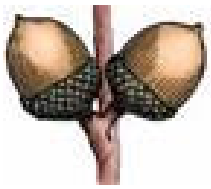

Project no. 034362

ACORNS

Acquisition of COmmunication and RecogNition Skills

Instrument: STREP
Thematic Priority: IST/FET

D5.1

PART A: Description of the ACORNS computational model version 2.0
PART B: Bootstrapping the lexicon – a model for first language acquisition

Due date of deliverable: 2007-11-31
Actual submission date: 2007-12-21 (final versions)

Start date of project: 2006-12-01 Duration: 36 Months

Organisation name of lead contractor for this deliverable: RUN

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

VERSION DETAILS
Version: 2.0
Date: Dec 21, 2007
Status: final

CONTRIBUTOR(S) to DELIVERABLE	
Partner	**Name**
RUN	Louis ten Bosch,
KUL	Kris Demuynck
KUL	Hugo Van hamme

DOCUMENT HISTORY			
Version	**Date**	**Responsible**	**Description**
0.1	30/11/07	Louis ten Bosch	writing
1.0	9/12/07	Louis ten Bosch	Added comments from Lou and Hugo

DELIVERABLE REVIEW			
Version	**Date**	**Reviewed by**	**Conclusion***
2.0	2007-12-18	Toomas Altosaar	Wording issues - approved
2.0	2007-12-21	Roger Moore	Formatting issues - approved

PART A

Description of the ACORNS computational model -

A system demonstrating the capacity for acquiring language and communication

Louis ten Bosch, Hugo Van hamme, Kris Demuynck, Lou Boves, ACORNS team

In this document (part A of D5.1) we provide an overview of the ACORNS computational model as an experimental platform. The aim of this part A is to provide a brief but explicit description of the learning platform. Part A is followed by the second part of D5.1 (part B) which discusses the word detection experiments.

The experimental platform

Introduction

ACORNS aims to investigate acquisition of language and communication skills by a young infant. To that end, a computational model is designed, built and tested. The computational model consists of two participating agents (a carer and a learner, abbreviated CA and LA, respectively) that communicate with each other. The learner (LA) models the acquisition of language by detecting word-like entities on the basis of multimodal stimuli that are provided by the carer (CA) in the dialogue between learner and carer. The CA simulates the person (or persons) who is looking after the infant. The manner in which the CA-LA interaction takes place is a simplified version of true human-human interaction (see part B). This simplification is evidently reflected in the computational model.

In the current version of the model that is used for the experiments, one of the main actions by CA is providing multimodal combinations of sound files and meta-tags to LA. The ordering in which CA presents the stimuli is determined outside of CA and is actually one of the design factors in the experiment. In the current experiments, the conceptual way in which CA deals with replies from LA is the same for all replies: CA provides correct/incorrect feedback to LA (which LA in turn is free to use if desired). CA also keeps track of the responses by LA, just as in real-life when a carer notes the reactions from a young infant.

In the dialogue, LA and CA are the genuine interacting parties. Apart from the modules modelling LA and CA, there is a third module DM that manages the CA-LA interaction by passing the messages between CA and LA. There is no intelligence in this passing mechanism, that is, the passing mechanism does not provide any contents into the CA-LA interaction. It is passive with one exception: in the very beginning of each series of CA-LA interactions, it signals CA and LA that the interaction may start. When the interaction exactly

starts is up to LA and CA. The message passing mechanism is flexible and, e.g., would not object if CA and LA would speak simultaneously or in the case of long periods of silence.
The knowledge about what is to be expected in carer-infant interaction is modelled in CA. For example, LA may take a rest to reset its memory representations ('sleep'). In such a sleep period, CA won't try to interact with LA.

In real life, when dealing with interaction between two adults or an adult and a more mature child, this knowledge about interaction would be distributed between all participants. In no case is this knowledge located outside of the participants.

In the beginning of the communication between LA and CA, that is, the start of the language acquisition, LA starts with an empty memory without any lexical representations and without any knowledge about phoneme-like entities. During the interaction between LA and CA, LA gradually detects more and different words (as acoustic forms) and their grounding. Grounding in this context indicates their 'meaning', the object that words refer to in the direct environment of the young learner, such as, e.g., *ball*, *diaper*, *Daddy*. This association is provided by the multimodal stimuli presented by CA. During interaction, CA and LA exchange multimodal messages, consisting of speech fragments, which may be accompanied with an abstract visual tag.

When presented an utterance such as "look at the ball" or "what a nice ball" in combination with an abstract tag 'ball', LA attempts to relate the abstract tag 'ball' (which does not represent the *word* 'ball' but instead refers to the *object* ball in a virtual scene) with an actual acoustic form that is common in both utterances.

The tag is representing abstract information that would otherwise be present with other modalities. In a real child-carer interaction, a round object could be presented through the visual channel in parallel with the audio information. In designing this computational model, however, we avoid the engineering implementation of recognisers in the visual or tactile modalities, and concentrate on the linguistic speech-based part of the word discovery process. It is therefore assumed that non-audio information can be represented by meta-tags, presented along with the audio. In other words, these tags idealise the input from other modalities and translate to the presence or absence of certain references in the audio stream.

The learner model LA

LA consists of four basic ingredients:

1. A memory (that is further divided into a sensory store, a short term memory (executive memory) and a long term memory). This includes a mechanism for storing and retrieving representations.

2. A perception module, transforming signals from the environment into primary sensory data to be processed further.

3. A method to derive more abstract from less abstract representations. In the current implementation of LA, virtually all representations are modelled by vectors and matrices (this is just a design choice at this stage of the project – other means such as

graphs, multigrams or dynamically expanding HMMs are also possible and are being investigated). In the current implementation, bottom-up and top-down processes are based on matrix manipulations. Abstraction (bottom-up) is represented by matrix factorisation, while the top-down process is exemplified by matrix multiplication. Computationally, matrix factorisation is performed by non-negative matrix factorisation (e.g., Hoyer, 2004; Van hamme, 2007).

4. A learning drive: a target function specifying the basic need to learn. The target function consists of two terms: a term related to LA's internal need to learn, and one related to its external need to learn (part B explains more details).

When LA perceives input, the speech input is processed by the feature extraction module. The outcome is stored into the sensory store (figure 1), from where it is directly copied to short-term memory (STM). In STM, a match takes place between the sensory information on the one hand and the stored representations on the other, and the best matching representation is searched for. The resulting representation is generated and sent to the carer in an abstract manner, e.g., in the form of a tag associated with a 'confidence' measure, comparable to the way in which a real baby uses a head turn to indicate its detection of a stimulus.

In LA, the information from the audio channel and from the non-audio channel is integrated by combining the corresponding representations (two vectors, one from audio and one from the meta-tag) into one augmented vector. In this way, cross-model information is made available in a natural way to the learning algorithm.

In the current experiments, LA uses STM as executive memory and there is no clear distinction between STM and LTM. But the current implementation of LA makes use of a very long term memory (VTLM) in which all observed utterances and tags are stored without any decay. Decay is relatively easy to introduce into LA and will be available in stage-2 experiments.

The carer model CA

Compared to the structure of LA, the structure of CA is not elaborate. Nevertheless, as in all learning situations, the behaviour of CA has an effect on the learning performed by LA. The main task of CA is to provide multimodal utterances to LA. The moment at which CA speaks to LA is determined by CA. The utterances used during training and their ordering are determined outside the carer, in this case by the experimenter. From an available pool of multimodal utterances, CA uses a list of utterances in the CA-LA interaction. The interaction stops when one of the parties, CA or LA, does not want to continue, or if all existing multimodal inputs have been presented to LA.

After a reply from the learner, CA provides feedback about the correctness of the reply. In the current implementation the feedback is limited to a yes/no (approval/disproval) variable. Finally, CA keeps track of the responses from LA, just as a real carer notes the responses from a young infant.

The dialogue manager DM

An entire dialogue between CA and LA consists of a potentially large number of CA-LA dialogue turns. When LA receives multimodal input from CA, a reply is sent to CA. In the next turn, CA provides LA with a feedback (indicated as 'ground truth' in figure 1) about the correct detection of the target word, after which LA is free to use this feedback information. If LA is ready, CA proceeds with the dialogue by presenting the next multimodal stimulus. During CA-LA interaction, the dialogue manager (DM) performs two tasks: (a) message passing, and (b) logging. There is no intervention from DM into CA-LA interaction.

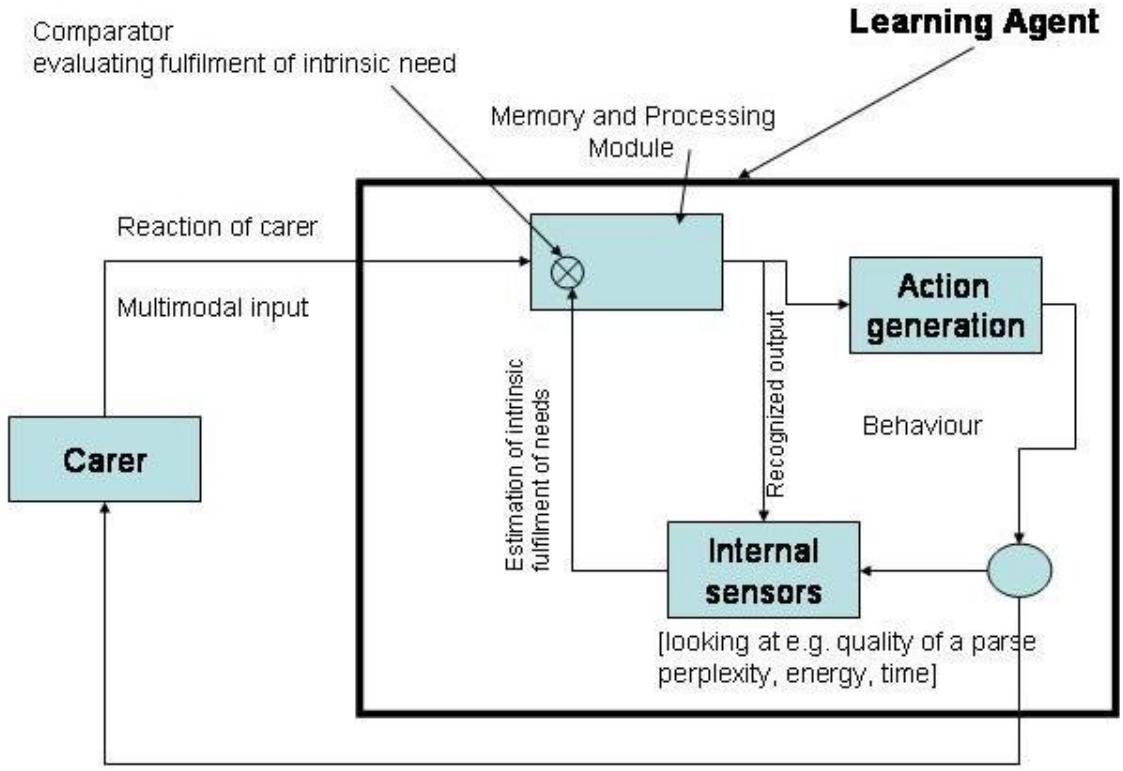

Figure 1. An overview of the overall interaction between the learner model LA (within the heavy-line box) and the carer environment CA.

Message passing

The communication between CA and LA is via messages. As noted above, the manager only organises and facilitates the messaging process in the dialogue and does not introduce linguistic or semantic knowledge into the dialogue.

Conceptually the dialogue between CA-LA is organised on the basis of events. These events are sent from CA to LA and vice versa. We now present a brief overview of the most relevant event types.

Whenever CA or LA receives an incoming event, it responds with an outgoing event. Events can have 'genuine' content or be a '*nop*' (no operation) event. Events are implemented as structures with fields specifying characteristic properties of the events. In the current implementation, an event is a structure containing at least the following three fields: id, time, and call back time. The 'id' is a string identifying the type of event (see below for a list of all possible events and a description of their use). The 'time' field is the virtual time, i.e., wall-clock time in the virtual world in which CA and LA live. The call back time is the (virtual) time LA or CA wants (or wanted) to be called back.

When the call back time of one of the parties is reached, a *nop* event will be sent to that party. Next to the three required fields there can be zero or more required or optional event specific additional fields – this depends on what information LA and CA agree to convey for particular events.

A 'nop' (no-operation) event is used in the following cases:

a. when a party doesn't want to respond on an incoming event, it answers with a 'nop' event,
b. when the 'call back time' of a party is reached, the discourse agent will send a'nop' event to that party to query for their reaction,
c. a 'nop' is returned in response to system administration events ('load','shutdown', ...) to signal that no error occurred.

Events can be divided into administrative events and dialogue events. Administrative events smooth the handling of the communication between LA and CA without interfering with the contents of the dialogue. For example, LA/CA can return an 'error' event if the incoming event caused some fatal problem(s). The interaction will be stopped after such an error. An 'error' event must have a 'msg' field describing the nature of the problem (a character string). Log-events can be used by LA/CA to write something into the log-file. These events are required to have a single 'msg' field. The power-up event is the very first event LA/CA will receive to start-up the internal process. The shutdown event is the last event LA/CA will receive and indicates that all services should be shut down.

The *save* event stores the internal memory status and the state of the dialogue. This allows continuation of a training run from an earlier point in time. The *load* event loads a previously saved state and these events organise the file handling (with a field 'fname'). If the fname-field in the event is present, LA/CA loads the state previously saved by a 'save' event with the same 'fname'. If the field is not present, LA/CA starts from the very beginning and loads the initial state (which is dependent on the configuration parameters given in the 'power-up' event).

Besides purely administrative events there are genuine dialogue events as well.
The audiovisual (av_data) event sends audio plus 'video' data from CA to LA. The event may contain the following additional fields:

a. duration (required) the length of the audio sample in seconds
b. data (required) the audio data
c. meta (optional) meta-information, i.e., addition multi-modal input

d. sample freq (required) sampling frequency of the audio signal

The reply event sends a reply from LA to CA. This event may contain the following additional fields:

a. answer (required) the recognised representation
b. assessment (optional) some internal measure (zero or positive) that quantifies the internal 'happiness' of LA (higher is better)

The feedback event sends feedback from CA to LA. It may contain the following additional field:

a. value (required) the value of the feedback function, i.e., a number in the range [0,1]

The assessment values are not directly comparable to a posteriori probabilities. The current learner can give assessment values larger than 1. The assessment measures to which extent the word is present. If the word is spoken twice in the utterance, the assessment value is theoretically equal to 2.

Thorisson three layer dialogue framework

In 2002, Thorisson proposed a framework for describing and investigating human-human (HH) dialogues. In HH-dialogues partners may interact on three levels. The lowest level contains the backchannels that are used to indicate the listener's presence and attention – they mainly serve to smooth the turn-taking in the dialogue. The highest level contains the well-thought conscious responses such as a deliberate answer to a question. The middle level contains the short phrases, such as the short content-based interaction of turns that are often heard in telephone conversations. All levels serve communication, albeit on three different levels.

The aim in ACORNS WP5 is to ultimately design a LA-CA interaction in which Thorisson's idea is reflected in the way the messages are sent and received. In the current ACORNS model, a multi-level messaging exchange is already possible: the message passing is flexible enough to cope with various types of messages, and to deal with messages on different time scales.

Practical issues

The software is available on the ACORNS wiki and can be run by all partners. A Matlab license and access to the ACORNS databases is required.

D5.1 Part B:

Bootstrapping the lexicon – a model for first language acquisition

Louis ten Bosch[1], Hugo Van hamme[2], ACORNS-team

[1] Dept. Language and Speech, Radboud University Nijmegen, The Netherlands
[2] ESAT, Katholieke Universiteit Leuven, Belgium

1 Introduction

In order to be able to effectively communicate, young infants must learn to understand speech spoken in their environment. They must learn that auditory stimuli such as stretches of speech are not just arbitrary sounds, but instead are sequences of reoccurring patterns associated with word-like elements. This word-discovery learning process is interesting since infants start without any lexical knowledge and the speech signal usually does not contain any clear acoustic cues about boundaries between words. So babies must 'crack' the speech code before being able to perform lexical acquisition, a complex task in which attention, cognitive constraints, and social-pragmatic factors all play an important role. Language acquisition mechanisms has been topic for discussion since a long time (see e.g.. Snow & Ferguson, 1977).

Cracking the speech code can be considered a pattern discovery problem, of which a solution has far-reaching advantageous consequences for the young infant. Not only the baby starts getting grip on the communication with its environment, also the discovery of word-like entities is the first step towards more complex linguistic analyses (cf. Saffran and Wilson, 2003).

The word discovery problem has been studied for decades, and several models and frameworks have been proposed. Recent studies have shown that the cognitive capability of word discovery might be supported by purely computational strategies (Saffran et al., 1996a; Aslin et al., 1998; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Kuhl 2004). By using computational mechanisms, young infants are able to find recurring patterns in the speech signal. Moreover, they appear to be able to use deviations in the statistical patterns as cues for segmentation. Infants of 7.5 months of age can use statistical regularities (such as the greater co-occurrence of syllables within words than between words) to successfully parse a continuous stream of speech (Saffran et al., 1996a; Aslin et al., 1998). Also prosodic information (i.e., longer duration, increased amplitude, and higher pitch on certain syllables) can guide word segmentation (e.g. Thiessen and Saffran, 2003).

The discovery of words requires the young infant to be able to perceive the speech signal, to store and retrieve representations, and to compute statistical patterns. By which mechanism

the brain does encode statistical patterns, and how does it detect changes in these patterns? The exact way in which changes in the input are detected is not known, but the fact that brain activity depends on changes in the stimulus has been convincingly shown for the visual domain. Neuroimaging studies have examined changes in the pattern of neural activity after a period of intensive training on a novel linguistic task (Friederici et al., 2002; Golestani and Zatorre, 2004), thereby showing the ability to spot previously unobserved patterns in (linguistic) data. Other studies explore the neural basis of language learning by examining changes in brain activity that occur that occur during learning (e.g. Hashimoto and Sakai, 2004).

A number of studies have provided indications about the increase of the number of items in the passive lexicon with age. The learning curve starts at 7 months at the latest: EEG scans show that at 7 months, children are able to detect words in utterances, also if they do not know the meaning of these words (Kooijman, 2007). In Kooijman's word segmentation experiments for Dutch, word stress appeared to play an important role as cue for this word spotting, in line with findings by Vroomen & De Gelder (1995), Johnson & Jusczyk (2001) and others.

It is estimated that young infants know on average about 100-150 words (or expressions) on average when they are 1 year and 4 months old. Especially between age 1 and 2 the size of the lexicon increases rapidly (word spurt). On average 8 to 10 words per day are acquired from birth to adulthood, but during the word spurt this number may increase by a factor of 3. During the second year of life, the rate at which children acquire new words accelerates dramatically. It has been hypothesised that this spurt is due to specialised cognitive mechanisms that leverage the few words learned in the initial slow phase for faster vocabulary growth later. The growth of the perceptual lexicon can indirectly be measured by head-turn experiments and is discussed in many language acquisition and behavioural studies.

Interestingly, McMurray (2007) demonstrates that the assumption about the 'efficient reuse of representations' are unnecessary to explain the vocabulary explosion. The word spurt phenomenon is guaranteed in any system that builds representations for multiple words simultaneously, but such that few words can be acquired quickly (easily) and a greater number of words take longer. This *distribution of difficulty* is explained by several factors such as word morphology and syntax. The word spurt is computationally shown by a model, in a simple short word is represented by a small jar, while more complex words are represented by a larger jar. Every time a particular word token is presented to the model, the corresponding jar is filled with a coin. Using the assumption that words occur with a Zipfian distribution and that words are actually learned as soon as the corresponding jar is full, a word spurt could be observed after an initial period of slow learning. This acceleration period was found to be very stable under varying word frequency distributions and jar size distributions.

In this paper, we will discuss a computational model for language acquisition with emphasis on the aspect of word discovery. In the next section, the relation with computational models of human speech processing is described briefly. Section 3 discusses the computational framework. Results are presented in section 4. Section 5 concludes with discussion and a conclusion.

2 Relation with psycholinguistic models of speech processing

Evidently, word discovery and language acquisition has an essential link to general (human) speech processing mechanisms. It is therefore interesting to investigate to what extent various psycholinguistic models of speech processing can provide a basis for a model of word discovery. Psycholinguistic theories and models (e.g. TRACE, McLelland & Elman, 1986; Shortlist, Norris, 1994; Luce et al., 1998; Goldinger, 1998; Scharenborg et al., 2005; Pisoni & Levi, 2007; Gaskell, 2007) have primarily been designed to simulate the speech processing as performed by humans, and are mostly put to a test on speech produced in controlled conditions, often by subjects participating in speech perception experiments.

The models developed in the eighties and nineties, of which TRACE and Shortlist are well-known relevant examples, are able to model the word decoding process given unknown input and a predefined lexicon. These models take *symbolic* representations of the speech as their input. As the input unrolls over time, it increases the activation of possibly multiple lexical candidates that locally match the input. These activated lexical candidates enter a competition in which a matching candidate receives more activation, while a less well matching candidate becomes inhibited. Finally, the hypothesized word sequence is the one that best accounts for the entire input. These models showed that in order to find words, the speech signal needs not to be 'segmented' into word-like entities a priori. The SpeM model (Scharenborg et al., 2005) shows that important aspects of human speech processing (such as word segmentation) can also be modelled on the basis of *real* speech as input, in combination with a competition between lexical entities in the form of a parallel search for best matching sequences. In SpeM, activation is not modelled on the word level, but dealt with on the level of entire word sequence hypotheses by using additive (log-likelihood-based) scores along each hypothesised path in the lexical search space. This idea can be found in a follow-up model for Shortlist, called Shortlist B (Norris & McQueen, submitted).

Virtually all models mentioned here need a predefined lexicon (in which the lexical items are represented by an abstract form in combination with a symbolic phonetic representation). This lexicon is used to span the search space within which word sequences are sought. The fact that a lexicon must be specified means that these models are not directly applicable for the modelling of word discovery (nor for language acquisition). The success of these models across many other tasks shows that word activation, competition and the dynamic search for pattern sequences are essential ingredients for a model aiming at the simulation of human speech decoding (cf. Pitt et al, 2002, for a discussion about these topics).

The framework that we propose in this paper builds on Boves et al. (2007) and combines the concepts of a multi-layered memory structure, competition between word hypotheses, and dynamic word sequence decoding. Simultaneously, it builds the lexicon in a dynamic way, starting with no words at the beginning of a training run. During training, new multimodal stimuli are presented to the model, and, depending on the internal need to do so, new lexical representations are hypothesized if existing representations fail to explain the input in sufficient detail. The experimental results obtained with this model will be discussed in the context of the findings as described in the literature on language acquisition.

3 Computational framework

3.1 Interaction between learner and carer

The computational framework involves two active participants ('agents') and simulates a learning agent (LA) that is involved in a dialogue with a caring agent (CA). The learning takes place within an interactive setting. The learner discovers words and word-like entities on the basis of the stimuli presented by the carer during the interaction. The entire framework actually consists of three sub models: the two active models (learner and carer) participating in the dialogue, and a passive 'dialogue facilitating model' (DM) which rules the interaction. The discourse model supervises the interaction between learner (LA) and carer (CA), by applying general world knowledge about dialogues, such as the turn taking mechanism.

LA starts with an empty memory, and during the interaction between LA and CA, LA gradually detects more and different words and their grounding. During interaction, CA and LA exchange multimodal messages, consisting of speech fragments, which may be accompanied with an (abstract visual) tag. The concept of word is not built-in a priori, but instead arises as an emergent property during learning.

When presented an utterance such as "look at the ball" and "what a nice ball" in combination with an abstract tag 'ball', LA attempts to relate the abstract tag 'ball' with an actual acoustic form that is common in both utterances. The tag is representing abstract information that would otherwise be available along other modalities. In a real child-carer interaction, a round object could be presented through the visual channel in parallel with the audio information. In designing this computational model, however, we want to avoid the engineering implementation of recognizers in the visual or tactile modalities, and concentrate on the linguistic speech-based part of the word discovery process. It is therefore assumed that non-audio information can be represented by meta-tags, presented along with the audio. These tags idealise the input from other modalities and translate to the presence or absence of vocabulary items in the audio stream. This approach is essentially different from the approach followed by Roy & Pentland (2002), who used image processing software and real images as input, and assumed the implausible availability of phone-like units to describe the audio signal directly from the start of the learning process.

A pictorial overview of the interaction between the learner and the carer is shown in figure 1a. The learner is depicted within the grey box, while the carer is indicated as the environment outside the box. An entire dialogue consists of a large number of interaction cycles, each cycle consisting of several turns. Per cycle, the learner receives multimodal input from the carer after which a reply is returned to the carer. In the next turn, the carer provides the learner with a feedback about the correct detection of the target word, after which it is up to the learner to use this feedback information.

3.2 The learner model (LA)

The learner consists of four basic ingredients:

- A memory (that is further divided into a sensory store, a short term memory and a long term memory). This includes a mechanism for storing and retrieving representations (figure 1a)
- A perception module, transforming signals from the environment into sensory data to be processed further
- A method to derive more abstract from less abstract representations. In the current model, this method is implemented by using matrix manipulations. Abstraction (bottom-up) is represented by matrix factorisation, while the inverse top-down process is represented by matrix multiplication.
- A learning drive: a target function specifying the basic need to learn. The target function consists of two terms: a term related to LA's internal need to learn, and one related to its external need to learn (see below).

When LA perceives input, the speech input is processed by the feature extraction module. The outcome is stored into the sensory store, from where it is copied to short-term memory (STM). In STM, a match takes place between the sensory information on the one hand and the stored representations on the other, and the best matching representation is looked for. The resulting representation is output and replied (sent back) to the carer in an abstract way (in the form of a tag associated with a 'confidence' measure, comparable to the manner in which a real baby uses a head turn to indicate its detection of a stimulus).

3.3 The carer model (CA)

The carer CA provides multimodal utterances to LA. The moment at which the carer speaks to the learner is determined by the messaging protocol. The utterances used during training and their ordering are determined by the carer. After a stimulus is presented to the learner, the learner performs a computation in which the goal is to formulate a reply. This will take some time. After a reply from the learner has been sent back to the carer, the carer provides feedback about the correctness of the reply. In the current implementation, the feedback is a yes/no (approval/disproval) variable. In the current set of experiments this feedback is not used by the learning system, which indicates that learning is possible without corrective feedback.

3.4 Learning drive

Infants require care, attention and energy in the form of food. A baby's drive to learn words is ultimately rooted in the desire to get to grips with its environment in order to have the basic needs for survival fulfilled and receive food, care and attention from the carers. For the current computational model, this boils down to the drive to make the carer understand and appreciate the learner's replies, which in turn boils down for the computational model to 'understand' as much as possible from the presented stimuli. The 'overall' need to learn is therefore based on two terms. The first term is related to the 'internal' need to understand the multimodal stimuli, that is, the ability to decode them in terms of known representations. The second term is connected to the communication between carer and learner and related to the desire to increase the perceived appreciation by the carer.

It is crucial to realise that the drive to understand the stimuli is bounded by the richness of the input – so learner and carer play a combined role in the learner's acquisition process.

Internal (intrinsic) factors:

- Intrinsic drive to understand the stimuli by reducing uncertainty and maximising predictability
- In the *model*: Quality of the parse and the fraction of the input explained by the stored representations

External (extrinsic) factors:

- Appreciation by the carer as perceived by the learner
- In the *model*: minimising error rates during the dialogue

Optimizing the appreciation of the carer is ultimately related to the optimisation of the accuracy of the replies given by LA during interaction between LA and CA. The optimisation of the accuracy can mathematically be expressed in terms of constraints on the minimisation between predicted reply (predicted by the learner model, in the current version based on non-negative matrix factorisation) and observed ground truth. This is done by combining acoustic and 'semantic' information into one single vector and applying non-negative matrix factorisation (NMF, Hoyer, 2004) on the augmented vectors (Stouten et al, accepted). The minimisation of the NMF cost function leads to the overall closest match between prediction and observation, and so to an overall minimisation of the recognition errors made by LA. Here, 'closest match' has a specific mathematical interpretation that depends on the precise cost function.

In the current implementation, the internal LA 'desire' to parse the stimuli can directly be related to the minimisation of the target function in NMF. The optimisation of the 'quality of the parse' can therefore be interpreted as a by-product of the optimisation of the externally-driven learning drive. This can be deduced from the NMF expressions. NMF attempts to factorize data matrix X into two smaller matrices W and H (X, W, H non-negative component-wise). In the case of minimisation of the Euclidean distance (that is, the Frobenius norm of the difference matrix) of the difference between X and WH, the cost function that is minimised reads

$$F(X, WH) = \tfrac{1}{2} \sum_{ij} (X_{ij}-[WH]_{ij})^2$$

while in case of the Kullback-Leibler divergence, this cost function reads

$$F(X, WH) = \sum_{ij} (X .* \log(X./WH) - X + WH)_{ij}$$

The structure of the expressions at the right-hand side indicates that in both cases the error between prediction and observation is an accumulated sum over all tokens that are processed during training. Splitting the 2-way sum in two separate one-way sums (over i and j, respectively), the terms $(X_j-[WH]_j)^2$ and $(X .* \log(X./WH) - X + WH)_j$ can be interpreted as the internal target function that is to be minimized on a token-by-token basis. In these token-

related expressions, X, WH and H are now column vectors rather than matrices; W itself is a matrix.

The second expression, related to the Kullback-Leibler distance between reference (X) and hypothesis (WH), can be regarded as the log-likelihood of the model (WH) predicting the observation (X). This, in turn, can be interpreted as a measure for the quality of the parse of the utterance associated to X, in terms of the words that are associated with the columns in matrix W.

The relation between the internal and external drive is loose, albeit in a mathematically well-defined manner. The applied mapping from the utterance to a column vector only loosely preserves the sequence structure in the utterance, in the sense that input patterns in which the acoustic objects are organised in a different order can lead to a column vector that is quite close to the original one. For instance, cyclic permutations will lead to minimal differences. Therefore, the interpretation of the parse is loose in exactly the same manner.

As a result, the learner attempts to explain from the presented stimuli as much as possible, given the constraints that are inherent in the set of presented inputs. Due to the combination of acoustic information and semantic information (Stouten et al., 2007), the resulting model stress is a measure for the fraction of the stimulus that can be explained by the learner. This also means that the eagerness to learn may result in a word spurt, of which the upper bound is determined by the richness of information available in the multimodal data presented to the learner.

3.5 Internal representations

Currently, internal representations are all vectors and matrices. At a high conceptual level, the process of abstraction is modelled as matrix factorisation. Factorisation of matrices is done by NMF.

NMF appears to be a very powerful tool for discovering structure in speech data. However, in the implementation we have been using so far we always need to transform blocks of data (i.e. blocks of [labelled] utterances). For this reason the results of the experiments are dependent on the way the training utterances are blocked. The order in which utterances occur in a block is immaterial, but across blocks, ordering is relevant. See also figure 1b.

3.6 Evaluation

We do not want to use word error rate ort concept error rate as the sole measure of learning performance. Instead, we also would like to investigate the internal representations, to see if these can speed up future learning. Therefore we will focus on the *effective reuse* of representations during learning. It appears that LA effectively reuses existing patterns during *testing*, as is clear in e.g. the multilingual experiment to be discussed below.

LA does not yet reuse (parts of) representations to build new representations during *training*. For example, one might expect the construction of the representation for a new compound

huisdeur to be facilitated by the existence of two representations for *huis* and *deur* [assuming of course these two are already trained]. One step more difficult for LA is to more rapidly build representations for '*taart*' if it already knows that '*paard*' and '*baard*' are different words.

In the experiments discussed below, the meta-tag associated with an utterance is always available during learning and it is always unambiguous. New experiments will investigate omitting the label in some proportion of the input utterances, and leave it up to LA to determine what to do.

LA makes use of three different memories: the sensory store, the STM (for storage and as executive buffer) and a long-term memory VTLM, in which all utterances and labels observed so far are stored. One would assume that part of the representations is static, sitting in long term memory. Activations caused by an incoming utterance are by definition short-lived, and much more likely to live (and die) in working memory. Right now, LA replies by providing a label in combination with an assessment value indicating how certain it is about the label. In the beginning of a training session, when LA does not have anything in memory, LA responds with nothing than 'huh?'. For each utterance, the activations are used for the formulation of the reply. In the current versions of LA, the activations are stored and kept in memory (VTLM).

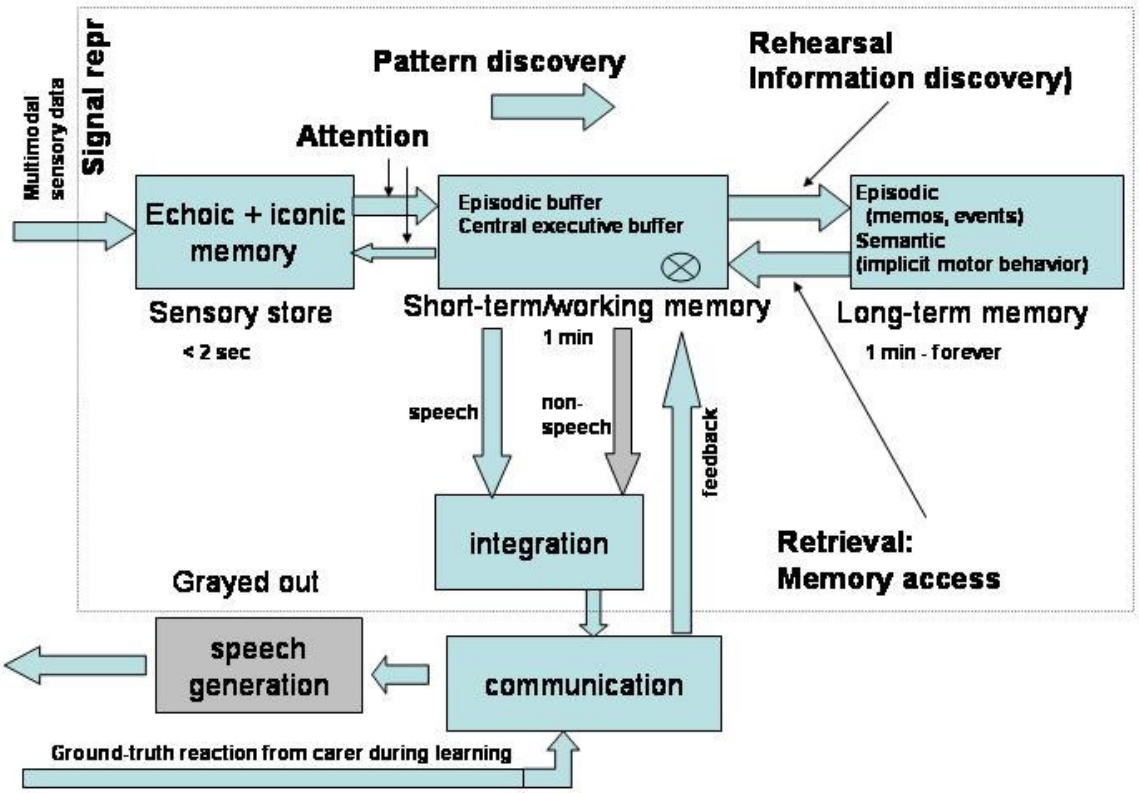

Figure 1a. *A picture of the overall interaction between learner model (within grey-line box) and the environment (carer, outside the box).*

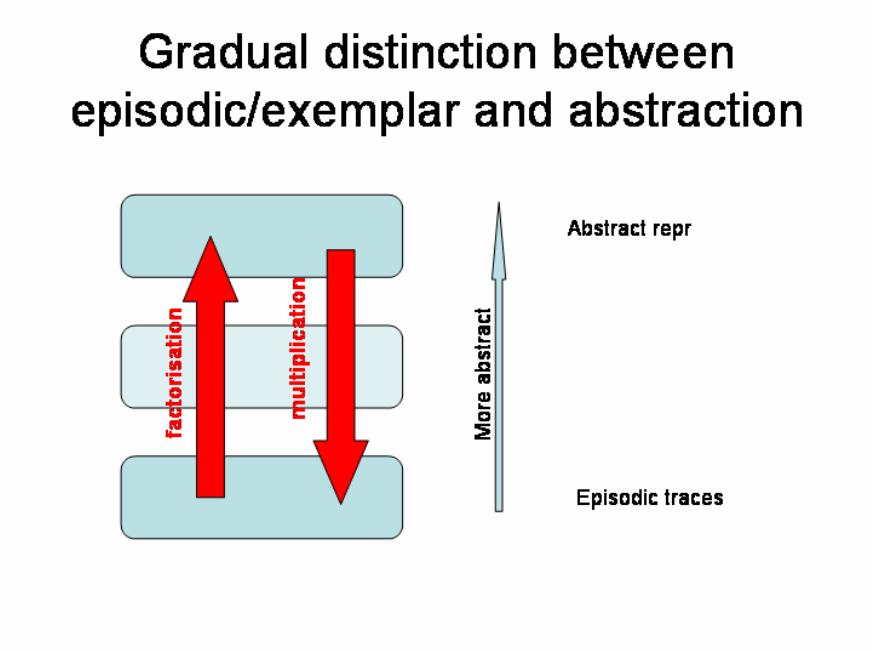

Figure 1b. *This figure shows a representation of the memory of the learner model LA. On the lowest level, data are represented in the most episodic form. Representations at this level include actual paths in the feature space spanned by the feature extraction module, as well as utterance-level vector characterizations of a fixed (large) dimension. The higher the level, the more abstract the corresponding representation becomes. The picture shows the general idea of having multiple different levels of abstraction. In the current computational model of LA, two levels are used, an episodic one, and an abstract one (basis vectors in a vector space representing words, in combination with activation strengths). The bottom-up process of abstraction is computationally translated into matrix factorisations, while the reverse bottom-up process is represented by matrix multiplications. Matrix multiplication will generate prototypical utterance-level representations, not actual paths. These top-down and bottom-up processes can interact in a natural way since they make use of the very same paradigm of algebraic matrix manipulation*.

4 Experiments

For experiments to be ecologically realistic a majority of all utterances in the first stage of the learning should come from one speaker (which then serves as the primary CA). In a normal family setting most of the remaining utterances would come from one other person. In addition, LA should hear a certain proportion on adult directed speech, which it should ideally ignore. The databases are and will be constructed in such a way that experiments investigating this mechanism can be carried out.

LA starts learning from the speech of the primary CA. Depending on the internal representations of the result of learning from the primary CA, learning to deal with speech

from other speakers might be accelerated (beyond some point in the development of LA). Apart from the learning curve itself, it is interesting to investigate the construction and activation of internal representations as a function of learning parameters and lexical variation of the multimodal input material.. The theory about episodic representations suggests that different representation may be formed for different speakers. Representations that conflate episodes pertaining to several speakers to adhere to the same semantic object may only form on higher levels in the hierarchy.

The experiments performed so far for the NL, FIN and SWE languages show that LA is able to do the following:

- -learn a limited set of concepts and classify a new stimulus in terms of one of these concepts
- -identify speaking style (instrumentally defined by the specs of the database), however not perfect
- rapidly adjust to a new speaker
- reuse already stored representations in testing
- reliably identify the speaker, almost perfectly, when speakers and speaking styles are presented in random order (but LA *may* use other coincidental artefacts in the database that might discriminate the speakers via recording condition, such as volume or spectral tilt. Unlikely but possible, and currently under investigation)

In the sequel, we will investigate the behaviour of LA in more detail.

4.1 Materials

Within the ACORNS project, three databases are available, one Dutch database (NL), a Finnish database (FIN), and a Swedish database (SW). A fourth database (UK) is under construction. Per language, each 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). The set of 1000 sentences contains 10 repetitions of combinations of 10 target words and 10 carrier sentences. (The content of the three databases differs in details that are not relevant for this discussion).

Within a database, not all target words are uniformly distributed. While all 4 speakers share the same target words, the proper name they use to address the learner is different per speaker. For example, the NL database (8000 utterances) contains 800 tokens of target words such as luier (diaper), auto (car), but only 200 tokens of each of the four proper names mirjam, isabel, damian, otto.

The database serves as a pool of stimuli that are used in all experiments that are discussed below. The ordering of the stimuli may differ from experiment to experiment. Ecologically relevant cases are:

- Random presentation
- Presentation speaker by speaker, random within speaker
- Presentation word-by-word, random within word (to investigate the creation of new representations at learner's side)

- Mixed effects, such as IDS/ADS distinctions

4.2 Short description of the experiments

A list of experiments is given below. The outcomes are discussed in the result section below.

Exp id	cochlea[1]	brief description of the experiment
001	**NL**	**test run, left out of discussion here**
002	**NL**	**NL data, random presentation**
003	**NL**	**NL data, blocked per speaker (Els, Henk, Margot, Peter)**
004	**NL**	**SWE data, blocked per speaker (Anna, Bjorn, Nancy, Olov)**
005	**NL**	**NL (Els, Henk), followed by SWE (Anna, Bjorn)**
006	**NL**	**NL, as 002, but blocked by word**
007	**NL**	**test run, left out here**
008	**NL**	**as 006, varying amounts of training, followed by same test**
009	**NL**	**FIN, randomly ordered data presentation**
010	**NL**	**as 002, NL, but also focussing in IDS versus ADS**

4.3 Experimental set-up

Per experiment, the entire training set is processed incrementally. The ordering in which the utterances are presented may differ and is one of the experimental parameters. Each *new* utterance is recognised from the acoustic input only and on the basis the stored representations of LA learned so far and excluding the current input. The internal model is updated on the basis of the multimodal information already observed during the training session. In the experiments reported here, the history-update window is infinitely long, that is, LA can reuse all previously observed utterances and tags.

4.4 Results

The result of each experiment is depicted in figures in which the horizontal x-axis represents the number of utterances presented during training. The vertical (y) axis presents the accuracy of LA's replies. The accuracy is defined as the number of correct responses (defined by comparing LA's reply with the ground truth in the multimodal stimulus by CA), divided by the total number of replies. The figures present two graphs, one representing the accuracy when measured from the beginning of the training run (indicated by the smooth curves in blue), the other one representing the accuracy over the most recent 50 utterances (spiky curve, in red). The latter representation is fair and appropriate if one would like to investigate the actual training process, while the former presents the accuracy comparable to an ASR-like measurement on the entire database.

[1] The model uses a codebook of spectral shapes. These were trained on Dutch (NL) data.

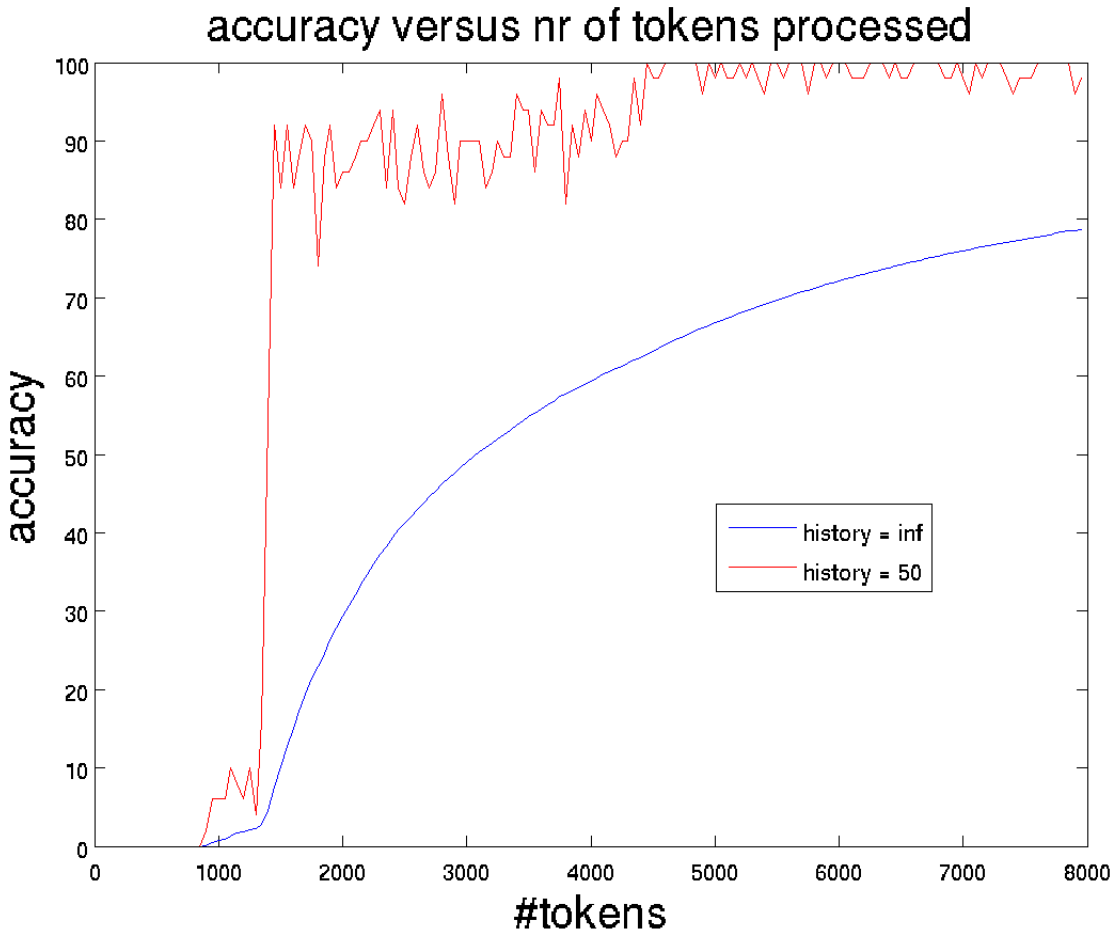

Figure 2. *Result of experiment 002 (Dutch, random ordering). The figure shows the concept accuracy (in percentage) as a function of the number of tokens in two ways (a) by showing the accuracy measured from the beginning of the training run (this moment is marked 0 on the horizontal axis). This plot is smoothly curved (in blue). (b) by showing the accuracy over the most recent fifty utterances (spiky curve, in red). Until x=4500, the internal vocabulary is not complete yet and lacks one single word.*

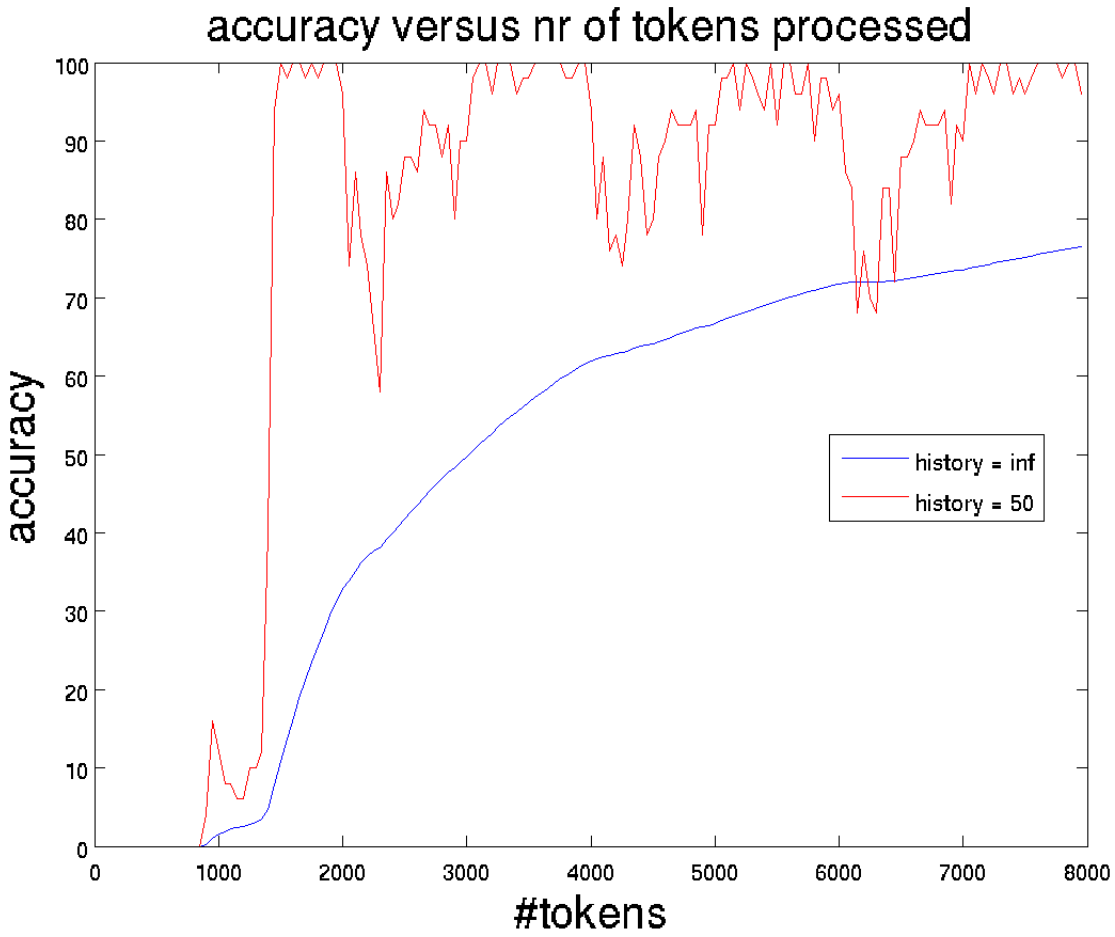

Figure 3. *Results of experiment 003 (Dutch, speaker-blocked). The axes are defined as in the previous figure. This experiment can be compared with experiment 002, the difference being that the multimodal data are now presented speaker-by-speaker. A drop in performance of about 20-30 percent is clearly visible every time when a new speaker starts (around #tokens = 0, 2000, 4000, 6000). Within about 1000 tokens (that is, approximately 100 tokens per word) the performance is back on its previous high level. The decrease in performance is due to two factors (a) each speaker introduces a new proper name, such that the vocabulary is not complete until at least 100 examples of the new word (and speaker) were offered (b) different speakers have different voice and speech characteristics which require an adaptation by the learning model.*

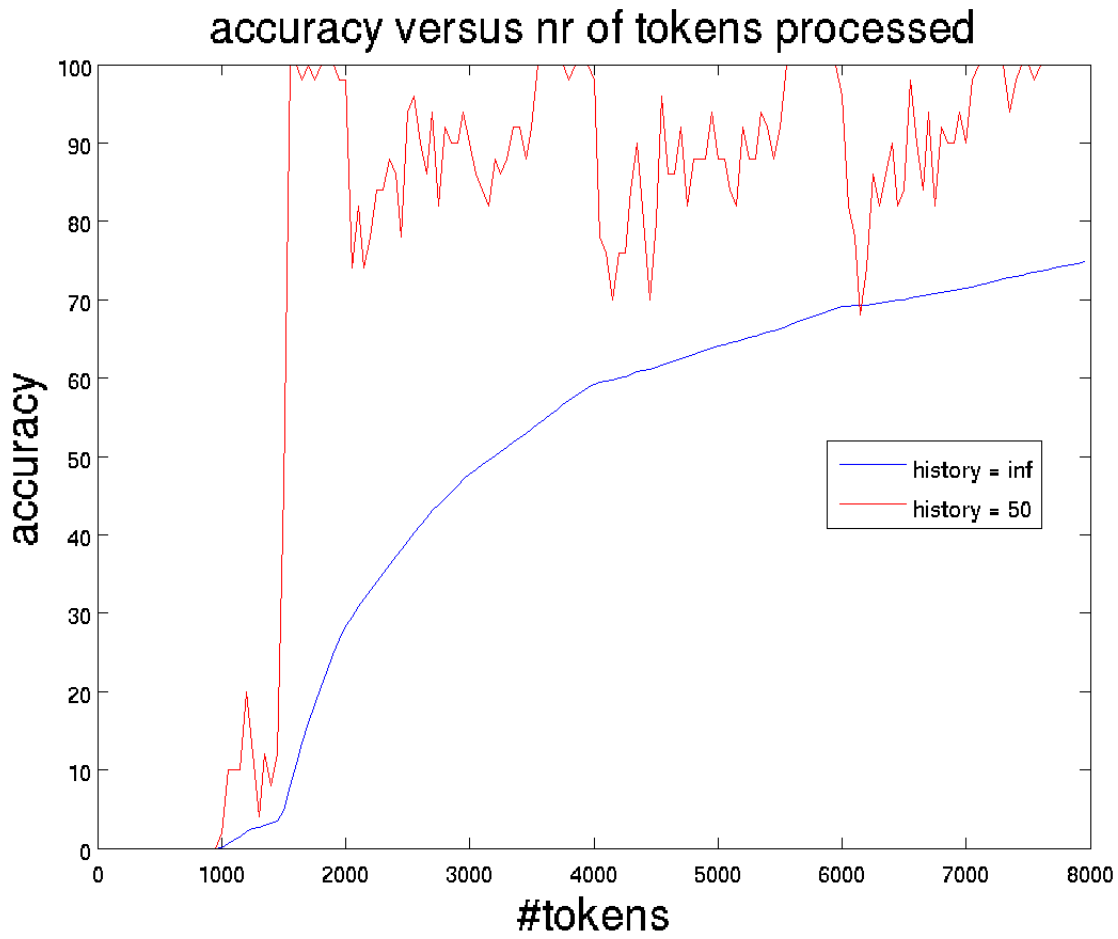

Figure 4. *Result of experiment 004 (Swedish, speaker-blocked). This figure presents the performance for Swedish. The data is presented speaker-by-speaker. The outcome can be directly compared to experiment 003 (Dutch). The eventual performance is about the same as for Dutch. The result of this experiment is remarkable for another reason: the 'cochlea' (i.e. codebook) used in this experiment is trained on Dutch utterances. This suggests that the codebook (used to encode the waveforms into an internal sensorial representation) is not very language sensitive.*

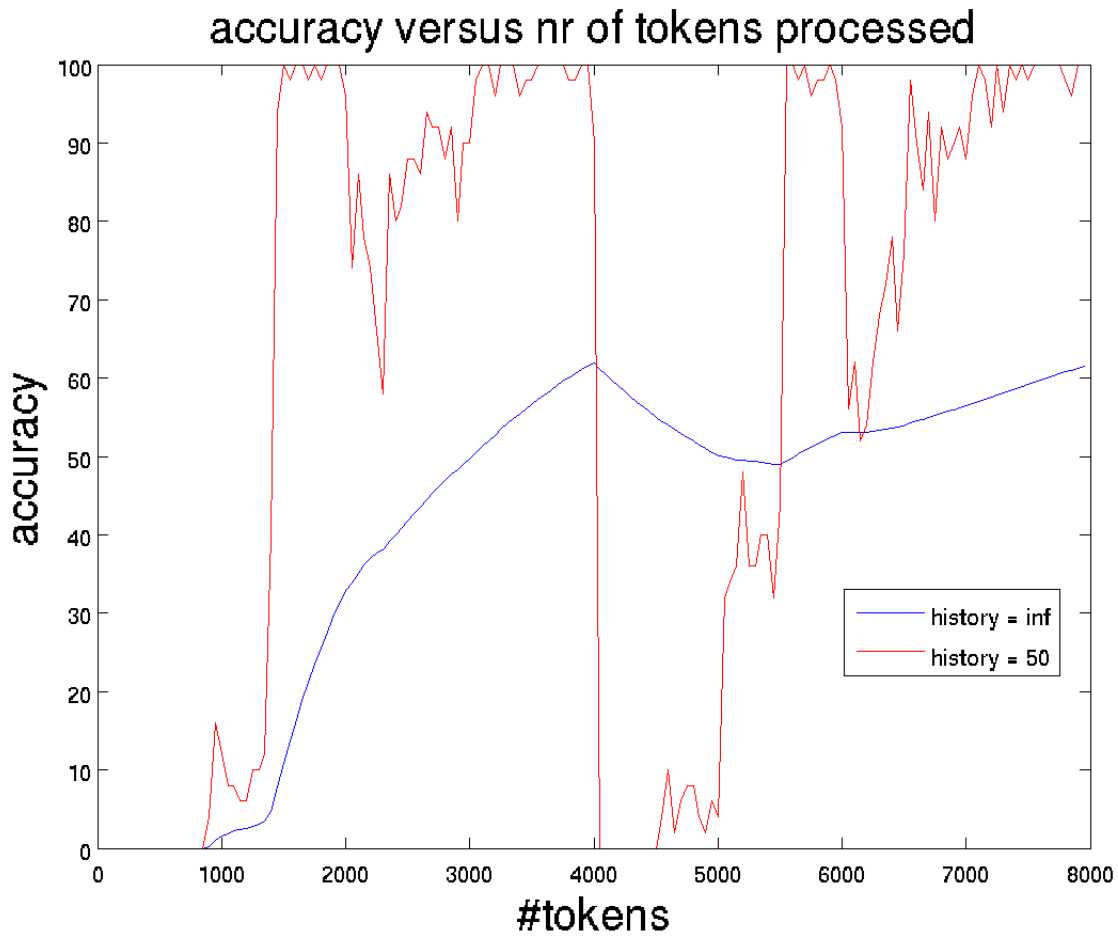

Figure 5. *Results of experiment 005 (multilingual, speaker-blocked). This figure presents results of a multilingual experiment, in which first two Dutch speakers are presented, after which two Swedish speakers follow. The speakers are Els (female), Henk (male), Anna (female) and Bjorn (male). Speaker changes drop the performance about 30 percent. A second language takes about the same learning time as the first language. The eventual model is able to both recognize Dutch and Swedish target words. Notice also that the total vocabulary involved is now doubled (the languages have separate tags). After 4000 utterances, the recognition rate drops to zero because none of the Swedish words is included in the vocabulary at that point in time.*

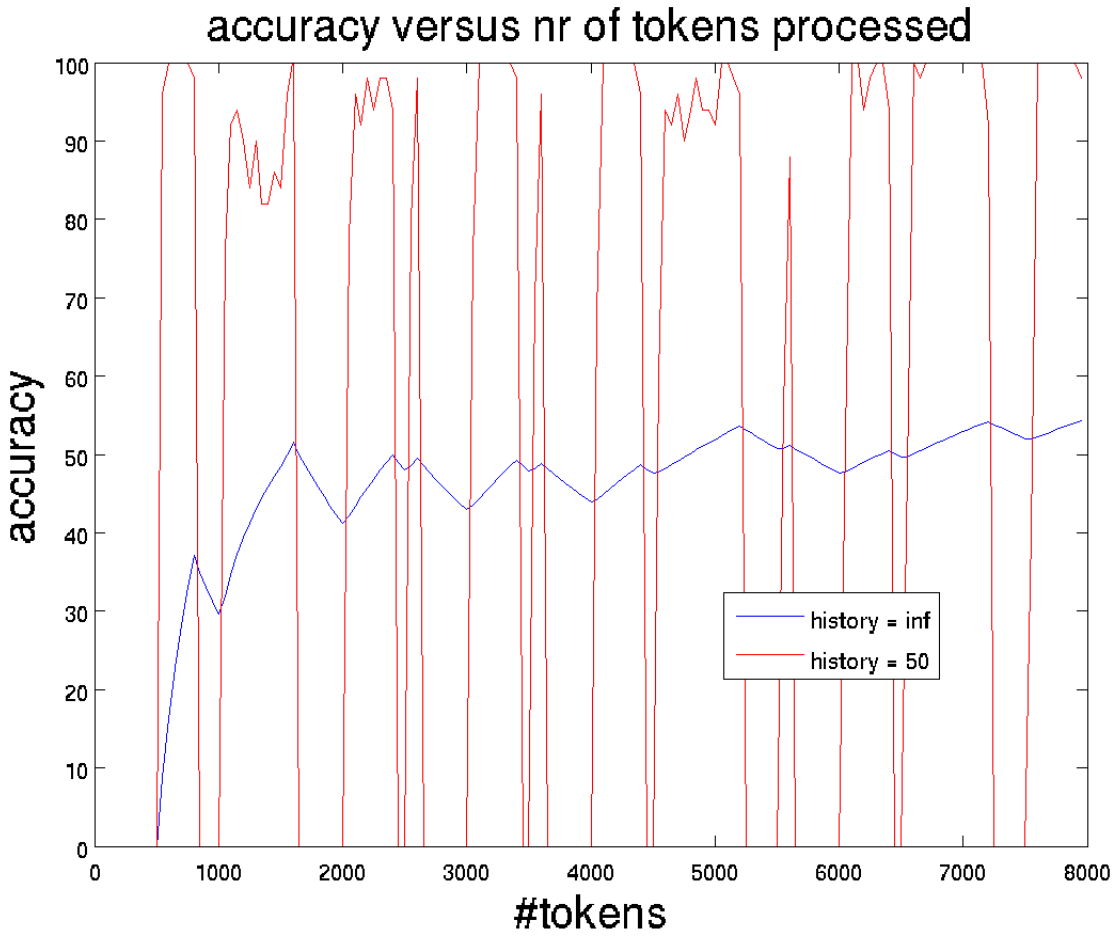

Figure. 6. *Results of experiment 006 (Dutch, word-blocked). This experiment is comparable to exp 002 (NL, random order) and 003 (NL, speaker-blocked). In this experiment, the data are presented word-blocked. While random order and speaker-blocked presentation gave average performance results close to 80 percent, word-block presentation shows a much poorer overall performance. This is a result of the specific manner in which the learner now decides to update its internal representations.*

Experiment 008 actually consists of a number of experiments. Each of these experiments is based on a full-blown training run. The single difference between these experiments is the exact amount of training tokens of *auto* (English *car*) at the *beginning* of each training run. One of the experiments, represented by the column '75' of table I, uses a training set in which first 75 tokens of *auto* appear, followed by all utterances with other target words (blocked by word). The final trail of 100 training utterances contains the target word *auto* again. So the list of utterances is (here N = 75)

- N utterances with *auto*, followed by
- utterances containing the other target words, followed by
- 100 utterances with *auto*.

All tokens of the target words are acoustically different. The first column in the table shows the recognition results for the 100 finally produced tokens *auto* in this particular case N=75. It appears that from the 100 tokens *auto*, none of these was correctly recognised.
The following column with head '100' presents the recognition results for the same set of 100 tokens of *auto*, but in the case where in the beginning of the training run 100 tokens have been presented (N = 100). In this case the recognition of *auto* is much improved: *auto* is recognized in 78 of the 100 test cases. The remaining columns show the behaviour for varying values of N, and thereby the effect of 'early training' on 'late test' for a couple of initial training sizes.

There is no evident process of forgetting: Forgetting could be shown by using a list of utterances with the structure

- a fixed number of utterances with *auto*, followed by
- a variable number of utterances containing the other target words, followed by
- a fixed number of (test) utterances with *auto*.

Forgetting is not explicitly modelled. Aspects similar to forgetting might result from competition between stored representations and a larger weight for those representations that receive more evidence from recently observed stimuli. Such effects are not explicitly modelled, but it cannot be excluded that these effects may play a role in certain cases.

Table I. This table shows the behaviour for varying sizes of training parts for *auto*, and thereby the effect of 'early training' on 'late test' for a couple of initial training sizes. There is no process of evident forgetting. For a description see the main text.

#tokens *auto* in beginning training set	75	100	125	150	250	350,450,550
100 test tokens recognised as:						
Auto	-	78	88	93	98	99
Bad	7	2	1	1	-	-
Boek	9	2	1	-	-	-
Damian	1	-	-	-	-	-
Isabel	7	2	1	1	1	-
Luier	6	2	3	1	-	-
Mama	4	-	-	-	-	-
Otto	35	11	5	4	1	1
Papa	17	3	1	-	-	-
Schoen	3	-	-	-	-	-
Telefoon	11	-	-	-	-	-
	100	100	100	100	100	100

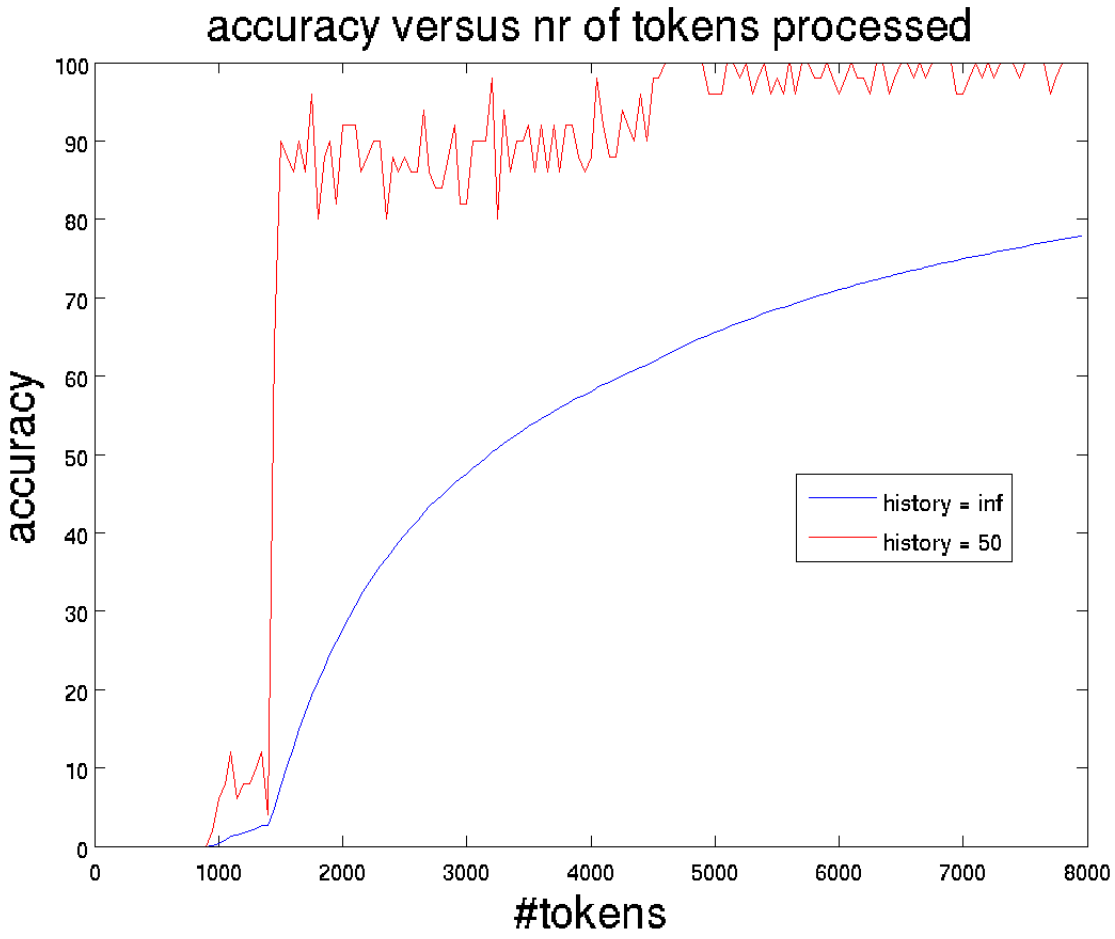

Figure 9[2]. *Results of experiment 009 (Finnish, random order). This experiment shows the result on the Finnish database. Data are presented in random order. This experiment can be compared with experiment 002 (Dutch data, random); the results for NL and FIN are nearly the same.*

[2] Figures 7 and 8 are omitted in this document, to maintain consistency with other ACORNS documents.

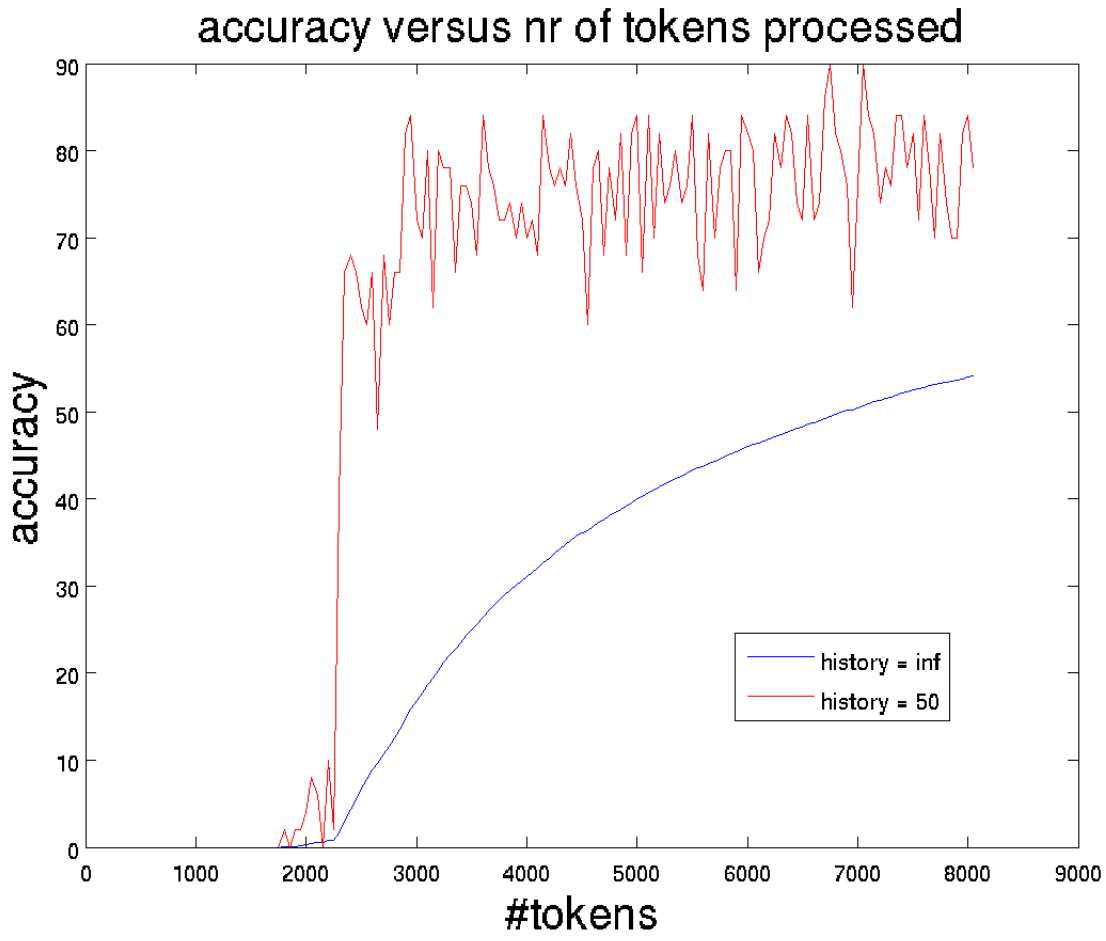

Figure 10. *Results of experiment 010 (Dutch, random order, LA trying to distinguish adult- and infant-directed speech). In contrast with experiment 002, the model must now detect words in combination with the speech style (IDS or ADS). If the speech style could not be detected at all, the IDS/ADS assignment would be random and the performance would theoretically drop to 50 percent. The actual performance of almost 80 percent shows a clear sensitivity of the model for speech style, on top of its ability to recognise the target words. This performance result must be based on abstraction since the number of tokens per style per word is too large for the model to represent individual tokens. The visual tags in this experiment are combinations of the genuine word tag (as used in e.g. exp 002) with the speaking style tag (IDS or ADS).*

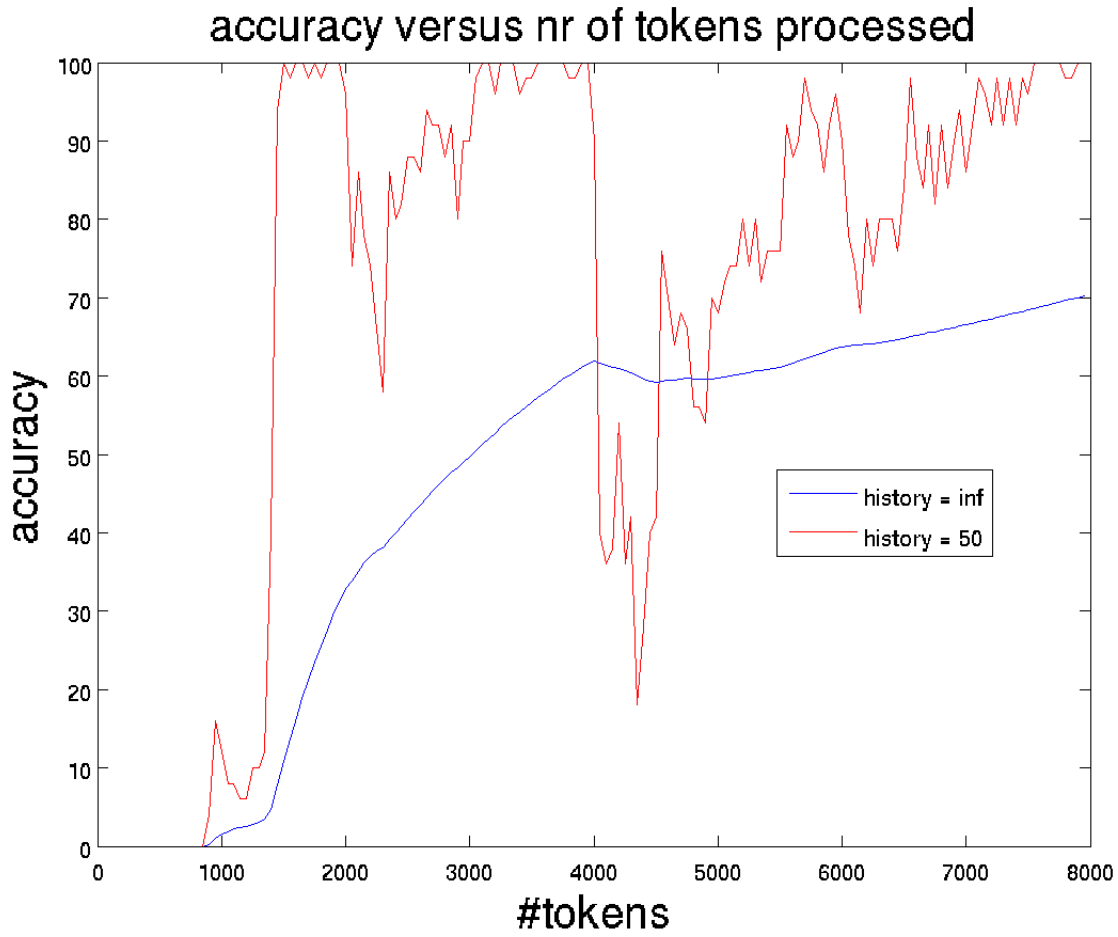

Figure 11. *This experiment (011) shows the effect of a bilingual experiment. First, two NL speakers are presented (female, male), followed by two Finnish speakers (female, male). Each speaker utters 2000 utterances. This experiment is very similar to exp 005 (figure 5), but one crucial difference: all original meta-tags in the NL and FIN database have now been replaced by a common set of tags that semantically make more sense: the new tags refer to the same extra-linguistic set of objects in a virtual scene, independent of language. Compared to experiment005, the leaner is able to reuse existing representations trained on the NL part for the Finnish utterances (this effect is visible between 4000 and 5000 utterances).*

5 Discussion

The computational model presented here shows that learning the relation between speech fragments and higher-order concepts can be established using a pattern discovery technique. The performance of the learner depends on a number of factors: the ordering of the data presented, the blocking per speaker and per word, changes of language. The effect of the ordering of the data, in particular of new speakers, is clearly visible in figure 3 (NL, speaker-blocked), when the performance is compared to the results is compared to random ordering (NL, random).

Although a preliminary result, the model shows that a second language is acquired with the same pace as a first language (exp 005, cf. fig. 5). We have investigated the behaviour of the

model just after the tokens of the second language are presented to the model. In this phase, during which the model did not yet adapt to the new language, the model clearly maps Finnish tokens unto Dutch representations as long it does not have created genuine Finnish representations. Once these are created, the performance goes up quickly (as is evidenced within the third quarter along the horizontal axis of figure 5). The multilingual experiment shows that already stored representations will be reused when necessary. When new data are presented, the learner can decide to create a new representation associated to a new concept. This is what happens in the second half of experiment 005, during which representations trained on Dutch are 'reused' to decode Swedish utterances. Only later in the training, when Swedish representations have been built, Dutch representations are not longer reused to decode Swedish utterances.

When monolingual data are presented in word-blocked ordering (experiment 006, figure 6), the model has overall more difficulty with the data compared to random-ordering (experiment 002, figure 2). Although the performance within one word-block rapidly increases to close to 95 percent, the overall performance (measured from the beginning) is substantially lower than in random mode. The assumption that will be investigated in subsequent work is that the main mechanisms of importance here are the criteria for triggering updates of the internal representations and the model of the non-relevant acoustics (carrier sentences). This means that for this particular computational model of word discovery, an equally evenly distributed presentation of tokens is favourable. (This does not say that the frequency of the target words must be uniform. As already observed, this is not the case in the databases.)

In the experiments presented, the input data are processed incrementally without forgetting, that is: each new datum is recognised with the current instantiation of available stored representations, and these stored representations get updated on the basis of knowledge about all previously observed utterances, without forgetting. It would be cognitively more plausible to improve the model in this aspect by introducing memory-dependent decays such that the retrieval from data in the past gets increasingly more difficult. Such a decay does not necessarily mean that these data will be entirely forgotten: relevant information could be accumulated in some other form and stored accordingly. Experiments with decay will be performed in follow-up experiments.

A common and interesting property of the word discovery approaches 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 utterance-dependent vector representation (a vector with a dimension independent of the utterance length or content). This mapping is structure-preserving in a very specific manner: The structure of an utterance (that is, the interpretation of an utterance as concatenation of lexical items) is transformed into the structure of a vector space (decomposability of a unknown vector as a weighted sum of given basis vectors). The experiments show that for speech decoding the actual segmentation of the input speech data is not required (nor in training, nor during decoding, nor as a necessary result of the decoding).

Since the model aims at word discovery with the findings about human language acquisition in a guiding role, its *cognitive plausibility* is one of the criteria along which the model can 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 specific word spurt between the age 1 and 2. This word spurt is generally attributed to cognitive factors: based on already trained representations of simple words, the learning of more difficult words gets increasingly faster. The McMurray model (2007) stipulates that the word spurt is a necessary effect of a combinatorial artefact. The 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 various 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. Figure 12 shows an example of this. It is obtained by running a Matlab implementation of the McMurray model and shows the number of stored representations versus number of word tokens observed over time (solid curve) for a particular chosen word frequency distribution and a word complexity distribution. The dashed curve represents the derivative of the solid curve, that is, the actual growth rate over time. The word spurt is the represented by the peak in this latter curve.

It is to be investigated to what extent word spurt can be modelled 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 more light on the relevance of various factors that are known to play a role in (models of) human speech processing. One of these factors deals with how words get activated (and to what extent), the second with the way how competition may act during the word search. In the current model, the activation of lexical items is separated from the actual competition. This is similar to what Shortlist does. 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 (as in TRACE) the entire lexicon, while in Shortlist the network in which competition plays a role is constructed from only those words supported by the input.

In the current model, competition is not explicitly implemented. It automatically emerges from the parallel search among multiple candidates. This is in line with earlier findings e.g. obtained with TRACE. TRACE showed that competition is not a necessary consequence of multiple processing in parallel. TRACE was implemented as a connectionist model using interactive activation techniques. Incoming input increased the activation of lexical candidates that it matched.

A crucial difference between this model and TRACE is the absence of inhibition. The more activation a candidate received in TRACE, the stronger the inhibition it could then exercise upon its rivals. Words which received ever more activation emitted ever stronger inhibition, and eventually the winning string of words should end up with higher activation than all competitor strings. In the current 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. In NMF, close competitors can completely inhibit each other. When two models have a common subvector which is activated, it can be explained by any combination of the two vectors. If in another subvector both models differ, the relative activation of each of the models will be determined by the data in this subvector.

But if one of the models already overestimates the values in this subvector, 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 weak but active mechanism of inhibition is in place.

The current model does not specifically focus on simulation of word activation patterns when the acoustic evidence unfolds over time.

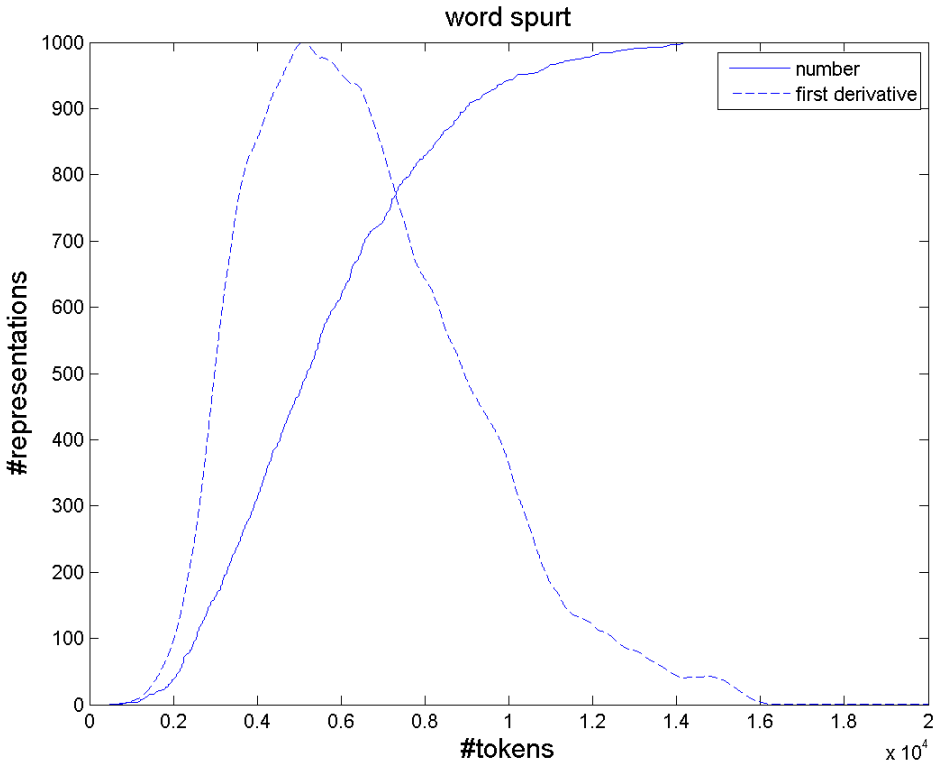

Fig 12. *Word spurt (lexical explosion). Simulation according to the model proposed by McMurray (2007). The word spurt is indicated by the dashed line. It shows the increase of number of representations as a function of observed tokens.*

Regarding the learning method, the model shows that non-negative matrix factorization allows the discovery of recurring acoustic patterns with an associated abstract tag. Uni-modal and cross-modal learning facilitates the separation of utterances into models of words or multiple words that best jointly explain the utterances. Unlike other unsupervised pattern discovery methods, a segmentation of the input in terms of the discovered units is not available and not required. During recognition, boundaries between word hypotheses are not aligned with the input, much like a human listener does not need to recognise all words to get the message or to recognize all phonemes to get the word. The absence of segmentation makes the detection of word onset, word order and word repetitions not trivial.

In the language acquisition literature, a few more characteristics of language learning are discussed of which the modelling will be a challenge for all models that are ultimately based on statistics. One of these characteristics is that babies need just a few examples to learn a new word. Apparently, a reliable representation can be built on the basis of a few tokens only. The current model is in principle able to do that, but to what extent this is dependent on other factors remains to be investigated.

The second characteristic of first language acquisition that we mention is a phenomenon referred to
as *fast mapping*. Fast mapping means that children learn that 'new' words refer to 'so far unobserved' objects. Apparently the formation of form-meaning pairs is a process that might be controlled by some primitive rules (in combination with statistically motivated updates of representations).

The third characteristic deals with the ability to perceive sound characteristics that do not support the distinctions of phonemes in the mother language. This ability gradually disappears from the age of 6 months. We believe that experiments involving minimal word pairs will reveal details about phone (and sub word) discovery as necessary step towards distinguishing ever more words.
The activation and deactivation of words with partial acoustic overlap with other words is another interesting characteristic of human speech processing. For example, for the stimulus 'inquiry', words like 'choir' may act among the candidates that are temporarily activated but get deactivated as time progresses. This phenomenon can be studied in detail by using specific groups of target words and to trace how word decoding results evolve during training. For example, by means of this and other experiments, SpeM has been proved to successfully model these aspects of human speech processing.

Another discussion issue is the exact meaning of the meta tag that associate the audio information. In the current database, this tag is theoretically an abstract representation of the object in the outside real world that the utterance relates to. In practice, it is a word that reflects this object in a linguistically neutral uninflected form (e.g. 'bok' in Swedish, even if this word occurs inflected in the utterance). However, in the cross-linguistic study, tags from both languages must be mapped onto 'supertags' that refer to objects in a virtual scene and go beyond the uninflected lexical sets of each language separately. So, in order to find a genuine tag system that is useful for multilingual experiments, the multimodal input should contain these supertags. How this must be done and what the effects are for the model is topic of current research.

Conclusion

We presented a computational model of word discovery as the first step in language acquisition. The word representations emerge during training without being specified a priori beforehand. Word-like entities are discovered without the necessity to first detect segmentations of sub-word entities.

Experiments show the behaviour of the model as a function of the ordering of the utterances during the training phase. The experiments provide clear indications about how the model can be improved for the next set of experiments on a database with a more complex linguistic

structure (such as a larger set of target words, more complex carrier sentences, and minimal word pairs).

The cognitive plausibility of the model (especially with respect to the use of memory and the reuse of already stored representations) is currently investigated in more detail.

Acknowledgements

This research was funded by the European Commission under contract FP6-034362 (ACORNS).

Website

The following web site provides updated info on lexical development (*communicative development inventories*) of infants of various ages between M8 and M16: http://www.sci.sdsu.edu/cdi/cdiwelcome.htm
This website also contains list of specific literature references for about 40 languages and language variants (including Dutch (Flemish), Swedish, Finnish).

References

Aslin RN, Saffran JR, Newport EL (1998). Computation of probability statistics by 8-month-old infants. Psychol Sci 9:321–324.

Boves, L., ten Bosch, L. and Moore, R. (2007). ACORNS - towards computational modeling of communication and recognition skills , in Proc. IEEE conference on cognitive informatics, pages 349-356, August 2007.

Cutler, A., Eisner, F., McQueen, J. M., and Norris, D. (in press) How abstract phonemic categories are necessary for coping with speaker-related variation, in C. Fougeron (Ed.), Papers from Laboratory Phonology 10. Mouton de Gruyter, Berlin, in press.

Friederici AD, Steinhauer K, Pfeifer E (2002). Brain signatures of artificial language processing: evidence challenging the critical period hypothesis. Proc Natl Acad Sci USA 99:529–534.

Gaskell, M. G. (2007). Statistical and connectionist models of speech perception and word recognition. In M. G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 55-69, Oxford University Press, Oxford, 2007.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, Vol. 105, 1998, 251-279.

Golestani, N., and Zatorre, R.J. (2004). Learning new sounds of speech: reallocation of neural substrates. NeuroImage 21:494–506.

Hashimoto, R. and Sakai, K.L. (2004). Learning letters in adulthood: direct visualization of cortical plasticity for forming a new link between orthography and phonology. Neuron 42:311–322.

Hoyer, P. O. (2004). Non-negative Matrix Factorization with Sparseness Constraints. Journal of Machine Learning Research 5 (2004). 1457–1469.

Johnson, E.K., and Jusczyk, P.W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. J Mem Lang 44:548–567.

Johnson, E.K., and Newport, E.L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. Cognit Psychol 21:60–99.

Kooijman, V. (2007). Continuous speech segmentation at the beginning of language acquisition. PhD thesis, 2007, Radboud University Nijmegen.

Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. Nat Rev Neurosci 5:831–843.

Lee, D., and Seung, H. (2001). Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems, vol. 13, pp. 556–562, 2001.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. Ear and Hearing, Vol. 19, 1998, pp. 1-36.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, Vol. 18, 1986, pp. 1-86.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science* 317(Aug. 3):631. Abstract available at http://www.sciencemag.org/cgi/content/abstract/317/5838/631.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In M. G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 37-53, Oxford University Press, Oxford, 2007.

Norris, D. and McQueen, J. M. (submitted). Shortlist B: A Bayesian model of continuous speech recognition. Submitted to Psychological Review.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, Vol. 52, 1994, pp. 189-234.

Park, A. and Glass, J. (2005). Towards unsupervised pattern discovery in speech, in Proc. ASRU, San Juan, Puerto Rico, 2005, pp. 53–58.

Pisoni, D. B. and Levi, S. V. (2007). Representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 3-18, Oxford University Press, Oxford, 2007.

Pitt, M.A., Myung, I. J. and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. Psychological Review, Vol. 109, 2002, pp. 472-491.

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

Saffran JR, Aslin RN, Newport EL (1996a). Statistical learning by 8-month-old infants. Science 274:1926–1928.

Saffran JR, Newport EL, Aslin RN (1996). Word segmentation: the role of distributional cues. J Mem Lang 35:606–621.

Saffran JR, Wilson DP (2003). From syllables to syntax: multilevel statistical learning by 12-month-old infants. Infancy 4:273–284.

Scharenborg O., Norris, D., ten Bosch, L., and McQueen, J. M. (2005). How should a speech recognizer work? Cognitive Science, Vol. 29, 2005, pp. 867-918.

Snow, C. and Ferguson, C. (1977). Talking to Children: language input and acquisition , Cambridge: Cambridge University Press, 1977.

Stouten, V., Demuynck, K. and Van hamme, H. (accepted) Discovering Phone Patterns in Spoken Utterances by Nonnegative Matrix Factorisation. IEEE Signal Processing Letters, 2007. Accepted for publication.

Thiessen ED, Saffran JR (2003) When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Dev Psychol 39:706–716.

Vroomen, J. and de Gelder, B., (1995). Metrical segmentation and lexical inhibition in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance, Vol. 21, 1995, pp. 98-108.

Zatorre RJ, Belin P, Penhune VB (2002). Structure and function of auditory cortex: music and speech. Trends Cogn Sci, 6:37–46.