BabyExp: Constructing a huge multimodal resource to acquire commonsense knowledge like children do

Massimo Poesio[1], Marco Baroni[1], Oswald Lanz[2], Alessandro Lenci[3], Alexandros Potamianos[4], Hinrich Schütze[5], Sabine Schulte im Walde[5], Luca Surian[1]

[1]University of Trento, [2]FBK, [3]University of Pisa, [4]Tech. Univ. of Crete, [5]University of Stuttgart
[1]Trento, Italy, [2]Trento, Italy, [3]Pisa, Italy, [4]Chania, Greece,[5]Stuttgart, Germany
massimo.poesio@unitn.it, marco.baroni@unitn.it, lanz@fbk.eu, alessandro.lenci@ling.unipi.it,
potam@telecom.tuc.gr, hs999@ifnlp.org, schulte@ims.uni-stuttgart.de, luca.surian@unitn.it

Abstract

The BabyExp project is collecting very dense audio and video recordings of the first 3 years of life of a baby. The corpus constructed in this way will be transcribed with automated techniques and made available to the research community. Moreover, techniques to extract commonsense conceptual knowledge incrementally from these multimodal data are also being explored within the project. The current paper describes BabyExp in general, and presents pilot studies on the feasability of the automated audio and video transcriptions.

1. Introduction

There is by now widespread agreement that the most realistic way to construct the large-scale commonsense knowledge repositories required by natural language and artificial intelligence applications is by letting machines learn such knowledge from large quantities of data, like humans do – see, e.g., Buitelaar and Cimiano (2008). A lot of attention has consequently been paid to the development of increasingly sophisticated machine learning algorithms for knowledge extraction. However, the nature of the input that humans are exposed to while learning commonsense knowledge has received much less attention. Thus, current knowledge extraction methods are mostly trained on huge amounts of raw text (e.g., from the Web or the Wikipedia), although this sort of input is hopelessly impoverished compared to the rich environmental stimuli available to humans when they learn about the world. For a variety of reasons – chief among which is the lack of appropriate resources – the majority of current work in this area must ignore the obvious consideration that a key part of human commonsense knowledge is acquired during childhood. Acquisition during childhood has three key aspects (Mandler, 2004):

1. *multimodal integration* – in human learning, nonverbal perceptual experience, and in particular visual experience, crucially complements – in fact, predates – verbal information, and has a dominant role in the acquisition of particular categories and aspects of our knowledge;

2. *incrementality* – human children are exposed to increasingly more varied stimuli as time and their learning capacities increase;

3. *full immersion in a "noisy" environment* – children learn how to carve knowledge not living in a controlled laboratory, in which stimuli are presented to them in a piecemeal and regular way. Instead, they are constantly immersed in an environment full of "noise", from which they learn how to distill the relevant pieces of information.

In this paper, we introduce BabyExp, a radically new kind of corpus that promises to be the first publicly available resource for training algorithms on input that is truly comparable to the one humans receive. The main aspect of originality of BabyExp consists in the fact that it will be the largest available resource of "ecological" data about the communicative and physical environment in which the baby is immersed during its first steps in concept and language learning. This will make BabyExp a unique resource among other existing datasets commonly used to model language and knowledge extraction. The latter, including also child-based corpora such as those in the CHILDES database (MacWhinney, 2000), are scarcely representative of the actual data used by the baby for concept learning for one or both of the following reasons:

1. they do not record the actual objects, events, etc., daily experienced by a baby;

2. they only provide a recording of a small sample of interactions (often recorded in controlled labs), actually a tiny fraction of the whole experiencing life of the baby.

The BabyExp corpus is based on continuous audio and video recordings of the full indoor waking hours of a single child in an English-speaking environment. Data collection started in September 2008 and it will end in August 2011, covering the first 3 years of life of the child under study. The audio and video streams will be automatically transcribed using state-of-the-art speech recognition and person and object recognition and attention tracking techniques. The resulting textually encoded corpus will capture not only the utterances heard by the child, but the trajectories and various visual properties of persons and objects surrounding the child, and that the child is paying attention to. The BabyExp corpus will be, as far as we know, the first resource of its sort that is open to the research community.
The BabyExp project is structured into the following main components:

1. Data collection in the child house;
2. Audio stream transcription;
3. Video stream transcription;
4. Corpus construction form the transcriptions;
5. Commonsense knowledge extraction from the corpus.

In this paper, after mentioning some related projects (section 2.), we describe the ongoing data collection initiative (section 3.), we report pilot studies in audio (section 4.) and video processing (section 5.), and we briefly discuss the motivation and general approach we intend to take to commonsense knowledge extraction (section 6.).

2. Related projects

Relevant literature that pertains to various aspects of the project is mentioned in the relevant sections of this paper. Here, we shortly review some related child data collection projects.

The seminal CHILDES project (`http://childes.psy.cmu.edu/`) has collected and transcribed naturalistic child-directed and child-produced speech since the mid eighties, and it also features video recordings of children. However, the fundamental goal of CHILDES is to collect data for the study of child language acquisition, and as such it does not provide (nor purports to provide) dense and continuous recordings of the child environment, of the sort we aim for. The densest corpora in the CHILDES database sample 2% of what a child hears, and just for short spans (Tomasello and Stahl, 2004).

Closer to BabyExp is the Human Speechome Project (HSP, `www.media.mit.edu/cogmac/projects/hsp.html`), that has recently concluded its first phase. BabyExp differs from HSP in several respects. HSP focused on data gathering, while the emphasis of BabyExp will be on the development of algorithms for learning from incrementally structured, multimodal input. We will take crucial advantage of important technical innovations in data collection, that will allow us to zero in on the relevant aspects of a scene while discarding the inessential, making the resource both more compact and of higher quality. The HSP data are not publicly available, whereas the full transcriptions of BabyExp data will be made available in standardized formats to the research community.

Another large scale child data collection initiative has been conducted by Infoture through the LenaPro wearable speech recording devices (`http://www.infoture.org/ProSystem/Overview.aspx`). This work - which has resulted in a large and growing collection of speech data from multiple children, controlled by parents, and only partially transcribed - again focuses on data collection only. We will go beyond Infoture by acquiring visual data as well as speech, and by adopting a strictly longitudinal, ultra-dense data collection methodology, instead of occasional recordings at times determined by parents.

Finally, several recent and ongoing projects in cognitive systems and robotics focus on robots that learn from their environment like children do. However, such projects (e.g., the ongoing CogX project: `http://cogx.eu/`) focus on self-learning robots that move in a relatively controlled environment, whereas we will use the BabyExp corpus to boost existing algorithms and methods for knowledge representation and extraction in NLP by feeding them with ecological multimodal data similar to those that lead a baby to acquire knowledge from the surrounding non-linguistic and linguistic input.

3. The BabyExp setup

The BabyExp corpus under construction is based on continuous audio and video recordings of the full indoor waking hours of a single child. Data collection started in September 2008 and it will end in August 2011, covering the first 3 years of life of the child of one of the researchers in the project (the child was born in August 2008, and data have been collected since his second month of life).

Data collection takes place in the researcher's apartment, with the collaboration of the researcher's wife (the mother of the child). Although the recording takes place in Italy, the baby is growing in an essentially monolingual English environment, in the English-speaking community of families and child-support structures of the University of Trento Center for Mind/Brian Sciences. The two rooms in which the child spends most of his time (child room and living room and kitchen area) are equipped with non-invasive cameras (LJD LJDNV15-101 FMC, 420 TVL resolution, infrared, CCD Sony SuperHAD), mounted at the 4 corners of the ceiling and with generic environmental microphones attached to one of the cameras. The recording equipment is turned on by a parent in the morning when the child wakes up, and turned off when the child goes to sleep and when the family goes out. Recording is also interrupted when it would pose privacy problems and in the presence of visitors that do not agree to take part in the study by signing the informed consent and privacy release forms (the parents can also, at any time, stop, rewind and watch previously recorded parts, and cut them). The recording equipment is controlled by an Apollo DVS server located in the living room of the apartment. The server stores the data temporarily in DAV (AVI-convertible) format. The recorded data are periodically transferred from the local server to a University of Trento server cluster, after the parents monitored them at high speed to filter out sensible data. Materials are transferred using a Portable Hard Disk approximately on a weekly basis.

As of March 2010, about 1.2 terabytes of raw data have been collected. Data collection and re-distribution after anonymization have been approved by the Ethical Committee of the University of Trento. Under the conditions of approval, we are allowed to share the full video and audio transcriptions with other researchers, as long as personal identifiers (proper nouns, locations) have been obfuscated to preserve anonymity.

4. Pilot study 1: Automated speech transcription

Recordings from a home environment contain a large variety of (often overlapping) signals that can be categorized

into: child speech, child-directed speech, adult-adult conversations, TV/Radio, noise or other sources. As a first step towards processing the audio data, we segment the audio and classify each segment into the aforementioned categories. Similar problems have been addressed in Xu et al. (2008) and Roy and Roy (2009). In Xu et al. (2008), home recordings from devices wearable by the child (ages 0-4 years) were segmented to categories similar to the ones proposed above. For this audio-only segmentation task, Gaussian Mixture Models with explicit duration modeling achieve 75% correct segmentation (at the frame level). In Roy and Roy (2009), the audio signal recorded via fixed microphones in a home environment was analyzed.

The most important cues for separating the audio data in the predetermined categories is short-time energy, followed by short-time spectral envelope and fundamental frequency. This is verified by the analysis in the following section. Given that, in our case, the position of the microphone is fixed, the distance between the speaker(s) and the microphone is variable and this significantly affects the short-time energy. Fortunately, the position of the speakers can be determined (with relative accuracy) from the multi-camera data.[1] However, for this preliminary analysis only the (unimodal) audio data were used to segment the audio stream.

4.1. Speech Analysis

As a first step we perform short-term analysis of the data to better understand the composition of the audio stream. In Fig. 1(a), the histogram of the short-term log energy is shown for a typical session (a full day of interaction). Analysis of short-time energy shows the following trends. High energy frames typically contain almost exclusively infant cries/crying (100-110 in the figure), high- to mid-energy frames (between 90-100) correspond mostly to infant vocalizations and child-directed adult speech (including motherese), mid-energy frames (70-90) contain mostly adult speech and low-energy frames (60-70) contain mostly noise/silence. Of course this distribution depends also on the distance between the speakers and the microphone.

Next, we choose to only analyze segments that have significant levels of energy. In order to obtain (relatively long) segments that have significant levels of energy we choose an energy threshold and then perform morphological closing and opening on the binary energy indicator function. The selected threshold based on the analysis above is 70, i.e., all frames with log energy higher than 70 are selected. This roughly corresponds to adult-adult speech, child-directed speech and infant phonations (including crying) frames. The three-phase operation (in sequence) for obtaining large continuous high-energy segments is as follows: i) do away with high-energy segments that are shorter than 50 ms (closing on the energy indicator function), ii) join together resulting high-energy segments that are at most 200 ms apart (opening), and iii) do away with resulting high-energy segments that are less than 1 second long (closing). The resulting segments contain 20-30% of the total number of frames for a typical session. The majority of the selected segments are a few seconds long (over 80%) and contain input from a single source (adult or child).

Next we perform pitch and formant tracking in the selected subset of the data. Our goal is to investigate the robustness of the existing speech analysis tools (for infant speech, especially) in this challenging acoustic environment. We have evaluated two pitch tracking algorithms: RAPT and the MDA pitch tracker. Next we report results for the MDA pitch tracker (Potamianos and Maragos, 1999) that performed better in this environment. The histogram of the estimated fundamental frequency for the selected segments is shown in Fig. 1(b). The distribution shows four peaks around 80 Hz, 125 Hz, 250 Hz, 350 Hz roughly corresponding to (two) male speakers, a female speaker (mother), and the infant, respectively. A careful analysis of the session shows a number of issues and processing difficulties that arise in this challenging environment. First, the infant distribution shows a long tail going all the way up to 475 Hz. This high-pitched sounds correspond to the baby crying. However, due to vocal fry, subharmonics are often present in infant's cry and the pitch tracking algorithm is often fooled into a pitch halving error (pitch halvings partially explain the asymmetry in the female voice distribution around 230-240 Hz). Another issue is the high-pitched sounds of the mother that often have a fundamental frequency of over 300 Hz. This is due to 'motherese', i.e., over-articulated speech with large peak-to-peak fundamental frequency variations and higher than average fundamental frequency. An example of such a sentence is shown in Fig. 1(c) where the mother utters: "Oh. Look at you". Note the jump in fundamental frequency in the beginning ("oh") and end ("you") of the utterance. Also the final word "you" is drawn out and over-articulated. Last but not least, the voiced-unvoiced decision is seriously impaired in this highly reverberant environment. The reverberated voiced speech is often of high-energy and masks unvoiced speech or silence leading to the labeling of a large percentage of frames as voiced. Overall, the speech analysis algorithms often fail due to the low signal to noise ratio (SNR) and, especially, due to the highly reverberant environment. Thus, existing pitch tracking algorithms have to be modified to cope with this challenging recording environment.

Next, we perform formant tracking on the high energy segments. Our goal is to investigate the robustness of formant estimation in the presence of noise, especially, for infant vocalizations. For this purpose we use the multiband demodulation formant tracking algorithm (Potamianos and Maragos, 1995). Overall, despite the noise and reverberation formant tracking works adequately well for adult speech, provided that the formants are above the noise floor. The MDA formant tracker had to be modified though for infant's speech (specifically the bandwidth of the Gabor filters were doubled) to avoid tracking harmonics as formant peaks. For both crying and infant vocalizations, the formant tracker was able to adequately track the first (F1) and second (F2) formant, with average values of 1500 Hz and 3200 Hz for F1 and F2 respectively. Higher formants were typically buried in the noise floor. In Fig. 1(d), the raw formant estimates are shown roughly corresponding to the first two formant tracks of an instance of the infant crying.

[1] Our goal is also to determine head/body pose in later stages in the analysis making it easier to determine the energy loss due to the distance between the speaker and the microphone.

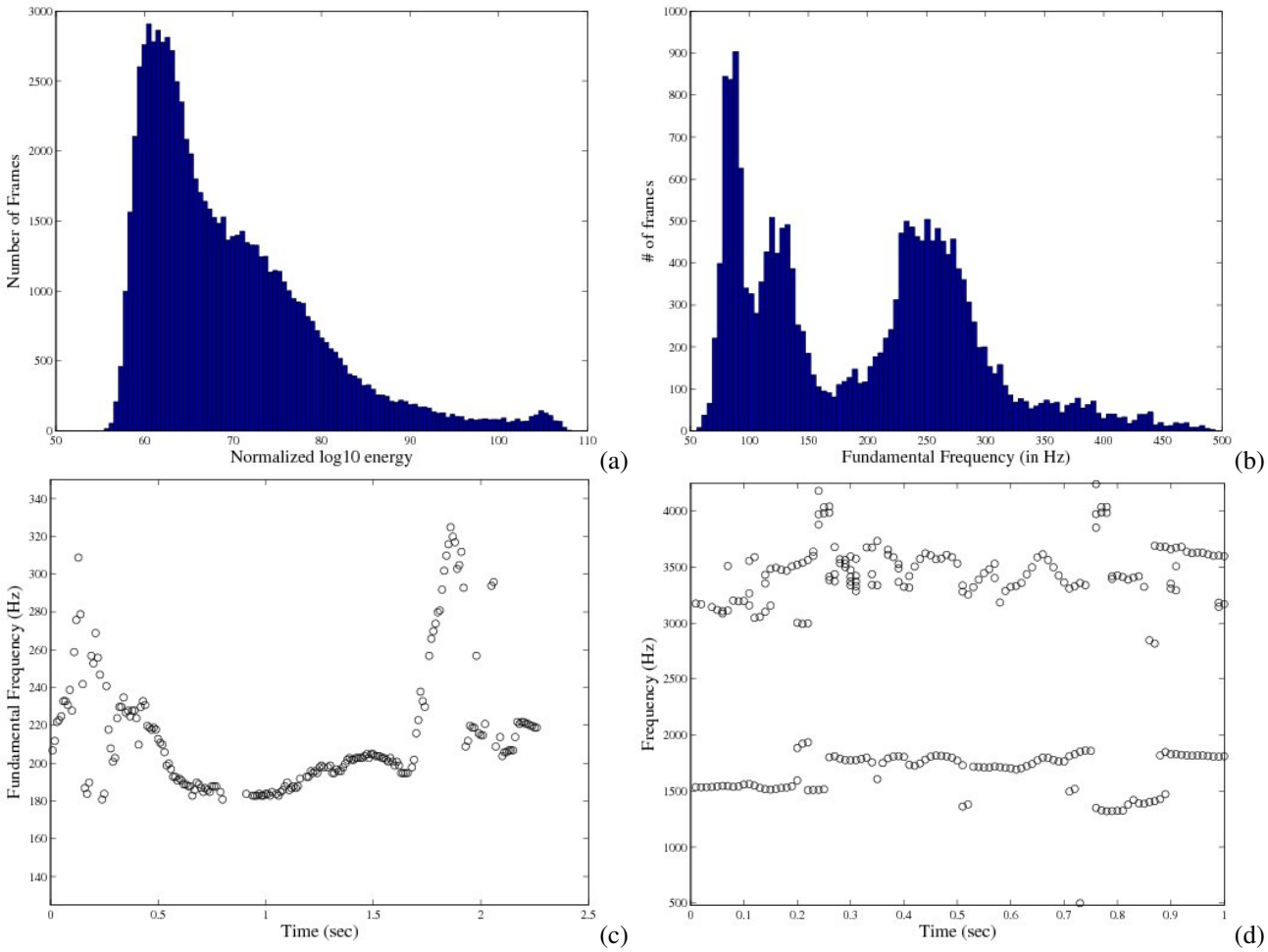

Figure 1: (a) Short-term energy histogram and (b) fundamental frequency histogram for a typical session. (c) Fundamental frequency contour (raw estimates) for the sentences "Oh. Look at you" (in motherese). (d) Formant frequency raw estimates for an example of the infant crying.

Note the discontinuities in the formant estimates that are due to the spectral estimate locking on harmonic frequencies. Overall, although there are issues in formant tracking for infant vocalization, the MDA formant tracker works adequately well.

4.2. Speech segmentation and speech recognition

Next we attempt to label the selected high-energy segments as infant speech, child-directed adult speech, and adult-adult conversations.[2] In order to classify segments we used the following features: energy (max, min, mean) and average fundamental frequency of a segment. Preliminary experiments on a single session of the audio data have shown that these two features alone can achieve accuracy of over 75%. This is to be expected from the speech analysis results above: infant speech is usually both high-energy and high-pitched, child-directed (motherese) speech is also high-pitched and high-energy (but less so that infant vocalizations), while adult-adult conversations are in the normal frequency register and have lower energy. We expect these results to improve when smooth spectral envelope features are incorporated (e.g., mel-frequency cepstrum coefficients). Further improvements are expected when the multimodal information (people detection from the video) is incorporated.

In this preliminary study, we perform pilot speech recognition experiments on child-directed speech and infant vocalizations. Speech recognition of adult-adult conversations is beyond the capabilities of today's ASR technology, due to the very low SNR and highly reverberant environment. Most of the adult-adult speech information is hidden in noise; in fact, transcription of these conversations is challenging even for human listeners.

Preliminary experiments of automatic speech recognition (ASR) of child-directed speech have shown a number of issues. First of all, the recording conditions are challenging. ASR using far-field microphones in low SNR reverberant environment is still an open-research problem.[3] In addition,

[2]Note that some segments (especially longer ones) contain multiple turns. However, for the purposes of this simple analysis we only determine a single label per segment. In future work, a Bayesian Information Criterion (BIC) may be used to further separate longer segments into turns see, e.g., Zhou and Hansen (2005).

[3]Fortunately, the SNR is adequately high to perform ASR experiments (significantly higher than adult-adult conversations).

child-directed speech (motherese) is over-articulated and contains large pitch variations that acoustic models trained on generic data cannot handle. Finally, the vocabulary and language used is different from that in adult-adult conversations or read speech. All these factors lead to poor word recognition performance (word accuracy well below 40%). Due to the large amount of data available, there is the possibility of retraining (or adaptation) of the acoustic and language models that can lead to large improvements in performance. Overall, a significant research effort has to be invested for the accurate automatic transcription of motherese. Note that child-directed speech is the most important verbal data source in constructing the BabyExp corpus.

Finally, as far as infant speech is concerned, it can be roughly grouped into four categories (roughly corresponding to the four speech development phases): i) initially crying and cries are the main way of communication, ii) next come quasi-vowels, squeals, growls and protophones, iii) then come babbles and syllables, and finally iv) words, phrases and sentences (Xu et al., 2008). Although the first words (typically consisting of two syllables, e.g., "baba", "dada") might appear around 12 months of age, the ability to form consistently phrases and sentences emerges around 24 months. In order to obtain robust and usable results, speech recognition of infant's speech will initially focus on vowel recognition, in order to automatic transcribe the phonetic F1-F2 vowel space as a function of age. We will also investigate the accuracy of word-spotting speech recognition technology for babbles and common words. Based on the preliminary results from formant tracking (see previous section) this mapping should be feasible.

5. Pilot study 2: Automated video transcription

A continuous transcription of the visual information the child has access to will be acquired from video recordings by means of i) tracking the spatial position and head orientation of the baby, ii) the position of the adults the baby interacts with, and iii) detection and localization of objects of interest to common sense acquisition. The visual focus of attention of the baby can then be logged in a post-processing step, by intersecting the viewing cone of the baby (rooted at the head position and oriented according to the estimated head pan and tilt) with the trajectories of the adults/objects collected in its surrounding environment. Extracting such information from far-field recordings in a home environment (BabyExp uses cameras mounted at the four corners of the ceiling of the apartment rooms) poses challenges in terms of low resolution (facial or object features may not be visible in the images), uneven lighting conditions (the same color may appear brighter near a window), and high ambiguity in the visual characterization of the baby (no hairs, non yet developed facial features, complex shapes and poses).

We adopt a particle filter based approach for head pose (Lanz and Brunelli, 2006) and multiple people tracking (Lanz, 2006). Such approaches have proven to work well in complex, but controlled settings with adults, where they achieve state-of-the-art results with less than 15cm and 25deg average tracking error even under significant occlusions (CLEAR evaluation 2006, 2007). By integrating work-in-progress on estimating color distortions due to uneven lighting, we expect that they scale to a home environment. Object detection and localization is instead addressed with a memory-based approach, i.e., using a set of visual exemplars for each object or object class to be detected.

Given the overall complexity of the task, in this paper we report first results on a properly designed experimental setup that matches the constraints of the adopted technologies at their current state but at the same time provides a proof-of-concept of the video transcription pipeline. We have acquired a multi-camera video sequence of about 10min in a laboratory equipped with four firewire cameras installed in the corners of the room at a height of about 2.7m. The cameras deliver jpeg-compressed RGB images at 15 frames per second with a resolution of 512×384 pixel. The sequence shows two children and an adult entering a room where a ball is placed on the floor. After a few seconds the younger child discovers the ball and invites the others to play with her. After two minutes the adult puts the ball on a chair and the children start to look around in the room. A minute later they start again playing, then the adult pictures a face on the ball. The older child looses his interest and goes to the PC desktop while the others start playing again after a while and he gets back to them to play again. Aim of the pilot study was to automatically transcribe the behaviour of the younger child (6 years age) and, in particular, its visual focus of attention, from the low resolution video footage. Fig. 5. shows the results on the sequence obtained with the *SmarTrack* system (`http://tev.fbk.eu/smartrack`) implementing Lanz (2006) and Lanz and Brunelli (2006).

The system tracks the 3D position of the ball as an attention object, as well as the movements of the adult and the older child (for each such target a color model has been extracted manually from the sequence which was then used for tracking). For the study subject, i.e. the younger child, it estimates the location and head and torso orientations. This information is then used to log i) the proximity of the study subject to the attention object (using the spatial distance between subject and target), and ii) an indicator for its visual attention towards the target (using the angular offset between the sight direction and the subject-to-target line). From the plot in Fig 5. it is easy to conclude when the child is paying attention to the ball. In the BabyExp setting, these two indicators will be exploited to infer when linguistic input about a target object is accompanied by appropriate visual experience of the baby, and thus is expected to be most effective, e.g., to fix the reference of a word or to extract properties that will enter into the baby's semantic representation for that word.

6. Commonsense knowledge extraction

The main goal of the BabyExp consortium is the development of algorithms that exploit the innovative properties of the BabyExp corpus. The BabyExp approach rests on two main assumptions, strongly supported by developmental psychology. One is that adult commonsense knowledge derives (in different degrees) from both *perceptual input*, i.e., our sensory-motor experiences with entities in the world,

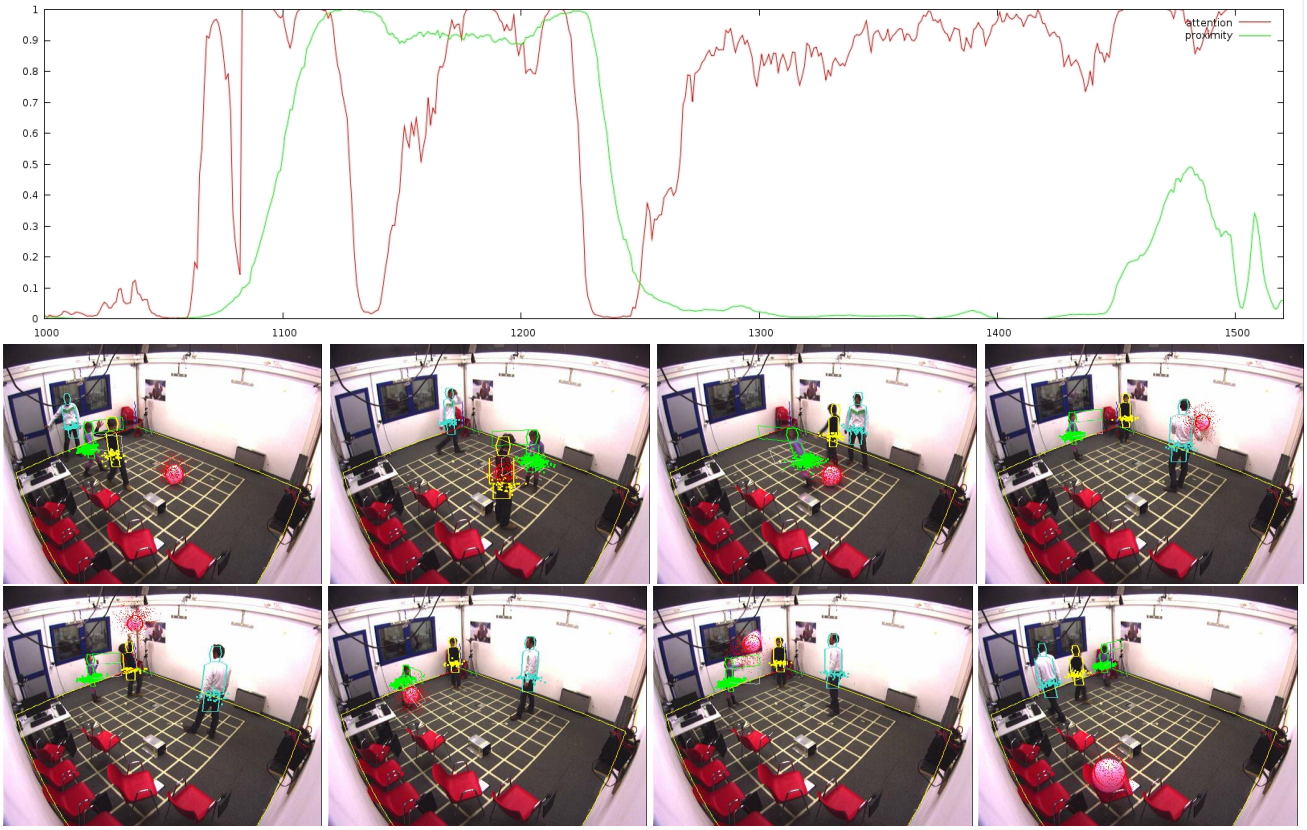

Figure 2: Automated video transcription of the pilot study sequence: visual focus of attention and proximity of the study subject towards the attention object over time (520 frames ≈ 35sec), and raw output of SmarTrack on frames 1085, 1160, 1230, 1370, 1535, 1785, 1860, 2240. The real time location and head orientation estimates of the study target are overlaid in green, and the locations of the attention object (the pink ball) and the actors are shown in red, yellow and blue.

and *linguistic input*, i.e., the information that we can extract from the linguistic structures used to talk about the world. These two knowledge sources play differential roles during conceptual development: Perceptual input dominates in the early stages of commonsense acquisition, but it is then rapidly integrated with other learning strategies that increasingly rely on information extracted from the linguistic input (Bloom, 2000). The second assumption is that children develop in time their ability to extract relevant information from the perceptual and linguistic environment and to use it for commonsense learning. As their cognitive maturation proceeds, they acquire more and more sophisticated abilities to exploit the input they receive. For instance, their still immature visual and attention systems limit and condition the types of conceptual categories they build in early stages (Mandler, 2004). When such limits are progressively overcome, deeper analyses of visual scenes take place allowing children to get at more fine-grained conceptual distinctions.

A multi-stage approach to commonsense learning will be adopted, to exploit incremental aspects of the information extracted from the BabyExp corpus. In each stage, we will learn increasingly complex pieces of commonsense knowledge. The computational models for more advanced learning stages will use the previously acquired knowledge as prior for new concept extraction. BabyExp intends to develop computational models able to learn commonsense knowledge by integrating experiential data extracted from visual recordings according to the methods that we exemplified in section 5., and data coming from shallow distributional analysis of the linguistic input, transcribed according to the methods presented in section 4., and then annotated at the morphosyntactic and syntactic level. The key factor here is the integration of distributional and visual information, under the assumption that distributional learning does not substitute experiential learning but complements it. With this goal in mind, we will expand various current algorithms for distributional semantics in computational psychology and linguistics, such as Semantic Vector Spaces (Landauer and Dumais, 1997; Lin, 1998; Padó and Lapata, 2007) and Bayesian models (Griffiths et al., 2007; Andrews et al., 2009).

7. Conclusion

The feasibility studies we have reported in this paper are a first, concrete step towards the possibility to transcribe the audio stream and significant parts of the video stream surrounding the baby by automated means. While automated multimodal transcription is at the present stage the main challenge faced by the project, we are also tackling the issue of the transcription target format, that should be such as to encourage and facilitate computational work that will use the corpus to incrementally mine commonsense knowledge from integrated linguistic and visual cues. Algorithms to

extract knowledge from developmental data are also being explored.

While the BabyExp project is clearly at a very preliminary stage, and some of the difficulties we face require bringing forward the state of the art in various fields (see for example the problems with identifying and transcribing child-directed speech), if the project succeeds, we hope that it will contribute a groundbreaking resource both for knowledge extraction for intelligent applications, as well as to understand the process of knowledge acquisition in human beings.

8. References

M. Andrews, G. Vigliocco, and D. Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

P. Bloom. 2000. *How Children Learn the Meanings of Words*. MIT Press.

P. Buitelaar and P. Cimiano, editors. 2008. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. IOS Press.

T. Griffiths, M. Steyvers, and J. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114:211–244.

Th. Landauer and S. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

O. Lanz and R. Brunelli. 2006. Dynamic head location and pose from video. In *IEEE International Conference on Multisensor Fusion and Integration*, pages 47–52.

O. Lanz. 2006. Approximate Bayesian multibody tracking. *IEEE Pattern Anal. Mach. Intell.*, 28(9):1436–1449, September.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, pages 768–774, Montreal, Canada.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk. 3d ed.* MIT Press, Mahwah, NJ.

J.M. Mandler. 2004. *The Foundations of Mind*. OUP.

S. Padó and M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

A. Potamianos and P. Maragos. 1995. Speech formant frequency and bandwidth tracking using multiband energy demodulation. In *ICASSP*, Detroit, MI, May.

A. Potamianos and P. Maragos. 1999. Speech analysis and synthesis using an AM–FM modulation model. *Speech Communication*, 28:195–209, July.

B.C. Roy and D.K. Roy. 2009. Fast transcription of unstructured audio recordings. In *Interspeech*, Brighton, UK.

M. Tomasello and D. Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31:101–121.

D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen. 2008. Signal processing for young child speech language development. In *Workshop on Child, Computer and Interaction*, Chania, Greece.

B. Zhou and J. Hansen. 2005. Efficient audio stream segmentation via the combined T2 statistics and Bayesian Information Criterion. *IEEE Speech and Audio Processing*, 13(4).