A joint model of word segmentation and meaning acquisition through cross-situational learning

Okko Räsänen[1] & Heikki Rasilo[1,2]

[1]Aalto University, Dept. Signal Processing and Acoustics, Finland

[2]Vrije Universiteit Brussel (VUB), Artificial Intelligence Lab, Belgium

Abstract

Human infants learn meanings for spoken words in complex interactions with other people, but the exact learning mechanisms are unknown. Among researchers, a widely studied learning mechanism is called cross-situational learning (XSL). In XSL, word meanings are learned when learners accumulate statistical information between spoken words and co-occurring objects or events, allowing the learner to overcome referential uncertainty after having sufficient experience with individually ambiguous scenarios. Existing models in this area 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 the meaning acquisition. In this paper, we argue that XSL is not just a mechanism for word-to-meaning mapping, but that it provides strong cues for proto-lexical word segmentation. If a learner directly solves the correspondence problem between continuous speech input and the contextual referents being talked about, segmentation of the input into word-like units emerges as a by-product of the learning. We present a theoretical model for joint acquisition of proto-lexical segments and their meanings without assuming a priori knowledge of the language. We also investigate the behavior of the model using a computational implementation, making use of transition probability -based statistical learning. Results from simulations show that the model is not only capable of replicating behavioral data on word learning in artificial languages, but also shows effective learning of word segments and their meanings from continuous speech. Moreover, when augmented with a simple familiarity preference during learning, the model shows a good fit to human behavioral data in XSL tasks. These results support the idea of simultaneous segmentation

and meaning acquisition and show that comprehensive models of early word segmentation should take referential word meanings into account.

Keywords: statistical learning, word learning, word segmentation, language acquisition, synergies in word learning

1. Introduction

Infants face many challenges in the beginning of language acquisition. One of them is the problem of word discovery. From a linguistic point of view, the problem can be posed as the question of 1) how to segment the incoming speech input into words and 2) how to associate the segmented words with their correct referents in the surrounding environment in order to acquire the meaning of the words. Many 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, 1996a), prosody (Cutler & Norris, 1988; Mattys, Jusczyk, Luce & Morgan, 1999; Thiessen & Saffran, 2003), or other properties of infant directed speech (Thiessen, Hill & Saffran, 2005), in order to find word-like units from speech (see also, e.g., Jusczyk, 1999).

Likewise, the problem of associating segmented words with 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 their 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 (e.g., Yu & Smith 2007; Smith & Yu, 2008; Smith, Smith & Blythe, 2011; Vouloumanos, 2008; Vouloumanos & Werker, 2009; Yurovsky, Yu & Smith, 2013; Yurovsky, Fricker, Yu & Smith, 2014; Suanda, Mugwanya & Namy, 2014) and that these statistics are accumulated and used incrementally across subsequent exposures to the word-referent co-occurrences (Yu & Smith, 2011; Yu, Zhong & Fricker, 2012).

Despite the progress in both sub-problems, a comprehensive integrated view on early word learning is missing. No existing proposal provides a satisfactory description of how word learning is initially bootstrapped without a priori linguistic knowledge, how these first words are represented in the mind of a pre-linguistic infant, how infants deal with the acoustic variability in speech in both segmentation and meaning acquisition, or how the acoustic or phonetic information in the early word representations interacts with the meanings of the words.

In order to approach the first stages of word learning from an integrated perspective, the early word learning problem can be also reformulated from a practical point of view: *How does the infant learn to segment speech into meaningful units?* When framed this way, there no longer is the implication that successful segmentation precedes meaning acquisition, but that the segment meaningfulness as such is the criterion for speech segmentation. Hence, the processes of finding words and acquiring their meaning become inherently intertwined, and the synergies between the two can make the segmentation problem easier to solve (see also Johnson, Demuth, Frank & Jones, 2010 and Fourtassi & Dupoux, 2014). One can also argue that segment meaningfulness should be the primary criterion in pre-lexical speech perception since the meaningful sound patterns (e.g., words or phrases) are those that have predictive power over the environment of the learner. In contrast, segmentation into linguistically proper word forms or

phonological units without meaning attached to them does not carry any direct practical significance to the child. The benefits of morphological or generative aspects of language only become apparent when the size of the vocabulary starts to exceed the number of possible sub-word units.

If infants are sensitive to statistical dependencies in the sensory input (e.g., Saffran et al., 1996a; Saffran, Newport & Aslin, 1996b; Saffran, Johnson, Aslin & Newport, 1999), it would be natural to assume that the earliest stages of word learning can be achieved with general cross-modal associative learning mechanisms between auditory perception and other representations originating from different modalities. Interestingly, recent experimental evidence shows that consistently co-occurring visual information helps in word learning from artificial spoken languages (Cunillera, Laine, Càmara & Rodríguez-Fornells, 2010; Thiessen, 2010; Yurovsky, Yu & Smith, 2012; Glicksohn & Cohen, 2013). This suggests that the segmentation and meaning acquisition problems may not be as independent of each other as they have been previously assumed to be.

Backed up by the behavioral findings, we argue in the current paper that XSL is not just a mechanism for word-to-meaning mapping, but that it can provide important cues for pre-lexical word segmentation, thereby helping the learner to bootstrap the language learning process without any a priori knowledge of the relevant structures of the language. We also put forward the hypothesis that cross-modal information acts as glue between variable sensory percepts of speech, allowing the infants to overcome the differences between realizations of the same word and thereby to form equivalence classes (categories) for speech patterns that occur in similar referential contexts. We follow the statistical learning paradigm for both segmentation and XSL, assuming that XSL is actually just a cross-modal realization of the same statistical learning

mechanisms observed within individual perceptual modalities, and operating whenever the representations within the participating modalities are sufficiently invariant to allow the discovery of statistical regularities between them.

The paper is organized as follows: Section 2 provides a brief overview of how the problems of statistical word segmentation and cross-situational learning have been explored in the existing behavioral and computational research. Section 3 presents a formal joint model of speech segmentation and meaning acquisition, describing at the computational level (c.f. Marr, 1982) why referential context and socially guided attention are relevant to the word segmentation problem, and why the two problems are solved more efficiently together than separately. Section 4 describes an algorithmic implementation of the ideal model by connecting the theoretical framework to the transition probability (TP) analysis used in many previous studies. The behavior of the model is then studied in six simulation experiments described in section 5. Finally, implications of the present work are discussed in section 6.

Before proceeding, it should be noted that much of the present work draws from the research on self-learning methods for automatic speech recognition (e.g., ten Bosch, van Hamme, Boves & Moore, 2009; Aimetti, 2009; Räsänen, Laine & Altosaar, 2008; Van hamme, 2008; see also Räsänen, 2012, for a review). One of the aims of this paper is therefore also to provide a synthesis of the early language acquisition research undertaken in the cognitive science and speech technology communities in order to better understand the computational aspects of early word learning.

2. Statistical learning, word segmentation and cross-situational learning

2.1 Statistical word segmentation

Statistical learning refers to the finding that infants and adults are sensitive to statistical regularities in the sensory stimuli and that these regularities can help the learner to segment the input into recurring patterns such as words. For instance, sensitivity to statistical dependencies between subsequent syllables can be already observed at the age of 8 months, enabling infants to differentiate words that have high internal TPs between syllables from non-words with low-probability TPs (Saffran et al., 1996a; Saffran et al. 1996b; see also Aslin & Newport, 2014, for a recent review). An increasing amount of evidence also shows that the statistical learning is not specific to speech, but operates across other auditory patterns (Saffran et al., 1999), and in other sensory modalities, such as vision (Fiser & Aslin, 2001; Kirkham, Slemmer & Johnson, 2002; Baldwin, Andersson, Saffran & Meyer, 2008) and tactile perception (Conway & Christiansen, 2005).

However, what the actual output of the segmentation process is and how it interacts with language learning in infants is yet to be established. One possibility is that infants use low-probability TPs surrounding high-probability sequences as candidate word boundaries, thereby performing segmentation of the input into mutually exclusive temporal regions, referred to as *bracketing*. Another possibility is that infants *cluster* acoustic events with high mutual co-occurrence probabilities (high TPs) together (Goodsitt, Morgan & Kuhl, 1993; see also Swingley, 2005; Giroux & Rey, 2009; Kurumada, Meylan & Frank, 2013), thereby forming stronger representations for consistently recurring entities such as words while clusters crossing word boundaries tend to diminish as they receive less reinforcement from the perceived input (low TPs; c.f., Perruchet & Vinter, 1998).

Following the behavioral findings, computational modeling of statistical word segmentation has been investigated from phonetic features or transcriptions (de Marcken, 1995; Brent & Cartwright, 1996; Brent, 1999; Goldwater, Griffiths & Johnson, 2009; Pearl, Goldwater & Steyvers, 2010; Adriaans & Kager, 2010; Frank, Goldwater, Griffiths & Tenenbaum, 2010) and directly from acoustic speech signals without using linguistic representations of speech (e.g., Park & Glass, 2005, 2006; McInnes & Goldwater, 2011; Räsänen, 2011; see also Räsänen & Rasilo, 2012). These approaches show that often recurring word-like segments can be detected from the input.

However, a significant issue in the models operating at the phonetic level is that the acquisition of the language's phonetic system hardly precedes the learning of first words (Werker & Curtin, 2005). Even though infants show adaptation to the native speech sound system during the first year of their life (Werker & Tees, 1984; Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992), the phonetic and phonemic acquisition are likely to be dependent on lexical learning (Swingley, 2009; Feldman, Griffiths & Morgan, 2009; Feldman, Myers, White, Griffiths & Morgan, 2013; see also Elsner, Goldwater & Eisenstein, 2012; Elsner, Goldwater, Feldman & Wood, 2013). This is primarily due to the immense variability in the acoustic properties of the speech, making context-independent bottom-up categorization of speech into phonological units impossible without constraints from, e.g., lexicon, articulation or vision (see Räsänen, 2012, for a review). This is also reflected in children's challenges at learning phonologically similar word forms during their second year of life (Stager & Werker, 1997; Werker, Cohen, Lloyd, Casasola & Stager, 1998), and in that phonological development seems to continue well into childhood (see Rost & McMurray, 2009, 2010, or Apfelbaum & McMurray, 2011, for an overview and discussion). Preceding or parallel lexical learning is suggested by the findings that 6-month-old

infants are already capable of understanding the meaning of certain high-frequency words although their phonetic awareness of the language has just started to develop (Bergelson & Swingley, 2012; see also Tincoff & Jusczyk, 1999). In addition, the sound patterns of words seem to be phonologically underspecified at least up to the age of 18 months (Nazzi & Bertoncini, 2003, and references therein). Sometimes young children struggle with learning new minimal pairs (Stager & Werker, 1997; Werker et al., 1998) while in other conditions they succeed (Yoshida, Fennell, Swingley & Werker, 2009) and show high sensitivity to mispronunciations of familiar words (e.g., Swingley & Aslin, 2000). However, since acoustic variation in speech generally affects word learning and recognition (e.g., Rost & McMurray, 2009, 2010; Houston & Jusczyk, 2000; Singh, White & Morgan, 2008; Bortfeld & Morgan, 2010), the overall findings suggest that the representations of early words are not based on invariant phonological units, but are at least partially driven by acoustic characteristics of the words (see also Werker & Curtin, 2005). Therefore, early word learning cannot be assumed to operate on a sequence of well-categorized phones or phonemes (see also Port, 2007, for a radical view).

Computational models of acoustic speech segmentation bypass the problem of phonetic decoding of the speech input (Park & Glass, 2005, 2006; McInnes & Goldwater, 2011; Räsänen, 2011). However, they show only limited success in the segmentation task, being able to discover recurring patterns that have only limited acoustic variation. As these approaches represent words in terms of frequently recurring spectrotemporal acoustic patterns without any compositional or invariant description of the subword structure, their generalization capabilities to multiple talkers with different voices or even different speaking styles by the same speaker are limited. Also, as will be seen in section 3, the referential value of these patterns is not known to the learning

algorithm, forcing the learning to use some heuristic that is only indirectly related to the quality of the discovered patterns, and more often biased by the algorithm designer's view of desired outputs.

2.2 Cross-situational learning

As for the word meaning acquisition, the operation of the XSL mechanism has been confirmed in many behavioral experiments. In their seminal work, Yu and Smith (2007; also Smith & Yu, 2008) showed that infants and adults are sensitive to cross-situational statistics between co-occurring words and visual objects, enabling them to learn the correct word-to-object pairings after a number of ambiguous scenarios with multiple words and objects. Later studies have confirmed these findings for different age groups (Vouloumanos, 2008; Yurovsky et al., 2014; Suanda et al., 2014), analyzed the operation of XSL under different degrees of referential uncertainty (Smith et al., 2010), and also shown with eye-tracking and other experimental settings how cross-situational representations evolve over time during the learning process (Yu & Smith, 2011; Yu et al., 2012; Yurovsky et al., 2013). There has also been an ongoing debate on whether XSL scales to the referential uncertainty present in the real world (e.g., Medina, 2011), and recent evidence suggests that the limited scope of an infant's visual attention may limit the uncertainty to a level that still allows XSL to operate successfully in real world conditions (Yurovsky, Smith & Yu, 2013).

In addition to studying XSL in human subjects, XSL has been modeled using rule-like (Siskind, 1996), associative (Kachergis, Yu & Shiffrin, 2012; McMurray, Horst & Samuelson 2012; Rasilo & Räsänen, 2015), and probabilistic computational models (Frank, Goodman & Tenenbaum, 2007; Fazly, Alishahi & Stevenson, 2010), and also through a purely mathematical analysis (Smith, Smith, Blythe & Vogt, 2006; see also Yu & Smith, 2012b). All these approaches

show that XSL can successfully learn word-to-referent mappings under individually ambiguous learning scenarios when the learner is assumed to attend to a limited set of possible referents in the environment, e.g., due to joint-attention with the caregiver, intention reading, and other social constraints (see Landau, Smith & Jones, 1988; Markman, 1990; Tomasello & Todd, 1983; Tomasello & Farrar, 1986; Baldwin, 1993; Yu & Ballard, 2004; Yurovsky et al., 2013; Frank, Tenenbaum & Fernald, 2013; Yu & Smith, 2012a). However, the existing models assume that the words are already segmented from speech and represented as invariant linguistic tokens across all communicative situations. Given the acoustic variability of speech, this is a strong assumption for early stages of language acquisition, and these models apply better to learners who are already able to parse speech input into word-like patterns in a consistent manner.

2.3 An integrated approach to segmentation and meaning acquisition

The fundamental problem in the "segmentation first, meaning later" approach is that the use of spoken language is primarily practical at all levels. Segmenting speech into proper words before attaching any meaning to them has little functional value for an infant. In contrast, situated predictive power (the meaning) of grounded speech patterns such as words or phrases provide the learner with an enhanced capability to interact with the environment (see also ten Bosch et al., 2009). As word meanings are acquired through contextual grounding, the word referents have to be present every time new words are learned at a level that also serves communicative purposes. The importance of grounding in early world learning is also reflected in the vocabularies of young children as a notable proportion of their early receptive vocabulary consists of directly observable states of the world, such as concrete nouns or embodied actions (MacArthur Communicative Development Inventories; Fenson et al., 1993).

Another factor to consider is that real speech is a complex physical signal with many sources of variability even between linguistically equivalent tokens, such as multiple realizations of a same word (Fig. 1). This makes discovery of regularities in the speech stream much more challenging than what can be understood from the analysis of phonetic or phonemic representations of speech. Also, typical stimuli used in behavioral experiments have only limited variability between word tokens. In contrast, pre-linguistic infants listen to speech without knowing what speech sounds should be treated as equivalent and what sounds are distinctive in their native language (c.f., Werker & Tees, 1984), making the discovery of functionally equivalent language units by finding matching repetitions of acoustic patterns an infeasible strategy.

In this context, consistently co-occurring visual referents, such as concrete objects, may act as the glue that connects uncertain acoustic patterns together as these patterns share similar predictions about these referents in the environment. Considering the clustering approach to word segmentation, contextual referents may actually form the basis for a word "cluster" due to their statistically significant correlation with the varying speech acoustics, while the acoustic patterns might not share a high correlation directly with each other (see Coen, 2006, for a similar idea with speech and mouth movements). Referents may also play a role in phonetic learning by providing indirect equivalence evidence for different pronunciation variants of speech sounds, thereby establishing a proto-lexicon that mediates these equivalence properties (e.g., Feldman et al., 2013) and helping infants to overcome the minimal but significant differences in phonological forms by contrasting the relevant and irrelevant dimensions of variation across word tokens in the presence of the same referent (Rost & McMurray, 2009, 2010). From this perspective, it would be almost strange if the systematic contextual cues would not affect the

word segmentation process since the communicative environment actually provides (noisy) labeling to the speech contents (c.f., Roy, Frank & Roy, 2012), and since the human brain seems to be sensitive to almost all types of statistical regularities in the environment within and across sensory modalities. The idea is also backed up by the fact that the development of basic visual perception is known to take place before early word learning, leading to a categorical perception of objects and entities and to at least partial object permanence already during the first 8 months of infancy (e.g., Spelke, 1990; Eimas & Quinn, 1994; Johnson, 2001). Moreover, McMurray et al. (2012) have argued for so-called slow learning in word acquisition where knowledge of word meanings accumulates slowly over time with experience (see also Kucker, McMurray & Samuelson, 2015). This "developmental-time" process is paralleled with dynamical competitive processes at a shorter time-scale that are responsible for interpreting each individual communicative situation using the existing knowledge. This framework extends naturally to gradual acquisition of word segments together with their meanings. More specifically, we believe that word segmentation also results from dynamic competition between alternative interpretations of the message in the ongoing communicative context.

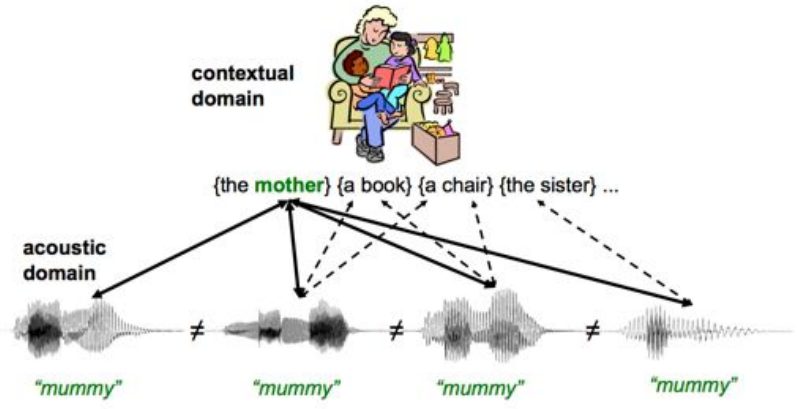


Figure 1: A schematic view of a contextual referent as the common denominator between multiple acoustic variants of the same word. Clustering of phones, syllables, or spoken words based on acoustic

similarity alone would lead to either under- or overspecified representations. Meaningful acoustic/phonetic distinctions and temporal extents of speech patterns are only obtained by taking into account the referential predictions of the signal (see also Werker & Curtin, 2005).

Behavioral evidence for referential facilitation of segmentation comes from the studies of Cunillera et al. (2010), Thiessen (2010), Glickson and Cohen (2013), and Shukla, White and Aslin (2011), who all performed experiments with parallel audiovisual segmentation and meaning acquisition in an artificial language. Cunillera et al. (2010) found out that when each trisyllabic word was deterministically paired with a meaningful picture, only the words that were successfully associated to their referents were also segmented at above-chance-level accuracy by adult subjects. Similarly, Glicksohn and Cohen (2013) found significant facilitation for learning of high-TP words when they were paired with consistent visual referents. In contrast, conflicting visual cues led to worse word learning performance than what was observed in a purely auditory condition. Thiessen (2010) tested adults and infants in the original statistical segmentation task of Saffran et al. (1996) and found that adult word segmentation was significantly higher with consistent visual cues and that segmentation performance and referent learning performance were correlated. However, 8-month-old infants did not show effects of visual facilitation (Thiessen, 2010), suggesting that the cross-modal task was either too hard for them, that the cross-modal associations simply need a more engaging learning situation (c.f., Kuhl, Tsao & Liu, 2003), or that the preferential looking paradigm simply failed to reveal different grades of familiarity with the words since the performance in the non-visual control condition was already at above-chance level (but see also experiment 2 of this paper).

Interestingly, evidence for concurrent segmentation and meaning acquisition already in 6-month-old infants comes from Shukla et al. (2011). In their experiments, infants succeeded in the

mapping of bisyllabic words of an artificial language to concurrently shown visual shapes as long as the words did not straddle an intonational phrase boundary. In addition to using real pre-recorded speech with prosodic cues instead of synthesized speech used in all other studies, Shukla et al. also used moving visual stimuli, possibly leading to stronger attentional engagement in comparison to the subjects in the study of Thiessen (2010). Unfortunately, Shukla et al. did not test for word segmentation with and without referential cues, leaving it open whether infants learned word segments and their meaning simultaneously, or whether they learned the segmentation first based on purely auditory information.

In addition, the studies of Frank, Mansinghka, Gibson and Tenenbaum (2007) and Yurovsky et al. (2012) show that adults are successful in concurrent learning of both word segments and their referential meanings when exposed to an artificial language paired with visual objects. However, unlike the above studies, Frank et al. did not observe improved word segmentation performance when compared against a purely auditory learning condition, possibly because the subjects were already performing very well in the task. Yurovsky et al. (2012) investigated the effects of sentential context and word position in XSL, showing successful segmentation and acquisition of referents for target words that were embedded within larger sentences of an artificial language. Unfortunately, Yurovsky et al. did not control for segmentation performance in a purely auditory learning task. This leaves open the possibility that the learners might have first learned to segment the words based on the statistics of the auditory stream only, and only later associated them to their correct visual referents (see Mirman, Magnuson, Graf Estes & Dixon, 2008, and Hay, Pelucchi, Graf Estes & Saffran, 2011, for evidence that pre-learning of auditory statistics helps in subsequent meaning acquisition).

Table 1 summarizes the above studies. Adult interpretation of acoustic input is evidently dependent on the concurrently available referential cues during learning. However, the current data on infants are sparse, and therefore it is unclear in what conditions infants can utilize cross-situational information, and whether the performance is constrained by the available cognitive resources, degree of attentional engagement in the experiments, or simply due to differences in learning strategies. Finally, it is important to point out that all above studies measure segmentation performance using a familiarity preference task, comparing high-TP and low-TP syllable sequences against each other. As will be shown in the experiments of section 5, a statistical learner can perform these tasks without ever explicitly attempting to segment speech into its constituent words.

Table 1: Summary of the existing studies investigating word learning with visual referential cues. *Learning of segments* refers either to successful word vs. part/non-word discrimination, or to learning of visual referents for words embedded in continuous speech. *Visual facilitation on segmentation* is considered positive only if the presence of visual cues leads to improvement in familiarity-preference tasks in comparison to a purely auditory baseline or to a condition with inconsistent visual cues.

study	natural speech	age	visual facilitation on segmentation	learning of segments	learning of referents	manipulation
Cunillera et al. (2010)	No	adults	yes	yes	yes	visual cue reliability
Frank et al. (2007)	No	adults	no	yes	yes	visual cue reliability, word position
Glicksohn & Cohen (2013)	No	adults	yes	yes	N/A	visual cue reliability
Shukla et al. (2011)	Yes	6 mo	N/A	yes	yes	prosodic phrase boundary location
Thiessen (2010)	No	8 mo	no	yes	no	visual cue reliability
Thiessen (2010)	No	adults	yes	yes	yes	visual cue reliability
Yurovsky, Yu & Smith (2012)	No	adults	N/A	yes	yes	carrier phrase word order

2.4 Existing computational models of integrated learning.

In terms of computational models, it was almost twenty years ago that Michael Brent noted that "*It would also be interesting to investigate the interaction between the problem of learning word meanings and the problem of segmentation and word discovery*" (Brent, 1996). Since then, a handful of models have been described in this area. Possibly the first computational model using contextual information for word segmentation from actual speech is the seminal Cross-channel Early Lexical Learning (CELL) model of Roy & Pentland (2002). CELL is explicitly based on the idea that "*a model acquires a lexicon by finding and statistically modeling consistent intermodal structure*" (Roy & Pentland, 2002). CELL assumes that the learnable words recur in close temporal proximity in infant directed speech while having a shared visual context. The model therefore cannot accumulate XSL information over multiple temporally distant utterances for segmentation purposes, but it still shows successful acquisition of object shape names from object images while the concurrent speech input was represented in terms of phone-like units obtained from a supervised classifier. CELL was later followed by the model of Yu & Ballard (2004), where phoneme sequences that co-occur with the same visually observed actions or objects are grouped together and the common structure of these phoneme sequences across multiple occurrences of the same context are taken as word candidates. Both CELL and the system of Yu and Ballard show that word segmentation can be facilitated by analyzing the acoustic input across communicative contexts instead of modeling speech patterns in isolation. However, the learning problem was simplified in both models by the use of pre-trained neural network classifiers to convert speech input into phoneme-like sequences before further processing, allowing the models to overcome a large proportion of the acoustic variability in

speech that is hard to capture in a purely bottom-up manner (cf. Feldman et al., 2013). Nevertheless, these models provide the first evidence that visual context can be used to bootstrap word segmentation (see also Johnson et al., 2010 and Fourtassi & Dupoux, 2014, for joint models operating at the phonemic level and Salvi, Montesano, Bernadino & Santos-Victor, 2012, for related work in robotics).

In parallel to the early language acquisition research, there has been increasing amounts of interest in the speech technology community to shift towards automatic speech recognition systems that could learn similarly to humans simply by interacting with their environments (e.g., Moore, 2013; 2014). This line of research has spurred a number of word learning algorithms that all converge to the same idea of using help from contextual visual information in building statistical models of acoustic words when no a priori linguistic knowledge is available to the system. These approaches include the TP-based models of Räsänen et al. (2008) and Räsänen and Laine (2012), the matrix-decomposition-based methods of Van hamme (2008) and ten Bosch et al. (2009; see also, e.g., Driesen & Van hamme, 2011), and the episodic-memory-based approach of Aimetti (2009). Characteristics of these models have been investigated in various conditions related to caregiver characteristics (ten Bosch et al., 2009), uncertainty in visual referents (Versteegh, ten Bosch & Boves, 2010), and preference for novel patterns in learning (Versteegh, ten Bosch & Boves, 2011). The common aspect in all of these models is that they explicitly or implicitly model the joint distribution of acoustic features and the concurrently present visual referents across time and use the referential information to partition ("condition") the acoustic distribution into temporal segments that predict the presence of the visual objects. This leads to the discovery of acoustic segments corresponding to the visual referents, thereby solving the segmentation problem without requiring any a priori information of the relevant units of the

language (see also Rasilo, Räsänen & Laine, 2013, for a similar idea in phonetic learning where learner's own articulatory gestures act as a context for a caregiver's spoken responses). Unfortunately, the above body of work seems to be largely disconnected from the other language acquisition research due to the highly technical focus of these papers. Also, the findings and predictions of these models have been only superficially compared to that of human behavior.

Building on the existing behavioral and computational modeling background, this paper provides a formal model of how cross-situational constraints can aid in the bootstrapping of the speech segmentation process when the learner has not yet acquired consistent knowledge of the language's phonological system. By simultaneously solving the ambiguity of reference (Quine, 1960) and the ambiguity of word boundaries, the model is capable of learning a proto lexicon of words without any language-related a priori knowledge.

3. A formal model of cross-situationally constrained word segmentation and meaning acquisition

The goal of section 3 is to show that simultaneous word segmentation and meaning acquisition is actually a computationally easier problem than separate treatment of the two, and that the joint approach directly leads to a functionally useful representation of the language. Moreover, this type of learning is achievable before the learner is capable of parsing speech using any linguistically motivated units such as phones or syllables—representations that are hard to acquire before some type of proto-lexical knowledge is already in place, as discussed in section 2.1.

We start by formulating a measure for referential quality of a lexicon that quantifies how well the lexicon corresponds to the observed states of the external world, i.e., things that are being talked about. This formulation is then contrasted against the sequential model of

segmentation and meaning acquisition where these two stages take place separately. The comparison reveals that any solution for the segmentation problem, when treated in isolation, is obtained independently of the referential quality of the resulting lexicon, and therefore the sequential process leads to sub-optimal segmentation with respect to word meanings. In other words, a learner, concerned with the link between the words and their meanings, should pay attention to the referential domain of words already during the word segmentation stage. We present a computational model of such a learner in section 3.2, showing that joint solving of segmentation and meaning acquisition directly optimizes the referential quality of the learned lexicon. Then we describe one possible algorithm-level implementation of the ideal model in section 3.3.

Schematic overviews of the standard sequential and the presently proposed joint strategies are shown in Fig. 2.

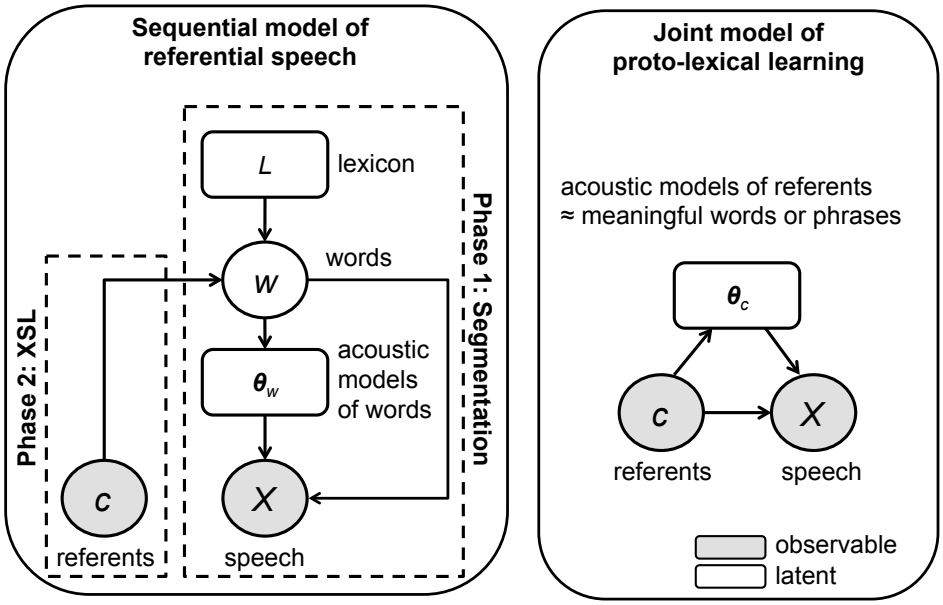


Figure 2: Sequential model of word learning including the latent lexical structure (left) and the flat cross-situational joint model (right). Note that both models assume that intentional and attentional factors are

implicitly used to filter the potential set of referents during the communicative situation, that both models neglect explicit modeling of subword structure, and that they avoid specifying the nature of speech representations in detail.

All following analyses are simplified by assuming that early word learning proceeds directly from speech to words without explicitly taking into account an intermediate phonetic or syllabic representation (cf. Werker & Curtin, 2005). However, this does not exclude the incorporation of native language perceptual biases or other already acquired subword-structures in the representation of speech input (see Fig. 2), although these are not assumed in the model. Instead, it is assumed that all speech is potentially referential, and the ultimate task of the learner is to discover which segments of the speech input have significance with respect to which external referents. Similarly to the model of Fazly et al. (2010), we assume that the set of possible referents in each communicative situation is already constrained by some type of attentional and intentional mechanisms and social cognitive skills (e.g., Frank et al., 2009; Frank et al., 2013; Landau et al., 1988; Markman, 1990; Yu & Smith, 2012a; Yurovsky et al., 2013; Tomasello & Todd, 1983).

The learner's capability of representing the surrounding environment in terms of discrete categories during word learning is also assumed similarly to all other models of XSL (e.g., Fazly et al., 2010; Frank et al., 2007; Kachergis et al., 2012; McMurray et al., 2012; Smith et al., 2006; Yu & Smith, 2012b). This assumption is justified by numerous studies that show visual categorization and object unity already at the age of 8–14 months (e.g., Mandler & McDonough, 1993; Eimas & Quinn, 1994; Bauer, Dow & Hertsgaard, 1995; Behl-Chadha, 1996; Marechal & Quinn, 2001; Oakes & Ribar, 2005; Spelke, 1990), the age for the beginning of vocabulary growth (Fenson et al., 1993). Although language itself may impact the manner in which

perceptual domains are organized (the Sapir-Whorf hypothesis), modeling of the two-directional interaction between language and non-auditory categories is beyond the scope of the present paper. In the limit, the present argument only requires that the representations of referents are systematic enough to be memorized and recognized at above chance probability, and that they occur with specific speech patterns with above chance probability. Under the XSL framework, potential inaccuracies in visual categorization can be seen as increased referential uncertainty in communicative situations, simply leading to slower learning with increasing uncertainty (Smith, Smith & Blythe, 2011; Blythe, Smith & Smith, 2014). Overall, the present model only assumes that the learner has access to the same cross-modal information as any learner in the existing XSL studies, but with the important exception that correct word forms are not given to the learner a priori but must be learned in parallel with their meanings.

Throughout this paper, the concept of a "referential context" is mostly used interchangeable with "visual referents". However, in any learning system, it is always the *internal representations* of the external and internal environment of the system that participate in learning and memory. This means that the internally represented state of a context is not equivalent to the set of auditory and visual stimuli presented by the experimenter, but is, at best, a correlate of the externally observable world. Besides the neurophysiological constraints of a biological system, this means that the system is completely agnostic to the source of the contextual representations, be they from visual or haptic perception or be they externally or internally activated (see experiment 6).

3.1 Measuring the referential quality of a lexicon

We start deriving the joint model from the definition of an effective lexicon. Assuming that a learner has already acquired a set of discrete words w that make up a lexicon L ($w \in L$), the

words have to be associated with their meanings in order to play any functional role. Further assuming that speech is referential with respect to the states of the surrounding world, a good word for a referent is the one that has high predictive value for the presence of the referent. According to information theory, and by using $c \in C$ to denote contextual referents of words $w \in L$, the mutual information (MI) between a word and a referent is given by

$$\mathrm{MI}(c,w) = P(c,w)\log_2 \frac{P(c,w)}{P(c)P(w)} \quad (1)$$

where $P(c,w)$ is the probability of observing the word w and referent c together while $P(w)$ and $P(c)$ are their base rates. MI quantifies the amount of information (in bits) that we know about the referential domain C ("the environment") given a set of words and vice versa (the word *informs* the listener about the environment, while the environment generates word-level descriptions of itself in the mind of the listener; see also Bergelson & Swingley, 2013). If MI is zero, nothing is known about the state of the referential domain given the word. The referent c^* of which there is most information conveyed by a word w is obtained by

$$c^* = \underset{c}{\mathrm{argmax}}\{\mathrm{MI}(c,w)\} \quad (2)$$

whereas the *referential value* (or *information value*) of the entire lexicon with respect to the referents is the total of information across all pairs of words and referents:

$$Q = \sum_{w,c} P(w,c)\log_2 \frac{P(w,c)}{P(w)P(c)} / \max\{\log_2 |C|, \log_2 |L|\} \quad (3)$$

where $|C|$ is the total number of possible referents and $|L|$ is the total number of unique words in the lexicon. Q achieves its maximum value of one when each word w co-occurs only with one

referent c ($|L| = |C|$)[1], i.e., there is no referential ambiguity at all. On the other hand, Q approaches zero when words occur independently of the referents, i.e., there is no coupling between the lexical system and the surrounding world. The logarithmic base normalization term $\max\{\}$ in Eq. (3) ensures that Q will be less than one if the total number of referents is larger than the number of words in the lexicon ($|L| < |C|$), even if the existing words have a one-to-one relationship with their referents, meaning that some of the potential referents cannot be addressed by the language. Similarly, if there are more words than there are referents ($|L| > |C|$), the quality of the lexicon decreases even if each word always co-occurs with only one referent, making acquisition of the vocabulary more difficult for the learner as more exposure is needed to learn all the synonymous words for the referents (at the limit, there are infinitely many words that go with each referent, making word recognition impossible). Overall, the larger the Q, the less there is uncertainty about the referential context c, given a set of words w. Although detailed strategies may vary, any XSL-based learner has to approximate the probability distributions in Eq. (3) in order to settle on some type of mapping from words to their referents across individually ambiguous learning scenarios (see Yu & Smith, 2012b).

Central to the thesis of the current paper, if the processes of segmentation and word-referent mapping were to take place sequentially, the word forms w would already have been determined before their meaning becomes of interest (c.f., Fig. 2). This means that the ultimate referential quality of the lexical system in Eq. (3) is critically dependent on the segmentation while the segmentation process is carried out independently of the resulting quality, i.e., without

[1] In this theoretical case, the state of the referential domain is fully determined by the currently observed words while any deviation from one-to-one mapping will necessarily introduce additional uncertainty to the system. In addition, a vocabulary with $Q = 1$ is the most economic one to learn because there are no multiple alternative words that might refer to the same thing, therefore requiring more learning examples.

even knowing if there is anything in the external world that the segmented words might refer to. For computational investigations of bottom-up word segmentation, this issue is easily obscured since the models and their initial conditions and parameters can be adjusted for optimal performance with respect to an expert-defined linguistic ground truth, steering the model in the right direction through trial and error during its development. Moreover, models operating on relatively invariant phone- or phoneme-level descriptions of the language bypass the challenge of acoustic variability in speech, having little trouble to determine whether two segments of speech from two different talkers correspond to the same word. Infants, on the other hand, do not have access to the linguistic ground truth nor can they process the input indefinitely many times with different learning strategies or initial conditions in order to obtain a useful lexical interpretation of the input, calling for robust principles to guide lexical development.

Appendix A describes a mathematical formulation of the word segmentation problem in isolation, showing that the problem is difficult due to multiple levels of latent structure. Given speech input, the learner has to simultaneously derive identities of words in the lexicon, the location of the words in the speech stream, and also how these words are realized in the acoustic domain. Of these factors, only the acoustic signal is observable by the learner, and therefore the problem has no known globally optimal solution. Yet, even a successful solution to this segmentation problem does not guarantee that the segments actually stand for something in the external world and are therefore useful for communicative purposes. In contrast, joint optimization of the segmentation and the referential system by maximizing Eq. (3) leads to optimal parsing of the input from the referential information point of view. As will be seen in the next section, the assumption of a latent lexical structure is unnecessarily complicated for this purpose and unnecessary for learning the first words.

3.2 Joint model of word segmentation and meaning acquisition using cross-situational word learning

The starting point for the joint model of word learning is the assumption of statistical learning as a domain-general mechanism that also operates not only within, but also across perceptual domains. This means that the statistical learning space is *shared* between modalities and driven by the regularities available at the existing representational levels.

In the context of this framework, the simplest approach to early word learning is to consider the direct coupling of the speech X with the referential context c through their joint distribution $P(X,c)$ (Fig. 2, right) and to derive its relation with respect to the joint distribution $P(w,c)$ of words and referents in Eq. (3). The joint distribution $P(X,c)$ captures the structure from situations where the speech content X co-occur with states c of the attended context with above-chance probability, i.e., where speech predicts the state of the surrounding world and vice versa. Our argument is that this distribution acts as the basis for learning of the first words of the language, or, following the terminology of Nazzi & Bertoncini (2003), learning of the first *proto-words,* i.e., words that can have practical use value but are not yet phonologically defined. Once the joint distribution $P(X,c)$ is known, it is straightforward to compute the most likely referents (meanings) of a given speech input. The main challenge is to model the so far abstract speech signal X in a manner that captures the acoustic and temporal characteristics of speech in different contexts c.

In order to do this, we replace the discrete words w of Eq. (3) with acoustic models $P(X \mid \theta_c)$ of speech signals X that occur during the concurrently attended referents c, where θ_c denotes the parameters of the acoustic model for c. In the same way, the probability of a word $P(w)$ is replaced by a global acoustic model $P(X \mid \theta_G)$ across all speech X, denoting the probabilities of

speech patterns independently of the referential context. Due to this substitution of words into referent-specific models of speech, there is now exactly one acoustic model θ_c for each referent c ($|C| = |L|$) and the overall quality of the lexicon in Eq. (3) can be written as

$$\begin{aligned} Q &= \sum_{w,c} P(w,c)\log_2 \frac{P(w,c)}{P(w)P(c)} / \max\{\log_2 |C|, \log_2 |L|\} \\ &= \sum_{X,c} P(X,c \mid \theta_c)\log_{|C|} \frac{P(X,c \mid \theta_c)}{P(X \mid \theta_{\mathrm{G}})P(c)} \\ &= \sum_{X,c} P(X,c \mid \theta_c)\log_{|C|} \frac{P(X \mid c,\theta_c)P(c \mid \theta_c)}{P(X \mid \theta_{\mathrm{G}})P(c)} \\ &= \sum_{X,c} P(X,c \mid \theta_c)\log_{|C|} \frac{P(X \mid c,\theta_c)}{P(X \mid \theta_{\mathrm{G}})} \end{aligned} \tag{4}$$

because $P(c)$ is independent of the parameters θ_c. Similarly to Eq. (3), the term inside the logarithm of Eq. (4) measures the degree of statistical dependency between referents c and speech patterns X. This means that $P(X \mid c, \theta_c) = P(X \mid \theta_{\mathrm{G}})$ if the speech signal and referents are independent of each other while $P(X \mid c, \theta_c) > P(X \mid \theta_{\mathrm{G}})$ if they are informative with respect to each other.

What these formulations show is that the overall quality of the vocabulary depends on how well the acoustic models θ_c capture the joint distribution of referents c during different speech inputs X and vice versa. There are two important aspects to observe here. Firstly, there is no explicit notion of words anywhere in the equations although quality of the referential communicative system has been specified. Secondly, the joint distribution $P(X, c)$ is directly observable to the learner. From a machine-learning perspective, learning of the acoustic model is a standard supervised parameter estimation problem with the aim of finding the set of parameters θ^* that maximizes Eq. (4):

$$\theta^* = \arg_{\theta}\max\{\sum_{X,c} P(X,c \mid \theta_c)\log_{|C|} \frac{P(X \mid c,\theta_c)}{P(X \mid \theta_{\mathrm{G}})}\} \tag{5}$$

Observe that now there are no other latent variables in the model besides the acoustic parameters. More importantly, the solution directly leads to useful predictive knowledge of the relationship between speech and the environment. In other words, optimizing the solution for Eq. (5) will also optimize the referential value of the lexicon. This shows that the direct cross-modal associative learning leads to effective representations of speech that satisfy the definition of proto-words (cf. Nazzi & Bertoncini, 2003).

Moreover, since the denominator is always smaller but approximately proportional[2] to the numerator for any X and c, and also assuming that θ_c are learned independently of each other, any increase in $P(X, c \mid \theta_c)$ for the observed X and c will necessarily increase the overall quality of the lexicon. Therefore, for practical acoustic model optimization purposes, we can make the following approximation[3]:

$$\Delta Q \tilde{\propto} \Delta \sum_{X,c} P(X,c \mid \theta_c) \qquad (6)$$

where Δ refers to a change in the values, i.e., improvement in the fit of the referent-specific joint distribution to the observed data will improve Q.

Eq. (5) and its approximation are easier to solve than the acoustic segmentation problem alone (see Appendix A) because the joint model has only one unknown set of parameters, the acoustic models θ, one for each referent and one for all speech. There are two mutually dependent latent variables L and θ_w in the sequential model (Fig. 2, left): the lexicon generating a

[2] For instance, if θ_G is interpreted as a linear mixture of referent-specific models θ_c, any increase in the referent-specific probability will also affect the global probability according to the mixing weight α_c of the referent-specific model, i.e. $\alpha_c \Delta P(X \mid c, \theta_c) = \Delta P(X \mid \theta_G)$, $\alpha_c = [0,1]$, $\Sigma_c \alpha_c = 1$.

[3] This was also confirmed in numerical simulations that show high correlation between the outputs from Eqs. (5) and (6).

sequence of words and the words generating a sequence of acoustic observations, neither of which can be learned without knowing the other. In contrast, speech X and referents c are all observable to the learner utilizing the joint learning strategy. Therefore the learner can use cross-situational accumulation of evidence to find the acoustic model θ_c that captures the shape of the distribution $P(c, X)$.

How does all this relate to the problem of word segmentation? The major consequence of the joint model is that word segmentation emerges as a side product of learning of the acoustic models for the referents (see Fig. 3 for a concrete example). The relative probability of referent (proto-word) c occurring at time t in the speech input is given simply by the corresponding acoustic model θ_c:

$$P(c,t \mid X_{0,\ldots},X_t) = P(c,t \mid X_{0,\ldots},X_t,\theta_c) \quad (7)$$

where $X_0,\ldots, X_t$ refer to speech observations up to time t. The input can be then 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 winning model changes, or 2) thresholding the probabilities to make a decision whether a known word is present in the input at the given time or not (detection task). The segmentation process can be interpreted as continuous activation of distributions for referential meanings for the unfolding acoustic content where word boundaries become points in time where there are notable sudden changes in this distribution (c.f., situation-time processing in McMurray et al., 2012, and Kucker et al., 2015). The nature of this output will be demonstrated explicitly in the experiments in section 5. What this all means in practice is that the learner never explicitly attempts to divide incoming continuous speech into word units. Instead, the learner simply performs maximum-likelihood decoding of referential meaning from the input, and this automatically leads to

temporal chunking into word-like units. Still, despite being driven by referential couplings, the learner is also capable of making familiarity judgments (see section 4), a proxy for segmentation performance in behavioral studies, for patterns that do not yet have an obvious referential meaning. As long as we assume that *c* is never an empty set, but stands for the current internal representational state of the learner, the statistical structure of speech becomes memorized even in the absence of the "correct referents".

4. Approximating cross-situational word learning with transition probabilities

In order to demonstrate the feasibility of the joint model of segmentation and meaning acquisition on real speech data, a practical implementation of the joint model was created in MATLAB by utilizing the idea of TPs to perform statistical learning on language input, often cited as a potential mechanism for statistical learning in humans (e.g., Saffran et al., 1996a), but now conditioned on the referential context. Our argument is not that humans would be actually computing TP statistics over some discretized representations of sensory input. Instead, the present analysis should be seen as a computationally feasible approximation of the ideal model described in Section 3, enabling estimation of joint probabilities within and across perceptual modalities with transparent mathematical notation while maintaining conceptual compatibility with the earlier statistical learning literature. The present section provides an overall description of the system, while step-by-step details of the algorithm are described in Appendix B. Fig. 3 provides a schematic view of the word recognition process in the TP-based model.

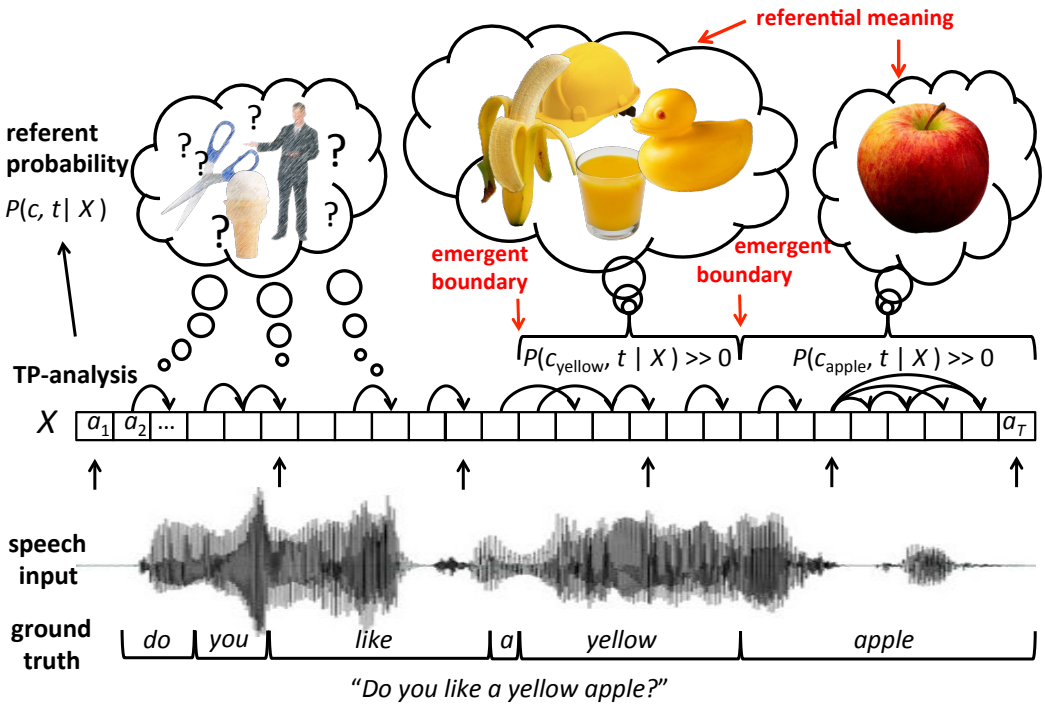


Figure 3: A schematic view of the word recognition process in the TP-based model. Incoming speech signal is first represented as a sequence of short-term acoustic events $X = [a_1, a_2, \ldots, a_T]$. Then the probability of observing the current sequence of transitions between these units in different referential contexts is measured as a function of time (only some transitions are shown for visual clarity) and converted into referent probabilities using the Bayes' rule. Finally, the hypothesized referential meanings exceeding a detection threshold are considered as recognized. Word boundaries emerge as points in time where the referential predictions change. In this particular case, the learner has already used XSL to learn that some of the current TPs occur in the context of visual representations of {yellow} or {apple} at above-chance level, leading to their successful recognition and segmentation. In contrast, "*do you like a*" has an ambiguous relationship to its referential meaning and is not properly segmented (but may still contain familiar transitions).

Let us start by assuming that speech input X is represented as a sequence of discrete units $X = [a_1, a_2, \ldots, a_T]$, where each event a belongs to a finite alphabet A ($a \in A$) and where subscripts denote time indices. These units can be any descriptions of a speech signal that can be derived in an unsupervised manner and they are assumed to be shorter or equal in duration to any meaningful patterns of the language. In the experiments of this paper, these units will correspond to clustered short-term spectra of the acoustic signal, one element occurring every ten milliseconds (see section 5.1). Recall that Eqs. (5) and (6) state that the quality of the lexicon is related to the probability that speech X predicts referents c. By substituting X with the discrete sequence representation, the maximum likelihood estimate for $P(c \mid X, \theta_c)$ is given as

$$P(c \mid X,\theta) = P(c \mid a_1,a_2,...,a_N,\theta) = \frac{F(a_1,a_2,...,a_N,c)}{\sum_c F(a_1,a_2,...,a_N,c)} \quad (8)$$

where $F(a_1, a_2, \ldots, a_N, c)$ is the frequency of observing the corresponding sequence $a_1, a_2, \ldots, a_N$ concurrently with context c, i.e., the acoustic model θ_c simply becomes a discrete distribution across the auditory and referential space. In other words, the optimal strategy to infer referents c from the input is to simply estimate the relative frequencies of different speech sequences co-occurring with the referent and contrasting them against the presence of the same sequences during all other possible referents. However, in the case of speech being represented using short-term acoustic events, this solution turns out to be infeasible since the distribution $P(c \mid a_1, a_2, \ldots, a_N)$ cannot be reliably estimated from any finite data for a large N.

The simplest approximation of Eq. (8) while still maintaining temporal ordering of the acoustic events is to model the sequences $a_1, a_2, \ldots, a_N$ as a first-order Markov process, i.e., to

compute TPs between the units and to assume that each unit a_t is only dependent on the previous unit $a_{t\text{-}1}$. In this case, the probability of a sequence of length N can be calculated as

$$P(a_1, a_2, ..., a_N) = \prod_{t=2}^{N} P(a_t \mid a_{t-1}) \tag{9}$$

where the TP of an event at time $t-1$ to the event at t is obtained from the corresponding transition frequencies F:

$$P(a_t \mid a_{t-1}) = \frac{F(a_t, a_{t-1})}{\sum_{a_t \in A} F(a_t, a_{t-1})} \tag{10}$$

This formulation aligns with the findings that humans are sensitive to TPs in speech instead of overall frequencies or joint probabilities of the events (see Aslin & Newport, 2014). However, the first-order Markov assumption does not generally hold for spoken or written language (see Li, 1990; Räsänen & Laine, 2012; 2013), making this approximation suboptimal. In order to account for dependencies at arbitrary temporal distances, an approximation of a higher-order Markov process is needed. Our approach here is to model the sequences as a mixture of first-order Markov processes with TPs measured at different temporal lags k (Raftery, 1985; see also Räsänen, 2011; Räsänen & Laine, 2012). The general form of a mixture of bi-grams is given as

$$P(a_t \mid a_{t-1,} a_{t-2}, ..., a_{t-k}) \approx \sum_k \lambda_k P_k(a_t \mid a_{t-k}) \tag{11}$$

where P_k are lag-specific conditional probabilities for observing a lagged-bigram $\{a_{t\text{-}k}, a_t\}$, a pair of elements at times t-k and t with any other non-specified elements in between, and λ_k is a lag-specific mixing weight that is typically optimized using the EM algorithm (Raftery, 1985; Berchtold & Raftery, 2002). Maximum-likelihood estimates for the lag- and referent-specific TPs are obtained directly from the frequencies of transitions at different lags:

$$P_k(a_t \mid a_{t-k}, c) = \frac{F_k(a_t, a_{t-k}, c)}{\sum_{a_t \in A} F_k(a_t, a_{t-k}, c)} \quad (12)$$

Assuming that the lag-specific weights λ_k are equal to all referents c (i.e., all speech has a uniform dependency structure over time), the instantaneous *relative probability* of each referent c, given speech X, can now be approximated as the sum of lagged bi-grams that occur during the referent contrasted against the number of the same bi-grams in all other contexts:

$$P(c,t \mid X) \propto \sum_k \frac{P_k(a_t \mid a_{t-k}, c)}{\sum_c P_k(a_t \mid a_{t-k}, c)} P(c) \quad (13)$$

In a similar manner, the instantaneous *familiarity* of the acoustic input in a given context c is proportional to the sum of the TPs across different lags in this context:

$$P(X,t \mid c) \propto \sum_k P_k(a_t \mid a_{t-k}, c) \quad (14)$$

Note that the conditional distribution $P(a_t \mid a_{t-k})$ approaches a uniform distribution for increasing k as the statistical dependencies (mutual information) between temporally distant states approach zero. At the acoustic-signal level, the time-window containing the majority of the statistical dependencies corresponds to approximately 250-ms, also corresponding to the temporal window of integration in the human auditory system (Plomp & Bouman, 1959; Räsänen & Laine, 2013), and therefore also setting the maximum lag k up to which TPs should be measured.

In principle, Eq. (13) could be used to decode the most likely referent for the speech observed at time t. However, since the underlying speech patterns are not instantaneous but are continuous in time, subsequent outputs from Eq. (13) are not independent of each other despite being based on information across multiple temporal lags. Therefore the total *activation* of referent c in a time window from t_1 to t_2 is obtained from

$$A(c \mid X_{t_1},...,X_{t_2}) = \sum_{t=t_1}^{t_2} P(c,t \mid X) \quad (15)$$

i.e., by accumulating the context-dependent TPs of Eq. (13) over a time-window of analysis. The integration in Eq. (15) can be performed in a sliding window of length W ($t_2 = t_1+W-1$) in order to evaluate word activations from continuous speech (c.f. word decoding in the TRACE model of speech perception; McClelland & Elman, 1986). Once the activation curves for referents have been computed, temporally contiguous above-chance activation of a referent c across speech input 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 summary, the learning process can be summarized with the following steps:

1) Start with empty (all zero) transition frequency counts.
2) Given a discrete sequence of acoustic events $X_i = [a_1, a_2, \ldots, a_T]$ corresponding to the speech input (e.g., an utterance) and a set of concurrent visual referents $\mathbf{c}_i = \{c_1, c_2, \ldots, c_N\}$ (e.g., observed visual objects), update the lag- and referent-specific transition frequencies $F_k(a_t, a_{t-k}, c)$ for all currently observed acoustic events in the input sequence: $F_{k,i+1}(a_t,a_{t-k},c) \leftarrow F_{k,i}(a_t,a_{t-k},c)+1$ for $a_t, a_{t-k} \in X_i$, $c \in \mathbf{c}_i$, $k \in [1, K]$.
3) Normalize frequencies into lag- and referent-specific TPs according to Eq. (12).
4) Repeat steps 2) and 3) for every new utterance and the corresponding referents.

During word recognition, the input is a speech signal $X_j = [a_1, a_2, \ldots, a_T]$ without referential cues. The probability of each referent (word) at each moment of time is computed using Eq. (13) by retrieving the probabilities of the currently observed transitions from the previously learned

memory. If speech pattern familiarity is measured instead of the most likely referent, Eq. (14) is used instead.

Note that all learning in the model is based on incremental updating of the transition frequencies $F_k(a_t, a_{t-k}, c)$, and the details of the input can be forgotten after K time units. The only free parameter in the model is the maximum lag K up to which transitions are analyzed. Also, note that if c is constant (the same context all the time), K is set to 1, and alphabet A corresponds to the syllables of the language, the model reduces to the standard TP-model used in the behavioral studies such as that by Saffran et al. (1996a).

In the experiments of this paper, W = 250 ms is always used as the length of the time integration window for Eq. (15). Also, the TPs are always computed over lags $k \in \{1, 2, ..., 25\}$ (10–250 ms). Finally, P(c) is assumed to be a uniform distribution in the absence of further constraints.

4.1 Implementing an attentional constraint

Preliminary experimentation indicated that the basic TP-based implementation leads to superior learning performance in comparison to human data. In addition, the model is invariant to the presentation order of the training trials, which is not true for human subjects (e.g., Yu & Smith, 2012; Yurovsky et al., 2013). In order to simulate the limited performance of human learners in tasks with multiple concurrent visual referents (experiments 3–5), a simple *attention-constrained* variant of the model was created where the basic update mechanism of simply counting the frequency F of transitions between acoustic events equally for all present referents was replaced with a rule that only updates the model with the most likely referent c in each situation (see also McMurray, Aslin & Toscano, 2009, for a similar mechanism in phonetic category learning):

$$F_{k,t+1}(a_t, a_{t-k}, c^*) \leftarrow F_{k,t}(a_t, a_{t-k}, c^*) + 1$$
$$\text{only if} \qquad (16)$$
$$c^* = \arg_c \max\{A'(c, t \mid X) \mid \forall c\}$$

where $A'(c, t \mid X)$ is the referent activation $A(c, t \mid X)$ computed using Eq. (15) and smoothed in time using a 250-ms moving average filter. The smoothing simulates "inertia" in attentional focus by limiting the maximum rate of attentional shifts. A small Gaussian noise floor was added to the instantaneous probabilities ($P(c, t \mid X)+N(0, 0.00001)$) to ensure that the attention was randomly distributed among visual referents during novel input.

The attention constraint effectively implements familiarity preference in learning, causing the learner to focus more on the referents that are already associated with the unfolding speech input. The constraint is in agreement with the eye-gaze data from human subjects in XSL learning tasks where longer looking times are observed towards the initial correct referents already after the second appearance of the referent for those learners who are more successful in the task (Yu & Smith, 2011; Yu et al., 2012; Yurovsky et al., 2013). The present constraint also converges with the foundations of the preferential looking paradigm used to probe infants' associative learning in behavioral learning tasks (e.g, Hollich et al., 2000) and with the fact that word comprehension is often reflected by visual search of the referent in the immediate surroundings (e.g., Bergelson & Swingley, 2013). Also note that the attention constraint does not deviate from the original joint model, but is a filtering mechanism that limits the original set of equivalently relevant referents *c* into the most likely referent at each moment of time.

5. Experiments

The joint model for cross-situational word learning was tested in six experiments. The first four aim to provide a behavioral grounding for the model, showing that it fits to human data on a

number of segmentation and cross-situational learning tasks. The last two show the real potential of the statistical learner when confronted with natural continuous speech, requiring joint acquisition of segments and their meanings across individually ambiguous learning scenarios and in the face of acoustic variability present in natural speech. The first experiment replicates the seminal study of Saffran et al. (1996) and shows that the model results fit to the behavioral data on segmentation when no contextual cues are available. The second experiment extends the first and shows how segmentation performance improves with referential cues, replicating the findings of Thiessen (2010) for adult subjects. The next two experiments investigate the compatibility of the model with human data on XSL by replicating the experiments of Yu and Smith (2007) and Yurovsky et al. (2013). The fifth experiment investigates concurrent segmentation and word-to-meaning mapping in real pre-recorded continuous speech when the speech is paired with simulated visual referential information. Finally, the sixth experiment focuses on acoustic variability and generalization across talkers in natural speech. All simulations require that raw speech input is first converted into a sequence of pre-linguistic acoustic events before processing in the model, and therefore these pre-processing steps are described first.

5.1 Speech pre-processing and TP-algorithm implementation

In order to simulate early word learning without making strong assumptions on the existing speech parsing skills, the currently used representation for speech signals does not make use of any a priori linguistic knowledge. Instead, the speech was first pre-processed into sequences of short-term acoustic events using unsupervised clustering of spectral features of speech (for similar pre-processing, see also Van hamme, 2008; ten Bosch et al., 2009; Driesen & Van hamme, 2011; Versteegh et al., 2010; 2011; Räsänen, 2011; Räsänen & Laine, 2012).

First, the instantaneous spectral shape of the speech was estimated in a sliding window of 25 ms using 10-ms steps for the window position. The spectrum in each window position was described using Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980) that represent the spectrum with a small number of decorrelated features. MFCCs are obtained by first computing the standard short-term Fourier spectrum of the signal, followed by Mel-scale filtering in order to mimic the frequency resolution of human hearing. Then, the logarithm of the Mel-spectrum is taken, and the resulting log-Mel spectrum is converted to the so-called cepstral domain by performing a discrete cosine transform on it. As a result, the first 12 MFCC coefficients, the signal energy, and their first and second derivatives were chosen as descriptors of the spectral envelope for each window of analysis.

In order to convert MFCC features into a sequence of discrete acoustic events, 10000 randomly chosen MFCC vectors from the training data were clustered into $|A|$ discrete categories using the standard k-means clustering algorithm (MacQueen, 1967). Cluster centroids were always initialized using $|A|$ randomly chosen MFCC vectors. All feature vectors were then assigned to their nearest cluster in terms of Euclidean distance, leading to a sequence of the form $X = [a_1, a_2, ..., a_T]$, where each discrete acoustic event is denoted with an integer in the range from 1 to $|A|$, and one event occurring every 10 ms. While the characteristics of these atomic units are dependent on the distributional characteristics of the speech spectra, they do not correspond to phones of the language but simply assign spectrally similar inputs to the same discrete event categories (see Räsänen, 2012). In the experiments of this paper, the number of acoustic categories $|A|$ in this "acoustic alphabet" can be considered as the amount of acoustic detail preserved by the representation: while a small set of acoustic categories may not be able to reliably differentiate between different speech sounds, a very large number of categories leads to

problems in generalization to new input as many of the TPs have never been observed before. The pre-processing stages[4] are illustrated in Fig 4.

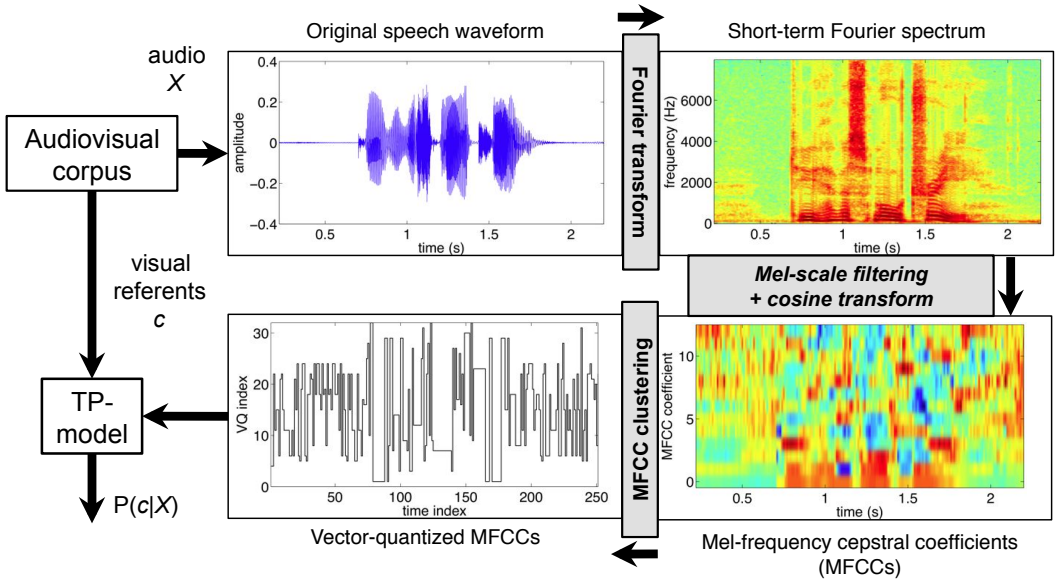


Figure 4: A block schematic illustrating the pre-processing stages used in the present study. The speech signal is first converted into spectral short-term features (Mel-frequency cepstral coefficients, aka. MFCCs), one MFCC vector occurring every 10 ms. The MFCCs are then clustered into $|A|$ discrete categories using the standard (unsupervised) k-means algorithm. As a result, the original signal is represented as a sequence of atomic acoustic events that are used as an input to the TP-based cross-situational learning model.

Results from all experiments are reported across several runs of the experiment where each individual run can be considered as a separate test subject. Variability across runs is caused by

[4] Note that all the pre-processing steps are standard procedures in digital speech signal processing. MFCCs can be replaced with the Fourier spectrum, wavelet analysis, a Gammatone-filterbank, or some other features, as long as they represent the short-term spectrum of the signal with sufficient resolution. Similarly, k-means clustering can be replaced with cognitively more plausible approaches such as OME (Vallabha, McClelland, Pons, Werker & Amano, 2007), as long as the method is able to capture distributional characteristics of the used feature vectors.

random initialization of the acoustic clustering, leading to learners with slightly different acoustic representations, and, depending on the experiment, by possible randomness in the allocation of training and testing stimuli for individual learners. We assume across-run variability of the overall performance levels to be normally distributed.

5.2 Experiment 1: Word segmentation without contextual referents

The goal of the first experiment is to first show that the model is compatible with the behavioral data on statistical word segmentation when no referential cues are present. In order to do this, the classical experiment of Saffran et al. (1996) was replicated.

5.2.1 Experimental setup. In this experiment, the learner is first exposed to a continuous stream of an artificial language consisting of four unique tri-syllabic CVCVCV words. All words are concatenated together in random order without intermediate pauses and so that a word cannot follow itself. Words are chosen so that the TPs between syllables within each word have TP = 1 while TPs between syllables from different words are TP = 0.33. Length of the familiarization stream is 45 tokens per word, corresponding to approximately 2 minutes of speech. After familiarization, the learner is exposed to two high-TP words ("true words") and two non-words (condition #1), or two true words and two part-words (condition #2), while the learner's familiarity with each of these tokens is measured. A non-word refers to a trisyllabic pattern whose syllables all occurred in the familiarization stream but not in the given order. Part-words refer to trisyllabic patterns that occurred in the training data but spanned across two subsequent true words (the first syllabic TP within the part-word is 0.33). Moreover, the familiarization and testing are counterbalanced so that the true words of one subject group are non-words (or part-words) of another group and vice versa.

Similarly to Saffran et al. (1996), the full words in condition #1 for learners in group A were *tupiro*, *golabu*, *bidaku*, and *padoti*, while familiarization in group B consisted of the words *dapiku*, *tilado*, *buropi*, and *pagotu*. Test words for both groups were *tupiro*, *golabu*, *dapiku*, and *tilado*, with the first two being full-words and the last two being non-words for group A, while the opposite was true for the subjects in group B. For condition #2, the full words for subject group A were *pabiko*, *tibudo*, *golatu*, and *daropi*, while subject in group B heard *tudaro*, *pigola*, *bikuti*, and *budopa*. The four test tokens were *pabiko*, *tipudo*, *tudaro*, and *pigola*.

In the original study, 8-month-old infants were exposed to the language synthesized in monotonic articulation and then tested in their knowledge by using the preferential looking-time paradigm of Jusczyk and Aslin (1995). In the present experiment, we synthesized the stimuli using the MBROLA synthesizer (Dutoit, Pagel, Pierret, Bataille & Van der Vreken, 1996), followed by acoustic pre-processing of speech as described in section 5.1. A TP-based model was then trained with referential context fixed to $c = 1$ as there were no contextual cues in the experiment. During test trials, the familiarity with each test token was estimated by computing the instantaneous probabilities of the test stimulus using Eq. (14) and then summing these probabilities across the stimulus duration and across the trials for each token type. The results were then converted into levels of surprisal by taking the negative value of the logarithm of the probabilities (a small-probability event being more surprising), a measure that should correlate positively with the infant's longer looking times in habituation tasks. Since there were no alternative contextual referents, no decoding of the most likely referent was attempted.

The experiment was carried out for three acoustic alphabet sizes of 4, 16, and 64 in order to ensure that there is a systematical relationship between acoustic discrimination capability and task performance. For each alphabet size, the experiment was repeated 12 times for both counter-

balanced sets to account for the random variability in the acoustic clustering stage, simulating 24 listeners with slightly different acoustic event categorization criteria. All results and their variations were measured across these runs.

5.2.2 Results. Figure 5 shows the mean levels of surprisal for true words and non-words in condition 1 (left) and for true words and part-words in condition 2 (right) as a function of the acoustic alphabet size $|A|$. In both conditions, true words are significantly less surprising than the non-words or part words for $|A| \geq 4$. Moreover, the difference in surprisal increases as the learner becomes better at differentiating between different speech sounds with an increasing acoustic alphabet size. As expected, non-words are perceived as more surprising than part words as they contain two previously unseen syllabic transitions instead of one high- and one low-probability transition of the part-words. Finally, the surprisal for true words is similar in both conditions despite different word-forms, indicating the results are not dependent on the phonological form of the stimuli.

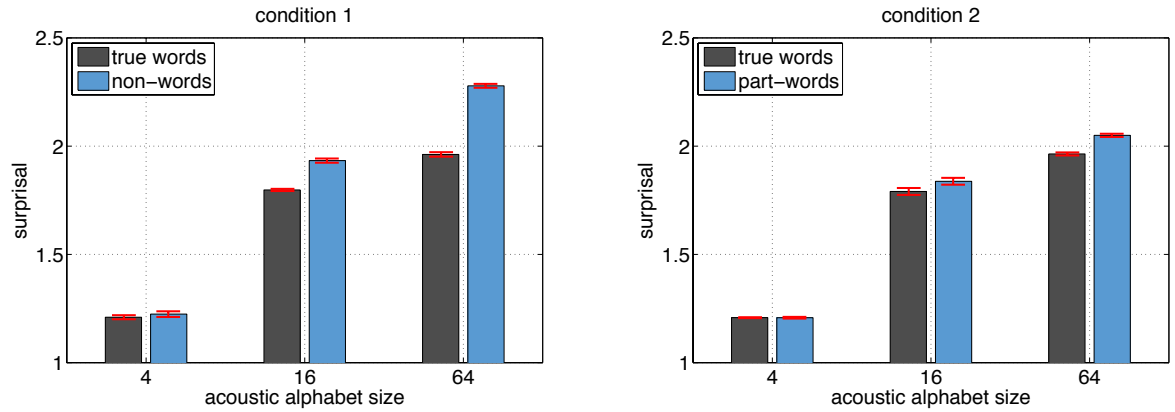


Figure 5: Surprisal (-log(probability)) for true words and non-words in condition #1 (left) and true words and part-words in condition #2 (right). The results are shown for different acoustic alphabet sizes $|A|$. The error bars correspond to ±1 SE.

5.2.3 Discussion for experiment 1. As noted in section 4, the syllable-based TP-analysis proposed by Saffran et al. (1996) is a special case of the present model. This is reflected in the current results where the model clearly differentiates between high- and low-TP syllable sequences in an artificial language after only 2 minutes of exposure, being in line with the finding that infants find low-TP sequences more surprising (Saffran et al., 1996).

The message of the first experiment is that there is no explicit mechanism for word segmentation in the joint model, yet its behavior still fits to the behavioral data without problems. Differences in the degree of familiarity with the stimuli are sufficient to capture human data in this type of habituation task without assuming that the listener is actually representing the speech in terms of distinct temporal segments. Moreover, the experiment shows that the listeners' preferences in the Saffran et al. (1996) type of task can be explained with a statistical learner incapable of performing syllabic segmentation or linguistically motivated categorization of the speech sounds. This is because all regularities at a linguistic level will be necessarily reflected also at the signal level.

5.3 Experiment 2: Word segmentation in an artificial language with visual cues

The goal of the second experiment is to build on the previous non-contextual segmentation task and to demonstrate that segmentation performance improves when consistent contextual cues are available in addition to the speech signal. In his work, Thiessen (2010) showed that when the tri-syllabic high-TP words of the artificial language of Saffran et al. (1996) (experiment 1) were paired with consistently co-occurring visual shapes, adult subjects performed significantly better in a forced-choice task between words and part-words after the familiarization period. Although such an effect was not found for 8-month-old infants, the result suggests an interaction between concurrent visual information and the learning of the auditory patterns in speech.

5.3.1 Experimental setup. We replicated the experiment of Thiessen (2010) by pairing the familiarization stream of condition #2 in experiment 1 with concurrent visual cues. More specifically, in the *regular visual* condition, a unique discrete visual label $c \in \{1, 2, 3, 4\}$ was paired with each of the trisyllabic words in the artificial language and was always presented synchronously with the auditory word. In the *random visual* condition the images appeared randomly with respect to auditory words while *no visual* condition had no visual cues at all, corresponding to the experiment 1. Otherwise the familiarization period and the test tokens were identical to the previous experiment with learners divided into two counter-balanced groups. Similarly to the adult subjects in Thiessen (2010), segmentation performance was then tested in a two-alternative forced-choice (2AFC) task between all possible pairs of true words and part-words. Referent learning performance was tested in a four-alternative forced-choice (4AFC) task between the four visual referents, given a true word. Word selection in 2AFC was performed by computing the overall familiarity for the tokens in each pair and choosing the most familiar token as the hypothesis. Since the statistics of the auditory stream were now distributed across multiple visual contexts c, the familiarity was determined as the largest one across all possible contexts (see also Appendix B):

$$Fam(X) \propto \max\{A'(X \mid c) \mid \forall c\} = \max\left\{\sum_t \sum_k P(a_t \mid a_{t-k}, c) \mid \forall c\right\} \quad . \tag{17}$$

The most likely referent for a word in the 4AFC task was obtained with Eq. (15).

Relevant to the present experiment, Cunillera et al. (2010) found in a series of similar word learning experiments that the facilitatory effect of the visual context is not dependent on the exact temporal synchrony between the words and referents. This type of flexibility in audiovisual timing would be beneficial for learning in natural environments, where attentional focus and object naming would be allowed to have some temporal jitter. Since Cunillera et al. (2010) found

that asynchrony in the range of 0–200 ms did not affect learning outcomes, we also created an asynchronous version of the present experiment, where the onset and offset of a visual referent was randomly shifted for ± 0–200 ms with respect to the corresponding word (shifts sampled from a uniform distribution for each token), allowing temporal overlap of two subsequent referents if they became shifted towards each other. The underlying assumption was that if the humans are insensitive to this type of asynchrony due to their temporal window of audiovisual integration, also the participants in the study of Thiessen (2010) should experience equivalent levels of uncertainty in the exact timing between acoustic content and the referents. In this case, the audiovisual jitter in the familiarization data will have the same effect as an equal amount of temporal smoothing in the audiovisual integration of our computational learner.

The experiment was repeated 30 times for both counter-balanced groups, for both synchronous ("sync") and asynchronous ("async") conditions, for acoustic alphabets of size |A| = 32 and 64, and with a new acoustic clustering performed on each run.

5.3.2 Results. Fig. 6 shows the mean results for the true word vs. part-word discrimination task together with the results of Thiessen et al. (2010) for adult subjects, measured across all 30 runs of the simulation.

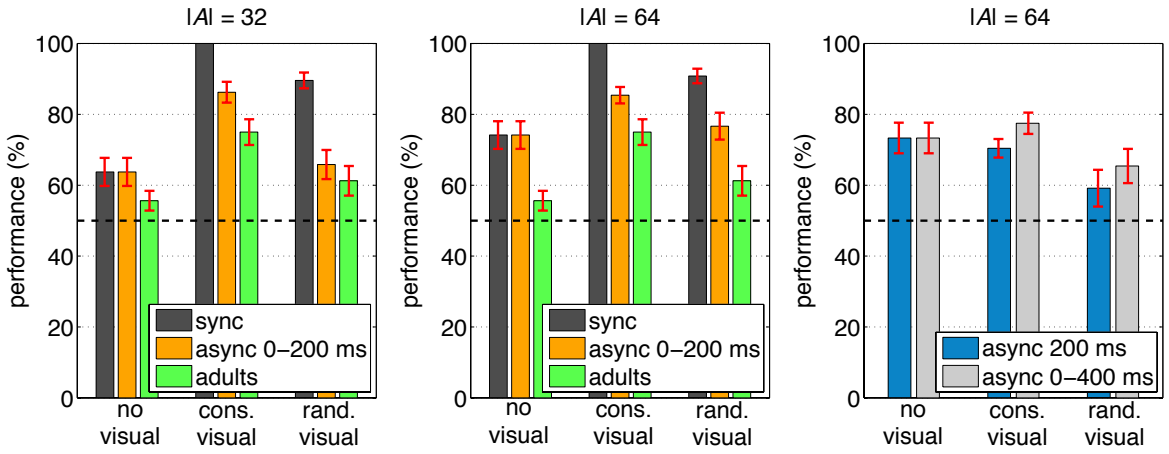


Figure 6: Model segmentation performance in experiment 2 for all three conditions. The left and middle panels show results for $|A| = 32$ and $|A| = 64$, respectively. The dark bars show the performance in "sync" condition where only transitions during the referent (and thereby during the word) are included in the acoustic model of the referent. The orange bars show the performance when there is ±0–200 ms random asynchrony between the words and their referents. Adult performance in the same task as reported by Thiessen (2010) is shown with the green bars. The right panel shows a re-run of the experiment with larger audiovisual asynchrony (the blue bars: fixed ±200 ms asynchrony; the grey bars: random ±0-400 ms asynchrony) in order to simulate infant performance in the task. The error bars correspond to ±1 SE.

In all three conditions and with both alphabet sizes, the segmentation performance is significantly above the chance level, similarly to the human subjects in Thiessen (2010). In addition, performance after access to visual cues during familiarization is significantly higher than without, also consistent with the human data. In the asynchronous case, the pattern of results across conditions resembles Thiessen's findings, showing no visual facilitation from random visual cues at $|A| = 32$ ($t(59) = 1.09$; $p = 0.28$) or at $|A| = 64$ ($t(59) = 1.63$; $p = 0.11$) with respect to the non-visual condition and clear facilitation when consistent visual cues are available. However, the overall performance is notably better in the case of the higher acoustic resolution of $|A| = 64$, suggesting that the ideal model is too powerful in comparison to human learners. Interestingly, the synchronous analysis leads to a clear improvement in segmentation also with random visual cues. This is because the synchronized analysis contains only within-word transitions for each referent c, making the statistical distinction between words and part-words clear and the model unrealistically powerful in the task. Finally, in the visual cue condition, the model learned correct referents for all words without errors in both synchronous and asynchronous conditions.

5.3.3 Discussion for experiment 2. The results from experiment 2 show that the segmentation performance, when measured in terms of pattern familiarity, is greatly improved when concurrent visual cues are available during the familiarization stage. This is because the statistical structure of the speech input becomes conditioned on the context in which the speech is heard, dividing the global TPs into more specific context-specific statistics that then compete during the word recognition process. Also, the asynchronous variant matches to the pattern of human data across different conditions while the synchronous variant benefits from the audiovisual synchrony also in the presence of random visual cues. This converges with the findings of Cunillera et al. (2010) by suggesting that humans may not benefit from the exact synchrony between auditory and visual cues of arbitrary form, but that the associations are learned as long as the information from different modalities is available within a finite time-window of audiovisual integration that spans some tens or hundreds of milliseconds in time.

Since infants, unlike the adults, were unable to benefit from visual context in the original study (Thiessen, 2010), it is also possible that the infants failed in the task simply because their time-window of audiovisual integration is known to be longer than that of adults (Lewkowicz, 1996)—a property that is not problematic in normal communication but leads to significant confusion in the present experimental setup due to the high temporal density and statistically balanced recurrence of words and referents. Since Kopp (2014) has reported behavioral indifference to audiovisual asynchrony of at least up to 200 ms in 6-month-old infants, we also ran a post-hoc version of the simulation 1) using random asynchrony of ±0–400 ms, or 2) having the onset of the visual referent always 200-ms before and offset 200 ms after the word, two subsequent referents always overlapping by 200 ms. The right panel in Fig. 6 shows the corresponding results. In both cases, the performance was again above chance level but no

facilitation from the visual cues was observed in comparison to the non-visual baseline ($t(59) < 1.28$; $p > 0.21$ for both cases), replicating the findings of Thiessen (2010) for infant subjects, and suggesting that the time-scale of audiovisual integration should be taken into account in future studies of this type.

5.4 Experiment 3: Cross-situational learning of word meanings

The goal of the following two experiments is to investigate whether the present framework of cross-situational word learning is also compatible with human performance in well-controlled XSL experiments. We start by simulating the original behavioral study on XSL by Yu and Smith (2007).

5.4.1 Experimental setup. In their experiments, Yu and Smith (2007) exposed adult subjects to a series of individually ambiguous learning trials, each consisting of a number of spoken nonsensical words and the same number of concurrently shown unfamiliar visual objects. The task of the subjects was to learn the mappings from words to their correct visual referents by accumulating co-occurrence evidence across multiple trials. From now on, the term *learner* is used indifferently to refer to the subjects of the behavioral experiment or to the present computational model unless further distinction is necessary.

Yu and Smith (2007) used five unique experimental conditions. In the first three, the stimuli consist of 18 words that always co-occur with a unique visual referent. Each learning trial consists of two (2 X 2 *condition*), three (3 X 3), or four (4 X 4) different spoken words and their corresponding visual referents so that each word-referent pair is presented a total of 6 times. This leads to 54, 36, and 27 trials in these three conditions, respectively. In the fourth condition, the number of unique words is reduced to nine and the number of repetitions is increased to eight (18 trials), whereas the fifth condition consists of nine words and twelve repetitions (27 trials) (Yu &

Smith, 2007). After training, the learners are subjected to a four-alternative forced-choice task, where they hear one word at a time and are asked to choose from among the correct referent and three foil referents also familiar from the training stage.

The main finding of Yu and Smith (2007) was that the human subjects were able to choose correct referents far above chance-levels in all five conditions, but with decreasing performance when the ambiguity of the individual trials was increased from 2 X 2 to 3 X 3 and 4 X 4. However, the reduction in the number of unique words or increase in the number of learning trials did not have a significant effect on the learning performance.

In order to test our model in the task, the experiments of Yu and Smith (2007) were replicated using the same experimental setup but using words extracted from continuous pre-recorded speech as the stimuli. The motivation for using real speech came from the desire to maintain natural variability of the stimuli, as one of the claims of the model is that it can overcome the variability in the speech acoustics. For this purpose, we used the Caregiver Y2 UK corpus (Altosaar et al., 2010) since it contains several repetitions of 50 different words, embedded within larger sentences and spoken in a child-directed manner (see section 5.6 for a detailed description of the corpus and Appendix C for the full list of words).

For each run of the experiments, eighteen words for conditions #1–#3 and nine for conditions #4 and #5 were randomly chosen from the list of 50 possible words in the corpus. A unique token for each occurrence of each word was then extracted from the utterances using word-level annotation of the corpus with the constraint that each word token had to be at least 250 ms in duration and spoken by the same talker. The extracted words were used to construct the training stimuli for all experiment conditions by concatenating the words in random order with 250-ms gaps of noise between words. During testing, six previously unseen tokens for each

word were used as the stimuli in the forced-choice task. Finally, in order to understand the impact of acoustic variability on the results, the experiment was repeated by using the same acoustic realization of each word during training and testing. All the experiment conditions were run 20 times, and all reported results are averaged across these runs.

5.4.2 Results. After pre-processing the speech as described in section 5.1 and quantizing the data into sequences of $|A| = 64$ unique acoustic events, the model was trained on the familiarization data as described in section 4. The only difference to the previous experiment was that the instantaneous referent activations were integrated over the duration of each test token during the recognition task in order to have the final overall activations for each competing referent. Fig. 7 shows the results across the five conditions together with the behavioral data from Yu and Smith (2007).

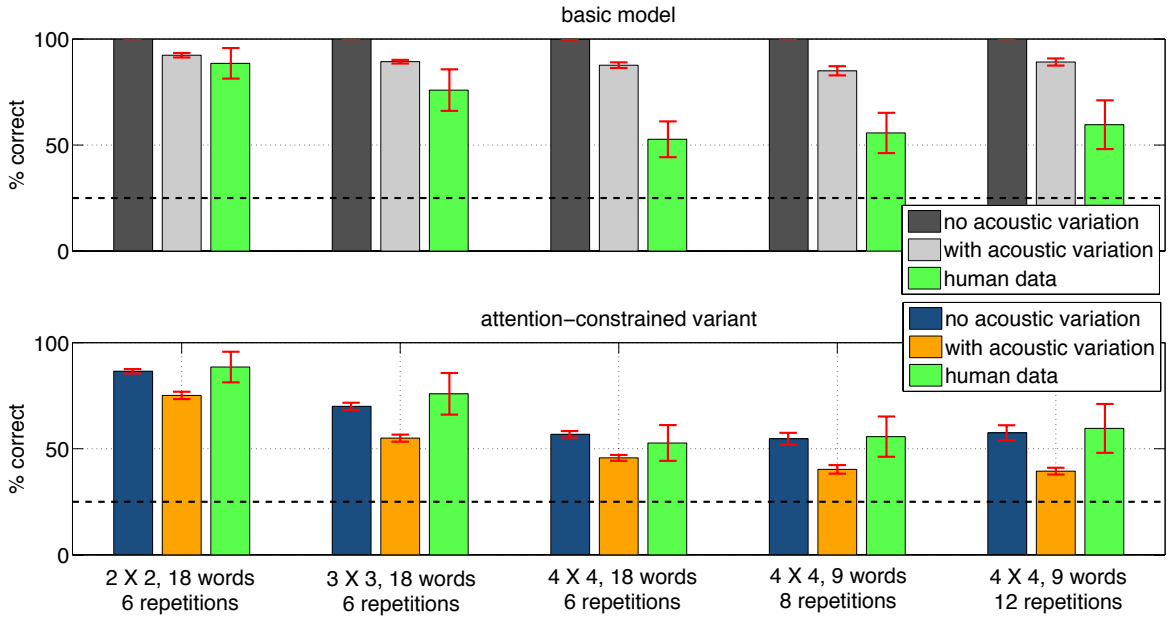


Figure 7: Results for all the five test conditions for the basic model (top) and the attention-constrained variant (bottom). The leftmost bars in each condition show the model performance without acoustic variability in the stimuli while the middle bars show the performance with realistic acoustic variation. The

rightmost bars show the corresponding behavioral results from Yu and Smith (2007). The error bars denote ±1 SE.

All variants in all experiment conditions lead to performance significantly above-chance level. As already noted in section 4, the basic model clearly outperforms human subjects in all conditions, leading to perfect performance if there is no acoustic variability in the word tokens. When all tokens are taken from different pronunciations of the words, the performance drops to 90% but is still invariant to the experiment conditions. In contrast, the attention-constrained model shows basically a perfect fit to the human data if acoustic variability is not present, being always within ±1 SEs from the human mean performance in each condition. When natural acoustic variability is present in the stimuli, the attention-constrained model is still successful in the task and follows the same pattern across different conditions as was observed in human subjects by Yu and Smith (2007). However, on the whole, the performance is somewhat lower in overall due to the increased complexity of the task that was not present in the original study. The word-learning accuracy drops when the referential ambiguity during learning trials is increased by increasing the number of simultaneous words and referents. On the other hand, reduction in the overall number of unique words (referents) and increase in the number of repetitions of word-referent pairs do not have notable effect on the results. In general, the model performance is always within ±1 SE from the human performance when the variability in the stimuli is similar to the original behavioral experiment.

5.4.3 Discussion for experiment 3. Results from experiment 3 show that the attention-constrained model follows the pattern of behavioral results, achieving this without having any a priori knowledge of relevant linguistic structures such as words. Importantly, the fit to human data requires that the information selection during each trial is driven by the already established

associations that become activated as the speech stimuli unfolds over time (see Yu & Smith, 2012a, for time-invariant word-level models). When typical acoustic variability across word tokens is included in the stimuli, the absolute level of model performance is consistently lower than that of human subjects. In this case, the source of ambiguity is not only the word-to-meaning mapping but also the variability across individual word tokens, requiring that the model is capable of generalizing towards novel tokens with only partially similar surface properties. When the acoustic variability is removed from the experiment, the model performance is similar to the human performance as measured using limited-variability synthesized stimuli. Note that acoustic variation was not present in the study of Yu and Smith (2007) and adult subjects are also likely to be insensitive to such variation due to their existing phonological knowledge.

5.5 Experiment 4: Competitive mechanisms in cross-situational learning

5.5.1 Experimental setup. Since the overall pattern of model performance was consistent with human performance in the previous experiment, the aim of the fourth experiment was to investigate whether the model is also capable of explaining subtler findings regarding competition between words during XSL. Therefore, we simulated the behavioral experiments performed by Yurovsky et al. (2013), where competition was analyzed in depth. Here we focus on experiments 1, 2, and 3 in the original study of Yurovsky et al. (2013), and *conditions #1–#3* will be used to refer to the experiments 1–3 of the original paper.

In all three conditions, the training stage consists of a series of 4 X 4 trials, as in the experiments of Yu and Smith (2007) described above. In condition #1, the stimuli consist of six *single words* that each always co-occur with a specific visual referent, six *double words* that always co-occur with both of their two different referents, and six *noise words* that only occur in the audio but have no referent. There are a total of 27 familiarization trials, with each having four

visual referents shown concurrently on a screen, paired with speech input containing four spoken words. Of these 27 trials, two contain four single words; 14 contain two single words, one noise word, and one double word; and 11 contain two double words and two noise words, each word and referent occurring together a total of six times.

In conditions #2 and #3, the stimuli consist again of six *single words* that each always co-occur with a specific visual referent, but now the six *double words* always co-occur only with one of their two possible referents in one trial. As in the condition #1, the familiarization period consists of 27 instances of 4 words x 4 visual referent trials, leading to a total of 6 exposures to each referent together with the corresponding word. The only manipulation between conditions #2 and #3 is the ordering of the double-word referents across the trials: The appearances of the two referents for each double word are interleaved across the trials in condition #2. In contrast, the ordering is changed in condition #3 so that each double word first occurs six times with the first of its referents, whereas the last six occurrences of the word occur together with the remaining referent.

The training stage in all three conditions is followed by a sorting task where the learners are asked to sort a list of four referents in the order of preference on hearing a word that had occurred during the familiarization period. In condition #1, the list of alternatives for each single word consists of the correct referent, a referent of another single word, and two referents for the double words. For each double word, the alternatives are both of its correct referents, a referent of a single word, and a referent from another double word. In conditions #2 and #3, the list of options for single words consists of the correct referent for the given word and three other single word referents. For each double word, the options consist of the both correct referents for the word and two other referents from other double words. A single word is considered as correctly

recognized if the correct referent is the first one in the order of preference. For double words, *either* of the word meanings is marked as scored if one of the two possible referents is chosen as the most likely, whereas *both* meanings are scored as correct if both correct referents are the first two in the list in the given trial.

When human subjects were exposed to these three tasks, Yurovsky et al. (2013) made the following main observations: Humans are able to learn word-referent mappings at above-chance level in all three conditions and for both single and double words. However, there is some sort of competition between word-referent associations since the single words were learned significantly better than both meanings of double words in conditions #1 and #2, even though the word-referent co-occurrence frequencies were equal for both word types. Since the single-word learning was also better in condition #2, where only one of the double word's two correct referents was present at a time, the competition seems to be *global*, i.e., not only dependent on the current input. Interestingly, both meanings of the double words were learned much better in condition #3 compared to condition #2 (mean 40% vs. 24% correct) when the same referents were presented in a different order, leading to a statistically insignificant difference between learning of the single and double words. When Yurovsky et al. probed human performance in a setup similar to condition #3 trial by trial (experiment #4 in their paper), they noticed that there was a significant preference for early referents over the later occurring referents for the double words, suggesting that the global competition prefers already established speech-to-referent associations over novel ones (Yurovsky et al., 2013).

In order to test the TP-based model in the task, the experiments of Yurovsky et al. (2013) were replicated using the same experimental setup but again using real speech from the Caregiver Y2 UK corpus (see section 5.4.1). For each run of the experiments, six single words, six double

words, and six noise words (condition #1) were randomly chosen from the list of 50 possible keywords. Twelve different tokens of each single and noise word and twelve (condition #1) or eighteen (conditions #2 and #3) tokens of each double word were then extracted from the utterances using word-level annotation of the corpus with the constraint that each word token had to be at least 250-ms in duration. Finally, the extracted words were used to construct the training stimuli for all experiment conditions by concatenating the words in random order with 250-ms gaps of white noise between words. During testing, six previously unseen tokens for each single and double word were used as the stimuli in the sorting task, each token tested one at a time (a total of 12 x 6 = 74 test tokens per condition). The same set of stimuli was always used for conditions #2 and #3 within the same run of the experiment, the only difference being the labeling order of the visual referents of the double words, as described above. Otherwise, the procedure followed that of experiment 3 and using exactly the same parameters as in the previous experiments. All reported results are accumulated across 20 runs of the experiment.

5.5.2 Results. Since the chance levels for correct responses are different for single and double words, we follow the procedure of Yurovsky et al. (2013) by first estimating the correct number of responses expected due to word knowledge before performing a statistical test on the data. This normalization is achieved by using a binomial distribution to model the number of responses that can be expected to be correct by chance, given the learner's non-normalized score in the three categories (single, either, and both correct) (for details of the normalization, see the appendix in Yurovsky et al., 2013). Also following the procedure of Yurovsky et al. (2013), Whitney-Mann U-test is used to compare token types since there is no reason to assume normality of these distributions while t-test is used to compare the non-normalized performance against the chance-level baselines.

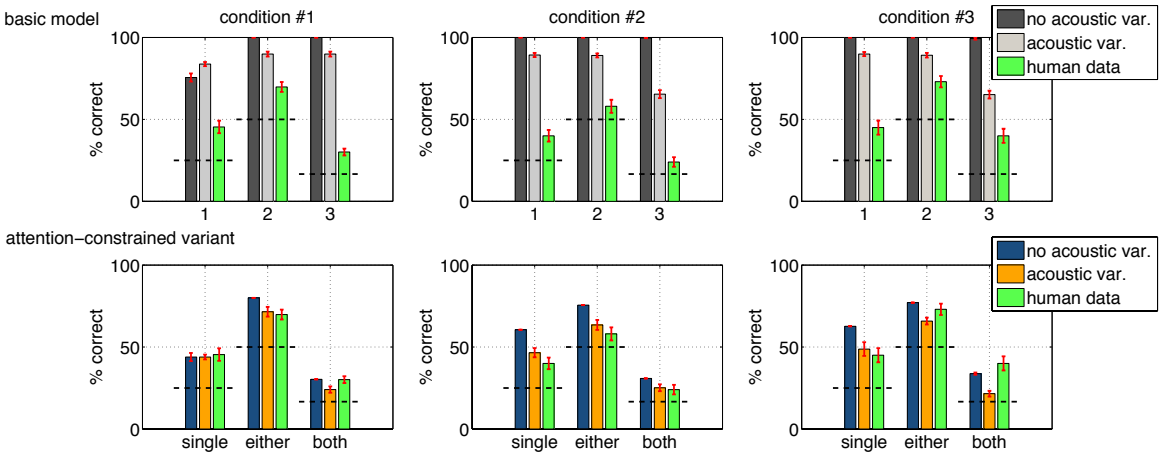


Figure 8: Results for the experiment 4. The top row shows the performance for the basic model in all three conditions whereas the bottom row shows the corresponding results for the attention-constrained model. The leftmost bars in each condition show the model performance without acoustic variability in the stimuli while the middle bars show the performance with acoustic variation. The rightmost green bars show the corresponding behavioral results from Yurovsky et al. (2013). "Single" refers to the learning accuracy of single word meanings whereas "either" and "both" imply that one or both of the two meanings of a double were learned, respectively. The error bars denote ±1 SE. Acoustic alphabet size is $|A| = 64$.

Fig. 8 shows the results for all three conditions (from left to right) for the basic TP-based model (top) and the attention-constrained model (bottom) together with the original behavioral results of Yurovsky et al. (2013).

All results are significantly above chance level in all three tasks with and without acoustic variability in the stimuli. Also, the basic model again clearly outperforms human subjects and the attention-constrained variant. Effects of global competition are observed for the basic model in conditions #2 and #3 with the performance on single words being significantly higher than on the both meanings of the double words with and without acoustic variability even after accounting for the chance-level differences between token types ($z > 5.24$, $p < 0.001$ for both; Mann-

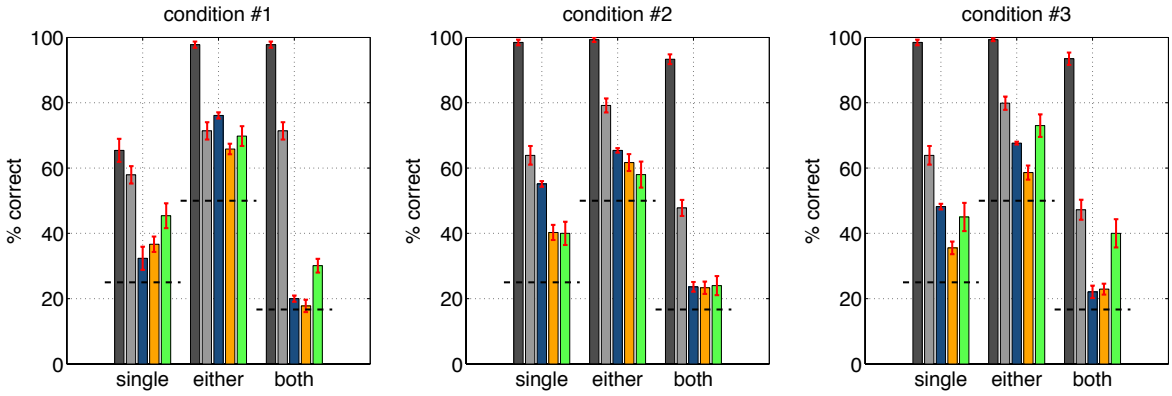
Whitney U Test). This competition is caused by the normalization of the TPs across competing models in Eq. (13). In contrast, both meanings of double words are learned better than the single word meanings in condition #1. Note that in condition #1 the basic model also achieves exactly the same performance level for learning either one or both meanings of the double words because the contents of the acoustic model are identical for both of the referents of each word, making them equally probable during each test trial. All in all, the basic model shows significantly above-chance performance in all three tasks. However, since it fits poorly to the behavioral data, the remainder of the section focuses on the results of the attention-constrained variant with the stimuli having normal acoustic variability across tokens.

In contrast to the basic model, performance of the attention-constrained model fits well to the behavioral data. In condition #1, the recognition rates for single words, either meaning of double words, and both meanings of double words are M_{single} = 43.89% (SD_{single}= 6.60%), M_{either} = 71.53% (SD_{either}= 13.33%) and M_{both} = 24.03% (SD_{both}= 8.75%), respectively. Statistical analysis, combined with the chance-level normalization of Yurovsky et al. (2013), confirms that the referents for single words are learned significantly better than both meanings of the double words in the condition #1 ($z = 4.44$, $p < 0.001$; Mann-Whitney U-test). In general, the attention-constrained model is almost always within ±1 standard error (SE) of the behavioral results in the first test condition (the difference is only significant for both meanings; $t(19) = -3.10$, $p = 0.0059$; one sample t-test).

Results for condition #2 follow those of condition #1. Recognition rates for the single, either and both meanings are M_{single} = 46.53% (SD_{single}: 12.68%), M_{either} = 63.47% (SD_{either} = 13.71%), and M_{both} = 25.14% (SD_{both} = 9.39%). Similarly to the data of Yurovsky et al. (2013), meanings of single words are again learned better than both meanings of double words (z =

3.944, p = 0.0001; Mann-Whitney U test), and all model results follow the pattern of the behavioral data, the difference in performance being marginally significant only for single words ($t(19) = 2.30$, $p = 0.0328$).

In condition #3, the recognition rates for single, either and both criteria are M_{single} = 48.75% (SD_{single} = 18.74%), M_{either} = 65.83% (SD_{either} = 9.46%), and M_{both} = 21.53% (SD_{both} = 7.95%), respectively. The model has again learned the referents of single words equally well as human subjects ($t(19) = 0.89$, $p = 0.3821$, difference not significant), and the difference in either meanings of double words is not large ($t(19) = -3.39$, $p = 0.0031$). In contrast, the increase in learning both meanings for the double words observed in human subjects is not replicated by the model, the learning performance of both meanings being significantly lower than the human performance in the same task (M_{both} = 21.53% SD_{both} = 7.95% versus M_{both_adult} = 40% SD_{both_adult} = 4.33%; $t(19) = -10.39$, $p < 0.001$). From the modeling point of view, this disparity is expected since the familiarity preference cannot account for enhanced learning of referents that are introduced at a late stage of the familiarization stage. Instead, a separate novelty preference mechanism would be needed to account for improved learning of the later referents (e.g., Rasilo & Räsänen, 2015).



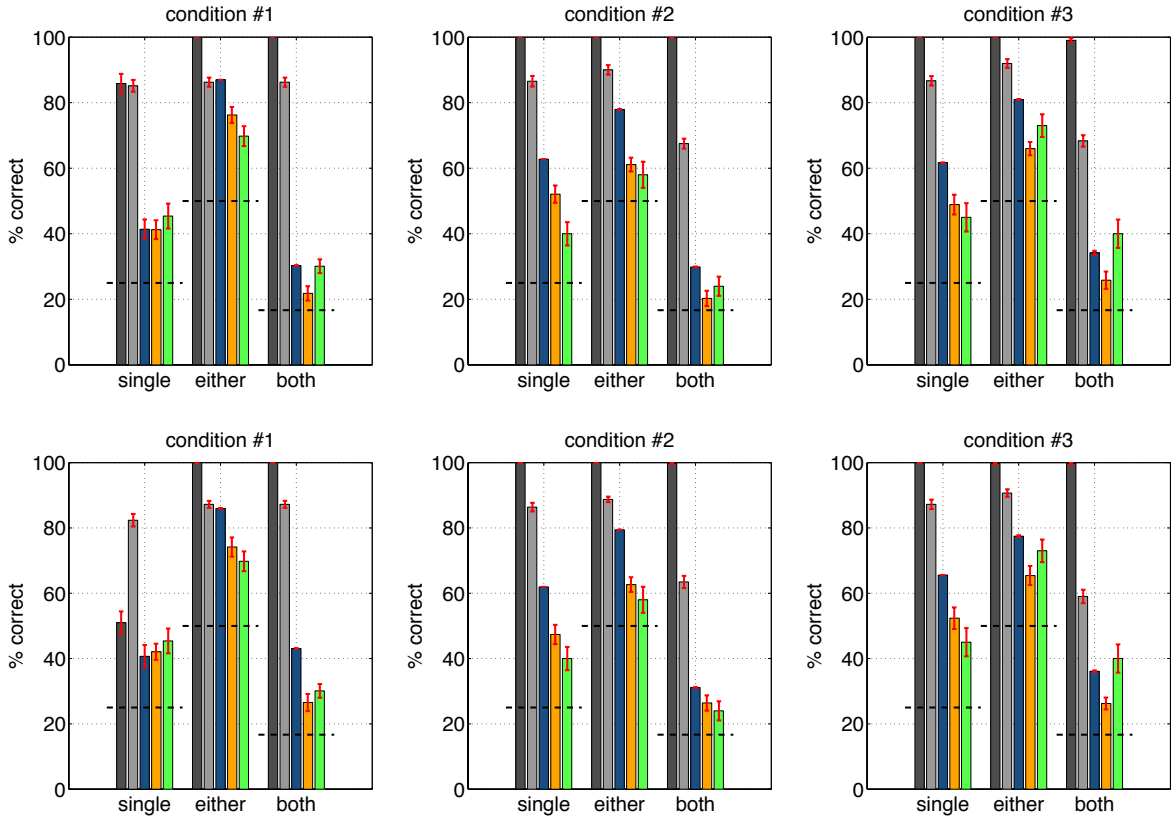


Figure 9: Results for all the three test conditions using an acoustic alphabet size $|A|$ of 8 (top), 32 (middle), and 128 (bottom). The dark grey bars (left): The performance for the basic model without acoustic variability in the stimuli. The light grey bars (second from left): The basic model with acoustic variability. The blue bars (middle): The attention-constrained model without acoustic variability. The orange bars (second from right): The attention-constrained model with acoustic variability. The green bars (right): The corresponding behavioral results from Yurovsky et al. (2013).

Finally, Fig. 9 shows the behavior of the two model variants as the resolution of the acoustic quantization is varied from the default value of $|A| = 64$. The overall pattern of the results is not greatly affected by the alphabet size for sufficiently discriminative acoustic representations ($|A| > 8$). Only when the representation becomes too coarse with $|A| = 8$, the attention-constrained model and the behavioral data start to diverge as the learning performance drops significantly in

condition #1. This excludes the learning of both meanings of the double words in condition #3, where human performance is never explained by the model.

5.5.3 Discussion for experiment 4. Similarly to the previous experiments, the results show that the TP-based joint model of speech segmentation and meaning acquisition is capable of learning word-referent mappings from a small number of exposures to individually ambiguous learning trials. Moreover, it is also capable of replicating the pattern of behaviorally observed effects of competition in word learning when a mechanism for familiarity preference is included in the model. The only exception is condition #3, where the model is unable to account for the improved learning of both meanings of double words observed in human subjects. The overall pattern of model outputs is also consistent across acoustic representations of different resolution. In addition, the model follows the mutual exclusivity principle (Clark, 1987; Markman, 1990) by learning single meanings for words consistently better than two alternative meanings.

The overall performance and error margins of the attention-constrained model are surprisingly similar to the human performance in the task, considering that the current study uses real speech stimuli instead of synthesized speech, and that the present model does not make use of any linguistic representations that are typically assumed to be central to speech perception in adult subjects. This suggests that this type of learning task may not necessarily require linguistic parsing of the input, but the behavioral results should also generalize to non-speech stimuli as long as they are otherwise acoustically equally distinct, like the words used in the study. Also, it is obvious that human performance is sub-optimal in the task as the basic TP-based model achieves superior performance in all conditions.

As the use of an attentional constraint leads to a nearly perfect fit to the human data in conditions #1 and #2, the results imply that humans may also employ a similar strategy of

attending to the most likely associations based on earlier experience. Whether this strategy is affected by explicit instructions to learn the audiovisual associations in the original study and whether similar results would be achieved through passive exposure to similar stimuli remains currently unknown. The attentional constraint also fits well with the findings of Yu and Smith (2011), Yu et al. (2012), and Yurovsky et al. (2013), who found increasing looking times and increasing preference towards the associative links established early in the learning process. However, the presently used attentional constraint should not be taken as an accurate model of attention but simply as an approximation to limited attentional resources or a focused learning strategy that fits well with the behavioral findings. For instance, the proper modeling of the effects observed in condition #3 likely requires some type of novelty preference, triggered, e.g., by the sudden appearance of the novel referents in the middle of the familiarization period (e.g., Räsänen & Rasilo, 2015). In addition, some learning is shown to take place also for the competing referents (Apfelbaum, 2013) whereas the present model only allows memory updates of the winning referent for each moment of time.

Overall, the fourth experiment provides further support for the feasibility of the present model as an efficient method for proto-lexical learning since it solves the same XSL problem that humans are known to be capable of solving and shows similar competitive effects when the exposure to word-referent statistics are manipulated. However, the present simulations cannot show that human infants or adults would apply a similar strategy in the word learning tasks. They simply show that a sufficient statistical structure is available in the sensory input and that a simple associative learning mechanism can capture it in order to create functionally relevant representations of the speech input. All this takes place without assuming that the learner is capable of parsing the speech input first into units such as phones, phonemes, syllables or words.

5.6 Experiment 5: Cross-situational learning in continuous speech

The goal of the fifth experiment was to study the behavior of the model in the case of natural continuous speech and to demonstrate how direct mapping of acoustic patterns onto their referential meanings leads to successful segmentation of the speech into word-like units.

5.6.1 Experimental setup. The Caregiver Y2 UK corpus (Altosaar et al., 2010) was chosen as the speech material since it readily contains utterances with referential information. More specifically, each utterance is paired with an unordered set of visual tags denoting the concurrent presence of visual referents corresponding to the so-called *keywords* (nouns, verbs, or adjectives) in the utterance, and where each keyword has its own unique referent. In addition to 1–4 keywords per utterance, each utterance consists of a surrounding carrier sentence with non-referential words, such as function words or pronouns (e.g., "*a* ***woman takes*** *the* ***yellow cookie***" or "*do you* ***see*** *the* ***red apple***", keywords emphasized). The main section of the corpus contains speech from four talkers (2 male, 2 female), each speaking the same set of 2397 utterances in enacted, child-directed speaking style. There are a total of 50 unique keywords and corresponding visual referents in the corpus, approximating the MCDI-based distribution of the receptive vocabulary of children during their second year of life (see Altosaar et al., 2010, for details), and the total size of the vocabulary, including the carrier sentences, is 81 unique words. All utterances are grammatically correct but semantically incoherent in order to ensure that there are no strong predictive dependencies between the keywords (e.g., "red apple", "square apple", and "dirty apple" are all equally likely). The corpus was recorded in a noise-insulated booth and the speaking style is acted UK English child-directed speech. All keywords and sentence structures are listed in Appendix C.

For our present purposes, all utterances with only one keyword were excluded from the data in order to have only referentially ambiguous learning trials. This led to an average of 6.1 words and 3.4 keywords per utterance (trial). For each run of the simulation, half of the utterances (N = 1000) from Talker-01 were randomly assigned as the training data whereas the remaining half (N = 999) were used to test the word-referent recognition performance of the model. Each keyword and the corresponding referent occurred on average 68 ± 32 times (min = 26; max 192) in the 1000 utterances. Speech was fed to the algorithm utterance-by-utterance, always accompanied by the set of referent labels $c_1, c_2, ..., c_N \in [1, 50]$ corresponding to the keywords in the utterance.

The experiment was performed separately with the original correct referential information and with two levels of additional uncertainty by randomizing 40% or 80% of the original visual referents to any of the 50 referents in the data (i.e., the correct referent for a spoken keyword is present, on average, in one fifth of the training trials at the 80% noise level). In order to recognize the meaning of an utterance X during the testing stage, the activation $A'(c, t| X)$ of each referent at each moment of time was first computed using Eq. (15) using a sliding window of W = 250 ms (see Fig. 10 for an example). Then the referents c*(t) with the highest activation and the corresponding activation levels $A^*(t)$ (the highest points of the activation curves in Fig. 10) were computed for each moment of time t.

$$\begin{aligned} c^*(t) &= \arg_c \max\{A'(c,t \mid X) \mid \forall c\} \\ A^*(t) &= \max\{A'(c,t \mid X) \mid \forall c\} \end{aligned} \tag{18}$$

Finally, only the referents included in c^* that had a maximum activation A^* higher than δ standard deviations above the mean of all winning activations were chosen as the word hypotheses for the utterance. Formally, the winning meaning $c^*(t)$ is added to the list of hypothesized utterance meanings only if

$$A^*(t) \geq \mu(A^*) + \delta\sigma(A^*) \quad (19)$$

where μ and σ are the mean and standard deviation of the winning activations across the utterance ($t = 1, \ldots, T$), respectively, and if the meaning $c^*(t)$ is not already in the list. In practice, Eq. (19) implements a dynamic word detection threshold that reacts only to patterns whose activation rises above the baseline activation that is observed during silence, transitions between words, and during speech that does not have clear referential predictions, i.e., when the meaning is not clear.

The overall recognition performance was measured in terms of precision (the proportion of correct referents from all hypotheses), recall (the proportion of referential meanings correctly detected from all true referents), and F-score (the harmonic mean of the two), all scaling between 0 and 1. As for the segmentation performance, the average temporal distances from the onset and offset boundaries of a hypothesized word were measured to the nearest word boundaries in the reference annotation. In addition, the average length of the correctly recognized word segments was measured in order to investigate how well the learned segments correspond to the durations of actual word segments in the corpus.

5.6.2 Results. Fig. 10 shows an example of the model output for an early stage of the learning and after processing of the full training set without added referential noise. In addition, global TPs from a standard non-contextual model computed using Eq. (14) are shown for comparison ($c = 1$ always; see experiment 1). As can be observed from the second panel, the activation of each referent, given the audio, is relatively noisy after observing only 60 utterances, and there is no clear winner except for the ending of "telephone". After full training, the words "*sad*" and "*telephone*" have been successfully associated to their corresponding referents, leading to clear activations that approximately correspond to the temporal extent of the underlying 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., "*the*") do not have distinct activation segments. Also, activation of the referent {*to see*} extends to across the entire phrase "*do you see*" as it almost always occurs within this phrase in the corpus.

In contrast, the output from context-free tracking of TPs shows no obvious cues to word boundaries (the bottom panel in Fig. 10). Quantitative results for this type of non-referential short-term acoustic TP-analysis is found in Räsänen (2014), where the minima in global acoustic TPs were shown to best correspond to phone boundaries instead of syllables or words (the three levels of representation achieved segmentation F-scores of 0.73, 0.35, and 0.25, respectively).

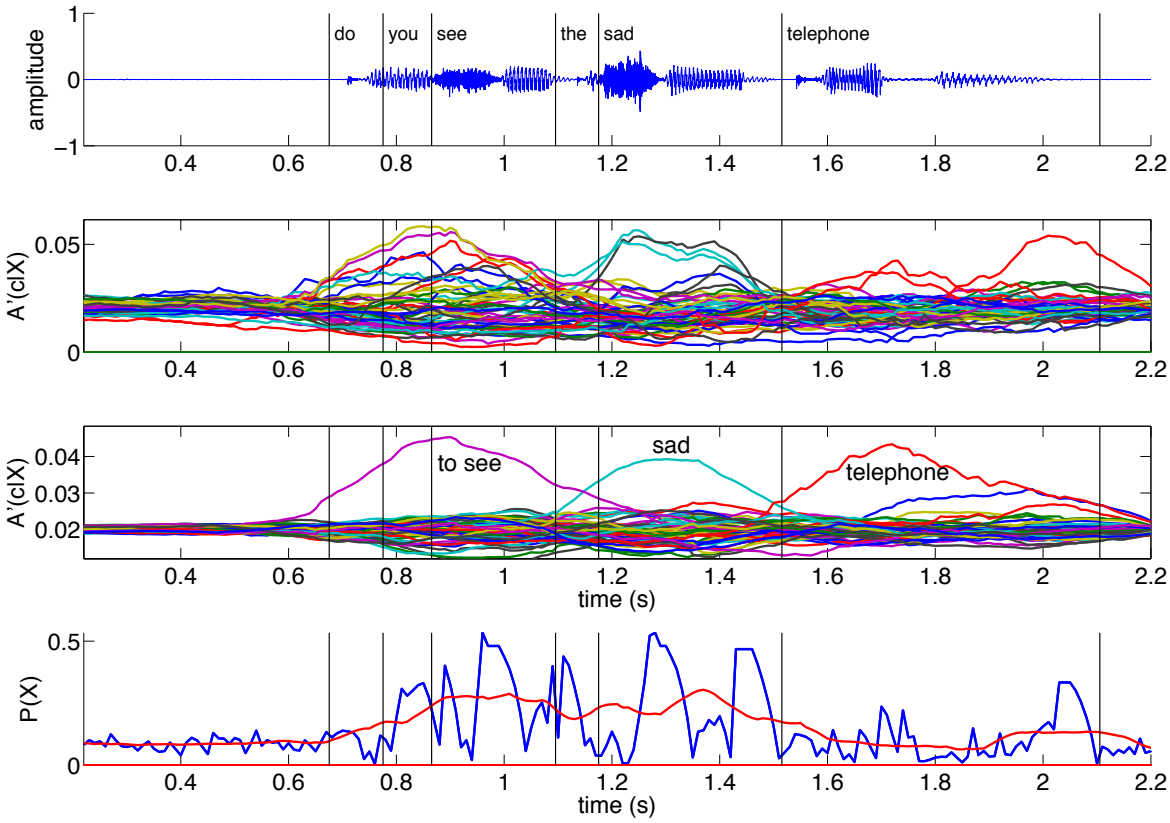


Figure 10: An example of model's recognition output for the sentence "*Do you **see** the **sad telephone***?" (referential words emphasized). Top panel: The original waveform. Second panel: The model output after

exposure to 60 sentences. Third panel: The model output after exposure to 1000 sentences. Bottom panel: The output of a non-contextual TP-based model with raw TPs (blue line) and temporally smoothed TPs (red line). In the top three panels, the different colored curves represent activations of all 50 different visual referents *c* as a function of time. The vertical lines show the true word boundaries extracted from the Caregiver corpus annotation.

Fig. 11 shows the word-referent recognition performance as a function of the number of perceived utterances. The detection threshold was set to $\delta = 0.6$ standard deviations (see Eq. (19)) since this was found to balance the precision and recall of the model. The result is shown for the original referential information where referents always correspond to the keywords in the spoken utterances. In addition, results with 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 speech, achieving an F-score of 0.85 with a precision of 0.88 and a recall of 0.82. In addition, learning is very fast initially, and the F-score is already above 0.75 after 250 trials, corresponding to on average of 16.5 presentations of each referential word.

The final results for the two noise conditions are $F_{0.4} = 0.82$ and $F_{0.8} = 0.54$ in the order of increasing uncertainty. This shows that the model copes well with referential uncertainty, and the precision and recall are very high even when almost half (40%) of the visual stimuli are not related to the speech contents. Even in the case of only 20% of referents being related to the actual contents of the utterances, the learner still gradually acquires consistent knowledge between word patterns and their referential meanings without signs of saturation in performance. Similarly, the attention-constrained model shows successful learning in the task. However, learning is much slower in this case as the learner has access to much less information during

each trial, achieving final performance levels of F = 0.57 without referential uncertainty and, interestingly, F = 0.60 with 40% of the referents randomized.

Fig. 12 shows the corresponding segmentation accuracy (left) and word-segment length (right) for all hypothesized words as a function of utterances perceived. In both cases, the model shows significant improvement in segmentation accuracy as more training data is observed. Also, the mean segment length approaches the true mean keyword length as the recognition performance improves. The Pearson correlation between the word-recognition F-score and average segmentation error is $r = -0.997$ ($p < 0.001$), further confirming that the accuracy of the temporal segments and the correct associations of speech patterns to their referential meanings develop hand-in-hand. The lengthening of the hypothesized segment also indicates that the learner first starts to discriminate different referential contexts based on short segments of speech that are especially prominent in these contexts and then gradually learns the overall extent of the word-like units as more evidence is accumulated with experience. The overshoot in word lengths after 500 utterances (Fig. 12, right panel) can be explained as follows: As the function words neighboring the keywords do not have referential meaning, they cannot win the competition in activation and therefore suppress the activation of a keyword. Since word activations are smoothed in time during the recognition (Eq. 15), this causes spreading of activation outside the actual temporal extent of the words. This effect is also seen in the activation curves of Fig. 10, where, e.g., word “the” becomes covered by the two neighboring keywords. All observations of the pattern length and segmentation accuracy are similar for different levels of referential uncertainty, with the performance curves simply being linearly shifted due to slower learning (not shown separately).

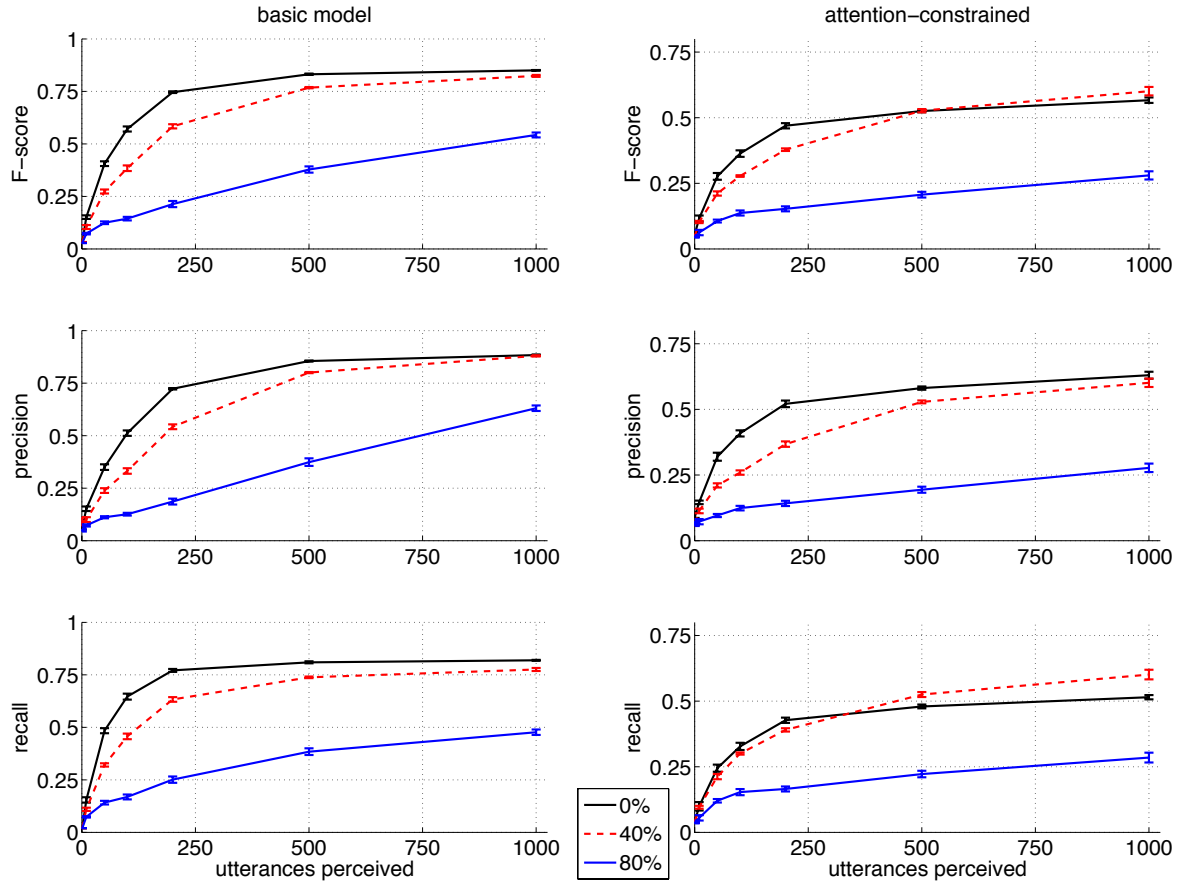


Figure 11: Word-referent recognition performance in terms of F-score (top), precision (middle), and recall (bottom) as a function of the number of sentences with which the model is trained. Left panel: The basic model. Right panel: The attention-constrained model. The top-most black line shows the result without added referential noise. The red and blue lines show the results with referential noise having 40% or 80% of the original referent labels randomized to any of the 50 possible referents. The error bars correspond to ±1 standard deviations across five runs of the experiment.

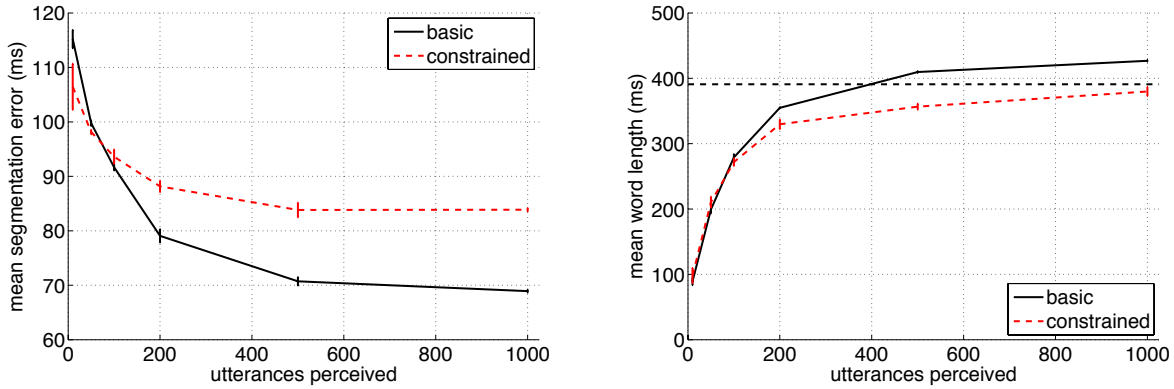


Figure 12: Left panel: The mean distance from real word boundaries to the algorithm-produced boundaries (points where the winning referent changes) as a function of utterances perceived. Right panel: The mean duration of the hypothesized word segments as a function of utterances perceived. The average keyword duration in the corpus is shown with a black horizontal dashed line. The basic model is shown with the black solid lines and the attention-constrained model with the red dashed lines. All results are from the experiments without referential noise. The error bars correspond to ±1 standard deviations.

5.6.3 Discussion for experiment 5. The fifth experiment shows that the model is capable of learning word-referent mappings and acoustic word-like segments simultaneously from real speech without a priori linguistic knowledge. Word learning is successful even though the referential and acoustic information are aligned only at the utterance level, containing significant uncertainty with respect to what parts of the acoustic signal correspond to which referent. Moreover, learning is successful even when the set of concurrently shown visual referents is only partially related to the speech contents, increased referential uncertainty simply being reflected as a slower learning rate. Finally, the learning is also successful when the learner can only attend to one referent at a time in the attention-constrained condition. In fact, this type of binary gating mechanism may be actually too limiting since there is evidence that human subjects are actually capable of learning about multiple competing referent candidates at the same time (Apfelbaum,

2013). Work on audiovisual integration (e.g., Kopp, 2014) and the results of experiment 2 also suggest that there is representational persistence of the previously perceived (attended) visual information even if the visual stimulus is no longer present. Still, it provides a useful lower bound for performance of a learner with very limited cognitive capacity. A learner with larger working memory and more flexible attentional span would perform at a level that is between the ideal model and the current attention-constraint variant.

In contrast to the words that occur with contextual referents, the present framework does not lead to segmentation of words that are statistically independent of the context in which they are used (no referential labeling). Measuring the learner's familiarity with these word forms similarly to the first two experiments would be still possible, leading to higher familiarity towards the words that occurred during the familiarization than any word that did not. Nonetheless, an actual word segment becomes defined only through competition between existing proto-lexical entities, i.e., acoustic patterns that have been associated to their (potential) referential meanings. What this also means from a statistical learning point of view is that a word can correspond to a set of transitions that have low probability in the overall space of all possible transitions between representational elements: As long as these rare transitions occur systematically in a specific communicative context, they have high predictive power over the referential domain and are therefore highly informative to the language user (e.g., compare contextual and non-contextual TP models in Fig. 10).

In all, experiment 5 shows that joint acquisition of word segments and their referential meaning is a feasible strategy and that word learning does not require an explicit segmentation stage before acquiring meaning. Note that the current results are obtained using a relatively simple discrete representation of the speech input, modeled by simply measuring the TPs

between them. The performance is expected to improve if the learner has access to speech representations that can overcome acoustic variability across different tokens of the same word.

5.7 Experiment 6: Generalizaton across talkers in XSL

5.7.1 Experimental setup. The final experiment investigated the generalization of learned words towards new talkers. Since all four main talkers (2 male, 2 female) of the Caregiver corpus speak the same set of sentences, the experimental setup was otherwise identical to the previous experiment but, instead of always training and testing with data from the same talker, the training was performed using utterances from M = 1, 2, or 3 talkers different from the testing stage. The number of training utterances per talker was split equally among all talkers present in the training. There were two basic conditions: 1) a "fully referential" condition, where the number of training trials was always N = 1000, divided equally among the M talkers in the training, and all utterances were paired with their correct referential cues, similarly to the previous experiment and without added noise in the referents, and 2) a "bootstrapping" condition, where the total number of training utterances was set to N = 1000 ("short bootstrap") or $N = M*1000$ ("long bootstrap"), but pairing the utterances from only one talker with referential cues for the keywords (the bootstrapping trials). As for the remaining training utterances, the learner had to use its lexical knowledge from the bootstrapping trials to recognize the keywords spoken in the utterances. The model was then trained using all words detected in these non-referential trials (or "self-labeling trials") as if they were actual referential cues. The detection threshold for words was set to $\delta = 0.6$, similarly to the experiment 5, and all detected words (meanings) were used in further training even if these were erroneous. During each run of the experiment, training and testing talkers were rotated so that each of the four talkers in the corpus became tested for each value of M. Training talkers were always randomly sampled from the set of three possible talkers

for M = 1 and 2. Acoustic alphabets of size $|A|$ = 64 and 128 were generated from the combined data from all four talkers as described in section 5.1. Table 2 summarizes the setup.

Table 2: The number of training trials with referential cues and trials where the learner had to recognize the word meanings autonomously. Each experimental condition and the number of talkers in the training stage listed separately. Bootstrapping was always performed with data from only one talker.

number of talkers	condition	referential trials	self-labeling trials	total trained
***M* = 1**	fully referential	1000	0	1000
	short bootstrap	1000	0	1000
	long bootstrap	1000	0	1000
***M* = 2**	fully referential	1000	0	1000
	short bootstrap	500	500	1000
	long bootstrap	1000	1000	2000
***M* = 3**	fully referential	999	0	999
	short bootstrap	333	666	999
	long bootstrap	1000	2000	3000

The goal was to investigate whether the acoustic differences between talkers impose a challenge to the learner, and whether the learner can overcome this variability by either associating acoustically varying word forms from different talkers to the shared referents in the fully referential trials or by improving its generalization skills by recognizing words in new sentences occurring in non-referential contexts and using the new data to update the lexical representations for the words. Should the word recognition performance be higher in the "fully referential" condition than in the bootstrapping conditions, even despite the smaller number of training trials, it would mean that the learner clearly benefits from cross-situational cues also when learning about talker-dependent variability. On the other hand, improvement due to exposure to new talkers in the bootstrapping conditions would mean that the learner has acquired sufficiently functional representations for the words so that it can keep improving the specificity of its lexical

representations without contextual support, and do so even in the face of novel talkers. Moreover, the higher performance would show that the present framework also scales to learning that is driven by internally generated representations of the referential meaning, even if these representations are not fully accurate with respect to the true meaning of the words (see Versteegh et al., 2010).

5.7.2 Results. Fig. 13 shows the resulting F-scores for word recognition across five runs of the experiment after the full training had been processed and for both acoustic alphabet sizes. The first finding from the fully referential condition is that generalization from one talker to another is much lower (F = 0.41 or F = 0.43 for $|A| = 64$ and $|A| = 128$, respectively) than when the training and testing utterances are from the same talker (F = 0.85 in the previous experiment). This indicates that there is a notable acoustic mismatch between words spoken by different talkers. However, if the variability in the training data is increased by introducing more talkers while keeping the amount of training data at the same level, the performance increases significantly with F = 0.50–0.51 and F = 0.54–0.55 for two and three unique talkers, respectively. This shows that the learner's capability to generalize towards new tokens increases when there is more acoustic variability in the training data, at least as long as there are referential cues that are shared across utterances from different talkers.

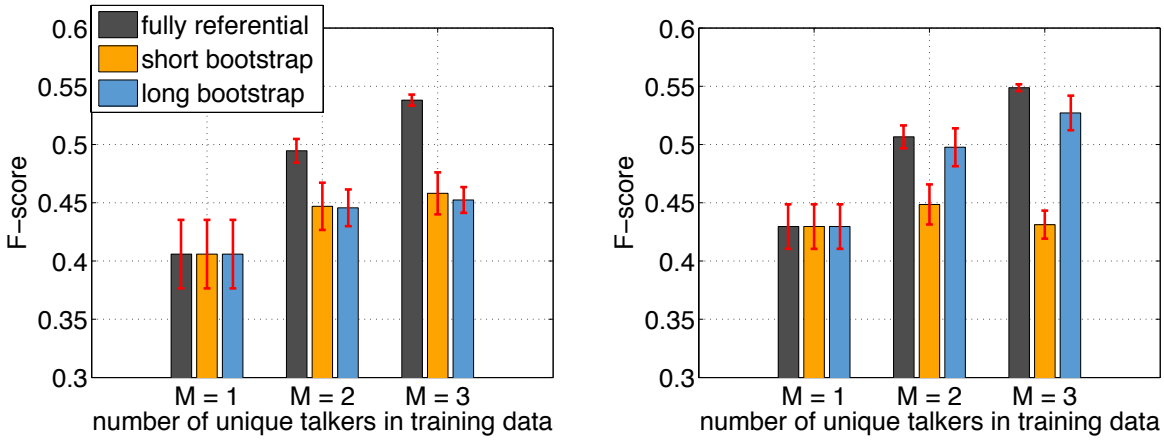


Figure 13: Word recognition performance on speech from a previously unseen talker when the vocabulary has been learned from M = 1, 2, or 3 different talkers with $|A|$ = 64 (the left panel) and $|A|$ = 128 (the right panel). The black bars show the results when a total of 1000 utterances with referential information are present in the training. The orange and blue bars show the results for short and long referential bootstrap periods with a single talker, respectively, followed by internally generated referential labeling of the remaining training utterances. The error bars denote ±1 SE.

In the bootstrap conditions, short bootstrapping with 500 (M = 2) or 333 (M = 3) referentially cued utterances does not reach the levels of performance in the fully referential condition, implying that the bootstrapped learner does not benefit from new talkers as much as a learner having access to concurrent referential cues. However, with a longer bootstrapping period with referential cues and an acoustic alphabet with more detail ($|A|$ = 128), there are clear improvements in word recognition accuracy when the learner has access to more speech data from new talkers, the F-score increasing from the initial F = 0.43 after the bootstrapping stage to F = 0.53 after interpreting and learning from 2000 new sentences without concurrent referents. This shows that the learner is no longer dependent on the availability of concurrent visual information in order to improve the accuracy of its lexical representations. However, perceiving a total of 1000 referentially cued utterances from three different talkers is still more effective than 1000 referentially cued utterances from one talker supplemented by 2000 non-referential utterances from the remaining two talkers (M = 3 in Fig. 13). It also appears that bootstrapping is not as successful with the smaller acoustic alphabet, leading only to marginal increases in performance with more data. Although it is hard to determine the exact origins for this difference, manual analysis revealed that the smaller alphabet led to the generation of more hypothetical meanings for utterances during the self-labeling stage when faced with a new talker. Due to this

increased ambiguity, the learning from new data is much slower because there are no external referential cues that would help to limit the ambiguity during each learning trial. It is likely that the performance after the bootstrapping stage could be improved somewhat by optimizing the detection threshold δ for different stages of the learning process. However, note that the bootstrapping performance can never become better than the fully referential performance, since the ideal output during the self-labeling stage corresponds to the correct referents of the keywords that are always shown in the fully referential condition.

5.7.3. Discussion for experiment 6. The experiment shows that there is large acoustic variability across talkers, making recognition of new word tokens from a familiar talker much easier than recognition of the same words spoken by another talker. However, the results also show that a cross-situational learner can improve its generalization capability by hearing speech from multiple talkers in similar referential contexts. Although the addition of new talkers does not reduce the referential ambiguity or add more training data in the fully referential condition, it provides a richer set of exemplars for each word, leading to a better understanding of the relevant acoustic characteristics of the words.

Importantly, the experiment also shows that word learning can continue in the absence of referential cues once the learner has acquired sufficiently detailed representations for the words during a referentially-driven bootstrapping stage. From computational perspective, this post-bootstrapping learning can be thought of as a process where the external referents and the associated referential ambiguity become replaced by internally generated hypotheses for the referents and the potential the inaccuracy of this recognition process. As long as the learner can come with a limited number of hypothesized meanings for an utterance, and as long as these hypotheses are correct at above chance level, cross-situational learning can proceed as if these

internally generated hypotheses would be equivalent to externally perceived referents. This makes an important prediction with respect to human learning, namely that the development of acoustic specificity for familiar words can continue as soon as the words (and their meanings) are recognized from the input, even if the recognition is not very accurate at first.

6. General discussion

For daily use of language, the meaningfulness of perceived speech patterns is the most important aspect of communication. Since language is all about sharing conceptual representations regarding the surrounding world, whether they are true or fictitious, the acquisition of the links between arbitrary sound patterns and these concepts, the referential aspect of language, should be the primary driving force in language learning.

In the first part of this paper, we presented a mathematical description of a joint model for word segmentation and meaning acquisition. The fundamental idea of the model is that statistical learning operates similarly not only in different perceptual domains, but also across them. We argued that the joint model provides a parsimonious solution to the early word-learning problem, this being easier to solve than the separate treatment of word segmentation and meaning acquisition problems, and ultimately leading to a proto-lexicon with its referential value directly dependent on the quality of the solution. According to the model, word segmentation does not occur before and independently of lexical access. Instead, word boundaries emerge from dynamic on-line competition between potential meanings for the unfolding acoustic input (c.f., TRACE; McClelland & Elman, 1986). The model essentially follows the clustering strategy of word segmentation discussed, but not formalized, in the earlier literature, but now integrating contextual referential information directly in the clusters in addition to purely auditory statistical

regularities. In the context of language acquisition, this type of mechanism allows the learner to capture statistical dependencies between auditory perception and other concurrent representational states, such as attended visual objects and events, concurrently performed motor activity, or, e.g., current emotional states, and these representations correspond to the concept of *proto-lexicon* as defined by Nazzi and Bertoncini (2003). The joint learning strategy also helps the learner to deal with acoustic variability in speech. This is because two acoustic patterns are considered to be lexically equivalent if their referential predictions are similar, even if the two signals do not share similar acoustic properties. This provides important anchoring when learning to differentiate relevant variation from irrelevant, providing a significant constraint for phonological learning where phoneme boundaries are, by definition, characterized by changes in the signal that lead to changes in the meaning of the word (cf. PRIMIR, Werker & Curtin, 2005).

In the second part of the paper we described a computational implementation of the model that makes a number of approximations to the ideal model. The experiments with synthesized and pre-recorded real speech and simulated visual input show that the model is able to learn word segments and their referential meaning from continuous speech without making use of any a priori linguistic knowledge. When augmented with a simple attentional mechanism that prefers already familiar audiovisual associations above novel ones, the model is also able to account for a large proportion of behavioral data observed in different XSL tasks, including the implicit realization of the mutual exclusivity principle (Clark, 1987; Markman, 1990) through competition between cross-modal associations. In addition, the model also succeeds in word learning under different degrees of referential uncertainty, showing that the model is not critically dependent on the reliability of the referential information.

Taken together, the theoretical treatment and the success of the model in the simulations argue for the central role of referential cues in proto-lexical learning and question the need for a separate segmentation stage preceding the meaning acquisition. The findings converge with recent behavioral evidence that shows concurrent learning of word-like patterns and their referents in audiovisual learning tasks (Cunillera et al., 2010; Thiessen, 2010; Yurovsky et al., 2012, Glicksohn & Cohen, 2013; Shukla et al. 2011; see also Wojcik & Saffran, 2013, and van den Bos, Christiansen & Misyak, 2012, for related findings). This suggests that speech perception should not be taken as a process independent of other concurrent cognitive processes or perceptual domains. In addition, this work shows that the inference of latent lexical structure may not be required for the bootstrapping of language acquisition, but that the proto-language can be acquired using feed-forward associative mechanisms. This is not to say that such an inference could not take place during language acquisition, but the point is that the estimation of model parameters in purely auditory segmentation is much more demanding than the joint account of segmentation and meaning acquisition.

A major drawback in the present study is that the behavioral data for context-driven speech segmentation in infants is so far very sparse, and therefore much of the experimental validation has been performed with respect to adult data. Although there is clear evidence for visual facilitation for word segmentation in adults (Cunillera et al., 2010; Glicksohn & Cohen, 2013; Thiessen, 2010), the questions of when and how this capability emerges in children is currently unclear, not least because probing of infants' learning processes in natural environments is far from trivial (but see Shukla et al., 2011, for initial evidence for joint learning already at the age of 6 months). This leaves much of the hypotheses presented by the current study to be validated in further studies. However, similarly to numerous other studies on statistical learning,

a reasonable working assumption is that implicit statistical learning in adults and infants is achieved with similar basic principles, the main differences being in the available cognitive resources such as span of the working memory, executive functions, plasticity (in the favor of infants), and the nature of the representations derived from the sensory input.

With this respect, the present model simply assumes that the learner is capable of 1) mapping acoustic input to short-term spectral representations in a consistent manner, 2) representing the temporal evolution of the short-term spectrum on the time-scale of some hundreds of milliseconds, and 3) memorizing these acoustic patterns in a context-sensitive manner by including information of the concurrent attended context in the memory trace. During word recognition, the learner simply uses the currently observed acoustic patterns to retrieve the contextual information that was stored earlier together with similar acoustic input, accumulating memory activations across the time-scale of a typical word-length. Finally, all these stages have to take place only at a representational fidelity that is sufficient to distinguish the words of the existing vocabulary from each other with only small but systematic updates to the memory across time (c.f., McMurray et al., 2012). None of these requirements seem unreasonable for an early word learner.

The main limitation in the current treatment is that it only accounts for word learning in terms of statistical learning at a very general level. In reality, additional contributions from the gradually and in-parallel emerging phonological and lexical knowledge (e.g., Eimas & Quinn, 1994; Kuhl et al., 2008; Feldman et al., 2009; Adriaans & Kager, 2010; Feldman et al., 2013; Elsner et al., 2013), earlier statistical learning (Graf Estes, Evans, Alibali & Saffran, 2007; Mirman et al., 2008; Hay et al., 2011) and the constraints imposed by stress patterns, prosody and syllabic structure (Cutler & Norris, 1988; Shukla, Nespor & Mehler, 2007; Shukla et al., 2011;

see also Jusczyk, 1999) are likely to influence the parsing of the speech input, making some speech patterns more likely candidates for words than others. Finally, the present model does not take into account speech production at all, although it can also provide useful constraints to the perceptual processing of words, especially during later stages of language acquisition.

However, none of these above factors are incompatible with the general model described in section 3, but can be seen as additional mechanisms and constraints affecting the type of representations that participate in the cross-modal learning process. As long as the representation of speech is still at a sub-word level, the input to the cross-modal learning can be phones, phonemes, syllables, or outputs from purely auditory statistical learning (cf., Graf Estes et al. 2007; Hay et al. 2011; Mirman et al., 2008) without a loss of generality. Similarly, different parts of speech may receive differential weights during memorization due to different prosodic saliency. A hybrid model utilizing all these different cues as per to their availability at different developmental stages would provide a more comprehensive description of the early word learning, but is beyond the scope of the present work. The present work simply argues that, according to theoretical referential optimality, computational simulations, and human adult data, referential information is a powerful cue for word segmentation. Therefore it is necessary to better understand how word segmentation and meaning acquisition interact in early language acquisition.

6.1 Relation to existing research

The present work is by no means the first to point out that cross-situational learning between perceptual modalities may be beneficial to the learning of the first word segments. For instance, the previous models by Roy & Pentland (2002), Yu & Ballard (2004), Räsänen et al. (2008), Van hamme (2008) and ten Bosch et al. (2009), Aimetti (2009), and Räsänen & Laine (2012) all make

use of this principle. However, the present work provides the theoretical motivation for such an approach and, to the best knowledge of the authors, is the first to explicitly and extensively compare the model performance to human behavior in similar tasks when no linguistic a priori knowledge is utilized in the speech pre-processing. Similarly, the cross-situational aspect of the model resembles other models of XSL (e.g., Fazly et al., 2010), since it is based simply on the joint probability distribution between words and their referents. The main difference is that there are no discrete pre-segmented words in the present model, but the word segments in XSL are actually replaced by direct associative models between referents and speech acoustics, directly optimizing the overall referential value of the learned lexicon (see sections 3 and 4).

With respect to the existing models of early language acquisition, the model is most closely related to the PRIMIR framework of language acquisition (Werker & Curtin, 2005) that states that proto-lexical word learning precedes the organization of the subword structure of the language. However, PRIMIR assumes that the learner is equipped with many evolutionary and epigenetic biases related to language perception whereas the present model mainly assumes that the learner is capable of learning statistical dependencies from sensory input, that the referential entities are perceived and represented as more invariant categories than the auditory input, and that the set of possible word referents in each communicative situation is already limited by some constraints, such as joint-attention between the learner and the caregiver.

The present model also takes an explicit position in the bracketing-versus-clustering discussion on statistical word segmentation (Goodsitt et al., 1993), since the proto-words of the model directly correspond to audio-visual clusters. From a statistical learning point of view, successful bracketing actually implies clustering: In order to evaluate the probability of a sensory input at a given point in time, the learner has to have some type of existing internal representation

for similar inputs so that the familiarity of the current input can be evaluated with respect to these representations. Since high-probability (frequent) patterns (e.g., words) are faster to learn than low probability (infrequent) patterns (e.g., transitions between words), whereas chance-probability events (noise) are not learnable at all, it is difficult to imagine a learning mechanism, artificial or biological, that specializes in the unreliable and rarely occurring transition points without learning the statistically significant patterns on both sides of these transitions.

In the present model, statistical learning on auditory speech in the absence of contextual grounding leads only to a familiarity effect. The output of the word learning stage is always a cross-modal association that binds together the acoustic pattern and the referential context in which the word was spoken. Temporal boundaries of this "word" emerge on-line from the probabilistic decoding of the present input (c.f., dynamic competition in the model of McMurray et al., 2012, or the TRACE model of speech perception; McClelland & Elman, 1986). This means that there are no explicit representations for distinct ungrounded word segments that would be somehow accessible to the learner. However, the present framework does not preclude the possibility for such representations. Instead, it shows that the existence of such units is not required for proto-lexical learning or replication of behavioral results on word learning.

Importantly, this work should not be confused as an argument against purely auditory statistical learning. Quite the contrary, the model is compatible with the idea that our perceptual system attunes to the distributional characteristics of the sensory input, leading to more efficient and consistent mapping from low-level input to stable internal representations, which then serve as more effective representational units in further statistical learning within and across perceptual domains. For example, this type of "pre-learning" is observed in the studies of Graf Estes et al. (2007), Mirman et al. (2008) and Hay et al. (2011), who show that pre-exposure to an artificial

language facilitates later mapping of high-TP words to their referential meanings, indicating that statistically coherent auditory patterns are more efficiently represented, and therefore more easily associated to visual representations. In the present experiments, the bottom-up clustering of acoustic features was used to approximate pre-lexical distributional adaptation to native speech sounds since clustering of larger units is challenging in the face of the acoustic variability present in natural speech. However, if the speech material were simple enough, e.g., the repetitive synthesized speech used in many behavioral experiments, clustering could be carried out directly at the signal level in order to discover the small number of syllables and words that make up the stream. Unfortunately, this type of invariance is not present in real speech where purely bottom-up clustering into phones, syllables or words is far more difficult. Even if the learner has access to syllabic boundaries, the discovery of word clusters based on acoustic similarity is extremely challenging (e.g., Räsänen, Doyle & Frank, 2015; see also Versteegh et al., 2015, for the state-of-the-art automatic discovery of words from speech in an unsupervised manner).

6.2 Assumptions and limitations of the TP-based implementation

The model connects the idea of TP-based statistical learning to the general early language acquisition framework by treating TP analysis as a practical approximation of an ideal mathematically abstract learning mechanism. TPs and the discrete units upon which the TPs are computed should not be taken as a ground truth for the underlying cognitive representations of speech. Instead, the TP analysis acts as a simple proxy for physical world regularities, providing researchers with practical tools to understand and model the structure of sensory input. Note that, in the general model of language bootstrapping described in section 3.2, no assumptions were made regarding how the speech input is represented by the learner. Therefore the general model should be taken as a *computational level description* of the word learning process whereas the

TP-based implementation used in the experiments and described in section 4 corresponds to *an algorithm level* implementation of this model (c.f., Marr, 1982). Neither should be taken as a description of the mental representations and mechanisms responsible for statistical learning in the human brain.

It is also worthy to note that the inference of the most likely referent from context-specific TPs leads to fundamentally different results than looking at the global and non-referential TP statistics across all speech input. For instance, Räsänen (2014) shows that the minima in global acoustic TPs correlate poorly with word boundaries in continuous American English whereas Kurumada et al. (2013) found that TPs at the syllabic level provide a poor fit to human data on word segmentation. So far, reasonable results with TP-based models have only been obtained when the statistics are analyzed using a set of mutually disjoint competing models, each focusing on specific temporally contiguous segments of speech (Räsänen, 2012) or on different referential contexts (this work).

With respect to the hyperparameters of the TP-based implementation, there are few that have notable impact on the results. The integration of TPs over temporal distances of 250 ms is not an arbitrary choice, but matches to what is known about temporal integration in the human auditory system (Räsänen & Laine, 2013) whereas integration of information at larger time scales (larger K) would not change the results due to the lack of statistical structure at those distances. Moreover, Räsänen & Laine (2013) have shown that the properties of the temporal integrator itself can be learned from auditory experience with statistical learning mechanisms similar to that in the present work, making the idea compatible with the statistical learning paradigm. This leaves the acoustic alphabet size $|A|$ as the most central free parameter in the system. However, the current study shows that the pattern of results is not critically dependent on the value of $|A|$, as

long as the number of unique events is large enough to differentiate between acoustic patterns with different referential meanings but small enough to allow generalization towards new speech input from a limited set of training samples.

The largest potential mismatch to existing behavioral data comes from the assumption that detailed co-occurrence statistics of low-level acoustic features (lagged bi-grams) always directly interact with potential referential representations. Graf Estes et al. (2007) and Hay et al. (2011) have shown that infants benefit from pre-exposure to a statistically structured artificial language before attempting to learn associative links between high-TP word forms and their visual referents. One possible explanation is that statistical word segmentation during the pre-exposure stage facilitates later word-referent mapping (Hay et al., 2011). However, another interpretation is that sufficient pre-exposure to statistical patterns of an unfamiliar language simply makes the pattern representations stable enough to participate in further cross-modal associations. In the present model, the learner is assumed to be able to already represent the evolution of the speech signal across time even though it may not know the actual word boundaries (see section 3.2). Infants possibly simply need to reach a certain degree of familiarity with the auditory input before properties of the input can be tracked across different referential contexts. In this case, representations involving high-frequency words will stabilize earlier than their low-frequency counterparts. It is also unclear how the requirement of a pre-exposure to high-TP syllable-patterns is related to the fact that a very large proportion of word tokens in speech are monosyllabic in many languages. Still, irrespectively of how the results of Graf Estes et al. (2007) and Hay et al. (2011) are interpreted, the present implementation does not account for such pre-exposure effects without a further mechanism for purely auditory learning.

Finally, the currently described hard allocation of TPs into disjoint referential contexts is likely to be unrealistic unless the referential context is taken to correspond strictly to the attended visual scene. However, it is a suitable simplification for demonstrating the power of referential cues in the present experiments with closed sets of words and referents and with well-controlled experimental manipulations. Learners are known to encode several aspects of word learning contexts that influence later performance (see, e.g., the discussion in Kucker et al., 2015). Therefore, in a general case, the referential context should be taken more broadly to include the aspects of the situation that are actively represented by the learner's cognitive system, whether it includes contents of the working memory, emotional state, spatial information, or the entire conceptual network primed by the situation and previously recognized words. This broad view is also more compatible with the fact that words of a language convey much richer meanings that can be possibly explained by simple word-to-object mappings. However, such "secondary" associations may require accumulation of cross-situational evidence across long periods of time, not unlike adults who gradually learn connotations of specific word choices in a foreign language. From the modeling point of view, a broader view of referential context would simply mean that each word becomes associated with a distribution of consistent meanings instead of just one. Upon hearing a word, all these different meanings would become activated to a varying degree based on the prior probabilities of different interpretations in the present communicative context.

6.3 Model predictions

An important test for the plausibility of any model of cognition is not just to be able to fit its behavior to the existing data, but to also provide predictions for human behavior in tasks for

which conclusive data does not yet exist. In this regard, the present approach provides a number of hypotheses:

1) Children's acquisition of word forms is aided by coherent contextual cues that co-occur with speech, a result that is already shown with adults (Cunillera et al., 2010; Thiessen, 2010; Glicksohn & Cohen, 2013). By incorporating consistent and attended cross-modal cues to a standard behavioral speech segmentation task similar to Saffran et al. (1996a), performance on the task should be better than in the case of familiarization without systematic cross-situational cues or when the visual cues are incoherent with respect to the regularities available in the speech stream. In order to probe this with young children, more engaging learning paradigms with a shared space of joint-attention between the learner and the caregiver might be required (e.g., novel object naming during play and use of full sentences). At least, the temporal properties of audiovisual integration should be taken into account when designing experiments with artificial languages (see experiment 2). Organizing auditory and visual stimuli into utterance-sized chunks could also improve learning outcomes by providing the learner with clearly bounded communicative situations and leaving processing time for interpretation of the multimodal scene.

2) Contextual facilitation of segmentation should also occur when there is referential ambiguity in the visual cues similar to the uncertainty present in XSL studies. In contrast, all the existing studies have either used deterministic visual cues with only one picture shown together with each word (Cunillera et al., 2010; Thiessen, 2010; Glicksohn & Cohen, 2013), or the learners have not been tested in their segmentation performance in a

non-visual control condition when they have shown successful segmentation and meaning acquisition with visual cues (e.g., Shukla et al. 2011; Yurovsky et al., 2012).

3) If segments and their meanings are acquired at the same time, learning should not be significantly slower in XSL studies even if the auditory words are not presented with silent gaps between them, as is done in the existing studies (e.g., Yu & Smith, 2007).

4) The importance of referential cues in word segmentation should increase with increasing variability in the acoustic domain (in children) and/or with decreasing reliability of the transition probability structure (children and adults). If the word tokens have very limited acoustic variability and they occur numerous times over a short time period (c.f., Saffran et al., 1996), clustering them into word-like units is not a problem for an efficient statistical learner operating purely on auditory information. However, if individual tokens have differing surface forms (e.g., spoken by different talkers), a learner without yet mastering the phonological categories of the language should benefit from visual referents as the common denominator between these tokens. Similarly, the model predicts that words of an artificial language do not necessarily have to have high TPs in order to be learned as long as they occur in consistent referential contexts.

5) As the entire model is based on the assumption of statistical learning as a general mechanism that operates within and across sensory modalities and internal states of the learner, the contextual facilitation effect in pattern learning from sensory input should not be limited to the speech domain. If symmetrical processing and learning is assumed

across perceptual domains, the model predicts that the words themselves start to act as "contexts" to other percepts as soon as the words become represented at a sufficiently invariant level to be recognized across situations. In this case, entities with varying surface properties but occurring in the context of a specific word become associated to this word label. This can lead to fine tuning of already established categorical boundaries (e.g., color naming), or to the emergence of entirely new conceptual categories consisting of entities that do not have an apparent shared surface form (e.g., "*furniture*"). This idea is consistent with the data on superordinate category words such as "*food*" or "*dish*" that become understood much later than words for the more concrete entities within these categories, such as "*apple*", "*banana*" or "*cup*" (MCDI data; Fenson et al., 1993), even if the words referring to the superordinate categories are also likely to be heard from early on.

6.4 Conclusion

Word learning is the most important first step in language acquisition, as words are the basic building blocks of any language that act as meaningful conceptual elements on their own. One hypothesis is that proto-lexical representations are an important intermediate stage in language development since the acquisition of the phonological system or grammar cannot take place without an intermediate and functionally operating representational system of the language (Werker & Curtin, 2005). In order to learn the first words of their native language, infants may use a variety of strategies to parse the continuous speech into word-like segments and then to map these segments to their referential meanings through contextual grounding.

In the present paper, we argued that infants might as well directly make use of their experience with referential contextual information in order to segment speech into units that are guaranteed to have referential significance. We have provided a theoretical description of how this process may occur, how it connects to the referential value of the learned vocabulary, and tested it in a series of word-learning simulations using real speech as the stimulus. The extent to which the model actually corresponds to the reality remains to be seen as more data on human learning is accumulated. Meanwhile, the model provides a viewpoint to early language acquisition with a minimal number of a priori linguistic assumptions and with the main focus on practical consequences of parsing speech input in a specific way.

Finally, by definition, the proto-lexical representation of spoken language is not the final one during normal language development. It cannot account for generative aspects of language since there is no morphological or grammatical structure beyond what is observed in the surface form of the data, nor does it account for the interaction between speech perception and production. Whether linguistic components such as phonology, morphology, or grammar can be acquired using similar statistical principles, whether they require statistical inference with unobservable variables and model constraints, or whether they require specialized rule-like knowledge is still open for debate. The current work only provides a simple strategy for solving the first stages in the complex process of language learning and meaningful language use.

Acknowledgements

This research was funded by the 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, Dan Yurovsky, Gabe Doyle, Lauri Juvela, Unto Laine, Sofoklis Kakouros, and the four anonymous reviewers for their insightful comments on the

manuscript. Many of the ideas in the current paper are based on the earlier work performed in collaboration with former members of the ACORNS consortium, including Roger K. Moore, Hugo Van hamme, Lou Boves, Louis ten Bosch, Guy Aimetti, Joris Driesen, Gustav Henter, and many others.

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, 62, 311–331.

Aimetti, G. (2009). Modelling early language acquisition skills: Towards a general statistical learning mechanism. *Proceedings of the Student Research Workshop at the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 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.

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative Mechanisms in Early Word Learning. *Cognitive Science*, 35, 1105–1138.

Apfelbaum, K. S. (2013). *Real time competition processes in word learning.* PhD (Doctor of Philosophy) thesis, University of Iowa.

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, 64: Suppl. 2, 86–105.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic action via statistical structure. *Cognition*, 106, 1382–1407.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, 29, 832–843.

Bauer, P. J., Dow, G. A., & Hertsgaard, L. A. (1995). Effects of prototypicality on categorization in 1- to 2-year-olds: Getting down to basic. *Cognitive Development*, 10, 43–68.

Behl-Chadha, G. (1996). Basic-level and superordinate-like categorical representations in early infancy. *Cognition*, 60, 105–141.

Berchtold, A. & Raftery, A. E. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17, 328–356.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109, 3253–3258.

Bergelson, E., & Swingley, D. (2013). Young toddlers' word comprehension is flexible and efficient. *PLoS One*, 8, e73359, doi:10.1371/journal.pone.0073359.

Blythe, R. A., Smith, A. D. M., & Smith, K. (2014). Word learning under infinite uncertainty. arXiv:1412.2487 [physics.soc-ph].

Bortfeld, H., & Morgan, J. L. (2010). Is early word-form processing stress-full? How natural variability supports recognition. *Cognitive Psychology*, 60, 241–266.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, 61, 1–38.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34, 71–105.

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

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coen, M. H. (2006). Self-supervised acquisition of vowels in American English. *Proceedings of the Twenty First National Conference on Artificial Intelligence*, Boston, MA.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology*, 31, 24–39.

Cunillera, T., Laine, M., Càmara, E., & Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language*, 63, 295–305.

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.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28, 357–366.

de Marcken, C. (1995). *The unsupervised acquisition of a lexicon from continuous speech*. AI Memo No. 1558, Massachusetts Institute of Technology, MA.

Driesen, J., & Van hamme, H. (2011). Modelling vocabulary acquisition, adaptation and generalization in infants using adaptive Bayesian PLSA. *Neurocomputing*, 74, 1874–1882.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vreken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-

Commercial Purposes. *Proceedings of the Fourth International Conference on Spoken Language Processing* (*ICSLP'96*), Philadelphia, PA, pp. 1393–1396.

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, 65, 903–917.

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics*, Jeju, Korea, pp. 184–193.

Elsner, M., Goldwater, S., Feldman, N., & Wood, F. (2013). A joint learning model of word segmentation, lexical acquisition, and phonetic variability. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, pp. 42–54.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34, 1017–1063.

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Austin, Texas, pp. 2208–2213.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J., (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, 127, 427–438.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J.S. (1993). *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. Baltimore : Paul H. Brokes Publishing Co.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12, 499–504.

Fourtassi, A. & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Baltimore, Maryland, pp. 191–200.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems, Volume 20* (pp. 1212–1222). Cambridge: MA: MIT Press.

Frank, M. C., Mansinghka, V., Gibson, E., & Tenenbaum, J. B. (2007). Word segmentation as word learning: integrating meaning learning with distributional cues to segmentation. *Proceedings of the 31st Annual Boston University Conference on Language Development*, Boston, MA, pp. 218–229.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20, 578–585.

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

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*, 9, 1–24.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, 33, 260–272.

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.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112, 21–54.

Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, 20, 229–252.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, 18, 254–260.

Hay, J. F., Pelucchi, B., Graf Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, 63, 93–106.

Hollich, G. J., Hirsh-Pasek, K, Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L, … Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65, 1–123.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1570–1582.

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.

Johnson, S. P. (2001). Visual development in human infants: Binding features, surfaces, and objects. *Visual Cognition*, 8, 565–578.

Jusczyk, P. W., & Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1–23.

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

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin and Review*, 19, 317–324.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83, B35–B42.

Kopp, F. (2014). Audiovisual temporal fusion in 6-month-old infants. *Developmental Cognitive Neuroscience*, 9, 56–67.

Kucker, S. G., McMurray, B., & Samuelson, L. K. (2015). Slowing down fast mapping: Redefining the dynamics of word learning. *Child Development Perspectives*, doi: 10.1111/cdep.12110.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255, 606–608.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100, 9096–9101.

Kuhl, P. K., Conboy, B. T., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions B of the Royal Society*, 363, 979–1000.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127, 439–453.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299–321.

Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 56–67.

Li, W. (1990). Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60, 823–837.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, pp. 281–297, University of California Press, CA.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, 8, 291–318.

Marechal, D., & Quinn, P. C. (2001). Categorization in Infancy. *Trends in Cognitive Sciences*, 5, 443–450.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science,* 14, 57–77.

Marr, D. (1982). *Vision: A computational approach*. San Francisco, Freeman & Co.

Mattys S. L., Jusczyk P. W., Luce P. A., & Morgan J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38, 465–494.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.

McInnes, F. R., & Goldwater, S. J. (2011). Unsupervised extraction of recurring words from infant-directed speech. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 2006–2011, Boston, Massachusetts.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12, 369–378.

McMurray, B., Host, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, 119, 831–877.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108, 9014–9019.

Mirman, D., Magnuson, J. S., Graf Estes, K., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, 108, 271–280.

Moore, R. K. (2013). Spoken language processing: Where do we go from here? In R. Trappl (Ed.), *Your Virtual Butler*, Lecture Notes in Artificial Intelligence, (Vol. 7407, pp. 111–125). Heidelberg: Springer.

Moore, R. K. (2014). Spoken language processing: Time to look outside? *Proceedings of the 2nd International Conference on Statistical Language and Speech Processing (SLSP 2014)*, pp. 21–36, Grenoble, France.

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

Oakes, L. M., & Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive familiarization. *Infancy*, 7, 85–98.

Park, A., & Glass, J. R. (2005). Towards unsupervised pattern discovery in speech. *Proceedings of the 2005 IEEE Workshop Automatic Speech Recognition and Understanding*, pp. 53–58, Cancún, Mexico.

Park, A., & Glass, J. R. (2006). Unsupervised word acquisition from speech using pattern discovery. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, pp. 409–412, Toulouse, France.

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

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39, 246–263.

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

Plomp, R., & Bouman, M. A. (1959). Relation between hearing threshold and duration for tone pulses. *Journal of the Acoustical Society of America*, 31, 749–758.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, 25, 143–170.

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Rabiner, L. R., & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3, 4–16.

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society*, B47, 528–539.

Rasilo, H., Räsänen, O., & Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, 55, 909–931.

Rasilo, H. & Räsänen, O. (2015). Computational evidence for effects of memory decay, familiarity preference and mutual exclusivity in cross-situational learning. *Proc. 37th Annual Conference of the Cognitive Science Society*, Pasadena, California, 2015.

Rost, G. & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12, 339–349.

Rost, G., & McMurray, B. (2010). Finding the signal by adding noise: the role of noncontrastive phonetic variability in early word learning. *Infancy*, 15, 608–635.

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

Roy, B. C., Frank, M. C., & Roy, D. (2012). Relating activity contexts to early word learning in dense longitudinal data. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Sapporo, Japan.

Räsänen, O., Laine, U. K., & Altosaar, T. (2008). Computational language acquisition by statistical bottom-up processing. *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pp. 1980–1983, Brisbane, Australia.

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

Räsänen, O., & Laine, U. K. (2013). Time-frequency integration characteristics of hearing are optimized for perception of speech-like acoustic patterns. *Journal of the Acoustical Society of America*, 134, 407–419.

Räsänen, O., & Rasilo, H. (2012). Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pp. 887–892, Sapporo, Japan.

Räsänen, O., (2012). Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, 54, 975–997.

Räsänen, O., (2013). *Studies on unsupervised and weakly supervised methods in computational modeling of early language acquisition* (Doctoral dissertation). Espoo, Finland: Aalto University Publication Series.

Räsänen, O. (2014). Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Quebec, Canada, 2014

Räsänen, O., Doyle, G., & Frank, M. C. (2015). Unsupervised discovery of words from speech using automatic segmentation into syllable-like units. *Proc. the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany.

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606–621.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.

Salvi, G., Montesano, L., Bernadino, A., & Santos-Victor J. (2012). Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, 42, 660–671.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, 54, 1–32.

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.

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4, 157–178.

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

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558–1568.

Smith, K., Smith, A. D. M., Blythe, R. A., & Vogt, P. (2006). Cross-situational learning: A mathematical approach. In P. Vogt et al. (Eds.): *Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, Lecture Notes in Artificial Intelligence, 4211, 31–44.

Smith, K., Smith, A. D. M, & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35, 480–498.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14, 29–56.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception

than in word-learning tasks. *Nature*, 388, 381–382.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, 126, 395–411.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76, 147–166.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86–132.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 3617–3632.

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. & 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.

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

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10, 172–175.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 57, 1454–1463.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language*, 4, 197–211.

Vallabha, G. K., McLelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104, 13273–13278.

van den Bos, E., Christiansen, M. H., & Misyak, J. B. (2012). Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *Journal of Memory and Language*, 67, 507–520.

Van hamme, H. (2008). HAC-models: A novel approach to continuous speech recognition. *Proceedings of the 9th Annual Conference of the International Speech Communication Association,* pp. 2554–2557, Brisbane, Australia.

Versteegh, M., ten Bosch, L., & Boves, L. (2010). Active word learning under uncertain input conditions. *Proceedings of the 11th Annual Conference of the International Speech Communication Association,* pp. 2930–2933, Makuhari, Japan.

Versteegh, M., ten Bosch, L., & Boves, L. (2011). Modeling novelty preference in word learning. *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pp. 761–764, Florence, Italy.

Versteegh, M.,Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The Zero Resource Speech Challenge 2015. *Proc. 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition,* 107, 729–742.

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, 45, 1611–1617.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence from perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, 34, 1289–1309.

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

Wojcik, E. H., & Saffran, J. R. (2013). The ontogeny of lexical networks: Toddlers encode the relationship among referents when learning novel words. *Psychological Science*, 24, 1898–1905.

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar sounding words. *Developmental Science*, 12, 412–418.

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. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414–420.

Yu, C., & Smith, L. (2011). What you learn is what you see: using eye movements to study infant cross-situational learning. *Developmental Science*, 14, 165–180.

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

Yu, C., & Smith, L. B. (2012a). Embodied attention and word learning by toddlers. *Cognition*, 125, 244–262.

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

Yurovsky, D, Yu, C., & Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: scaffolding from child-directed speech. *Frontiers in Psychology*, 3, 1–9.

Yurovsky, D., Smith, L. B., & Yu, C., (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, 16, 959–966.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, 37, 891–921.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin and Review*, 21, 1–22.

Appendix A: Bottom-up model of word segmentation

A bottom-up approach to word segmentation from continuous speech starts with a probability distribution $P(X)$ over all speech input $X \in \mathsf{R}^d$ (uni- or multivariate time-series). The goal of the learner is to decompose this distribution to a non-overlapping sequence (parse) W of words $w \in L$ from an unknown lexicon L given the X, i.e., $X \rightarrow W = [w_1, w_2, \ldots, w_N]$, where the subscripts denote the relative position of the word in the word sequence and N is the total number of word tokens in the input X. The probability of each word is defined as the relative frequency of occurrence of the word in comparison to the other words in the lexicon, and therefore parsing of the speech input into a most likely sequence of words will simultaneously solve the lexicon. Using Bayes' formula, the joint probability $P(W, X)$ of a lexical parse and speech input can be then factorized as

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)} = \frac{P(X \mid w_1, w_2, ..., w_N)P(w_1, w_2, ..., w_N)}{P(X)} \quad \text{(A1)}$$

In order to solve the most likely parse, $P(X)$ in Eq. (A1) can be neglected, as it will not affect the relative likelihood of different possible parses. This leads to

$$P(W \mid X) \propto P(X \mid w_1, w_2, ..., w_N)P(w_1, w_2, ..., w_N) \quad \text{(A2)}$$

Eq. (A2) shows that the probability of the parse of a given input is relative to the probability of observing the speech signal, given the sequence of words, times the probability of observing the words themselves. Assuming that the words are mutually independent (no grammar) and non-overlapping in time, Eq. (A2) can be formulated as

$$P(W \mid X) \propto \prod_{n=1}^{N} P(X(t_n), X(t_n + 1), ..., X(t_n + d_n) \mid w_n)P(w_n) \quad \text{(A3)}$$

where t_n and d_n are the onset time and the duration of the nth word, respectively, and $P(w_n)$ is the probability of word w_n occurring in the data. In order to compute this probability, an acoustic model $P(X \mid w, \theta_w)$ is needed that connects each latent word w in the lexicon to the corresponding acoustic features X using θ_w as the model parameters. This expands the model to

$$P(W \mid X) \propto \prod_{n=1}^{N} P(X(t_n), X(t_n+1), ..., X(t_n+d_n) \mid w_n, \theta_{w_n}) P(w_n) \quad \text{(A4)}$$

with the simplifying assumption that the acoustic model parameters of each word are independent of the other words in the lexicon. The segmentation problem is then to find the most likely parse W without knowing the number of unique words in the lexicon, the number of word tokens in the input, or how each word is realized in the acoustic signal. Given some choice for an acoustic model, e.g., a hidden-Markov model (Rabiner & Juang, 1986), the best parse W^* and the corresponding acoustics-to-word mapping θ^* is the one that maximizes

$$W^*, \theta^* = \arg\max_{W,\theta} \{P(W \mid X)\} \quad \text{(A5)}$$

where $P(W|X)$ is computed using Eq. (A4). However, the problem is analytically intractable since the W and θ are mutually dependent stochastic variables. A solution attempt requires numerical inference from some initial parameters and model-based constraints on structure (e.g., Dirichlet process for lexicon, c.f., Feldman et al., 2009, combined with an HMM for the acoustic modeling of each lexeme) in order to find a locally optimal solution (see, e.g., Elsner et al., 2013, for solving a similar problem in the joint estimation of lexicon and phonetic pronunciation variants of words). Non-optimal solutions are also available by using different heuristic methods and/or assuming linguistic representations of the speech input (e.g., de Marcken, 1995; Brent & Cartwright, 1996; Brent, 1999; Goldwater et al., 2009; Pearl et al., 2010; Adriaans & Kager,

2010; Frank et al., 2010; Räsänen, 2011). In all cases, the main issue is that there is no principled way to define a learning criterion that would ensure that the obtained lexicon is *meaningful*.

Appendix B: Step-by-step description of the TP-based algorithm

The TP-based implementation of the joint model consists of two basic mechanisms: memory update ("learning") and memory recall ("word recognition"). The basic idea is to model the temporal evolution of a discrete input sequence X = [a_1, a_2, …, a_T] (e.g., a quantized spoken utterance) in the context of co-occurring referents *c*. The sequential structure is modeled using a mixture of bi-grams at different temporal lags *k* (see Räsänen & Laine, 2012).

The core of the algorithm is a set of three-dimensional frequency arrays $\mathbf{F}_k(a_t, a_{t-k}, c)$, a corresponding set of bi-gram probability arrays $\mathbf{B}_k(a_t, a_{t-k}, c)$, and a set of referent probability arrays $\mathbf{P}_k(a_t, a_{t-k}, c)$, each of size $|A|$ x $|A|$ x C, and with a separate array for each temporal lag k. During the learning stage, the co-occurrence frequencies of acoustic bi-grams and the concurrent referents are stored into the frequency arrays **F**. The raw frequencies are then first normalized to context-specific bi-gram probabilities **B** at different lags k, and then to conditional probabilities of referents **P** given the bi-grams. The only hyperparameter in the model is the set of lags $k \in K$ at which the bi-grams are estimated.

Learning

1) Given an acoustic sequence $X = [a_1, a_2, \ldots, a_T]$, $a_i \in A$, and an unordered set of concurrent referents $\mathbf{c} = \{c_1, c_2, \ldots, c_N\}$, $c_j \in C$, update the co-occurrence frequencies

$$\mathbf{F}_{k,t}(a_t, a_{t-k}, c) \leftarrow \mathbf{F}_{k,t-1}(a_t, a_{t-k}, c) + 1 \qquad \text{(B1)}$$

for all referents $c \in \mathbf{c}$, lags $k \in K$, and times $t \in [1+k, T]$.

2) Normalize frequencies in **F** to lag-specific transition probabilities (TPs) in **B:**

$$\mathbf{B}_{k,t}(a_i, a_j, c) = \frac{\mathbf{F}_{k,t}(a_i, a_j, c)}{\sum_{a_i \in A} \mathbf{F}_{k,t}(a_i, a_j, c)} \quad \text{(B2)}$$

for all a_i, a_j, c and k.

3) Compute conditional probabilities of contexts c, given the bi-grams:

$$\mathbf{P}_{k,t}(a_t, a_{t-k}, c) = \frac{\mathbf{B}_{k,t}(a_t, a_{t-k}, c)}{\sum_c \mathbf{B}_{k,t}(a_t, a_{t-k}, c)} \quad \text{(B3)}$$

for all a_i, a_j, c and k.

4) Repeat steps 1–3 for each new speech input X and the concurrent set of referents.

Estimation of stimulus familiarity (experiments 1–2)

1) Given an acoustic sequence $X = [a_1, a_2, \ldots, a_T]$ (e.g., a word) and an already learned array **B** of bi-gram TPs, compute the sum of bi-gram probabilities across all lags k to obtain the instantaneous *activation* of the memory trace for each possible c at each moment of time t:

$$A_{\text{fam}}(X, t \mid c) = \sum_k \mathbf{B}_k(a_t, a_{t-k}, c) \quad \text{(B4)}$$

2) Compute the total activation for each c across all t:

$$A'_{\text{fam}}(X \mid c) = \sum_{t=1}^{T} A_{\text{fam}}(X, t \mid c) \quad \text{(B5)}$$

3) Select the largest activation across all possible c as the final familiarity measure:

$$\text{FAM}(X) = \max\{A'_{\text{fam}}(X \mid c) \mid \forall c\} \quad \text{(B6)}$$

Recognition of the most likely referent (word decoding)

1) Given an acoustic sequence $X = [a_1, a_2, \ldots, a_T]$ (e.g., an utterance) and an already learned array **P** of bi-gram-specific referent probabilities, compute the sum of referent

probabilities across all lags *k* to obtain the instantaneous *activation* of the memory trace for each possible *c* at each moment of time *t* (P(*c*) is assumed to be uniform):

$$A(c,t \mid X) = \sum_k \mathbf{P}_k(a_t, a_{t-k}, c) \quad \text{(B7)}$$

2) Compute the total activation of each *c* across the entire signal (forced-choice tasks):

$$A'(c \mid X) = \sum_{t=1}^{T} A(c,t \mid X) \quad \text{(B8)}$$

OR compute the cumulative activation of each *c* at time *t* in a sliding window of length *W* (experiments 5 and 6, and all experiments with attention-constrained learning)

$$A'(c,t \mid X) = \sum_{x=t-W+1}^{t} A(c,x \mid X) \quad \text{(B9)}$$

3) Select the referent with the largest activation as the referent hypothesis for the trial:

$$c^* = \arg_c \max\{A'(c \mid X)\} \quad \text{(B10)}$$

or, in case of continuous word recognition, get the most likely referent at time t as

$$c^*(t) = \arg_c \max\{A'(c,t \mid X)\} \quad \text{(B11)}$$

After decoding a sequence of the most likely referents c*(*t*) across time, a word segment boundary can be defined for each *t* that satisfies c*(*t*-1) ≠ c*(*t*), i.e., there is a change in the winning referent.

Attention-constrained learning (experiments 3–5)

In the attention constrained-learning used in experiments 3–5, the learner first recognizes the most likely referent *c**(*t*) for all *t* in the speech input using Eqs. (B7), (B9), and (B11). Then the memory is updated according to Eqs. (B1)-(B3) using only the winning referent for each *t* in place of the full set of referents present in that training trial.

Appendix C: Details of the Caregiver Y2 UK corpus

Table C1: A list of referential keywords used in experiments 5 and 6.

'airplane'	'car'	'duck'	'lion'	'small'
'animal'	'cat'	'eagle'	'looks'	'smiling'
'apple'	'clean'	'edible'	'man'	'square'
'baby'	'cookie'	'fish'	'mummy'	'to take'
'ball'	'cow'	'frog'	'Porsche'	'telephone'
'banana'	'crying'	'give'	'red'	'toy'
'big'	'daddy'	'happy'	'robin'	'tree'
'bird'	'dirty'	'to have'	'round'	'truck'
'blue'	'dog'	'horse'	'sad'	'woman'
'bottle'	'doll'	'like'	'to see'	'yellow'

Sentence structures of the corpus are shown below (only the sentences with more than one keyword were included in the experiments). Each <word> was randomly sampled from the corresponding category of keywords during the corpus generation.

Structure	**Examples (keywords emphasized)**
Do you <verb> <noun>?	*Do you **like** an **animal**?*
Where is <adj> <noun>?	*Where is the **red car**?*
The <adj> <noun> <verb> <noun>?	*The **crying woman likes** the **animal**.*
She/he <verb> <adj> <noun>.	*He **has** the **dirty robin**.*
Where is <adj> <noun>?	*Where is the **happy airplane**?*
Where is <adj> <adj> noun> and <noun>?	*Where is the **happy big dog** and a **ball**?*
Here/There is <adj> <noun> and <noun>.	*Here is a **small apple** and a **toy**.*
Here/There is <adj> <adj> <noun> and <noun>.	*There is a **small clean Porsche** and a **robin**.*
<noun> <verb> <adj> <noun>.	***Mummy gives** the **clean fish**.*