009; Räsänen, Laine & Altosaar, 2008; van Hamme, 2008) assuming phonemic representation of speech. Recent behavioral evidence also shows that consistent visual cues help in word segmentation (Thiessen, 2010; Glicksohn and Cohen, 2013; Shukla, White & Aslin, 2011). The goal here is to formalize the joint-problem from referential point of view and to show with concrete simulations how this leads to successful word segmentation under varying degrees of referential uncertainty.

We will start by defining the *referential quality* of a lexicon. By making a simplifying assumption that there is no grammar (i.e., all words are independent of each other), the referential quality (or *information value*) of the lexicon can be measured using mutual information:

$$Q = \sum_{w,c} P(w,c)\log_2 \frac{P(w,c)}{P(w)P(c)} / \max\{\log_{|C|},\log_{|W|}\} \qquad (1)$$

In the equation, $w \in W$ are the words known by the learner and $c \in C$ are discrete referents (states of the world) that the language attempts to describe. $P(w,c)$ is the probability of observing word w and referent c in the same context while $P(w)$ and $P(c)$ are the probabilities of observing them individually.

What Q quantifies is that, given a set of words (e.g., an utterance), how much information we know about the state of the world. Q achieves its maximum value of one when each word w co-occurs only with one referent c, i.e., there is no referential ambiguity at all and all referents have been named, each word having deterministic consequence to the state of the world. On the other hand, Q approaches zero when words of the lexicon W occur independently of the referents, i.e., there is no coupling between the lexical system and the surrounding world. The max{}-term normalizes the base of the logarithm, ensuring that Q decreases if the number of words is larger than the number of referents and vice versa, indicating ambiguity or redundancy in referential capability of the vocabulary.

From learning point of view, a *referentially optimal language learner* wants to discover a vocabulary of words W that maximizes Eq. (1), i.e., considering those speech patterns as words that are maximally coupled to the concurrent environment in each communicative situation.

Let us assume that speech input to the learner is represented as a sequence of observations $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ (e.g., short-term spectra of speech or neural firing patterns of the auditory nerve) with subscripts denoting time indices. These observations can be uni- or multivariate, or they can also be discrete, but here we will use vectors $\mathbf{x}_t$ for the purpose of generality. Moreover, each observation $\mathbf{x}_t$ is paired with a set $\mathbf{c}_t$ describing the present communicative context attended by the learner. In this case, the goal of an referentially optimal learner is to map X into a sequence of words $f(X) \rightarrow [w_1, w_2, \ldots, w_M]$ so that the Q in Eq. (1) becomes maximized. This can be seen as a segmentation and classification problem: how to find sequences of acoustic observations that consistently co-occur in specific communicative contexts.

The first step in solving the optimization problem is to consider the direct coupling of the speech X with the referential context c through their joint distribution $P(X, c)$. Importantly, unlike a generative latent lexicon W that would be responsible for generating sequences of words, and each word generating an acoustic realization, the distribution $P(X, c)$ is directly observable to the learner. The challenge is to model the signal X so that it compactly and discriminatively captures the acoustic and temporal characteristics of speech in different referential contexts c.

From Eq. (1) we can infer that, in order to maximize Q, the co-occurrence matrix $P(w,c)$ should be a sorted diagonal matrix, i.e., there would be one word for each unique referent and they only occur with each other as this minimizes the referential uncertainty. The easiest way to ensure that the size of vocabulary equals to the number of referents ($|W| = |C|$) is to have a separate model for speech X occurring in the context of each possible referent c, capturing the cross-modal statistical dependencies between the two. Formally, an acoustic model with parameters θ_c is introduced for each referent c,

$$P(c \mid X) \propto P(X \mid c,\theta_c)P(c), \qquad (2)$$

capturing the probability of observing referent c given speech input X and therefore constituting the meaning part of the model. This model θ_c can be any algorithm or rule that maps from X to c, but the overall quality of the lexicon will depend on the accuracy and consistency of these mappings. Learning of the words is then a parameter estimation problem with the aim of finding the set of acoustic model parameters θ^* that maximize the joint probability of the concurrent referents across all speech X:

$$\begin{aligned}\theta^* &= \arg_\theta \max\{P(X \mid c,\theta)P(c) \mid \forall X,c\} \\ &= \arg_\theta \max\{P(c \mid X,\theta)P(X) \mid \forall X,c\}\end{aligned} \qquad (3)$$

From referential quality point of view, by replacing the discrete words w with the probabilistic models θ_c of referent c during speech X, i.e., setting $P(w,c) = P(c \mid X,\theta_c)P(X)$, we get

$$Q = \sum_{X,c} P(c \mid X,\theta_c)P(X)\log_{|C|}\frac{P(c \mid X,\theta_c)P(X)}{P(X)P(c)}$$
$$= \sum_{X,c} P(c \mid X,\theta_c)P(X)\log_{|C|}\frac{P(c \mid X,\theta_c)}{P(c)} \quad (4)$$
$$= \sum_{X,c} P(X \mid c,\theta_c)P(c)\log_{|C|}\frac{P(X \mid c,\theta_c)}{P(X)}$$

where $P(X)$ replaces $P(w)$, being the probability of observing the speech-signal state X. Now, since $P(X)$ and $P(c)$ are independent of the model parameters θ, *optimizing the solution for Eq. (3) will also optimize the referential value of the lexicon*. Informally put, the overall quality of the lexicon depends on how well the model θ_c discriminates different referents c in different speech inputs X, giving more importance to referents occurring more often.

As for the segmentation, the major consequence of the above formulation is that word segmentation emerges as a side product of learning of the acoustic models $P(X \mid c, \theta_c)$ for the referents. The relative probability (or familiarity) of word w occurring at time t in the speech input is given simply by the corresponding acoustic model θ_c:

$$P(w,t) = P(c,t \mid \mathbf{x}_{0,\ldots},\mathbf{x}_t) \propto P(c,t \mid \mathbf{x}_{0,\ldots},\mathbf{x}_t,\theta_c) \quad (5)$$

where $\mathbf{x}_0, \ldots, \mathbf{x}_t$ refer to speech observations up to time t. Input can be parsed into contiguous word segments by either 1) assigning each time frame of analysis into one of the known referents (proto-words) with word boundaries corresponding to points in time where the most likely referent changes, or 2) thresholding the probabilities in order to make a decision whether a known word is present in the input at the given time or not.

Note that the learner never explicitly attempts to segment the incoming speech into words as a separate stage from meaning acquisition. Instead, the learner simply performs maximum-likelihood decoding of referential meaning from the input, and this dynamically leads to the emergence of word boundaries in time.

In contrast, if the processes of segmentation and word-referent mapping are to take place independently of each other, the ultimate referential quality of the lexicon cannot recover from potential errors in the segmentation without further corrective mechanisms. Also, the "pre-segmented" vocabulary W is of no practical value before meaning is attached to the words, making at least explicit attempts to solve the segmentation problem alone questionable for an infant that doesn't even know what kind of entities words are. Language has functional value only when the words carry some significance with respect to the states of the world as perceived by the language user.

Approximating the ideal model with TPs

In order to demonstrate the feasibility of the joint model of segmentation and meaning acquisition, a simple computational implementation of the model was created by utilizing the idea of using transition probabilities (TPs) to perform statistical learning on language input (c.f., e.g., Saffran et al., 1996), but now conditioned on the referential context. This is not to suggest that humans would actually utilize TPs over discrete representations of speech. Instead, the discrete domain analysis should be simply seen as a practical tool for analyzing statistical regularities of speech that, in the general case, reside in a much more complex multidimensional acoustic or perceptual space.

Let us start by assuming that speech input X is represented as a sequence of discrete acoustic events $X = [a_1, a_2, \ldots, a_L]$, where each event a belongs to a finite alphabet A ($a \in A$). These events can be any descriptions of a speech signal that can be derived in an unsupervised manner, and they are assumed to be shorter in time than any meaningful patterns of the language. Eq. (4) states that the quality of the lexicon is proportional to the probability that speech X is observed during referent c. By substituting X with the discrete sequence representation, the maximum likelihood estimate for $P(c \mid X, \theta_c)$ is given as

$$P(c \mid X,\theta) = P(c \mid a_1,\ldots,a_N,\theta) = \frac{F(a_1,a_2,\ldots,a_N \mid c)}{\sum_c F(a_1,a_2,\ldots,a_N \mid c)} \quad (6)$$

where $F(a_1, a_2, \ldots, a_N \mid c)$ is the frequency of observing the corresponding sequence $a_1, a_2, \ldots, a_N$ concurrently with referential context c. In the general case, this solution is infeasible since the distribution $P(a_1, \ldots, a_N \mid c)$ cannot be reliably estimated from any finite data for $N >> 0$ in the presence of variability characteristic to normal speech. However, Eq. (6) can be approximated as a mixture of TPs between adjacent and non-adjacent states (see Räsänen & Laine, 2012, for details):

$$P(c,t \mid X) \propto \frac{\sum_k P(a_t \mid a_{t-k},c)}{\sum_{c,k} P(a_t \mid a_{t-k},c)} = \frac{\sum_k F(a_t,a_{t-k},c)}{\sum_{c,k}\sum_{a_t \in A} F(a_t,a_{t-k},c)} \quad (7)$$

Eq. (7) also makes a further simplifying assumption that $P(c)$ is a non-informative uniform distribution. Note that with $k = 1$, c = constant, and A being the set of syllables in the language, this model becomes equal to the basic TP-model used by Saffran et al. (1996).

In order to decode model information in terms of contiguous patterns instead of doing it frame-by-frame, the *activation* $A(c,t)$ of a referent (word) c at time t is given as

$$A(c \mid X_{t_1},\ldots,X_{t_2}) \approx \frac{1}{t_2 - t_1 + 1}\left(\sum_{t=t_1}^{t_2}\frac{\sum_k P(a_t \mid a_{t-k},c)}{\sum_{c,k} P(a_t \mid a_{t-k},c)}\right) \quad (8)$$

i.e., by simply integrating the context-dependent TPs over the time-window of analysis from t_1 to t_2 (see also Räsänen & Laine, 2012). Once the activation curves for referents have been computed, temporally contiguous above-chance activation of a referent c can be seen as a candidate word segment, or *cluster*, that is both familiar to the learner and that spans across both auditory and referential representational domains. In the experiments of the current paper, decoding in Eq. (8) is always performed in a sliding window of 250 ms. TPs are always estimated from lags $k = \{1, 2, \ldots, 25\}$, corresponding to temporal distances of 10-250 ms, as this time-scale captures the statistical regularities

of speech available at the low-level acoustic features (see Räsänen & Laine, 2012).

Simulations

Data and evaluation

The model was tested on pre-recorded continuous speech by using the Caregiver UK Y2 corpus (Altosaar et al., 2010). The material contains spoken utterances paired with visual tags denoting the concurrent presence of visual referents for the "keywords" (nouns, verbs, adjectives) in each sentence. In addition to the 1–4 referential keywords per utterance (mean 2.9), the utterances also contain additional words, such as function words (e.g., "*a* ***woman takes*** *the* ***yellow cookie***", referential keywords emphasized), leading to an average utterance length of 5.4 words. The main section of the corpus contains 2397 utterances for each talker, spoken in enacted, child-directed speaking style. There are a total of 50 unique keywords and corresponding visual referents in the corpus.

For each run of the simulation, half of the corpus (N = 1199 utterances) from Talker-01 was randomly assigned as the training data while the remaining half (N = 1198) was used to test the word-referent recognition performance of the model. The experiment was performed separately with the original referential information and with varying degrees of additional referential uncertainty by randomizing 20%, 40%, or 80% of the original visual referents to any of the 50 referents in the data. During the training stage, referents of the spoken keywords were always shown to the algorithm.

For each test utterance, the M words with the highest non-concurrent maxima in activation (Eq. 8) were chosen as the referent hypotheses, where M is the true number of referents associated with the utterance. The overall recognition performance was measured as the proportion of correct hypotheses across all test utterances and across five independent runs of the simulation.

Speech pre-processing

In order to represent speech in terms of short-term discrete events, Mel-frequency cepstral coefficients (MFCCs) representing the short-term spectrum of the speech were first computed from the speech signals using 25-ms sliding window with 10-ms steps. 10,000 randomly chosen MFCC-vectors were then clustered into 64 unique categories in an unsupervised manner using the standard k-means algorithm. Finally, each MFCC vector was assigned to the nearest cluster, leading to a discrete sequential representation of X = [a_1, a_2, …, a_N] with $a \in [1, 64]$ with one element a_t occurring every 10 ms (see, e.g., Räsänen, 2011 for more details).

Results

Fig. 1 shows an example of the model output for an early stage of the learning and after processing of the full training set and without added referential noise. As can be observed from the middle panel, the activation of each referent, given the audio, is relatively noisy after observing only 60 utterances (recall that there are 50 different referents and 1–4 referents per utterance in the dataset). In the bottom panel, the words "*small*" and "*tree*" have been successfully associated to their corresponding referents after training,

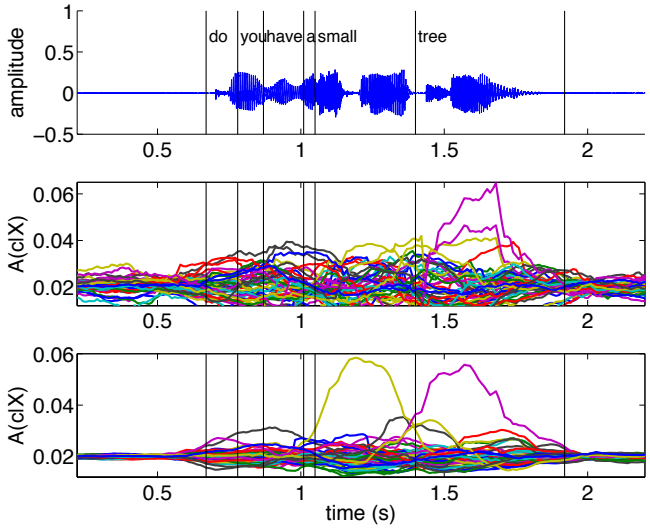

Figure 1: An example of the basic model output for the sentence "*Do you* ***have*** *a* ***small tree***?" (keywords with visual referents emphasized). Top: The original waveform. Middle: The model output after exposure to 60 sentences. Bottom: The model output after exposure to 1199 sentences. The different colored curves represent probabilities of different visual referents. The vertical lines show the true word boundaries extracted from the corpus annotation.

leading to clear activations that approximately correspond to the temporal extent of the underlying linguistic word forms, thereby also leading to segmentation of the input into word-like units. On the other hand, words without a visual referent (e.g., "*a*") do not have distinct activation segments. Also, activation of the referent {*to have*} extends to across the entire phrase "*do you have*" as it almost always occurs within this phrase in the corpus.

Top panel in Fig. 2 shows the word-referent recognition performance as a function of the number of utterances perceived by the learner. The result is shown for the original referential information where referents always correspond to the keywords in the spoken utterances. In addition, results with 20%, 40% and 80% of the original referents randomized are also shown in order to analyze model behavior under varying degrees of referential uncertainty. As can be seen from the results, the basic model successfully learns the word-referent mappings from the continuous utterances, achieving a mean referent recognition rate as high as M = 89.5% (SD = 0.4%) across all 50 keywords in the data. The final results for the three noise levels are $M_{0.2}$ = 89.1% ($SD_{0.2}$ = 0.8%), $M_{0.4}$ = 88.3% ($SD_{0.4}$ = 0.5%), and $M_{0.8}$ = 59.4% ($SD_{0.8}$ = 1.9%) in the order of increasing uncertainty. This shows that the model copes well with referential uncertainty since nearly 90% of the word tokens are associated to their correct referents even

when almost half (40%) of the attended referents are not related to the speech contents during learning. Even in the case of only 20% of referents being related to the words in the utterances, the performance is still 59.4% and would likely keep increasing with more training data.

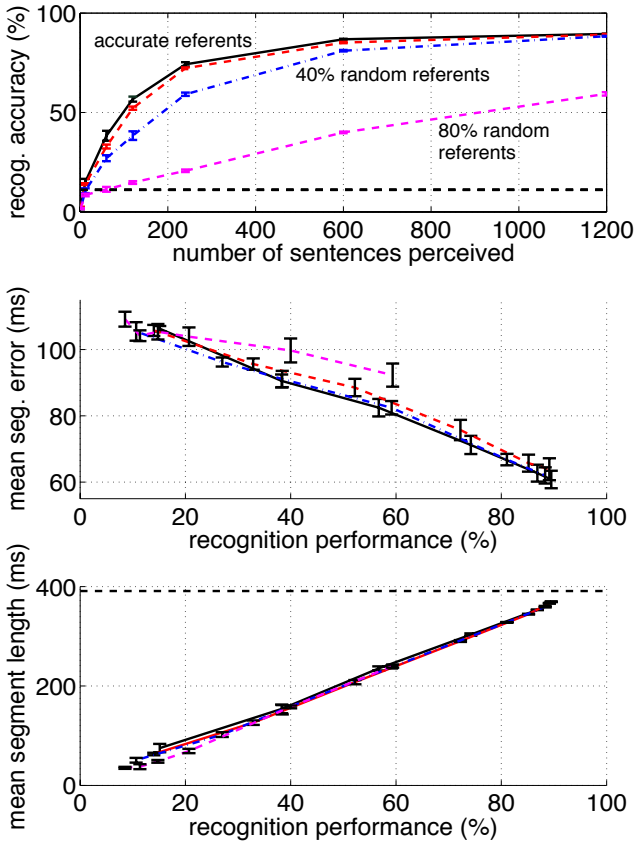

Figure 2: Top: Word-referent recognition performance as a function of the number of sentences with which the model is trained and for different levels of referential noise. Middle panel: The mean distance from annotated word boundaries to the model-generated boundaries (points where the winning referent changes) as a function of word recognition performance. Bottom panel: The mean duration of the model-generated and correctly associated word segments as a function of word-recognition performance. The black, red, blue, and magenta colored lines show the results with referential noise of 0%, 20%, 40%, or 80% of the original referent labels randomized to any of the 50 possible referents, respectively. The horizontal dashed line shows the chance level performance in the top panel and the true mean word length in the bottom panel. The error bars correspond to ±1 SE.

Middle panel in Fig. 2 shows the corresponding segmentation accuracy for all hypothesized word segments with respect to underlying word-level annotation and the bottom panel shows the segment length for correctly recognized words as a function of word-referent recognition performance. In all noise conditions, the model shows improvement in segmentation accuracy as more training data is observed and the final error of approximately 60 ms is small in comparison to the typical word durations.

Finally, bottom panel in Fig. 3 shows the mean segment length that approaches the true mean keyword length of ~400 ms as the recognition performance improves. In addition, the segment lengths of the correctly learned words are nearly identical at all referential uncertainty levels. This suggests that the learner first starts to discriminate different referential contexts based on short snippets of speech that are acoustically prominent in these contexts and then gradually learns the overall extent of the word-like units as more evidence is accumulated. In all, the model distinguishes different referents based on segments that are distinct in different referential contexts, ultimately converging to words or phrases that have referential meaning (also seen in Fig. 1).

Discussion and conclusions

In the present paper, we argued that language learners could utilize referential cues in communicative contexts by segmenting speech into units that are guaranteed to have referential significance. We provided a mathematical framework for joint-model of word segmentation and meaning acquisition by connecting referential value of the learned lexicon to the segmentation task. We tested the model in a word-learning simulation, showing that the model can successfully learn words from continuous speech.

The present results converge with earlier modeling studies using visual referential information for perceptual grounding of acoustic patterns (e.g., Räsänen et al., 2008; van Hamme, 2008; Aimetti, 2009; ten Bosch et al., 2009). All these models exhibit successful word learning after sufficient exposure to the language without any a priori linguistic knowledge, and the present mathematical framework explicates why this is the case, i.e., why the cross-modal strategy is valid for early word learning.

The idea of learning a referentially meaningful proto-lexicon without any phonological decoding of speech converges with the definition of proto-lexicon by Nazzi and Bertoncini (2003). Also, according to PRIMIR framework of language acquisition (Werker & Curtin, 2005) and recent work on learning of phonological categories (Feldman et al., 2013), it is likely that language learners have to acquire lexical knowledge before or in parallel with phonological representations instead of learning the sound system of the language before word learning. The present model provides one possible approach for bootstrapping the learning process by starting from proto-lexical learning that already results in meaningful representations of the language and thereby enables (receptive) language use before more sophisticated language skills emerge.

From a machine learning point of view, the present model can be characterized as weakly-supervised learning. The referential context provides labeling for the speech input, but the labels are noisy and inaccurate due to the referential ambiguity in each communicative situation. Efficiency of the learning is dominated by the learner's ability to limit the

number of possible referents in each communicative situation, possibly driven by attentional and social cues in case of real infants.

In general, it is likely that learners use a number of different strategies to bootstrap their word learning. This also involves the use of purely bottom-up cues to words and word boundaries (see the introduction). However, the essence of language is in the word meanings. An optimal language learner will therefore take the meanings of potential word segments into account when trying to make sense of the auditory world.

Acknowledgments

This research was funded by Academy of Finland, the ETA Graduate School of Aalto University, Finland, and the ERC starting grant project ABACUS, grant number 283435. The authors would like to thank Mike Frank and Unto K. Laine for their insightful comments on the paper.

References

Aimetti, G. (2009). Modelling early language acquisition skills: Towards a general statistical learning mechanism. *Proc. EACL'09,* Athens, Greece, pp. 1–9.

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A speech corpus for modeling language acquisition: CAREGIVER. *Proceedings of the International Conference on Language Resources and Evaluation*, Malta, pp. 1062–1068.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93–125.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–121.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psych. Review*, 120, 751–778.

Fourtassi, A. & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. Proc. *18th Conf. on Computational Natural Language Learning*, Baltimore, Maryland, pp. 191–200.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117, 107–125.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1), 1–55.

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin and Review*, 20, 1161–1169.

Johnson, M., Demuth, K., Frank, M. C., & Jones, B. K. (2010). Synergies in learning words and their referents. *Advances in Neural Information Processing Systems (NIPS 2010)*, 23, 1018–1026.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science*, 6, 136–142.

Pearl, L., Goldwater, S., & Steyvers, M., (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, 8, 107–132.

Pinker, S. (1989). *Learnability and cognition*. Cambridge, MA: The MIT Press.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science*, 26, 113–146.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, 120, 149–176.

Räsänen, O., Laine, U. K., & Altosaar, T. (2008). Computational language acquisition by statistical bottom-up processing. *Proc. Interspeech'08*, pp. 1980–1983, Brisbane, Australia.

Räsänen, O., & Laine U. (2012). A method for noise-robust context-aware pattern discovery and recognition from categorical sequences. *Pattern Recognition*, 45, 606–616.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, 108, 6038–6043.

ten Bosch, L., Van hamme, H., Boves, L., & Moore, R. K. (2009). A computational model of language acquisition: the emergence of words. *Fundamenta Informaticae*, 90, 229–249.

Thiessen, E. D. (2010). Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, 34, 1092–1106.

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, 706–716.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, 7, 53–71.

Van hamme, H. (2008). HAC-models: A novel approach to continuous speech recognition. *Proc. Interspeech'08,* pp. 2554–2557, Brisbane, Australia.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1, 197–234.

Yu, C., & Ballard, D. H. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions*, 1, 57–80.

Yu, C. & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, 119, 21–39.

Yu, C., Zhong, Y., & Fricker, D. (2012). Selective attention in cross-situational statistical learning: evidence from eye tracking. *Frontiers in Psychology*, 3, 1–16.