A Speech Corpus for Modeling Language Acquisition: CAREGIVER

T. Altosaar[1], L. ten Bosch[2], G. Aimetti[3], C. Koniaris[4], K. Demuynck[5], H. van den Heuvel[2]

[1]Aalto Univ. School of Science and Tech., Dept. of Signal Proc. & Acoustics, P.O. Box 3000, FI-02015 TKK, Finland
[2]Radboud University Nijmegen, Language and Speech unit, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
[3]Univ. of Sheffield, Speech & Hearing group, Dept. of Computer Science, 211 Portobello Street, Sheffield, S1 4DP, UK
[4]School of Electrical Eng., Sound and Image Processing Lab, KTH, Osquldas väg 10, SE-100 44 Stockholm, Sweden
[5]K.U.Leuven - ESAT/PSI, Kasteelpark Arenberg 10 bus 2441, B-3001 Heverlee, Belgium
E-mail: Toomas.Altosaar@gmail.com, {l.tenbosch, h.vandenheuvel}@let.ru.nl, g.aimetti@sheffield.ac.uk, koniaris@ee.kth.se, Kris.Demuynck@esat.kuleuven.be, h.vandenheuvel@let.ru.nl

Abstract

A multi-lingual speech corpus used for modeling language acquisition called CAREGIVER has been designed and recorded within the framework of the EU funded Acquisition of Communication and Recognition Skills (ACORNS) project. The paper describes the motivation behind the corpus and its design by relying on current knowledge regarding infant language acquisition. Instead of recording infants and children, the voices of their primary and secondary caregivers were captured in both infant-directed and adult-directed speech modes over four languages in a read speech manner. The challenges and methods applied to obtain similar prompts in terms of complexity and semantics across different languages, as well as the normalized recording procedures employed at different locations, is covered. The corpus contains nearly 66000 utterance based audio files spoken over a two-year period by 17 male and 17 female native speakers of Dutch, English, Finnish, and Swedish. An orthographical transcription is available for every utterance. Also, time-aligned word and phone annotations for many of the sub-corpora also exist. The CAREGIVER corpus will be published via ELRA.

1. Introduction

Building and testing computational models of the speech understanding component of first language acquisition requires the availability of relevant speech corpora. While it can be argued that such corpora are available in abundance (e.g., the CHILDES database (MacWhinney, 2000)) it appears that there are few –if any– corpora that are suitable for the task. Most of the existing corpora focus on speech produced by babies, rather than on the speech of the caregivers. In addition, most existing corpora only provide verbatim or a phonetic transcription of the speech, rather than the speech signals proper. For a computational model of language acquisition that is embodied in the sense that learning proceeds on the basis of actual speech produced in a specific situational context, access to the speech and some representation of the context are indispensable.

ACORNS was a three-year Future and Emerging Technology project that aimed at building a computational model of language acquisition that would avoid the frame-of-reference error that is well known in Artificial Intelligence (i.e., modeling a meta-level description of the output of some process instead of the process proper) (ten Bosch et al., 2009, www.acorns-project.org). For this purpose a corpus was needed that would contain utterances that approximate the speech that caregivers address to babies. Also, for each utterance an accompanying representation of a situational context would be required. With such a corpus a computational model should be able to acquire word-like units, by making use of the repetitions in the speech modality in combination with cross model associations (refer to Newman, 2008, and Smith & Yu, 2008), in other words, learn associations between portions of speech utterances and visual representations of the situational context. To make sure that the computational models can learn different languages on the basis of the same, language independent assumptions related to cognitively plausible processing and representations, the corpus should contain utterances in several different languages referring to the same scene.

2. Corpus Design

Since infants learn incrementally, we designed a corpus that contains a fair proportion of 'simple' utterances, i.e., utterances with a simple syntactic structure and referring to a single well-defined object in the environment. In addition, we included utterances that, despite having a relatively simple syntactic structure, still contained references to up to two objects, each with one or two relevant properties (shape, color, etc.).

In real life, infants hear very similar utterances over and over again spoken by their caregivers, and it is reasonable to assume that this repetitiousness is a necessary requirement for discovering systematic associations between sounds and references (Smith & Yu, 2008). For this reason it was decided to record all utterances several times spoken by the same speakers. Moreover, a large proportion of the speech that is addressed to babies is characterized by a slow speaking rate combined with exaggerated intonation, and it is not known to what extent the special characteristics of this 'Infant Directed Speech' are important in helping infants to discover associations between speech and referents. For this reason it was decided to record the simple utterances in the corpus both in "Infant Directed" (IDS) and in "Adult Directed" (ADS) modes. Last but not least, it is not known whether language acquisition is facilitated by interaction with several different speakers. For that reason it was decided

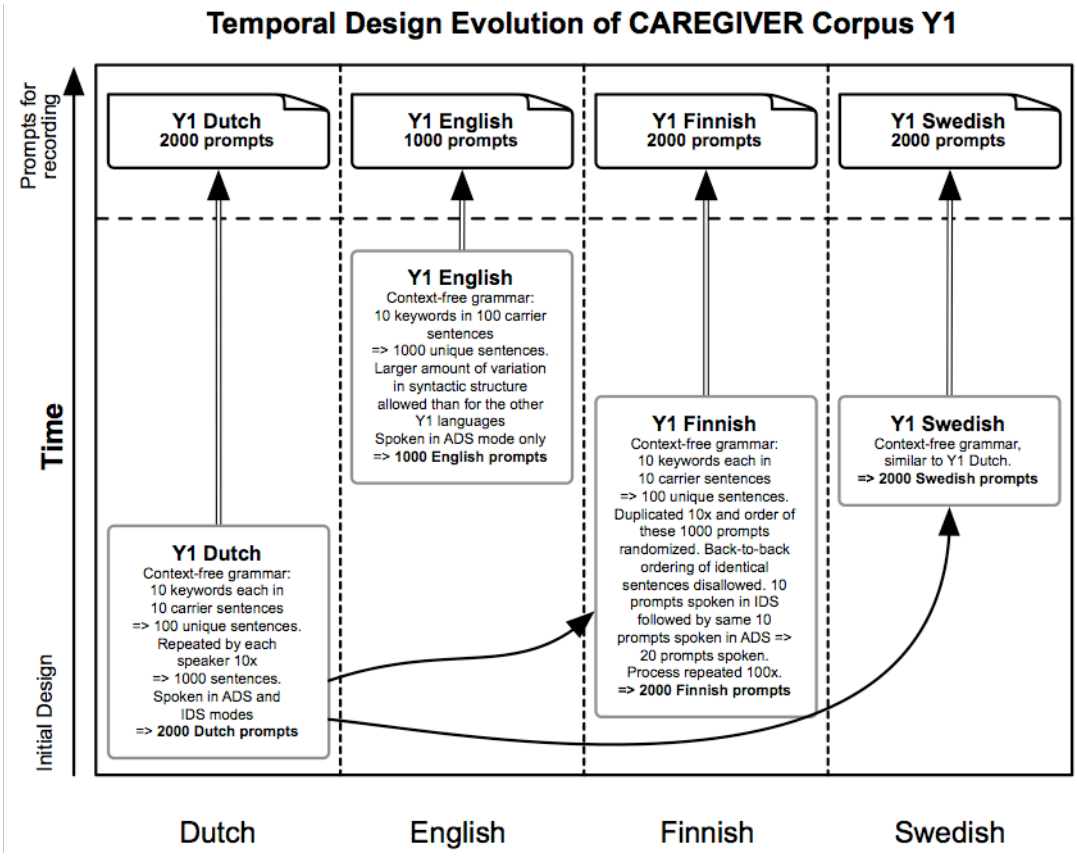

Figure 1 The prompts existing in Y1 Finnish and Y1 Swedish are direct translations of the Y1 Dutch prompts while Y1 English follows a different format that contains more syntactic variation.

to record a large number of utterances from four speakers in each language (two males, two females, typically the *primary caregivers*) and a smaller number of utterances from six additional speakers (three males and three females for each language, i.e., the *secondary caregivers*). Speakers were selected on two criteria: (1) having experience with infants, e.g., being parents, and (2) being available for re-recordings during the lifetime of the ACORNS project.

The contents of the utterances in the corpus have been designed with Dutch as the point of departure. For that end two unweighted context-free grammars were designed that generated syntactically correct sentences. The first grammar used a set of carrier phrases and combined these with one of 10 keywords (9 concrete nouns and 1 proper name). Each of the 9 keywords occurred 100 times in the corpus for each language in both IDS and ADS modes. The tenth keyword was different for each pair of speakers, e.g., their child's name. The set of sentences (or prompts) to be spoken by the speakers that this grammar produced is referred to as Year 1 speech (*Y1*). Utterances for Y1 Finnish and Swedish were then generated in a similar manner and verified for correctness. Y1 English was designed only after the other Y1 languages had been recorded and a comparable set of sentences was generated with a slightly larger amount of variation in syntactic structure. Y1 English was restricted to ADS mode only and thus contained 1000 utterances per speaker. Figure 1 reveals the design evolution of the Y1 prompts for each of the four languages.

The expanded vocabulary of the second grammar contained 50 keywords (nouns, adjectives and verbs), and produced prompts referred to as *Y2*. *(Note that the Y1 and Y2 nomenclature reflects more on the project years during which the corpus material was designed and recorded rather than on an infant's language acquisition capabilities during its first two years of development).*

Following a similar approach, the Y2 grammar was first applied to Dutch and the same grammar was then used to generate the English prompts. Applying the same grammar to produce viable Finnish prompts was deemed too complicated due to inflections so a native phonetician was employed to translate the 2000 Y2 English prompts into Finnish by hand. This allowed taking into account word choices and forms that strove for an equivalent degree of complexity across languages. During the translation process 397 additional utterances were added to improve the natural flow of speech and to reduce speaker weariness by the use of dialog word queries and answers. For example, to break up long monotonous stretches of prompts, triples such as (1) "Here is a small toy and a dog", (2) "Cat?", (3) "No, I meant dog!" were distributed evenly within the prompt sets. These additional Y2 Finnish prompts were reflected back into Y2 English so as to achieve identical corpora between English and Finnish in terms of prompt ordering and complexity. Note that these changes were not

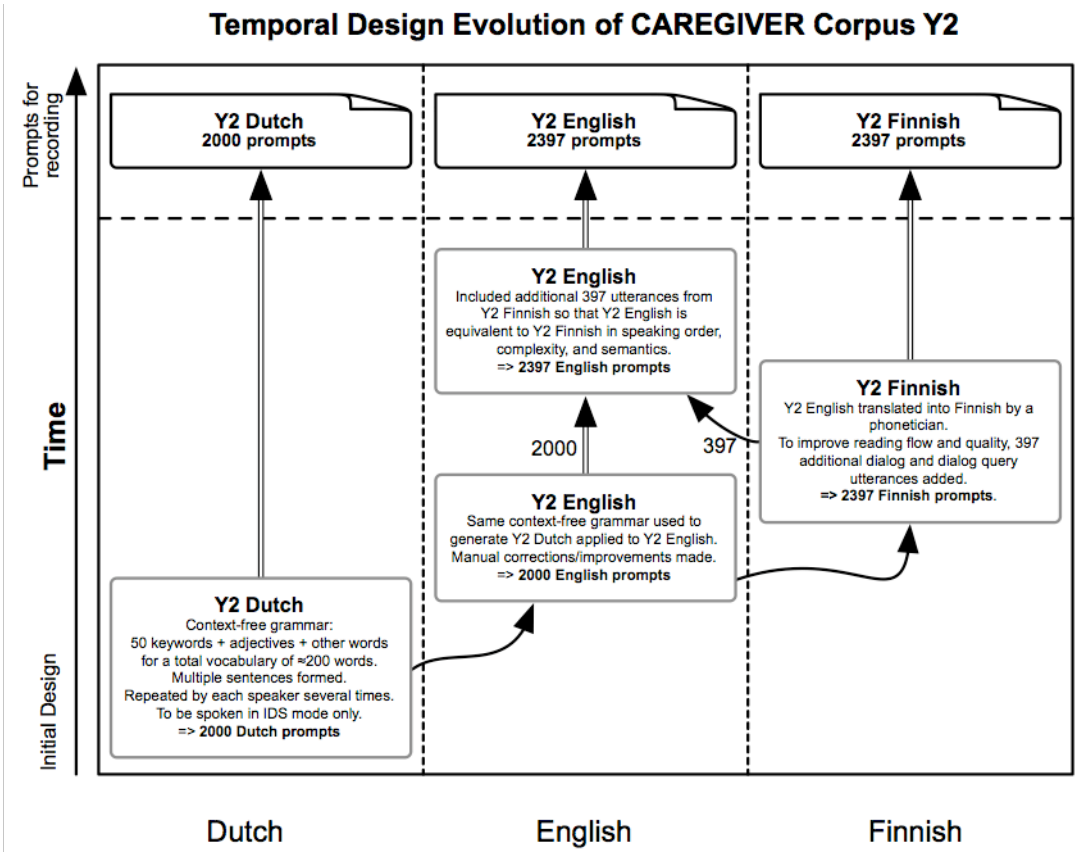

Figure 2 The context-free grammar used for Y2 Dutch was used to generate Y2 English which was then manually translated into Finnish. Nearly 400 additional prompts were added to improve speech flow in Finnish and these were subsequently reflected back into Y2 English.

incorporated into Y2 Dutch since these recordings had already been completed. Figure 2 displays the evolution of the prompts that were generated for the three Y2 languages. Swedish was not included in the Y2 corpus due to project budgetary constraints.

Some of the multi-keyword utterances in Y2 speech contain <adjective> <noun> or <adjective> <adjective> <noun> constructions. In generating these utterances no attempt was made to impose naturalistic semantic restrictions. We decided to keep the context-free grammars that generated possible utterances as simple as possible. Specifically, we did not attempt to attach attributes to adjectives and nouns that could be used to prevent semantic clashes. Therefore, noun phrases such as 'square frog' or 'clean round cow' were accepted (if only because such expressions might occur in a fairy tale). However, we deleted contradictory expressions such as 'little big car' from the output of the generative grammars.

For Y2 Dutch the smaller 600 utterance set spoken by the secondary caregivers was selected from the primary caregiver's randomly. For Y2 English and Y2 Finnish the 600 utterances were selected from the primary caregiver's prompts by including half of the dialog triples (72 cases), all of the keywords exactly once (50), isolated verbs once (3), and the remaining 475 prompts selected in the same order as the context-free grammar generated prompts appeared in the larger 2397 set.

Both Y1 and Y2 lexicons contained additional words such as auxiliary verbs, prepositions, articles, etc., that are needed to generate complete sentences. In total all Y1 and Y2 corpora have comparable syntactic complexity across the languages. The choice of the keywords is largely based on the content of the on-line available Communicative Development Databases (Fenson, et al., 2003) which is available for English and several other languages. Moderate effort was exerted to make all sentences comparable in phonetic structure (by, e.g., not including many minimal pairs). The context-free grammars used to generate the utterances are also provided as part of the corpus package.

3. Recording Procedure

All read utterances were 'acted' from prompts and the audio captured in recording studios. This meant that neither an infant (for IDS) nor another adult (for ADS) was present thereby greatly simplifying recording logistics. Speakers produced sets of utterances in blocks, after which they could relax and recover. Block lengths for Y1 utterances had no fixed size since speakers were permitted to break freely between any prompt and continue after resting. This was also the manner in which Y2 Dutch was recorded.

For Y2 English and Y2 Finnish the prompts were arranged into pre-determined sets that varied in length, e.g., 50, 100, or 150 prompts and took anywhere from approximately 4 to 12 minutes to record. After such a fixed length block the speaker could rest for as long as

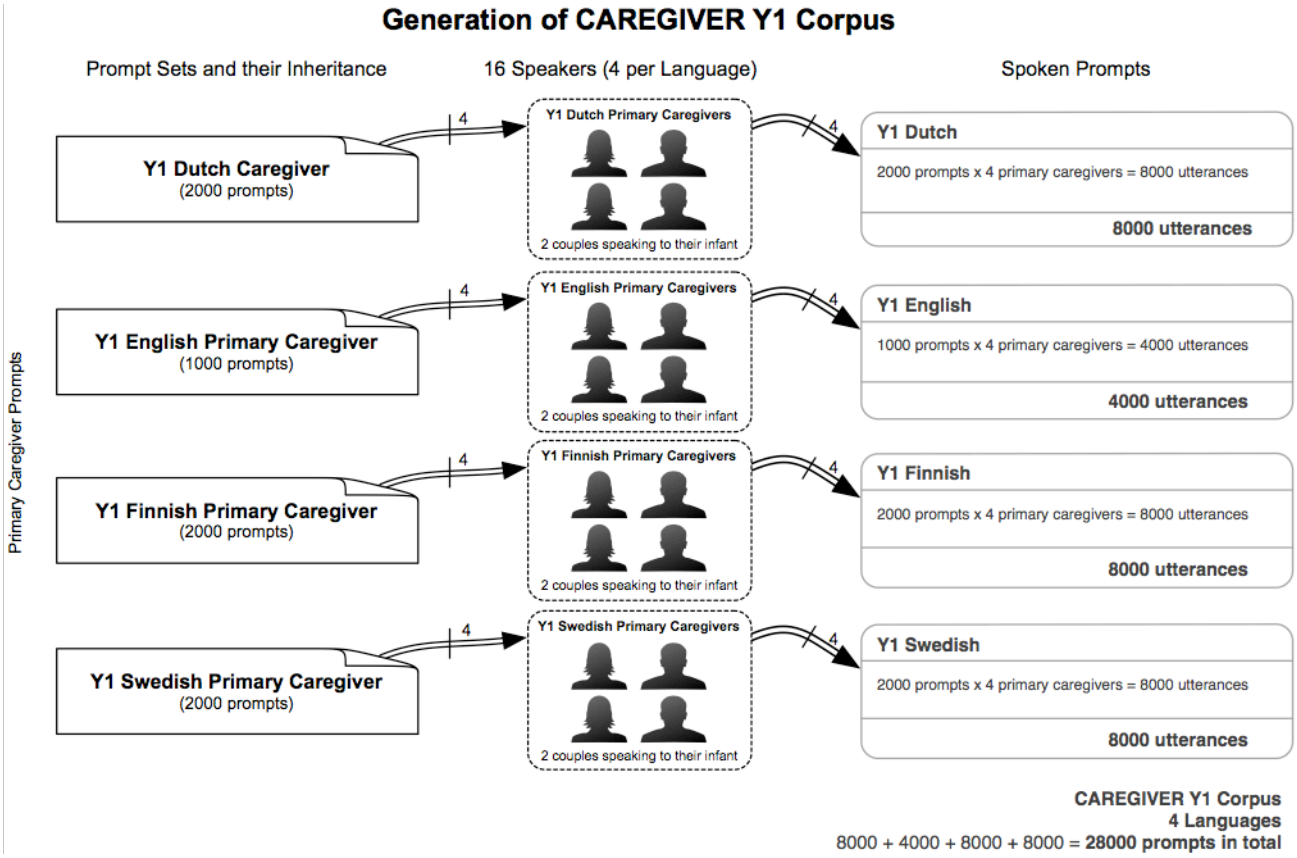

Figure 3 The Y1 prompts were spoken by four native speakers (typically the infant's mother and father) for each language.

was desired. Typically primary caregiver recordings (2000 for Dutch and 2397 for English/Finnish) took a minimum of two days to complete and were broken up into numerous recording sessions. Secondary caregiver recordings for Y2 (600 prompts/speaker) could usually be completed in less than two hours.

To help a speaker produce the different required modes of speech (IDS and ADS), an image of a familiar infant (e.g., a speaker's own child) or adult (e.g., spouse) was made available along with the textual prompt. All speakers reported that the presence of their infant's or partner's picture made the task more comfortable.

Two different audio recording platforms were used, both operating at a 44.1 kHz sampling rate. All recordings for Dutch and Swedish were made with a speaker reading the prompts from a printed page of paper and the audio saved as a single file for each session. The beginnings and ends of the individual utterances were marked after the completion of the recording using a separate process. For improved realism, an image of the infant and adult whom the speaker was addressing was placed within viewing distance.

For the Y1 and Y2 recordings of English and Finnish, a computer-based prompt recording platform was implemented utilizing dual displays. One display was for the speaker and the other for a technician who supervised the recording procedure. The speaker's display contained an image of the addressed infant (or adult for ADS) in close proximity with the current textual prompt. This arrangement provided a more lifelike environment to capture read speech: the speaker could look directly at the child while speaking since the current prompt's text was located in the visual periphery. This system freed up speakers from reading thousands of prompts from paper thus reducing speaker fatigue by minimizing eye, head and hand movements. Figure 5 shows the Y2 Finnish recording setup that was located in an anechoic chamber.

Once a prompt was spoken correctly, the technician overseeing the recording process would trigger the next prompt to be activated. This semi-automatic approach allowed misspeaks to be detected in-situ and gross errors to be re-recorded immediately, thereby improving corpus quality and lowering post-processing costs. Furthermore, the approach proved to be effective in avoiding the "prompt-reading-runaway" phenomena, i.e., when a speaker left on their own begins to race through the prompts to complete the (repetitive) task more quickly and ends up compromising corpus quality.

For the Y2 set of English and Finnish recordings a novel technique was used to investigate whether priming the speaker's mind subconsciously with the words or concepts of the upcoming utterance would help to retain speaker attention throughout long recording procedures. This was done by incorporating an iconic representation of the text that was synthesized using 1 to 4 primitive images of the concepts within a prompt. This collection of images was then displayed concurrently for an amount of time that was dependent upon the complexity of the prompt (e.g., the number of concepts included). Therefore, the combined images would appear

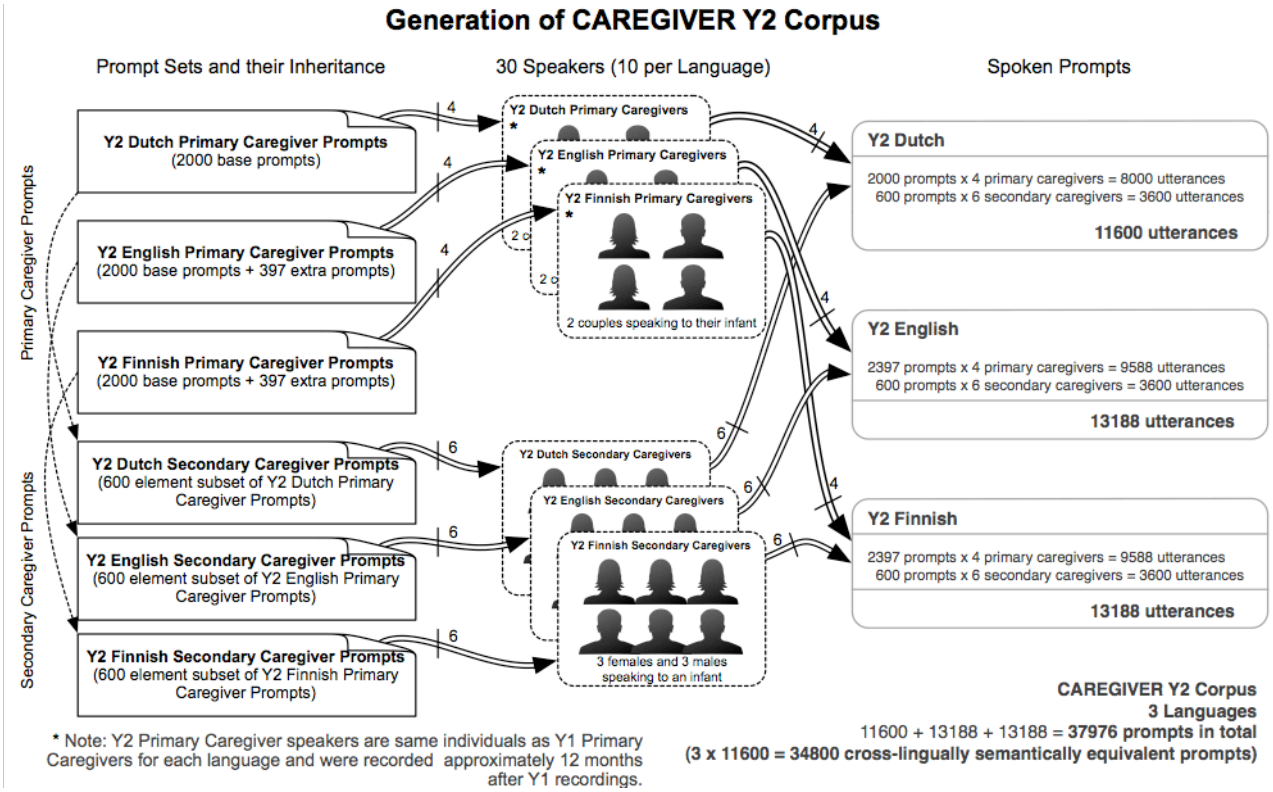

Figure 4 Y2 prompts were read by 30 speakers (10 per language, of which four were the same primary caregivers as in Y1). The other six speakers per language (secondary caregivers) spoke a 600 utterance subset of the prmary caregiver's prompt set.

simultaneously on the computer display typically from 400 to 1100 ms prior to the appearance of the textual prompt. Since the images were abstract in nature (they had all been drawn by young children) and since there were multiple icons for each concept, neither the speakers nor the technicians were able to relate a sequence of icons with the exact textual representation of the upcoming prompt even after 10 exposures to the same prompt. Figure 6 shows one possible set of iconic prompts for the sentence "The woman sees the round frog." Since misspeaks were logged by the system the effectiveness of this priming method could be measured by the relative change in elicitation errors, phonation rate, etc., all potential indicators of speaker alertness.

Each recording session was saved as one long audio file and post-processed later. Since the absolute times of the visibility of each prompt had been recorded by the prompt-recording system, segmenting long block-length audio files into utterance length recordings was greatly simplified. This also enabled the study of speaker fatigue as a function of prompt rate, speaking rate, prompt block lengths, and the effectiveness of iconic prompts as was already mentioned above.

The technician's display indicated, in addition to what the speaker had visible on their display, the current location within a prompt set, the number of misspeaks for each prompt recorded so far, and other recording-centric information. With this information at hand the technician could judge whether it was fruitful to continue recording or to come back at some later time once the speaker had rested. The prompt-recording system also automatically checked for correct amplitude levels, notifying the technician immediately if audio levels were too high or low. The peripheral data captured during the recording sessions related to the state of the speaker is planned to be included in the CAREGIVER corpus. By analyzing the data in more detail the extracted knowledge will hopefully be able to provide corpora designers and collectors with realistic guidelines. Figure 6 shows typical displays for a speaker and a technician.

The segmented utterance-based recordings of both recording systems were verified manually through efficient mass visual inspection of temporal waveforms and/or HMM forced alignment. An utterance error rate of 0.25 % was detected and for these cases either the audio was re-recorded in a future recording session, the prompt changed to reflect the audio, or the utterance removed from the corpus altogether.

4. Data

Altogether approximately 12 Gb of audio data is available in nearly 66000 segmented utterances divided over 4 languages, and 34 unique speakers. Audio data is represented using the WAV file format and the files are situated in a file hierarchy that is partitioned according to the year the recordings took place (Y1 or Y2), language, speaker, and recording session or prompt set (depending on the type of blocking that was implemented).

The six additional Y2 speakers per language who only produced 600 utterances each were designed to be test speakers, i.e., speakers that can be used to investigate the

Figure 5 Recordings for Y2 Finnish took place in an anechoic chamber. A technician was seated by the table on the left controlling the issuing of prompts for the speaker as well as monitoring all recordings for accuracy and quality of elicitation.

extent to which a computational model of language acquisition can deal with speech from other speakers than the daily caregivers. Except for this obvious division in training and test data based on speakers, there is no fixed structure imposed on the data. This is in line with one of the tenets in the ACORNS project, namely that a cognitively plausible simulation of language acquisition cannot be based on the training-testing dichotomy characteristic for conventional automatic speech recognition research.

5. Annotations

ASR systems, trained for the different languages, were used to discover discrepancies between the prompt texts and the speech that was actually produced. In the (rare) cases where obvious mispronunciations were observed, the prompt texts were changed accordingly and the discrepancies were noted in the annotation files. Every audio WAV file has associated with it a parallel XML encoded annotation file residing in the same directory that contains the utterance‘s orthography. Several of the sub-corpora, e.g., Y2 English, also have a time-aligned word and phone level description of the audio and are included in the annotation files as well.

6. Experiments

CAREGIVER has been used for numerous experiments within in the ACORNS project. For example, a study that compared the prosodic aspects of elicited keywords for IDS and ADS speech between Finnish and Swedish was performed (Räsänen et al., 2008).

To enable other researchers to repeat experiments it is necessary to publicly define the training and test sets that were used in ACORNS. For example, the corpus package contains a specification of the data that have been used for comparing three different learning methods (Concept Matrices, Non-negative Matrix Factorization, DP-Ngrams) developed in the ACORNS project (ten Bosch et al., 2009) when using speech from the Y1 and Y2 English

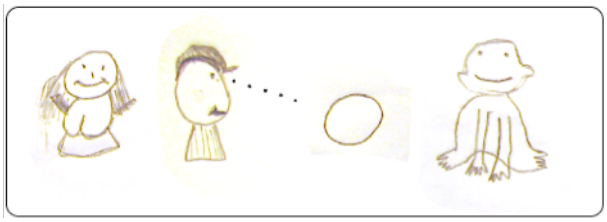

Figure 6 To retain speaker attentiveness throughout the recording process an iconic representation of the prompt was displayed for a fraction of a second prior to the text being shown to the speaker. Displayed here are the icons for the prompt “A woman sees the round frog.”

portion of CAREGIVER. In this experiment, training and test sets were kept apart. The learning system processed the training utterances only once. The test utterances, on the contrary, were used repeatedly to probe the developing competence of the learning system. The training pool was taken from Y1 English that consisted of 4000 English utterances spoken by two female (F1, F2) and two male (M1, M2) speakers (1000 utterances per speaker). Each of these utterances contains a single keyword, chosen from the following set: ‘bath’, ‘book’, ‘bottle’, ‘car’, ‘daddy’, ‘mummy’, ‘nappy’, ‘shoe’, ‘telephone’ and either of the proper names ‘Angus’ or ‘Ewan’ depending on which couple was addressing their child. Each utterance was accompanied by an abstract symbolic tag (that represented visual information).

From the Y1 English corpus, five different training sets were created. These five different trainings sets are: F1, F1+F2, F1+M2, M1+M2, and F1+F2+M1+M2, the notation indicating the speakers present in the training set. The ordering of the stimuli (480 spoken prompts in F1, 520 in the others) within each training set was controlled appropriately so that testing would be meaningful. The number of examples per keyword in each training set was the same for each keyword and balanced per speaker. Each learning method was applied to a combination of training and test sets. Word representations were built during training, and after each 20 utterances the model was probed by measuring its accuracy on the full test set. Each training set was combined with 10 different test sets: 4 test sets (F1, F2, M1, M2) that contained only held-out data, and 6 sets from the additional Y2 English speakers numbered from 5-10. Other sub-corpora can be readily created for other experimental designs from the complete corpora.

In learning associations between speech and visual representations of objects in a scene, decisions must be made with respect to the representation of the scenes. For this purpose a set of visual and semantic features has been developed that can be used to specify the objects, attributes and actions represented in the scenes. Specifically, the feature representations make it possible to investigate learning of individual objects (e.g., ‘dog’ and ‘cat’) as well as more general concepts (such as ‘animal’). Examples of the feature coding are provided as part of the corpus package.

Figure 7 Speaker (top) and technician (bottom) displays used in the prompt-recording system that was used to capture English and Finnish caregiver speech.

7. Availability

The complete corpus will be made available through ELRA. The software needed to repeat the experiments conducted in the ACORNS project is available through the public project website http://www.acorns-project.org.

8. Conclusion

This paper described the motivation and need for a corpus that can be applied to the study of language acquisition. Its design called for capturing meaningful utterances generated by the caregivers of an infant. The evolution of the sub-corpora, the recording procedures, and the final contents of the CAREGIVER corpus that was recorded over two years and covering 34 native adult speakers from the Dutch, English, Finnish, and Swedish languages, spoken in both infant-directed and adult-directed speech modes, was covered.

9. Acknowledgments

This corpus described in this paper was created in the EU FP6 FET project *Acquisition of Communication and Recognition Skills* (ACORNS), contract no. FP6-034362. The authors would like to thank Prof. Lou Boves for his guidance and help in preparing this publication.

References

Fenson, L., Marchman V.A., Thal D.J., Dale, P.S., Reznick, J.S., & Bates, E. (2003) *MacArthur-Bates Communicative Development Inventories (CDIs)*, Second Edition, Baltimore, MD: Brooks Publishing

MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs. Volume 2: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates.

Newman, R.S. (2008). The level of detail in infant's word learning. Current directions in Psychological Science. Vol. 17 (3). 229-232.

Räsänen, O., Altosaar, T. & Laine, U.K. (2008) Comparison of prosodic features in Swedish and Finnish IDS/ADS speech, In *Proc. of Nordic Prosody X*, Helsinki, Finland.

Smith, L.B., & Yu, C. (2008). Infants Rapidly Learn Word-Referent Mappings via cross-situational Statistics. *Cognition,* 106, 333-338.

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

ten Bosch, L., Räsänen, O., Driesen, J., Aimetti, G., Altosaar, T., Boves, L., Corns. A. (2009) Do Multiple Caregivers Speed up Language Acquisition? *Proc. Interspeech 2009,* pp. 704-707.