Fundamenta Informaticae 90 (2009) 229–249
DOI 10.3233/FI-2009-0016
IOS Press

A Computational Model of Language Acquisition: the Emergence of Words

Louis ten Bosch[*†]
Centre for Language and Speech Technology
Radboud University
P.O. Box 9103, 6500 HD Nijmegen, the Netherlands
L.tenBosch@let.ru.nl

Hugo Van Hamme
ESAT
Katholieke Universiteit Leuven
Leuven, Belgium

Lou Boves
Centre for Language and Speech Technology
Radboud University, Nijmegen
the Netherlands

Roger K. Moore
SPandH
Dept. of Computer Science
Sheffield University, U.K.

Abstract. In this paper, we discuss a computational model that is able to detect and build word-like representations on the basis of sensory input. The model is designed and tested with a further aim to investigate how infants may learn to communicate by means of spoken language. The computational model makes use of a memory, a perception module, and the concept of 'learning drive'. Learning takes place within a communicative loop between a 'caregiver' and the 'learner'.
Experiments carried out on three European languages with different genetic background (Finnish, Swedish, and Dutch) show that a robust word representation can be learned in using less than 100 acoustic tokens (examples) of that word. The model is inspired by the memory structure that is assumed functional for human cognitive processing.

Keywords: I5 pattern recognition, J5 arts and humanities, F5 theory of computation, miscellaneous (cognitive modelling)

[*]Centre for Language and Speech Processing, Radboud University, P.O. Box 9103, 6500 HD Nijmegen, the Netherlands
[†]This research was funded in part by the European Commission, under contract number FP6-034362, in the ACORNS project www.acorns-project.org

1. Introduction

Language processing is a complex cognitive capability of humans. Speech production, speech perception, and speech understanding in the communication between a speaker and a listener seem to be performed effortlessly, but on closer inspection appear to involve as yet ill-understood top-down and bottom-up processes. Over the last few decades research in different disciplines such as psycholinguistics, linguistics, biology, neuroscience, psychology, phonetics and speech technology has resulted in the design of theories and models that account for *parts* of the chain that links the speaker's intentions and the listener's comprehension. However, we are still far from a coherent and comprehensive model or theory of speech communication.

The ability to understand speech is difficult to envisage without the capability to decode continuous speech signals into a sequence of meaningful entities. These entities may be words or, more generally, word-like elements. Infants learn to discover word-like elements in speech without any prior knowledge about lexical identities, while information about the segmentation of the continuous speech signal in such word-like units is often rather implicit. The capability to discover words (and word-like units) in continuous speech without using a pre-specified lexicon is one of the intriguing aspects of language acquisition.

The question how children learn language is closely related to the question what the properties of natural languages are that make them learnable in the first place. A closely related challenge is to develop formal descriptions of complex cognitive processes such as learning and using language. The first question is addressed by Kaplan et al. [16], who discuss various computational approaches towards the modeling of language learning. They distinguish five different 'stances', i.e. possible perspectives, covering the generative viewpoint, a statistically-based approach, a social/embodiment based approach, a child-based developmental approach, and the 'language evolution' viewpoint. They strongly advocate the development of computational models since "Computational models [...] help refine our intuitions, suggest novel lines of empirical investigation with humans, and build concepts that shed a reinvigorating light on childrens fantastic learning capacities." [16]. The issue of modeling cognitive processes in formal terms is addressed in [40]. In this paper, the authors develop a formal scheme in the theoretical framework of 'Cognitive Informatics' (CI), in which the processes involved in learning (such as perception, hypothesizing, reinforcement, storage, forgetting, retrieval) are described using so-called process algebras. The CI framework allows for a specific description of hierarchy and abstraction (called 'Layered Reference Model') and of object-attribute relationships in knowledge representation (see also [39]).

Both computational models and formal models are necessary for developing a theory of language acquisition and to suggest experiments to investigate details of such a theory. Especially the availability of a computational model of language acquisition, focussing on the very first steps in acquiring a set of sound-meaning correspondences, is of considerable interest for several reasons. The literature about language acquisition contains numerous observations that seem to be difficult to reconcile. For example, newborns are able to discriminate phonetic properties of all natural languages, by when they acquire their mother tongue, they loose the capability of distinguishing phonetic contrasts that do not play a role in the native language (cf. [41] for additional phenomena in language acquisition that seem to be counter-intuitive). A comprehensive computational model that can account for seemingly contradictory phenomena would be a giant step towards understanding language acquisition.

Of course, we must ask ourselves to what extent computational models must account for the details of a physical or biological process. We understand bird flight in part from what we learned in building airplanes that fly without flapping their wings. Yet, we argue that the cognitive and ecological plausibility of a computational model of language acquisition is extremely important, since attempts at understanding cognition on the basis of computational models that do not adhere to basic knowledge about the underlying neuro-anatomical and physiological systems are not likely to advance our understanding significantly.

Following Marr [22], three levels of modeling can be distinguished. The highest Marrian level (called 'computational' level) describes the model in terms of a (usually limited) number of states, which are connected by (usually a limited number of) processes. This level addresses the conceptual way of how the problem is to be solved. The second level, the 'algorithmic' level, deals with the strategies that can be followed to build representations of the states and perform the processes. Details on this level are to be specified by the theory underlying the model. This mid level does not need to contain the implementation details that are not part of the theory itself. Such details are dealt with on the lowest Marrian level (the 'implementation level'). Usually, these implementation details are not specified by the underlying theory, but they must nevertheless be chosen carefully for the sake of plausibility of the resulting model. Existing models of speech processing are mainly on the computational level, leaving many potentially crucial choices as yet unspecified [29].

Automatic Speech Recognition (ASR) and Computer Vision can be considered as computational models of the corresponding cognitive skills. Both deal with the process of mapping continuous signals (speech, images) onto discrete representations (a sequence of words, or a label for an image). At the Marrian computational level ASR can serve as a computational model of human speech recognition (HSR). Therefore, it is not surprising that attempts have been made to reuse computational techniques developed in ASR for modeling Human Speech Recognition (e.g. [35], [3]). Both the HSR and ASR-based approach of decoding speech have important conceptual aspects in common, such as the dynamic evolution of hypotheses over time, and the competition of words during the word search. However, these commonalities are on Marr's computational level. Therefore, there may well be fundamental differences between human and automatic speech recognition processes at the algorithmic (and certainly at the implementation) level. ASR systems are designed as statistical pattern recognizers, and in all extant systems the patterns are based on a linguistic description of speech signals, rather than on a description of the HSR process proper. As a consequence, a straightforward mapping between HSR and currently available ASR approaches is likely to fall in the trap that is known as the error of frame-of-reference in (embodied) artificial intelligence [32].

The statistical approach to ASR has met with limited success. Advances in hardware, algorithms and and an ever increasing amount of training data have enabled the implementation of quite useful large vocabulary, continuous speech recognition (LVCSR) systems and a range of voice-enabled dialogue systems. However, all conventional ASR systems perform substantially worse than humans [19, 37, 42]. On the basis of an analysis of the development of the ASR performance over the last decades, Moore [26] argues that all attempts to close the performance gap between ASR and HSR by adding ever more training data will be futile. Nowadays there is general agreement that fundamentally new training and matching paradigms must be explored, and a potential fruitful direction is informed by knowledge about human speech processing. This amounts to narrowing (and perhaps eventually closing) the gap between HSR and ASR at Marr's algorithmic level.

Recently, cognitive science has witnessed the development of new directions in modeling language acquisition and speech understanding. Roy and Pentland [33] focused on machine learning of words; Werker and Curtis [41] presented a comprehensive model of human language acquisition, while Maloof and Michalski [21] focus on incremental learning. However, these approaches still suffer from the well-known symbol-grounding problem: conventional pattern recognizers are trained to discriminate predefined patterns that are invariably based on a (mostly symbolic, often human-crafted) meta-description of the phenomena under investigation. For ASR this means that systems are trained to recognize 'words' that are represented in the form of a sequence of discrete sounds. However, while such a linear representation of words may be very convenient for the purpose of linguistic description, it does not properly reflect the fact that speech production is fundamentally a continuous process [30]. This essentially means that conventional ASR systems make an error of frame-of-reference [32]. In contrast, for biological agents 'patterns' occur in the sensory input as emergent properties [14] that are learned because of the need to associate (inherently variable) sensory inputs to meaningful objects and behaviors in the environment. Because the sensory signals corresponding to ecologically relevant entities are so variable, it is essential that biological agents are able to *adapt* and *generalize* known patterns quickly and effortlessly to recognize new variants that were not previously encountered.

In this paper, we propose a novel computational model for language acquisition. The model has similarities with the Cross-channel Early Lexical Learning (CELL) model [33]. CELL is trained with audio recordings of play sessions between care givers and seven-to-eleven month-old infants. During these sessions, caregivers and infants played with toys from seven categories (balls, shoes, keys, cars, trucks, dogs, and horses). Pictures taken of each toy from various angles were used for building a visual model of each toy. CELL learned to discover words by listening to the speech, while simultaneously looking at the visual representations.

The approach in the ACORNS project differs from CELL in an essential aspect. CELL makes the assumption that infants represent speech signals in the form of a symbolic representation of pre-defined phonemes. From the perspective of human language acquisition that basic assumption is clearly unwarranted. Our novel approach that is presented here addresses the issues of cognitive plausibility and frame-of-reference by avoiding the use of such pre-defined representations for decoding the information in the input signals. Instead, the representations in the model emerge from the multimodal stimuli that are presented to the model. This is in line with growing evidence that speech and language skills are *emergent* capabilities of a developing communicative system [14, 20] and that the way in which linguistic patterns are stored and used during language acquisition changes constantly as these patterns become more numerous and fine-grained, and as the methods needed for processing the patterns become correspondingly more complex [41].

In the ACORNS model learning takes place in an interaction loop between learner and 'caregiver'. To that end, caregiver and learner are implemented as two artificial agents that interact by exchanging messages. The caregiver provides multimodal stimuli to the learner, and the learner hypothesizes and reinforces internal representations during the interaction with the caregiver.

In the next section of this paper we introduce the main components of the ACORNS model. Section 3 discusses the cognitive architecture of the ACORNS model. In section 4 we report a number of experiments that show that the new model is indeed able to perform basic language acquisition processes. In the final section we put our results into perspective.

2. A computational model for word discovery

A comprehensive model of language acquisition and speech communication must integrate different aspects of signal processing and learning, viz. sensory front-end processing of presented stimuli, pattern discovery (hypothesizing, reinforcing and updating representations), memory access and organization (storage, retrieval), and interaction between caregiver and learner that is conducive to learning.

2.1. Front-end processing

The learner takes multimodal stimuli as input. In our model, these stimuli have an auditory and visual component. The auditory component processes actual acoustic recordings of spoken utterances. To that end the model uses an auditory front-end processor, i.e., a module that converts acoustic signals into a rich internal representation that can be used for learning new patterns and for recognizing known patterns. There is growing evidence that this internal representation must account for features of the input signals in multiple simultaneous temporal resolutions, with a lower limit in the order of 0.5 to 2 ms, and an upper limit of about 250 ms [11]. The representations must be suitable to characterize and process essentially all ecologically relevant sounds, from various non-speech sounds such as approaching footsteps to infant and adult directed speech. Since all these sounds can occur simultaneously, the representation must be suitable for the different sources to be processed independently [4].

For the time being the visual input modality is heavily simplified. Instead of implementing complex image processing (such as in the CELL model) we limit the visual input to a vector of (semantic) features that characterize the objects that are referred to in the speech utterances.

2.2. Pattern discovery

In conventional pattern recognition systems the patterns to be recognized, as well as the primitive elements from which complex pattern can be formed, are defined *a priori*. For example, in the conventional approach to speech recognition the patterns to be recognized are almost invariably words, while the primitives are related to the phonemes of the language (i.e., the speech sounds that distinguish between one word and another, such as *big* and *pig* in English). By doing so most ASR systems sidestep the task of detecting suitable basic units – because these are pre-defined by the developer. However, the representation of words as sequences of phonemes like beads on a string is far from adequate [30]. 'Episodic' theories of speech processing (e.g. MINERVA 2, [12]) deal with speech patterns in the form of episodes (low-level representational traces of speech), spanning syllables or complete words, if not multiword expressions [7, 9].

In normal speech, and even in infant-directed speech, words are not separated by silences. Rather, words blend and merge at their boundaries. This makes it necessary for a 'newborn' speech acquisition system to discover patterns in the continuous input stream that correspond to meaningful speech events. This word detection task is probably simplified at least to some extent by the fact that infant-directed speech often consists of several repetitions of the same words and phrases [36, 27], and that in stress-based languages the location of stressed syllables may help to hypothesize word starts in utterances.

Pattern discovery is closely related to the way information is stored in memory. Storing patterns in memory is only useful if there are efficient and effective techniques for retrieving them. All available behavioral data strongly suggest that memory for speech and language is organized in an associative

manner. Therefore, we apply computational methods that are able to extract structure from data and represent this structure for future use. The method applied in the most elaborate implementation of the model today is based on the decomposition of (large) matrices. Low-level sensory information, obtained from the multimodal stimuli, is transformed into a feature vector and stored in a large matrix (V). To that end utterances of different length are encoded in fixed-length vectors using the procedure explained in [10]. The relevant structure is then extracted by means of a factorization of the matrix V as a product of two much smaller matrices W and H such that the dissimilarity between the observed matrix V and the reconstructed matrix $W \cdot H$ is as small as possible (equation 1). The matrices W and H represent the basic acoustic units in speech and the (positive) weights with which units must be combined to reconstruct an arbitrary utterance. As the matrix V grows (or changes, if we implement forgetting) as a result of observing ever more multimodal stimuli, the matrices W and H are updated to reflect the results of the learning process.

$$\text{dissimilarity}(V, W \cdot H) \text{ is minimal} \tag{1}$$

The matrix V can be considered as an 'episodic' representation of the input stimuli.

The decomposition implied by equation 1 can be accomplished by means of the well-known as non-negative matrix factorization (NMF, [18, 13]). NMF is member of a family of computational approaches for structure mining that are based on matrix decomposition. It is a powerful tool for discovering structure in speech data [38]. NMF has similarities with Latent Semantic Analysis (e.g. [2]), which was developed for document retrieval, but also proposed as a model for representing the semantic content of documents.

In equation 1 several measures can be used for defining the dissimilarity between V and the matrix product $W \cdot H$. In the literature iterative optimization schemes are described for the L2-metric and Kullback-Leibler divergence ([13]). Since each dissimilarity measure may yield a different factorization, the choice of measure may affect cognitive plausibility. It was opted to use the (asymmetric) Kullback-Leibler divergence because its minimization can be interpreted as the mathematical translation of 'learning drive' (cf. section 3.3).

In the implementation of the model used in the experiments reported in section 4, the 'episodic' data matrix V is updated with each new multimodal stimulus, and the decomposition into W and H is done after blocks of N stimuli. Here, N is a parameter related to the storage capacity of the working memory and therefore to the cognitive plausibility of the algorithm.

By using the matrix concept, all internal representations are vectors or matrices, and *processes* receive a clear and explicit interpretation in terms of linear algebraic *operations*. In addition, the approach in the ACORNS model is related to the formal approach in [40]. The *Layered Reference Model* is a platform for dealing with abstraction as a process to elicit a 'subset of objects that share a common property from a given set of objects and to use the property to identify and distinguish the subset from the whole in order to facilitate reasoning.' (p. 265). From this viewpoint, the approach taken here (factorization of V) is an example of generalization: after the NMF step, W exactly contains those common parts of V that prove useful in the minimization of dissimilarity$(V, W \cdot H)$.

2.3. Memory organization and access

Cognitive theories of memory distinguish at least three types of memory: a sensory store in which all information is captured only for a very short time (in the order of seconds), a short-term memory (also called working memory) that holds representations of sensory inputs and serves as a processing system

that is able to compare new sensory inputs to previously learned patterns that are retrieved from a long-term memory ([1]). Ecologically plausible implementations of memory organization and access must be able to account for low-latency perception-action loops and dynamic construction of abstractions.

An important aspect of memory processes is how representations of novel patterns can form and be stored (see also [40]). In addition to storing a representation of the input signal, short-term memory must also be able to form and hold 'codes' derived from these representations. These codes are then used to activate patterns that are already present in the long-term memory. Activated patterns are compared to a more complete representation of the input signal in the working memory, by verifying the amount of activation of internal representations on the basis of the full signal input. If the match between the novel input and stored patterns is not good enough for the input to be recognized, the novel input becomes a candidate for storage in the long-term memory. Details of the memory architecture used in ACORNS are presented in section 3.1.

2.4. Interaction and communication

Speech and language acquisition happen as the result of purposeful interaction between an infant and its environment. Learning is not an isolated process but takes place in a communication loop. Therefore, it is essential to integrate all processing to realistically simulate speech acquisition driven by the intrinsic desire of the learner to communicate with its environment. In the beginning an infant interacts with only a limited number of 'biological' agents. This will result in learning patterns that are strongly biased towards the personal voice characteristics of the caregivers. However, the infant will increasingly be addressed by other persons, thereby forcing the representations to update and generalize. From the very first days of its life, successful communication will contribute to fulfilling the most basic needs of the infant, as specified by Maslow's hierarchy [23]. In the case of an infant acquiring speech and communication skills it is difficult to map Maslow's hierarchy of needs onto concrete behaviors, if only because Maslow's formulations address relatively abstract and high level needs. For an artificial agent it is even more difficult to map Maslow's hierarchy (see [34] for an attempt to adapt Maslow's hierarchy of needs to the situation where an individual is replaced by a team of software experts). We have implemented the learning drive by means of a target function that is to be optimized during learning. The target function is based on the incentive (by the learner) to optimally interpret the incoming stimuli in terms of what the learner already knows. That means that each stimulus that cannot be sufficiently interpreted may lead to

(a) a new internal representation, to improve or complete the explanation of the input, or

(b) the update of an existing internal representation.

The current algorithm explicitly simulates the intention to learn a continuously growing vocabulary in order to maximize the appreciation it receives from its environment. This is done by translating the appreciation from the caregiver into an attempt to responds to the caregiver with the correct interpretation of the stimulus. If the caregiver's feedback shows that the interpretation was indeed correct, the result may be a reinforcement or an update of the internal representations. The 'quality of interpretation' is mathematically expressed by the extent to which basis vectors in W can explain the observed data in V (equation 1). The concept of learning drive and its mathematical implementation is discussed in more detail in section 3.3.

3. Implementation Details of the ACORNS Model

In this section, we discuss a number of relevant implementation details of the computational model under development in ACORNS.

3.1. Memory architecture

The computational model of the learner that we have implemented is based on the architecture depicted in Figure 1. This architecture combines two aspects (a) it represents a structure of human memory that is based on experimental research and (b) the way in which the learner model is embedded in a communication loop with the caregiver. The memory structure comprises a sensory store, a short-term memory and a long-term memory and is directly based on recent psycholinguistic research on the organization of memory in connection to speech and language processing (e.g. [15], [1]). The sensory store, short-term/working memory and long term-memory form the entire memory, each with different decay times.

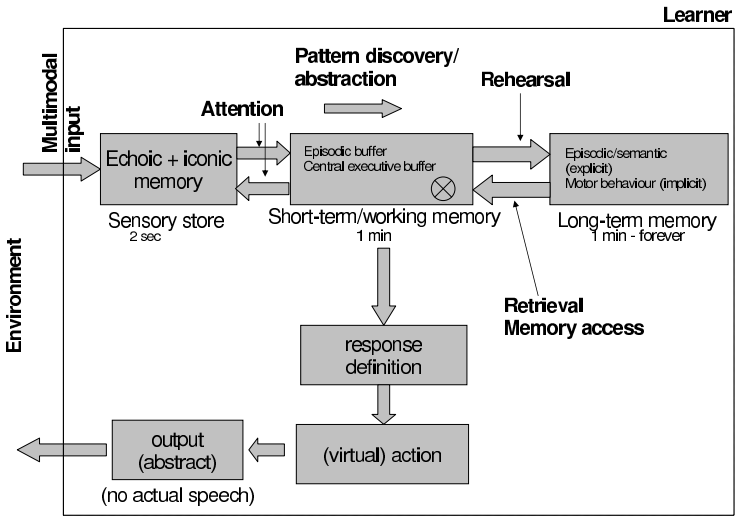

Figure 1. Global layout of the architecture used in the computational model. Multimodal input is presented to the model (upper left corner).

Multimodal input is presented to the model (upper left corner) and processed by the feature extraction module. The resulting low-level feature representation is put into the sensory store. This part of the memory can store data for a few seconds (which is enough for the storage of one utterance). The sensory information is copied to the short-term (working) memory. The way how this is done exactly is determined by a process called 'attention'. In human sensory processing, attention determines to what extent and which information is actually copied from the sensory store to the short-term (working) memory. In the current computational model, attention is implemented in a default way such that *all* sensory information is available in the working memory.

This short-term/working memory can store data for up to about a minute. One of its functions is to store full episodic traces, suggesting that the memory capacity is large enough to accumulate essential in-

formation obtained from the observed stimuli for the initialization and update of internal representations. For the computational model, this means that the storage of a matrix in which spectro-temporal information is accumulated is cognitively defendable. This is further supported by considering two other main functions of the short-term memory. Short-term memory serves as a central execution platform, which allows to compare tokens and representations stored in long-term memory. In addition, it can perform dedicated tasks (e.g. for visual information it serves as a sketch pad, for speech it may support phone detection tasks in dedicated experiments). However, the capacity of the short-term/working memory is limited, which means that new data may overwrite older data.

The representations available in short-term memory can be stored in long-term memory. For humans this storage is enhanced by rehearsal; it may be facilitated by the repetition of intrinsic and extrinsic stimuli presentation. Once stored in long-term memory, items can be available for decades. In the computational model, the long-term memory is used for the storage of all persistent internal representations.

The attention and rehearsal processes operate on representations stored in memory. The precise relation between these processes in human cognition is not yet known. Some authors interpret experimental results as proof that attention is *maintained through rehearsal* in order for information to be stored in short-term memory. Rehearsal of an *extrinsic* presentation may be forced by the frequent occurrence of a specific entity (e.g. a target word) in the input speech stream. Another way of rehearsal is *intrinsic*, in which the rehearsal is result of internal reflection on a certain representation. Attention is a process that reduces the part of the input stream that must be analyzed in detail and is therefore indispensable for managing time, space, effort and in the end for being successful: to keep the computation load manageable, to reduce the storage into short-term (working) memory, to reduce the ambiguity to be resolved during the search, and to keep promising input features within the attention 'beam'.

In the current computational model, rehearsal is implemented by an internal learning loop to update the presentations learned so far on the basis of the current stimulus (see section 3.2). As already mentioned above, attention is implemented in such a way that all information that arrives in the sensory store is moved into short-term memory. The module 'response definition' in figure 1 reads in representations from the short-term memory – the modules 'virtual action' and 'output' subsequently provide (coded) output to the caregiver. The response of the learner is a code that corresponds to the internal representation with the highest activation.

3.2. Feedback and learning loops

Apart from the multimodal stimuli offered by the caregiver, the learner may also receive *feedback* from the caregiver. This feedback informs the learner about the appropriateness of the reply to the previous stimulus, on an utterance by utterance basis.

The feedback is provided in the external learning loop. In the current implementation of the model there are two learning loops: an external and an internal one (see Figure 2). The external loop involves the interaction with the caregiver. By the feedback mechanism, the external loop supports the optimization of appreciation that the learner receives from the caregiver. The *internal* loop is different: this loop takes into account the *learner-internal* findings during the learning process, such as the amount of disambiguation of the input in terms of what the learner knows ('quality of the parse'), the time it takes to perform a certain action, the perplexity during the search, or the amount of resources required to disambiguate a certain input. In the current implementation, the 'quality of the parse' is the only performance measure in the internal learning loop.

The use of an internal and an external loop explicitly means that learning takes place over both loops at the same time. These loops do not necessarily have the same cycle-time: a short-cycle loop make take place several times per utterance, and a long-cycle loop may take place on an utterance-by-utterance basis or only after a block of N utterances has been received.

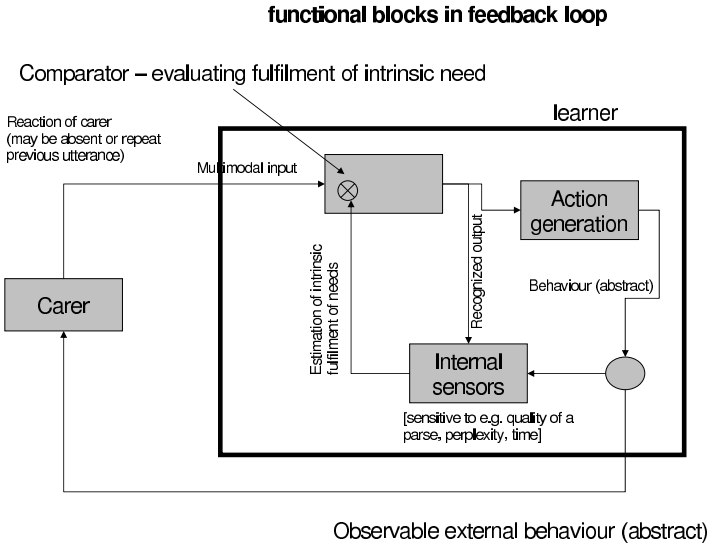

Figure 2. Overview of the external and internal loop. In the external loop the caregiver interprets the output of the model and provides feedback on the correctness of the response to the latest input stimulus. This feedback takes place on an utterance by utterance basis. The internal feedback deals with the optimization of the learner-internal target function (equations 1, 2).

3.3. Learning drive

As mentioned above, the learning drive is interpreted as the drive to 'interpret' the incoming stimuli in terms of the learned representations. It thereby directly connects the internal loop (which aims at improvement of the quality of the parse) and the external loop (which aims at correct replies to the caregiver). The learning drive of 'optimal parse' is mathematically implemented as a target function that operates in a computational learning scheme. This implementation is globally characterized by equation 1, but can now be made more precise.

Denote $V = v_{ij}$ and $W \cdot H = (WH)_{ij}$. The Kullback-Leibler divergence $KL()$ between V and $W \cdot H$ is determined on a column-by-column basis (that is, utterance-by-utterance):

$$KL(V, W \cdot H) = \sum_i (\sum_j v_{ij} \log(v_{ij}/(WH)_{ij}) + (WH)_{ij} - v_{ij}) \quad (2)$$

where i and j run over the columns and rows in V (and WH), respectively.

Equation 2 implies that the improvement of the quality of the parse is translated into the search for an optimal basis W such that the column vectors in W explain the columns in V. Equation 2 specifies the precise error made between observed V and predicted $W \cdot H$.

4. Experiments

4.1. General setting

Using NMF as the algorithmic implementation of the structure discovery (learning) algorithm we conducted a large number of experiments to investigate the capability of our model to acquire a basic set of 10 'words'. In actual practice this amounts to the capability to build internal representations of some 10 different acoustic signals that correspond to keywords embedded in carrier sentences. From the set of experiments that we have conducted we select some that clarify the impact of further implementation details on the cognitive plausibility of the operation of the learner model.

In their first month infants interact with a very small number of caregivers. Thus, one would expect that internal representations of speech will be speaker-dependent. Later on, infants will interact with a much larger number of persons, which will require adaptation of the internal representations to make these speaker-independent. To investigate this, we conducted experiments in which we offered training utterances in two ways: blocked per speaker versus randomized.

A convincing computational model of language acquisition should be able to learn any language. For this reason we compared the performance of our model for three different languages. In addition, we conducted experiments with bi-lingual learning.

Infants hear essentially two types of speech: utterances that are addressed to them and utterances exchanged between others in the scene that the infant overhears. Especially in the first stages of language acquisition it is important to distinguish these two speech types (or styles), if only because the probability that infant directed speech refers to some easily observed object or event in the scene is much higher than for so called adult directed speech.

4.2. Materials

For the investigation in the ACORNS project three databases have been recorded, one Dutch database (NL), a Finnish database (FIN), and a Swedish database (SW). Finnish is a member of the Finno-Ugric language group, which is very different from the Indo-Germanic language group to which Dutch (West-Germanic) and Swedish (North-Germanic) belong. All three languages differ in several aspects, such as the type of inflection of nouns, adjectives, pronouns, numerals and verbs. The typological diversity of these languages makes a comparison very useful and potentially interesting.

For each language the database contains utterances from 2 male and 2 female speakers. Each speaker utters 1000 sentences in two speech modes (adult-directed, ADS, and infant-directed, IDS), for a total of 2000 utterances per speaker. Per speaker, the set of 1000 sentences contains 10 repetitions of combinations of 10 target words and 10 carrier sentences. The set of target words consists of one proper name (for the learning agent) and nine nouns that refer to common objects that were chosen on the basis of literature on language acquisition. All speakers share the same target words, but the proper name they use to address the learner is different per speaker. For example, the NL database contains 800 tokens of target words such as *luier* (Eng *diaper*), auto (*car*), but only 200 tokens of each of the four proper names 'mirjam', 'isabel', 'damian', 'otto'.

Compared to adult-directed speech, infant-directed speech is characterized by a greater affect in intonation and prosody, a simpler syntax, and a more careful and slower pronunciation. The recordings are as realistic as possible; we asked the speakers for the database to speak *as if* they were addressing an adult or a child. The adult-directed version consists of the sentences read aloud in a neutral manner.

The databases also contain meta-information. In addition to codes for speaker and language a tag is available that codes the object that is referred to in the speech utterance. In part of the experiments reported below the tag was used as an unambiguous reference to one of the objects (or the name). However, it is possible to replace an unambiguous tag with more fuzzy references, somewhat similar to what was done with the CELL model [33].

The databases serve as a pool of stimuli that are used in all experiments that are discussed below. The ordering of the stimuli is an experimental design factor and differs from experiment to experiment. These effects will be discussed in more detail below.

4.3. Results

The result of a learning experiment is summarized in a figure in which the horizontal axis represents the number of utterances processed by the learner. The vertical axis presents the accuracy of the learner's replies. The accuracy is defined as the proportion of correct responses in the last 50 replies (where 'correct' is defined by comparing the reply with the ground truth as defined by the multimodal input stimulus). As a result, the plots show the accuracy that comes close to the 'instantaneous accuracy' of the learner.

Basic learning capability Figure 3 (Dutch) and 4 (Finnish) show how the ACORNS model learns a limited set of word representations and classifies a new stimulus in terms of one of these representations. During the learning, the model gradually improves the quality of its internal representations, by minimizing the Kullback-Leibler distance (equation 1) between the observed data and the internal representations. The model's memory is quasi-infinite: in the update of its internal representations, the model uses all utterances that it heard so far. In this experiment utterances from one language database were offered in random order and no difference was made between ADS and IDS. The figures show the concept accuracy as a function of the number of utterances processed. Until about x=1000, not enough tokens of any 'word' has been observed to construct any internal representation; as a result, the recognition rate is 0%. From x=1000 until x=4500, the internal vocabulary is being built. It starts unstable, is not complete yet and lacks one single word. After x=4500, the internal lexicon is complete and the performance ranges between 95 and 100% correct. The learning curves for Dutch and Finnish are similar, despite the very different morphological structure of the two languages (Finnish has many more different word forms than Dutch).

Speaker-dependent learning Figures 5 and 6 show that the learner's learning behavior in case of speaker-blocked stimulus presentation, for Dutch and Swedish, respectively. Again, no distinction is made between ADS and IDS. This experiment can be compared with the experiment displayed in figures 3 and 4. A drop in performance of about 20-30 percent is clearly visible every time when a new speaker starts (around #tokens = 0, 2000, 4000, 6000). After about 1000 tokens (that is, approximately 100 tokens per word) the performance is back on its previous high level. The decrease in performance when a new speaker is introduced is due to two factors (a) each speaker uses a different proper name, such that the vocabulary is not complete until enough examples of the new word were offered (b) different speakers have different voice and speech characteristics which require an adaptation by the learning model. This experiment shows the speaker dependency of the internally stored representations. However, it also

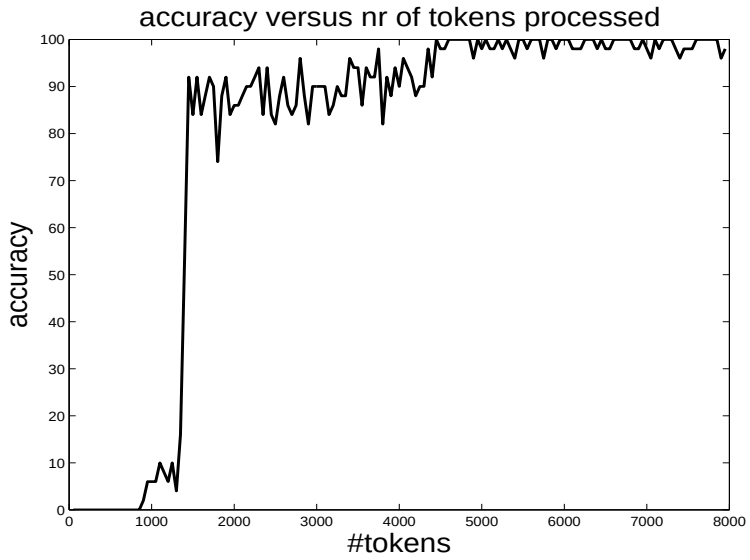

Figure 3. Result of a learning experiment (Dutch, random ordering of stimuli).

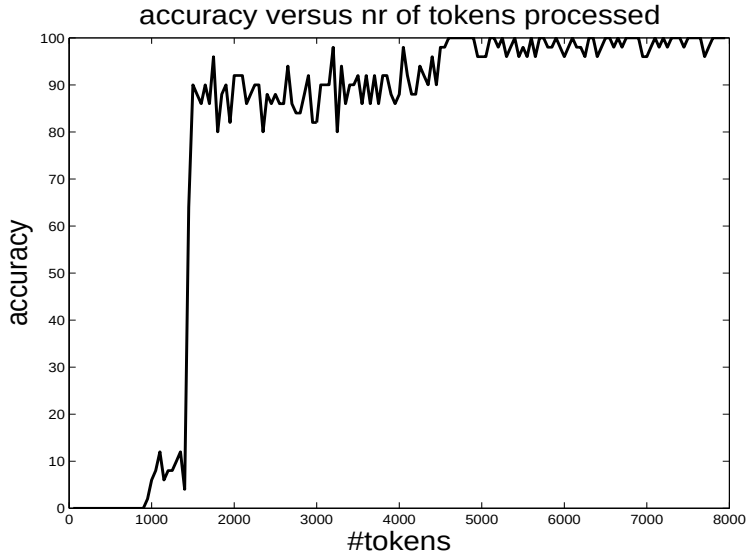

Figure 4. As figure 3, but for Finnish.

shows that the representations built for one speaker can be re-used at least in part for processing a new speaker: when a new speaker comes in, performance does not drop to 0%.

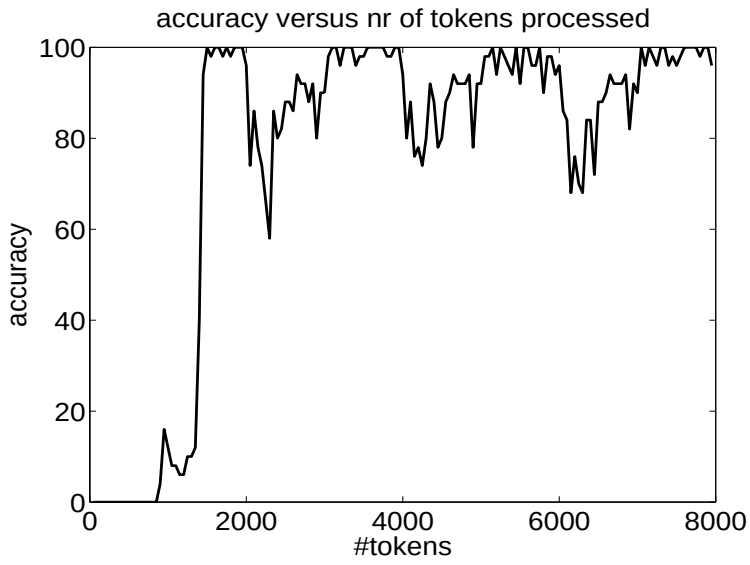

Figure 5. Dutch, speaker-blocked. To be compared to figure 3.

Bi-lingual language acquisition While the experiments described above show that the ACORNS model can learn different languages with equal success, we now turn to bi-lingual learning. Experiments with bi-lingual learning require some relaxation of the 'crispness' of the tags that accompany speech utterances. We compared two situations: one in which we assume that words in Dutch are completely different from words in Swedish, and another in which the tags allowed for sharing representations between the two languages. The results are shown in Figure 7. It shows two plots made on the basis of two learning experiments with a *bi-lingual* stimulus presentation. In both experiments, first two Dutch speakers are presented, after which two Swedish speakers follow. The speakers are a Dutch female, followed by an Dutch male, a Swedish female and a Swedish male. The *dashed* curve shows the result in case of *language-dependent tags* that do not allow sharing representations between the two languages. As could be expected, switching from Dutch to Swedish causes performance to drop to 0%. Also, the time it takes to 'learn' Swedish after having learned some Dutch is essentially equal to the time it took to learn the first Swedish words in the experiment in Figure 6. The eventual model (at x=8000) is able to both recognize Dutch and Swedish target words. Also the total vocabulary is now doubled (since the languages make use of language-dependent non-overlapping tag sets).

The *solid* curve in Figure 7 refers to the case in which the original language-dependent tags in the NL and SWE database were replaced by a common set of tags that semantically make sense in a language-independent way. This experiment shows that the learner is able to reuse existing, already built representations. The learner reuses existing representations trained on Dutch to decode the Swedish utterances. This reuse effect is particularly visible between 4000 and 5000 utterances. The accuracy does not drop to 0 percent due to the reuse of the already existing word representations trained on Dutch.

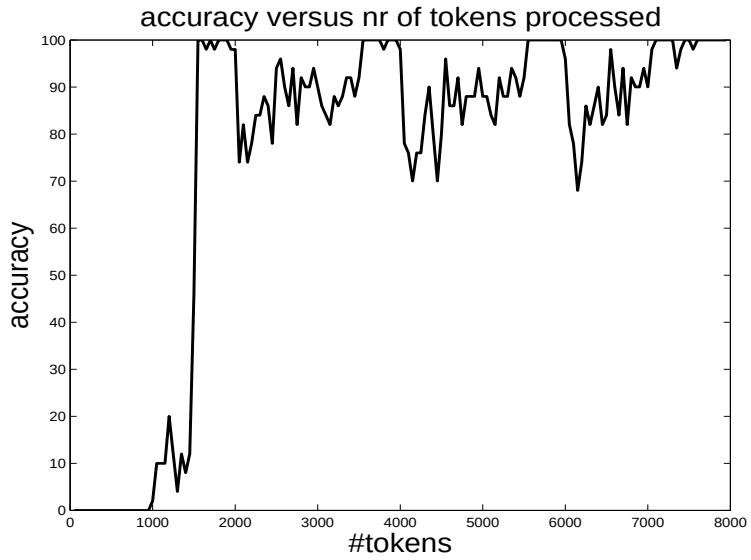

Figure 6. As figure 5, but for Swedish.

Know if one is being addressed Figure 8 shows the results of an experiment on Dutch, with random stimulus order, while the learner attempts to distinguish both the target words and the style (i.c. ADS and IDS). It shows that the leaner is able to identify speaking style, in combination with words. While the accuracy of word recognition is about 97-98%, the accuracy of the style recognition is close to 80%. If the speech style could not be detected at all and the word detection were error free, the IDS/ADS assignment would be random and the performance would theoretically be equal to 50%. The actual performance of almost 80 percent shows that the learner is able to identify the speaking style with an accuracy of about 80%. That is, the experiment shows a clear sensitivity of the learner for speech style, on top of its ability to discover the actual target words. Interestingly, this is the first sign of the learner to be able to form a certain abstraction, since the number of tokens *per speech style* is far too large for the model to represent individual tokens.

Summary These experiments presented in this section show the following results:

1. The learner is able to learn a limited set of target words and classify a new stimulus in terms of one of these words. The learner needs a number of tokens before it can build a reliable representation. During the learning, it gradually improves the quality of its internal representations, by minimizing the Kullback-Leibler distance between the observed data and the predictions based on internal representations. In the experiments we made the implausible assumption that the learner's memory is infinite: in the update of its internal representations, the learner assumes all utterances that it heard so far to be available for training (figures 3 and 4). However, later experiments have shown that it is easy to relax the infinite memory constraint.

2. The learning results on the three databases (NL, SWE, and FIN) are not essentially different (figures 3 to 6). Thus, this suggests that the ACORNS model is able to learn any language.

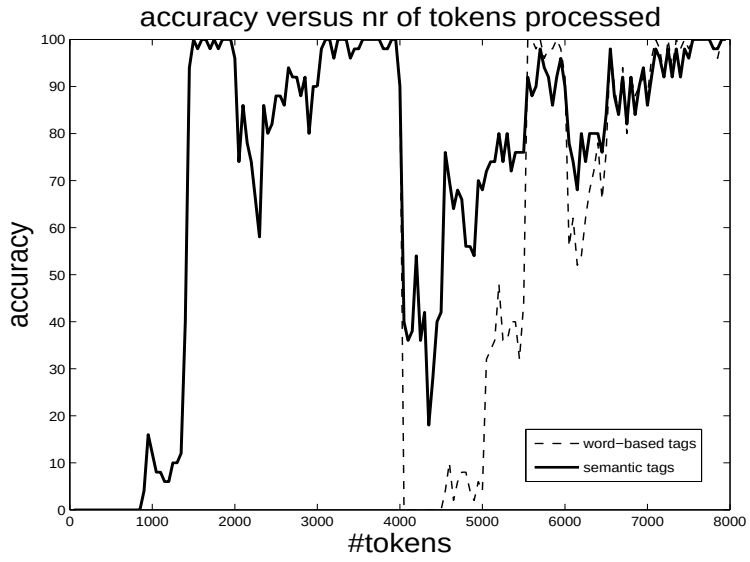

Figure 7. Results of a bilingual experiment, based on the use of a common tag set (solid line) or a language-dependent tag set (dashed line).

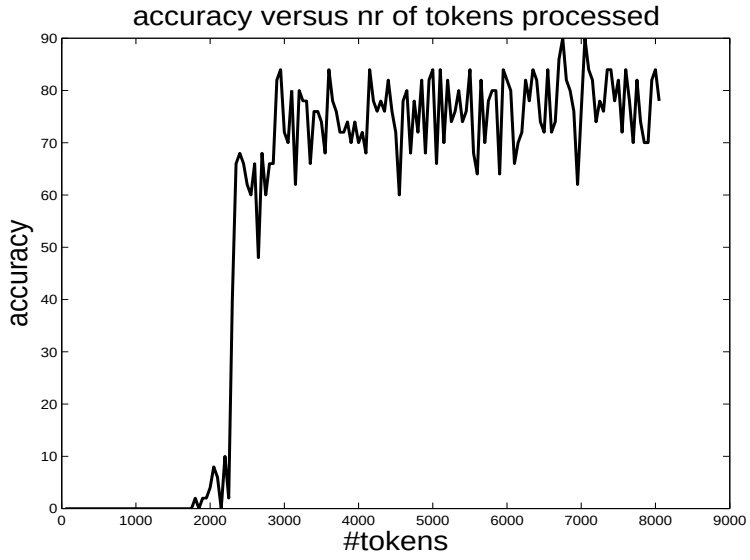

Figure 8. Results of an experiment on Dutch, random stimulus order, the learner trying to distinguish *both* words *and* speech style (adult- and infant-directed speech).

3. The learner rapidly adjusts to a new speaker. When the role of the caregiver is fulfilled by one speaker, the learner's internal representations will be speaker-dependent. As soon as a new speaker starts interacting with the learner, the internal representations will be adapted to accommodate the speaker characteristics (figures 5 and 6).

4. The learner reuses already stored representations whenever possible. This is particularly interesting in the bi-lingual experiment based on the semantic tags (figure 7).

5. The learner is able to identify speaking *style* (fig 8). The accuracy of the style recognition is close to 80%.

5. Discussion

The computational model presented in this paper shows that cross-modal cross-situational learning of the association between speech fragments and references to objects can be established using a general purpose pattern discovery technique. The performance of the learner depends on a number of factors – in this paper we investigated the ordering of the data (stimuli), the blocking per speaker, the effect of speaker changes, and multi-lingual training. All experiments showed that our model is able to build internal representations that can be used as stepping stones when new things must be learned (a new speaker or a new language).

The NMF implementation used in the experiments reported in this paper processed the input data incrementally, but without forgetting, that is: each new utterance is recognized with the current W and H matrices, and these were updated after blocks of N new input stimuli by decomposing a matrix V with N additional columns. Although this does not necessarily imply that all previously observed stimuli must be kept in memory in all detail, from a cognition point of view it would be more plausible to introduce some kind of forgetting. The cognitive plausibility of the NMF algorithm would also increase if it were feasible to do a new decomposition after essentially each individual input stimulus. In recent research we have made significant advances in both directions.

An interesting property of the word discovery approach discussed above is the *absence of segmentation*. The learning model does not use segmentation in order to hypothesize the target words in an utterance. Instead, the learning model makes use of structure in another way. Each utterance is mapped into a fixed-length vector [10]. This mapping is structure-preserving in that the sequence of words is transformed into the vector space that supports decomposability of an unknown vector as a weighted sum of given basis vectors (some of which might represent words in the linguistic sense of the term. Our experiments have shown that utterances can be interpreted correctly without a segmentation of the input speech into words.

The method for building associations between speech utterances and references to objects some (virtual) world as laid out in this paper can be compared with other approaches. The comparison with conventional HMM-based ASR approaches is meaningless, since the entire set-up and the assumptions underlying HMM-based approaches are incompatible with the set-up necessary to simulate the language acquisition process. The approaches that come closer are the CELL model [33] and the segmental dynamic time warping approach [31]. However, the CELL model lacks the symbol-grounding for the representation of its speech input as a lattice of phoneme symbols. In that respect, the segmental DTW model developed by Park and Glass [31] avoids the use of symbolic representations of the speech signal.

This technique allows to find a common stretch in a pair of acoustic utterances. Their next step involves graph-based clustering to build groups of common 'chunks' and to relate these clusters to 'words'. As the authors admit 'Although the inspiration for our methods is partially derived from experiments in developmental psychology, we make no claims on the cognitive plausibility of these word acquisition mechanisms in actual human language learning.' In ACORNS we are investigating a somewhat similar approach, based on DP-Ngrams [17].

Our model continuously retunes categorical boundaries on the basis of stimuli, and thereby shows interesting conceptual parallels with empirical results from psycholinguistic experiments (cf. [5]). When (adult) listeners are confronted with acoustically ambiguous sounds (e.g. sounds on the /f-s/-continuum in non-words where only one of /f/ or /s/ was phonotactically legal, e.g., frul/*srul or *fnud/snud), their category boundaries had shifted in subsequent categorization tasks involving stimuli along a phonetic continuum from /f/ to /s/. The stimulus '-rul' led to expanded /f/ categories, while '-nud' led to the expansion of /s/. This and similar findings imply that lexical access is not required for inducing perceptual retuning of category boundaries; phonotactic sequence information alone is sufficient. A similar retuning process is active in our model. Once internal representations have been built, the boundaries between them are updated continuously on the basis of new example stimuli. That means that competition between representations takes place, since a certain number of representations must be accommodated within a restricted volume in the 'representation space'. In the current model, this representation space is modeled as a vector space endowed with the (statistically motivated) asymmetric Kullback-Leibler divergence as dissimilarity measure.

Since the model aims at word discovery informed by the findings about human language acquisition, its *cognitive plausibility* is one of the criteria with which the model can (and must) be judged. In the literature on language learning and word acquisition, a number of characteristics of language acquisition by young children are highlighted. Firstly, the number of words that young infants understand increases over time, with a 'word spurt' between month 16 and 24. This word spurt is generally attributed to cognitive factors: based on already existing representations of words, the learning of more words gets increasingly faster. The McMurray model [25] holds that the word spurt is a necessary effect of a combinatorial artifact. That model shows that the word spurt phenomenon is guaranteed in any system that builds representations for multiple words simultaneously, and in which few words can be acquired quickly and a greater number of words take longer. Under reasonable conditions (saying that words occur with a Zipfian distribution; and that each word needs a word-dependent number of tokens in order to be stored into some representation) a word spurt can be observed after an initial period of slow learning. It is to be investigated to what extent such a word spurt can be modeled by the current computational model, and whether the existence of this word spurt phenomenon is stable across various model parameters that can be cognitively explained.

The computational model presented here sheds light on the relevance of various issues that are known to play a role in (models of) human speech processing. One of these issues is how words get activated (and to what extent), the second is how competition happens during the word search. In our model the activation of lexical items is separated from the actual competition. This is similar to what Shortlist, one of the widely used computational models for human word processing does [28]. Shortlist is a two-stage model in which activation of words by incoming speech input is separated from competition between the activated words. In contrast with Shortlist, however, the current model plays out the entire lexicon, while in Shortlist the network in which competition plays a role is constructed from only those words directly supported by the input (which, in the case of Shortlist, is a symbolic representation of the speech signal).

In the current model, competition is not explicitly implemented. Instead, a form of competition automatically emerges from the combination of (a) parallel search among multiple candidates (b) the concept of 'best match' given a certain dissimilarity. This is in line with earlier findings e.g. obtained with another model of human word processing TRACE [24]. TRACE showed that competition is not a necessary consequence of multiple processing in parallel. TRACE was implemented as a connectionist model based on interactive activation. Incoming input increases the activation of lexical candidates that it matched. A crucial difference between the ACORNS model and TRACE is the absence of inhibition. The more activation a candidate receives in TRACE, the stronger the inhibition it will exercise upon its rivals. Words which receive ever more activation wield ever stronger inhibition, and eventually the winning string of words will end up with higher activation than all competitor strings. In our model there is no such effect: the activation of a winner will not completely eliminate its closest competitors. However, there is a stronger inhibition process than for instance in the computation of posterior probabilities, where the normalization to unity implies that if a candidate model receives a larger probability, its competitors must scale their probabilities down. Using a decomposition-based method such as NMF, close competitors can completely inhibit each other. When two internal representations have a common sub-vector which is activated, it can be explained by any combination of these two representations. If in another sub-vector both representations differ, the relative activation of each of the representations will be determined by the data in this second sub-vector. But if one of the representations already overestimates the values in this sub-vector, activation of the second one is not necessary (a better fit would require a negative activation, which is not allowed) and it is hence inhibited. Hence, a form of inhibition is in place.

References

[1] Baddeley, A.: *Working Memory*, Clarendon Press, Oxford, 1986.

[2] Bellegarda, J.: Exploiting Latent Semantic Information for Statistical Language Modeling, *Proceedings of the IEEE*, **88**, 2000, 1279–1296.

[3] ten Bosch, L., Kirchhoff, K.: Bridging the gap between human and automatic speech recognition, *Speech Communication*, **49**, 2007, 331–335.

[4] Cooke, M., Ellis, D.: The auditory organization of speech and other sources in listeners and computational models, *Speech Communication*, **35**, 2001, 141–177.

[5] Cutler, A., McQueen, J., Butterfield, S., Norris, D.: Prelexically-driven perceptual retuning of phoneme boundaries, *Proc. Interspeech 2008*, ISCA, Brisbane, Australia. September 2008, 2008.

[6] den Os, E., Boves, L., Rossignol, S., ten Bosch, L., Vuurpijl, L.: Conversational Agent or Direct Manipulation in Human-System Interaction, *Speech Communication*, **47**, 2005, 194–207.

[7] Ernestus, M., Baayen, R., Schreuder, R.: The recognition of reduced word forms, *Brain and Language*, **81**, 2002, 162–173.

[8] Goldinger, S.: Words and voices: episodic traces in spoken word identification and recognition memory, *Journal Experimental Psychology Learning Memory Cognition*, **22**, 1996, 1166–1183.

[9] Goldinger, S.: Echoes of echoes? An episodic theory of lexical access, *Psychological Review*, **105**, 1998, 251–279.

[10] Van hamme, H.: HAC-models: a Novel Approach to Continuous Speech Recognition, *Proceedings Interspeech*, ISCA, Brisbane, Australia. September 2008, 2008.

[11] Hermansky, H.: Auditory modeling in automatic recognition of speech, *ESCA Workshop on the Auditory basis of speech perception*, Keele University (UK), 15-19 July, 1996.

[12] Hintzman, D.: Schema-abstraction in a multiple-trace memory model, *Psychological Review*, **93**, 1986, 411–427.

[13] Hoyer, P.: Non-negative matrix factorization with sparseness constraints, *Journal of Machine Learning Research*, **5**, 2004, 1457–1469.

[14] Johnson, S.: *Emergence*, New York: Scribner, 2002.

[15] Jones, D., Hughes, R., Macken, W.: Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory, *Journal Memory and Language*, **54**, 2006, 265–281.

[16] Kaplan, F., Oudeyer, P.-Y., Bergen, B.: Computational Models in the Debate Over Language Learnability, *Infant and Child Development*, **17**, 2008, 55–80.

[17] Kuhn, R., Nowell, P., Drouin, C.: Approaches to Phoneme-Based Topic Spotting: An Experimental Comparison, *Proceedings ICASSP*, 1997.

[18] Lee, D., Seung, H.: Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems*, **13**, 2001.

[19] Lippmann, R.: Speech Recognition by Human and Machines, *Speech Communication*, **22**, 1997, 1–14.

[20] MacWhinney, B.: Models of the emergence of language, *Annual Review of Psychology*, **49**, 1998, 199–227.

[21] Maloof, M., Michalski, R.: Incremental learning with partial instance memory, *Artificial Intelligence*, **154**, 2004, 95–126.

[22] Marr, D. C.: *Vision: A computational investigation into the human representation and processing of visual information*, W.H.Freeman & Co Ltd., 1982.

[23] Maslow, A.: *Motivation and Personality*, New York: Harper & Row, 1954.

[24] McClelland, J., Elman, J.: The TRACE model of speech perception, *Cognitive Psychology*, **18**, 1986, 1–86.

[25] McMurray, B.: Defusing the childhood vocabulary explosion, *Science*, **317**, 2007, 631.

[26] Moore, R.: A comparison of the data requirements of automatic speech recognition systems and human listeners, *Proceedings EUROSPEECH'03*, Geneva, 2003.

[27] Newport, E., Aslin, R.: Learning at a distance: I. Statistical learning of non-adjacent dependencies, *Cognitive Psychology*, **48**, 2004, 127–162.

[28] Norris, D.: Shortlist: A connectionist model of continuous speech recognition, *Cognition*, **52**, 1994, 189–234.

[29] Norris, D.: How do computational models help us develop better theories?, *A. Cutler (Ed.), Twenty-first century psycholinguistics: Four cornerstones*, 2005, 331–346.

[30] Ostendorf, M.: Moving beyond the 'beads-on-a-string' model of speech, *Proc. IEEE ASRU-99*, Keystone, Colorado, USA. Dec 12-15, 1999.

[31] Park, A., Glass, J.: Unsupervised pattern discovery in speech, *Transactions of the ASLP*, **16**, 2008, 186–197.

[32] Pfeifer, R., Scheier, C.: *Understanding Intelligence*, Cambridge, Mass.: MIT Press, 1999.

[33] Roy, D., Pentland, A.: Learning words from sights and sounds: a computational model, *Cognitive Science*, **26**, 2002, 113–146.

[34] Sarma, A., van der Hoek, A.: *A Needs Hierarchy for Teams*, ISR Technical Report: UCI-ISR-04-9, 2004.

[35] Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.: How should a speech recognizer work?, *Cognitive Science: A Multidisciplinary Journal*, **29**(6), 2005, 867–918.

[36] Snow, C., Ferguson, C.: *Talking to children: language input and acquisition*, Cambridge, New York: Cambridge University Press, 1977.

[37] Sroka, J., Braida, L.: Human and machine consonant recognition, *Speech Communication*, **44**, 2005, 401–423.

[38] Stouten, V., Demuynck, K., Van hamme, H.: Automatically Learning the Units of Speech by Non-negative Matrix Factorisation, *Proceedings Interspeech-2007*, Antwerp, 2007.

[39] Wang, Y.: Cognitive Informatics: A new transdiciplinary research field, *Brain and Mind*, **4**, 2003, 115–127.

[40] Wang, Y.: On cognitive informatics foundations of knowledge and formal knowledge systems, *Proc. ICCI 2007*, IEEE, Lake Tahoe, California, USA. August 2007, 2007.

[41] Werker, J., Curtis, S.: PRIMIR: a developmental framework for infant speech processing, *Language Learning and Development*, **1**, 2005, 197–234.

[42] Wesker, T., Meyer, B., Wagener, K., Anemueller, J., Mertins, A., Kollmeier, B.: Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines, *Proceeding of Interspeech-2005*, Lisboa, 2005.