User Model User-Adap Inter (2016) 26:393–423
DOI 10.1007/s11257-016-9176-8

Fostering parent–child dialog through automated discussion suggestions

Adrian Boteanu[1] · **Sonia Chernova**[1] · **David Nunez**[2] · **Cynthia Breazeal**[2]

Received: 16 July 2015 / Accepted in revised form: 19 June 2016 /
Published online: 22 July 2016

Abstract The development of early literacy skills has been critically linked to a child's later academic success. In particular, repeated studies have shown that reading aloud to children and providing opportunities for them to discuss the stories that they hear is of utmost importance to later academic success. CloudPrimer is a tablet-based interactive reading primer that aims to foster early literacy skills by supporting parents in shared reading with their children through user-targeted discussion topic suggestions. The tablet application records discussions between parents and children as they read a story and, in combination with a common sense knowledge base, leverages this information to produce suggestions. Because of the unique challenges presented by our application, the suggestion generation method relies on a novel topic modeling method that is based on semantic graph topology. We conducted a user study in which we compared how delivering suggestions generated by our approach compares to expert-crafted suggestions. Our results show that our system can successfully improve engagement and parent–child reading practices in the absence of a literacy expert's tutoring.

✉ Adrian Boteanu
aboteanu@wpi.edu

Sonia Chernova
soniac@wpi.edu

David Nunez
dnunez@media.mit.edu

Cynthia Breazeal
cynthiab@media.mit.edu

[1] Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, USA

[2] Media Lab, Massachusetts Institute of Technology, Cambridge, USA

Keywords Context-aware computing · User models · Dialog analysis · Recommender system

1 Introduction

Early literacy is a term used to describe the stage of literacy development that occurs before children are able to read and write. During this stage, important abilities develop, including vocabulary development, phonemic awareness and letter knowledge, all of which ultimately influence general cognitive skills. The development of early literacy skills through early experiences with books and stories has been critically linked to a child's reading and academic success. Repeated studies have shown that reading aloud to children and providing opportunities for them to discuss the stories that they hear is of utmost importance to later academic success (Wells 1985; Bus et al. 1995; Burns et al. 1999; Duursma et al. 2008).

This area has been studied extensively from an educational standpoint, with theories such as dialogic reading suggesting that, in order for the child to develop good language skills, it is important for parents to engage with their child in focused conversation that is driven by shared reading (Arnold and Whitehurst 1994). Having enough exposure to language, both in terms of hearing words spoken by adults and learning new words, has been indicated as crucial for future academic success (Hilbert and Eis 2014). Children in this age group go through a rapid learning period during which it is particularly important for them to have verbal interactions with adults, especially their parents, from which they can learn. Therefore the learning process is partially conditioned by the parents' ability to steer these interactions toward learning goals (Whitehurst and Lonigan 1998). There exists evidence that parents receiving professional coaching on dialogic reading are more capable to lead joint reading sessions from which the child gains literacy skills (Pillinger and Wood 2014). However, such training may not be available for all families due to additional expenses and unawareness of its necessity.

To the best of our knowledge, there have been no contributions towards helping parents have better conversations with their children through algorithmic means. Existing adaptive tutoring systems, which track students' progress and provide customized feedback, have focused on teaching and verifying mathematical or scientific knowledge, with the aim to enhance and supplement classes taught in school (Feng et al. 2006; Weld et al. 2012; Wenger 2014). Such systems are designed for students that are in school and that already have a set of language and reading skills. Instead, we target a younger age group, of children that either have yet to learn how to read and write, or are very early in the process of learning to do so.

The main goal of this work is to provide an interactive parent–child reading experience which would help parents talk more with their children. Our work leverages the fact that electronic books and tablet readers have become increasingly prevalent in recent years. In many cases, these devices seek to promote early literacy and increase child engagement by including animation and sound effects in the stories. Scientific evaluations of these technologies have found that, although engaging, such devices do not effectively achieve educational goals when used alone (Korat and Shamir 2008). Instead, recent studies highlight the importance of joint parent–child reading, show-

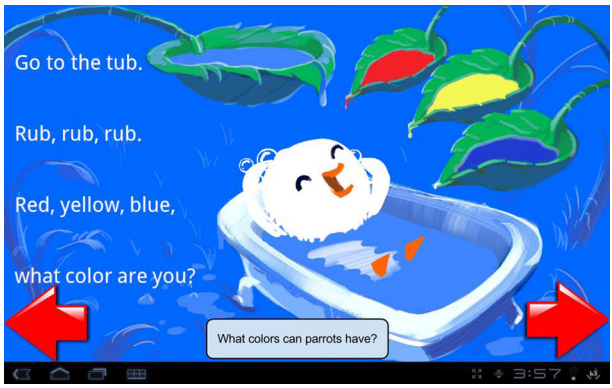

Fig. 1 Simulated in-story discussion suggestion via a dialog box

ing that learning gains are achieved by combining the use of digital media with adult interaction (Segal-Drori et al. 2010). Interactive stories tend to distract readers with multimedia features without increasing the dialog between parents and children during reading, which results in lower learning rates (Parish-Morris et al. 2013).

Taking these findings into account, we have developed CloudPrimer, an interactive reading primer to support parents in these discussions by offering suggestions for broadening the dialog and enriching the verbal interaction. The CloudPrimer application (1) records discussions of parent–child pairs engaged in a reading activity, (2) builds discussion topic models based on data gathered from across the community of readers, (3) generates English suggestion phrases using the topic models, and (4) delivers the suggestions at appropriate times during new interactions. The primer is based on an existing multimedia tablet application, in which a narrative is delivered through text, images, animations and sound, shown in Fig. 1 (Chang and Breazeal 2011). The tablet application blends in simple tasks such as color mixing in order to provide learning opportunities for children. Our system delivers prompts to support parents in starting and conducting discussions that revolve around elements from the story. In this work, we introduce a semi-supervised method for generating prompts without expert input. The traditional method for obtaining these prompts is for them to be authored by a literacy expert, which we consider to be the gold standard in evaluating our method.

Our approach consists of leveraging an initial community of readers to derive the topics of discussion parents and children approach while using the story. We then use these topics to generate suggestion prompts. We choose to derive topics from the dialog transcriptions instead of using only the text contained in the application because, compared to more traditional printed forms, interactive stories are rich in visual media and contain few words, both printed and spoken. Otherwise the potential discussion topic suggestions would be limited by the relatively rigid and simple story structure. Our main assumption is that, within this body of initial readers, some parents will have more developed and engaging discussions with their children, which our system can use to generate suggestions that will benefit future readers. It is important to note that our goal, to get parents to talk more with their children, is not equivalent to a dialogic

reading strategy, since the suggestions generated through our method are formed by single statements that do not closely follow the development of the story.

Our goal is to add breadth to the interaction and to expand on what the story provides. Our method is not designed for more specific goals that a literacy expert might target, such as prompts that specifically elicit dialogic reading. Creating suggestions with specific targets within the story narrative requires the expert to have a holistic understanding of both the story narrative and child literacy as an educational area. We consider such analysis to be an open question in narrative understanding, and are not aware of any other work that attempts to automatically generate suggestion prompts targeted for advanced techniques such as dialogic reading by using audio recordings captured with the on-board tablet microphone in uncontrolled environments.

Our approach primarily focuses on generating prompts using data captured in environments familiar to the users, such as their homes or places that they visit (e.g. museums). We take a topic modeling approach instead of attempting to gain a deeper understanding of the discussions because it is unfeasible to automatically derive complex models for this use setting. The unstructured nature of parent–child dialog, the noise introduced in the audio recording from either ambient noise or from manipulating the table computer, together with the unreliable nature of transcribing speech from children, does not provide sufficient sentence integrity in order to build more complex models. Because, due to the deployment settings, we cannot rely on full and grammatically correct sentences that could parsed reliably, we designed our approach to use bags of words as inputs. We leverage semantic networks to expand the information content of these collections of words extracted from the annotation, by identifying relations between these words within the semantic network. Our topic modeling approach uses these relations to group the input vocabulary into topics.

Informal discussion differs from other types of dialog, presenting a particular set of challenges. The context and development of such interactions contrast with formal settings, such as written articles and news broadcasts. The latter have a distinct, professional, approach in handling a subject and describing it. Speakers in this context try to meet the expectations of a large public who does not offer direct feedback, so stating opinions and facts accurately is important. This is accomplished through a crisp discourse which uses names and jargon to anchor the readers' or viewers' focus of attention. Conversely, in free form discussions the goals are not agreed upon in advance, the speakers do not announce detailed intentions on how they expect the interaction to evolve, and thus abrupt changes of subject are frequent. For the same reasons, the discourse may not necessarily consist of fully formed sentences, and feedback is richer. Furthermore, casual conversation uses improvised references, which may be developed during the conversation; since the dicussion does not address a broad audience, references used in informal discussions are only required to be relevant to the participants, and are thus richer as well (Linell 1998).

To evaluate the relative impact of prompting readers with suggestions, we present an end-to-end user study conducted in a lab setting that evaluates the impact our suggestions have on parent–child dialog in comparison to two other conditions. We used the same story narrative and design throughout data collection and evaluation. The results of the study indicate that the overall improvement in parent–child communi-

cation resulting from delivering suggestions generated by our method is comparable to exposing users to prompts created by child literacy experts.

We argue that our system offers an automated, semi-supervised method for generating suggestions. In order to build topic models and to generate suggestion prompts, separate dialog collection, annotation and processing is required for each new story. The core components, i.e. topic modeling and suggestion generation, are entirely unsupervised once the model parameters have been established. In Sect. 5.4 we detail the thresholds we used for generating the results presented in this paper. The system requires human input for dialog transcription, as current automated speech transcription technology is not sufficiently robust, especially for the speech of small children. Likewise, the suggestions need to be filtered by the crowd for appropriateness, as the application is sensitive. However, our approach of using crowdsourcing offers an arguably higher degree of automation than contracting literacy experts – in Sect. 2 we review recent advances in near-real-time crowdsourcing, which could be potentially used to decrease the response time for the stages at which our system requires human input. As the speech transcription and sentiment analysis technology matures, it may be feasible to replace the crowdsourced elements with algorithmic equivalents. Nonetheless, we maintain that the golden standard from a perspective solely focused on literacy benefits are the suggestions authored by experts. Our method represents an algorithmic alternative to this standard.

The rest of the paper is organized as follows: we first review relevant literature, including works on literacy primers, topic modeling and recommendation engines. We then present details on how we use ConceptNet and other external resources in Sect. 3. The main topic modeling algorithm, together with an evaluation study on its performance, is shown in Sect. 5. We then discuss multiple strategies for generating suggestions starting from topics in Sect. 6, and evaluate the effectiveness of these suggestions through a user study, presented in Sect. 7. We conclude with a discussion of the applicability of our method and possible extensions which could increase its degree of automation.

2 Related work

The CloudPrimer suggestion engine uses the TinkRBook tablet application (Chang 2011; Alonso et al. 2011; Chang et al. 2012) as the story for which suggestions are generated. TinkRBook was developed for educational interactive story-telling, enabling readers to interact directly with story elements through a tablet touch interface. Figure 1 shows a screen-shot of one of the pages in TinkRBook, including our suggestion prompt. All text in the application can be tapped, in response to which the application plays back speech utterances of the tapped word. Readers navigate the story using the large arrows on the screen. The application uses a relatively low number of words, of 117 including stop words (as mentioned in the introduction, existing research has described media rich interactive story applications to use fewer words than printed stories). In addition, there are interactive elements in the environment designed to enable game-like learning interactions. For example, tapping each leaf in Fig. 1 will paint the protagonist in its respective color in an additive manner, thus allowing the readers

to explore mixing colors; tapping the leftmost leaf resets the scene. The application collected data, including audio recordings, using a version of the GlobalLit framework (Nuñez 2015).

Several intelligent systems which learn from users have been designed in numerous areas. Such solutions include automated office managers (Modi et al. 2005), personified assistants (Rich and Sidner 1998) and smart house applications (Bouchard et al. 2006). These systems either interact directly with the user, through messages or an avatar, or can change their behavior without explicitly notifying the user. The common approach between all these numerous applications is to create a personalized model by observing an individual user. More broadly, recommender systems have been applied in a variety of domains, such as for restaurant recommendations (Boteanu and Chernova 2013b), music suggestions (McFee et al. 2012), e-commerce (Schafer et al. 1999) or media websites (Bennett and Lanning 2007). We will briefly review the main types of recommender systems as defined in existing literature (Ricci et al. 2011):

- *Collaborative filtering* these methods use rating systems to rate and rank items. The rating systems use scales of varying complexity, such as positive-only, which are common in online social networks (Gerlitz and Helmond 2013), or five star scales, which are common in online stores (Linden et al. 2003), or other systems. What all these systems have in common is that the recommendation engine only takes into account the relative rank of an item, and not its properties, when promoting it to a user.
- *Content-based* instead of using an external scale of measuring the merit of each item, content based methods directly compare items based on their properties, for example the text of a review or the tags associated with multimedia (Pazzani and Billsus 2007). New suggestions are generated from the pool of items that are similar in content between a user's previous selections and choices made by other users;
- *Hybrid-methods* these methods combine the previous approaches in some degree, proposing criteria of comparing users and items that take into account both evaluation scales and the content associated with items. For example, a user's preferences may be detected using information about the content, while the final recommendation may be based on review scores (Boteanu and Chernova 2013b).

Within this taxonomy of recommneder systems, the work presented in this paper can be considered a content-based recommender. Our method uses vocabularies collected from a large population of users to generate suggestions, without taking into account any rating system that the users may provide with respect to their interaction. The main difference between our work and recommender systems is that our system does not prompt suggestions based on user feedback, instead using data collected during interactions of previous users with the system to provide to generate suggestion prompts.

Topic models have been used to classify or cluster a broad variety of text corpora, such as e-mail (McCallum et al. 2005), community-generated online text such as cooking recipes (Krestel et al. 2009), news reports (McCallum 1999), or social network posts (Hong and Davison 2010; Zhao et al. 2011). Other applications of topic models include recommendation engines (Haruechaiyasak and Damrongrat 2008), word

sense disambiguation (Boyd-Graber et al. 2007), and sentiment analysis (Sommer et al. 2011). All these approaches have in common the use of a flavor of probabilistic topic models (Blei and Lafferty 2009). The simplest topic model, Latent Dirichlet Allocation (LDA), works as follows: (1) a fixed number of topics are initialized as random distributions over the vocabulary in the corpus, (2) the word probabilities in the topic set are refined iteratively through expectation-maximization (EM) steps (Blei 2012). LDA makes a number of significant assumptions: (1) the text is ignored (bag-of-words assumption), and (2) that the order of the documents can be ignored. Another popular algorithm is latent semantic indexing (LSI) (Deerwester et al. 1990), predating LDA. LSI makes the same assumptions described above, and creates topics by computing the singular value decomposition (SVD) of the matrix formed by document-word occurrences. Latent semantic analysis (LSA) (Hofmann 1999) improves on LSI by introducing an EM step in place of SVD.

Recent work has attempted to address some of the limitations all these probabilistic topic models exhibit. Models that identify both topics in data as well as the intensitiy at which they are expressed was proposed in order to overcome the assumption that the order of documents does not matter (Krause and Guestrin 2006; Wang and McCallum 2006). In order to address the bag-of-words assumptions, hierarchical generative models have been proposed in order to learn word groupings as part of topics (Wallach 2006). Other work on probabilistic topic models has proposed using topics to produce linear segmentations of documents instead of assigning individual words to topics, also generalizing beyond the bag-of-words assumption (Eisenstein and Barzilay 2008).

The key difference between our topic modeling method and latent statistical approaches is that, because we are using a readily available semantic network, we do not require a large corpus, and results do not reflect word associations as expressed in that corpus alone. Instead, topics are constructed from a vocabulary using the semantic network. By doing so, we separate the process of relating words from the process of modeling the topics encountered in a single document. Although our topic models are dependent on the semantic network, we argue that, in particular for small or highly specific corpora, our method reduces the training bias present in topics. Furthermore, our method is not restricted to common words since semantic networks can incorporate information about entities (for example, ConceptNet has nodes about countries, geographical points, cities, etc). In addition, semantic networks can include information derived statistically from co-occurrence in documents, which in the case of ConceptNet is represented as *RelatedTo* edges. To illustrate the need for an approach different than probabilistic topic modeling for our approach, we will first list a two topics (out of ten in total) generated using the LDA algorithm, after removing stop words: (1) *{david, duck, like, adult, baby, want, tap}*; (2) *{adult, david, want, child, researcher, red, duck}*. These results may be improved through manual fine-tuning, however, we speculate that significant improvements would be limited by the size of the corpus which can be collected for our applications. The first stage of our algorithm for grouping words into topics also has some similarity with K-Means approaches for clustering words (Steinbach et al. 2000). The raw topic algorithm can be viewed as clustering in a non-Euclidean space using Divisi similarity to measure distance between words. However, the key differences are the absence of a cluster center and having the same word belong to multiple clusters.

Finally, recent work in crowdsourcing has focused on providing fast response times on tasks (Bernstein et al. 2011; Lasecki et al. 2013), with focus problems such as quality control (Mashhadi and Capra 2011). Thus, we consider that one key advantage our work has over expert-sourced suggestions is that it provides shorter response times, with a significant part of the suggestion process automated by topic and suggestion generation algorithms. Of particular relevance to our work, crowdsourcing has been used for dialog generation (Bessho et al. 2012). In this work, the authors use crowdsourcing to select answers from a corpus of utterance pairs in order to provide replies in dialog. Since in our application domain generating a sufficiently large discussion corpus is not feasible due to the number of participants, we use semantic networks to expand the vocabulary used in discussions.

3 Methods

In this section we will provide details on the design of the semantic network we use in our evaluation, ConceptNet, as well as the metric we use for evaluating pairwise concept similarity, Divisi. ConceptNet is a freely available commonsense knowledge base and natural-language-processing tool-kit which supports many practical textual-reasoning tasks over real-world documents, including topic-listing, analogy-making, and other context oriented inferences (Havasi et al. 2009). ConceptNet forms a large graph of concepts connected through relations. It is derived from a number of both authored and mined linguistic sources, including DBPedia, WordNet, and VerbNet. We chose to use ConceptNet over other semantic networks because of its breadth of concepts and relations. While the number of edge types in the graph is relatively small compared to ResearchCyc, of 48, ConceptNet covers a broad spectrum of relations, most of which are easy to understand and thus suitable for our application. For example, relation types include type hierarchies (*IsA*), properties (*HasProperty*), uses (*UsedFor*), abilities (*CapableOf*) and intents (*Desires*). Figure 2 shows a small portion from ConceptNet, exemplifying some of these relations connecting nodes related to the interactive application we use. From the ConceptNet project we also use the Divisi toolkit (Speer et al. 2010) to approximate pairwise concept similarity. Divisi produces similarity values by a spectral graph method, using the singular value decomposition (SVD) of the ConceptNet graph. These values are in the interval $[-1, 1]$, representing similarities ranging from entirely opposite to identical.

In addition to ConceptNet, we use the Natural Language Toolkit (NLTK), which is a collection of natural language processing tools; we use the Python packages for NLTK (Loper and Bird 2002). We used common text pre-processing methods offered in NLTK for segmenting text into sentences and words, and to remove punctuation, as well as stop words, which are words that occur with equal probability in all texts and thus are not useful in discriminating a given piece of text against others (e.g. prepositions) (Wilbur and Sirotkin 1992). We process transcriptions by separating text into words and removing stop words using the NLTK stopword list (Bird 2006). To account for typing errors in the annotation, we then perform spellchecking using a large US English dictionary and the *enchant* Python library (Perkins 2010). In addition, we used the WordNet interface that NLTK provides for converting words into lemmas.

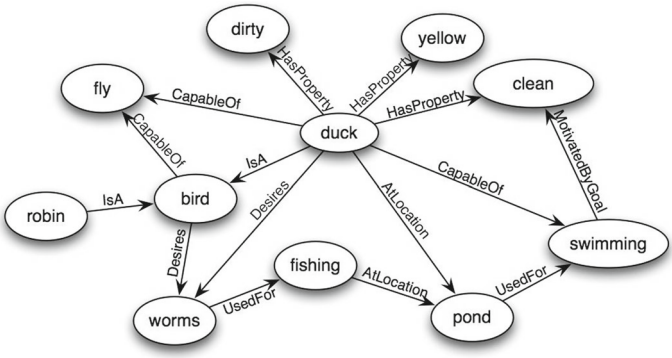

Fig. 2 A small example of nodes and relations in ConceptNet

WordNet is a semantic network that focuses on describing relations between words using two structures: synonym sets, or synsets, which represent equivalent meanings of a concept, and a taxonomical tree of hypernyms and hyponyms that contains broader concepts towards the root and more specific meanings towards the leaves (Miller 1995; Fellbaum 1998). The lemma form of a word is the root for of a word, for example *run* is the lemma of *running* (Plisson et al. 2004). We include a lemmatization step as part of pre-processing, as described in Sect. 4, in order identify the same concept in any of its various syntactic forms (Levelt 1999).

Our topic modeling method is not limited to ConceptNet and does not use any features unique to ConceptNet. We anticipate that it could be used with any other semantic network that provides a measure for pairwise concept similarity—for example, WordNet provides such features as well. However, our method of generating suggestions relies on edges denoting "common sense" notions which are present in ConceptNet and not in WordNet. The main drawback of WordNet with respect to our application is that it includes a fairly limited set of relations, focusing on taxonomical dependencies between concepts. Our choice of ConceptNet is primarily motivated by its availability and richness in both concepts and relations.

4 Overview of system architecture

In this section, we give a high level overview of our system, which consists of the modular pipeline shown in Fig. 3. In Table 1 we present a running example that illustrates how the data is processed at each step before being interactively delivered during reading sessions. The stages described in these figures are the following:

1. *Dialog data acquisition* this step includes recording and annotating the interactions from which we generate suggestions. We collected 45 reading sessions from a preschool in Worcester, MA, and from the Boston Museum of Science (5 and 40 final usable reading sessions after discarding unusable ones, respectively). The data collected through the preschool was obtained by giving parents the tablets to take home for a week. The data collected at the Boston Museum of Science was

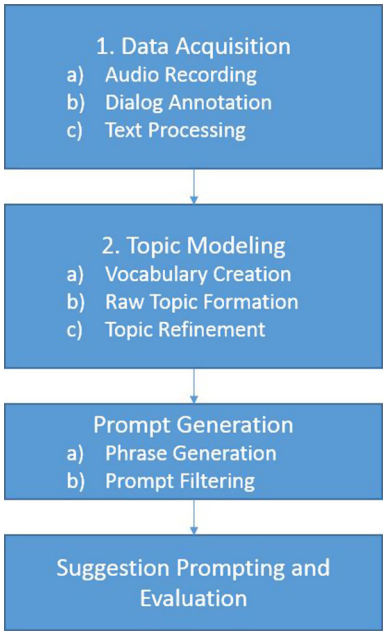

Fig. 3 System block diagram

from from readers (parent–child pairs) recruited during their visit at the museum. For the latter, the interaction took place on the museum premises. All collected data included a full audio recording of the interaction, which was recorded using the on-board tablet microphone. In order to setup the collection process with the existing primer, the TinkrBook application was modified to interface with a data collection framework (Aharony et al. 2011) that would record and upload audio recordings. These recordings were annotated by a professional text transcription service with one annotator per reading session. Each annotation included all utterances that were recorded, along with the corresponding time-stamp and presumed speaker (either parent, child or speech produced by the tablet application).
Recording the sessions in a casual setting resulted, we believe, in more natural interactions, but at the expense of noise being present in the recording. In addition to background noise, the main source of distraction for the readers were interactions with other people not actively involved in using the tablets. Furthermore, the transcription process itself introduced noise such as misheard words or typos. A short example of this transcription is presented on the first row of Table 1;
To prepare the transcription texts for topic modeling, we pre-processed them by removing all stop words and other very common English words using NLTK. We then converted all words into lemmas form using the NLTK WordNet Lemmatizer library (Loper and Bird 2002). The second row of Table 1 shows the text after pre-processing.

2. *Topic modeling* we model topics from pre-processed text by first creating the start vocabulary, then forming raw topics, and finally refining topics. Table 1 shows the output of each of these steps on rows 3, 4 and 5, respectively:

Table 1 Example of the results after each processing stage

Algorithm stage	Data
Dialog data collection	Here you go, you have a purple duck. What's your favorite color? Um, purple. So baby duck is hungry, eat two beetles. The beetles? Yeah, that's the beetles. More, more. Those are ladybugs
Preprocessing	purple duck favorite color purple baby duck hungry eat beetle beetle yeah beetle ladybug want beetle need beetle need beetle want feed
Start Vocab. Creation	{purple, tap, duck, beetle, yeah}
Raw topic formation	{blue purple, green, yellow, different}
	{ant, firefly, cricket, ladybug, beetle}
	{owl, bird, duck}
	{need let want}
Topic refinement	{blue, purple, green, yellow}
	{ant, firefly, cricket, ladybug, beetle}
	{owl, bird, duck}
	{need, want}
Suggestion generation	What are green and yellow?
	An owl is a bird, what other birds do you know?

(a) *Start vocabulary creation* In this step, we corroborate all words from all discussions and filter out words that occur only in one session in order to filter out unrepresentative words. We name this set the *start vocabulary*, because it is the input for our topic modeling approach.
(b) *Raw topic formation* In the first stage of topic modeling, we construct *raw topics* based on the start vocabulary using the Divisi module of ConceptNet (Speer et al. 2010). The output of this module presents the starting point for refining topics.
(c) *Topic refinement* The raw topics produced by the previous step are generated using approximate similarity. In this step, we *refine* these topics by directly exploring ConceptNet's graph structure. This process involves directly traversing the graph identify the connected components within a topic's set of words.

3. *Suggestion generation* We use topics and edge-information to generate question phrases. By using topics, we reduce the exploration space in which word tuples are tested against the template defined for each question type. In addition to the end-to-end study, we evaluate the output of this module separately in Sect. 6.1. Two suggestion examples are shown on the last row of Table 1;
4. *Suggestion prompts* We deliver suggestions during conversations and evaluate their impact via metrics such as the number of words spoken by the participants.

5 Topic modeling

The topic modeling method used in this work builds on our previous work, incorporating common sense reasoning in evaluating topics occurring in free-form the discussions (Boteanu and Chernova 2013a). By using semantic network topology, our method can include infrequent words in the resulting topic models, provided that these words are present in the semantic network. The main assumption of this method is that topological distance between concepts in a semantic network represents similarity. As mentioned before, because the semantic network can incorporate data produced from a variety of mechanisms (expert authored, statistically derived, crowdsourced information—as in the case with ConceptNet), our topic modeling method is only limited by what is represented in the semantic network. Our method works at the graph level of the semantic network, which implies that any type of concept (node in the graph) and any type of relation (edge in the graph) can be potentially used to form topics—all parts of speech and proper names, which ConceptNet represents.

One limitation of our topic modeling approach is that it can not directly assign unknown words to topics: if a word is not present in the semantic network, or if the corresponding node is not connected in a representative manner to other concepts in the network, our algorithm would not include it as part of the output topics. This would be the case of rare words or jargon, which normally are less represented in ConceptNet. However, our method can use such words provided they are first connected to existing concepts through an external method (expert knowledge or co-occurrence in documents, for example) and then added through appropriate edges to the semantic network, allowing statistical information to be combined with other data sources. Probabilistic topic models can potentially include such rare words into their output, provided that the words are sufficiently expressed in the corpus, but the resulting topics only superficially represent these words and lack information on their relation to other concepts.

5.1 Start vocabulary creation

The first step in grouping words by topic is to identify which words are the most relevant to the discussion. Written text and speech are very different in terms of phrasing, word selection and connectors. In particular, parents talking with their children have other goals beside communication, such as teaching new words (Hausendorf and Quasthoff 1992). We observed that parents often have to restate the goals and important concepts to keep their children on track. Another characteristic of dialog is that the density of topic-relevant words, such as nouns, verbs and adjectives, is low compared to a written document.

In prior work, we introduced an interest metric heuristic to determine which words may be of interest to the readers at any point during the session (Boteanu and Chernova 2013a). This method was designed to select words from a very small number of reading sessions, potentially producing user-specific vocabularies from single reading sessions. In that context, the limited number of separate dialog recordings made it unreliable to filter out incidental words based on word commonality between reading

sessions. These incidental words are normally spoken by participants in relation to some other factor in the environment and not as part of reading the story. Since we collected a relatively large corpus of 45 discussions, it enabled the use of a frequency-based heuristic, in which we include into the start vocabulary all words that occur in at least two reading sessions, except for stop words. We compared the resulting vocabulary with the union of all vocabularies resulted from applying the interest metric and found no significant difference. Therefore we use a frequency heuristic to create the start vocabulary for all results presented in this paper, removing all words that do not occur in at least two sessions.

5.2 Raw topic formation

In this section, we describe the first stage of grouping the vocabulary of interest into discussion topics. This step produces rougher topics that are later refined. An example of a set of raw topics is given in Table 1. We define a *topic* as a set of words relating to a common theme, without any particular order. Topics do not have names themselves and are defined only by the words that belong to them. This allows us to model each topic through the common sense relations between its constituent words. One key characteristic is that a word can belong to multiple topics at once. We allow this since words usually have multiple meanings.

Algorithm 1 Pseudo-code for raw topic creation.

```
topics = φ
v = readVocabulary()
for each w in v do
  for each t in topics do
    similar = 0
    for tw in t do
      if similarity(w, tw) > similarityThreshold then
        similar = similar + 1
      end if
    end for
    if similar > size(t) * minSimilarityVote then
      t = t ∪ w
    end if
  end for
  if w was not added to any existing topic then
    topics = topics ∪ w
  end if
end for
```

The algorithm, shown in pseudo-code in Algorithm 1, starts constructing topics by removing a word from the vocabulary of interest, v, and creating a new topic containing only that word. Then, for each of the remaining words, w, we compute the Divisi similarity between it and each word present in all existing topics, tw. If the absolute Divisi similarity value, $similarity(w, tw)$, is above a threshold (*minSimilarityVote* in Algorithm 1), the result counts as a positive vote from tw. We use the absolute value of the Divisi similarity since we are interested in identifying related words and not concepts aligned in meaning.

After comparing w with all words in each topic, we sum all votes per topic, and compute the *voting ratio* by dividing the sum by the number of words in the topic. If the this ratio is higher than a set threshold, *minSimilarityVote* in Algorithm 1, the new word is included in the topic. The process repeats itself until there are no more words in the vocabulary of interest. Note that a word can thus be part of multiple topics. We can control the specificity of each topic by adjusting both the minimum similarity threshold as well as the minimum percentage of votes. For example, by using a high minimum similarity (0.5) but a low voting ratio (0.50) loose topics are generated, such as the following:

- ant, owl, bird, ladybug, duck;
- noise, tap;
- color, blue, purple, green, yellow;
- need, say, let, want;
- ladybug, bird, beetle;
- happen, let, pass;
- cricket, bird;
- purple, different, green;
- yummy, hungry;
- firefly, ladybug;
- push, angry.

Using a lower similarity threshold (0.3) but a higher minimum voting ratio (0.85) produces tighter topics, mostly because a newly introduced word has to have some association with most other words present in the topics, such as the following example:

- owl, bird, duck;
- push, tap;
- blue, purple, different, green, yellow;
- ant, firefly, cricket, ladybug, beetle;
- push, say, let, pass;
- need, let, want;
- push, happen, let, pass;
- yummy, hungry.

Setting both thresholds high produces raw topics that are very conservative and contain very few words. We do not include these results for the sake of brevity, but we do not consider such results practical. Similarly, setting both thresholds low produces looser and noisier topics. Since the goal of the raw topic stage is to restrict the search space of the refinement algorithm, and not to provide high quality topics itself, choosing either extreme is detrimental to the final result.

The raw topic formation step provides a starting point for the refinement step to test graph connectivity. It is thus preferable to allow broader topic at this stage, which would be then improved by the latter step, because the topic refinement step only removes words from topics without exploring other possible associations. Throughout the remainder of this paper we used the thresholds shown in the second example above (0.3 and 0.85).

5.3 Topic refinement

Since it is based on Divisi, a similarity measure derived from the semantic network's SVD, the topic generation method we introduced in the previous section produces imprecise results, in which words may be erroneously associated into the same topic if they are present in a dense region of the graph. However, the previous algorithm has the advantage of speed over directly exploring the highly connected ConceptNet graph, segmenting the initially very large search space of possible topic assignments into smaller regions; this space is particularly large since our method does not set a fixed number of topics in advance, as probabilistic methods do. Directly searching for connected components is unfeasible not because identifying connected components in a graph is a hard problem, but because semantic networks are very large and linear time algorithms are not sufficient. In ConceptNet, nodes with a degree of 30 or higher are common. The large size of data makes retrieval time significant as well. The key idea of refining topics is to find connected components of the subgraph represented by the raw topic in ConceptNet. Starting from raw topics makes this problem tractable by reducing the number of concept pairs that need to be tested. In this section we introduce a method of refining those results by directly exploring the graph structure of ConceptNet.

We consider two concepts to belong to the same topic after refinement if there is a path between the two respective connected components of at most the length of the search depth. For example, in Fig. 2, concepts "robin" and "worms" have no direct edge connecting them, but are both connected to "bird" by "Is A" and "Desires" relations, respectively. A search with the depth of 1 will separate them into different topics, while a search depth of 2 will group them into the same topic. Algorithm 2 shows in pseudo-code for refining topics.

This approach eliminates spurious associations introduced in the raw topic formation step by efficiently searching for connected components within groups of similar words, which limits the exploration space. We show an example of the effect of search depth on refining a small topic in Table 2. All types of relations are taken into account for these topic refinement results.

Algorithm 2 Pseudo-code for refining topics. t is the topic that is being refined, p is the resulting set of topics after separating the words from t and d is the search depth.

```
p = φ
for w in t do
  candidates ← nearest_neighbors(w, d)
  split ← True
  for q in p do
    if q ∩ candidates = φ then
      q ← q ∪ {w}
      split ← False
    end if
    if split = True then
      p = p ∪ {w}
    end if
  end for
end for
```

Table 2 Topic refinement results for increasing search depths

Raw topic	{deer wing frog duck}	{owl bird duck}	{push happen let pass}
Depth = 1	{deer}{wing} {frog}{duck}	{owl}{bird} {duck}	{push}{pass} {happen}{let}
Depth = 2	{deer}{wing}{frog duck}	{owl}{bird duck}	{push}{pass} {happen}{let}
Depth = 3	{deer}{wing}{frog duck}	{bird duck} {owl}	{push}{pass}{happen let}
Depth = 4	{wing duck}{frog duck} {deer}	{owl bird duck}	{happen let pass} {push}

5.4 Evaluation of topic quality

We conducted two surveys to evaluate our topic refinement algorithm. The main goal of both surveys was to compare the output of our topic refinement algorithm at different refinement depths with selections made by the crowd, allowing us to determine an appropriate value for the d parameter in Algorithm 2. To obtain the topics used in both surveys, we constructed topics from the 45 reading session transcriptions from the training corpus and then sampled uniformly from the resulting set of topics. The topics presented to workers were un-refined topics, and the crowd's choices were compared against the selections made by our topic refinement algorithm. For both evaluations, we crowdsourced workers through the Crowdflower platform. We used the default task distributions options (tasks are also relayed to other platforms such as Amazon Mechanical Turk), but we selected only workers from countries with English as the majority language.

The first survey required participants to read a raw topic, presented as a list of words, and select a subset that forms a common topic through check boxes corresponding to each word selection. For example, when presented with the set of words *{brown, old, long, hello, thing, green, yellow, okay, yes, whole, white, red,}* a possible response would be to check *{brown, green, yellow, white, red}*. We generated survey questions from twelve topics and collected 10 answer for each topic, for a total of 120 responses. To compute the inter-worker agreement that a word from the list was part of the topic, we used the proportion answers that marked that word. The mean inter-worker agreement value for all tasks was low, of 19.25 %, with a total of 17 participants in the survey. To obtain topic selections from the agreement values, we binarized these values per word via clustering (fitting two clusters using the k-means algorithm or a bimodal Gaussian Mixture Model produced identical results), and selected the cluster corresponding to the majority of selections as the final topic. For example, selection answer agreement values for the raw topic *brown, thing, green, orange, white, whole, hello, red* were 18, 3.6, 18, 18, 18, 3.6, 3.6, 18 %, respectively; for these selection values the cluster assignments were *1, 0, 1, 1, 1, 0, 0, 1*. Cluster 1 corresponds to a higher mean selection, thus the refined topic selected by the crowd is*brown, green, orange, white, red*.

We then computed the agreement between the topic refined by the crowd with the output of our algorithm at different search depths as the proportion between the number of selection matches per word and the size of the unrefined topic. The results of this comparison were not conclusive: the average agreement between our algorithm and

the crowd was 58, 52, and 56 % for respective refinement depths of 1, 2 and 3, with values distributed uniformly in a wide interval, from 25 to 89 %. This result, together with the low inter-worker agreement of 19 %, do not indicate a strong correlation between the crowd and our results. We speculate that the task of selecting a subset of words from a raw topic was too complex for the relatively small number of responses we collected per question.

As a result, we designed a second survey, which consisted of a set of "odd word out" problems in which respondents were given a short list of words and instructed to select the one that did not fit with the others. We produced these lists by refining raw topics using our method, and then randomly adding a back word that was excluded by the refinement process back to the topic. The option "None" was also available in the case the workers considered that all words were similar. In total there were 26 topics. On average, each topic received 12 different evaluations, with 312 judgments in total.

The average inter-worker agreement was 73.5 %, calculated per task as the percentage of judgments that the most selected option had, out of the total number of response per task – 12 on average. The survey answers are divided into two groups by agreement, high and low. For the nine topics which contained mostly verbs, the agreement ranged between 30 to 60 % with an average of 45.4%. For the rest of 17 topics, mostly formed by nouns, the agreement ranged between 75 and 100 %, with a mean of 88.4 %. This high agreement group of topics contains words that are interpreted similarly by the reviewers. In contrast, topics with low agreement contain words with multiple meanings, thus subjective to evaluate. Using a search depth of 2, the topic refinement algorithm matched the dominant decision of the survey answers for 47 % of the topics – it either eliminated the same word or kept the topic unchanged. For a search depth of 3, the percentage is 29 %. Thus, we selected a refinement depth of 2 for our results. Table 3 presents a few examples of the survey questions, showing the raw topics, the words selected as outliers by the topic refinement algorithm, and words selected as outliers by the survey participants.

Based on these results, we can conclude that our system best matches human respondents when analyzing topics composed of nouns. An exception to this is the example in the last row of Table 3, in which the algorithm was unable to differentiate animals (deer, frog, duck) from a limb (wing). We attribute such errors to currently missing edges in the constantly expanding commonsense knowledge network. The most significant disagreement, both between our system and the respondents, and between the respondents themselves, occurs on topics consisting of verbs, such as in lines 3 and 4 of Table 3. These collections are more difficult to interpret, and refining the topic would imply adopting a specific angle. For example, on line 3 of the table, the consensus in selecting 'Say' might be that it is the only action that produces speech, but similar classifications can be found to eliminate other words.

This evaluation shows that, for the situations in which human respondents reach consensus on the constituency of a topic, our approach successfully matches that consensus. While our topic modeling approach requires setting thresholds, these present a significantly smaller space (two real values in the interval *(0, 1)* for the raw topic formation stage and an integer for the topic refinement stage). We determined empirically that exploring a subsection of this space is sufficient, for example the topic refinement distance should not exceed 3, with greater values resulting in no effective

Table 3 Comparison between our approach and survey responses for refining raw topics

Raw topic	Outliers—topic refinement	Outliers—survey responses
Blue, purple, different, green, yellow	Depth 2, 3: Different	**Different (13/15)** Purple (1/15) *None* (1/15)
Ant, Firefly, Cricket, Ladybug, Beetle	Depth 2: Cricket Depth 3: *None*	Ant (2/15) Firefly (1/15) Cricket (2/15) Beetle (1/15) ***None (9/15)***
Push, Say, Let, Pass	Depth 2, 3: Push, Say, Let, Pass (No common topic found)	Push (1/12) **Say (8/12)** Let (1/12) *None* (2/12)
Need, Want, Let	Depth 2, 3: Let	Need (1/15) Let (7/15) *None* (7/15)
Deer, wing, frog, duck	Depth 2, 3: Deer, Duck	**Wing (12/13)** Duck (1/13)

For the survey responses, the number of agreeing responses is shown as a fraction of the total responses for that particular question (note that the number of responses per topic varies). If there is one, the predominant decision is in bold

pruning. Varying these parameters allows for topics to be varied in generality depending on the desired outcome. Following these surveys, we did not change the topic generation algorithm, but used the results to establish parameters which we used in the remainder of this work. Specifically, for generating topics for our evaluation, we set *similarityThreshold* to 0.3, *minSimilarityVote* to 0.85 and used a refinement depth of 2.

6 Generating suggestions

In this section we describe the method we introduce to generate prompts for our end-to-end user study. The goal of these prompts is to enhance and foster the conversation parents have with their children via questions and other suggestions. Therefore, we are not interested in whether the child is able to correctly answer the questions and do not model or provide any input method for answers.

As mentioned before, since our corpus contained sufficient data, totaling 45 reading sessions, we used a frequency heuristic to obtain the start vocabulary, selecting all words that occurred in at least two separate reading sessions. Using all discussion transcriptions recorded from participants, we created a vocabulary of 1009 containing all spoken words that occurred in more than one discussion, which is a significantly larger number of words compared to the 117 present in the TinkrBook application. We then created topic models starting from this vocabulary, producing a single set of twelve topics for the entire narrative. Keeping a unified vocabulary allowed us to have the richest connectivity between words, while at the same time enabling topics to cross between pages.

We designed a number of heuristics to generate suggestions from these topics, such that they would use a broader range of commonsense relations present in ConceptNet. Although not specifically designed for literacy education, the language and world knowledge present in ConceptNet is arguably relevant and related to expanding on the information in the story. While some of this information may be initially considered too abstract for a child to assimilate directly, it is the parent that is the target of our suggestion system. Thus, a suggestion that may seem initially very abstract (e.g. "Why do balls roll?") could be adapted to a discussion about round objects, even exemplified with other objects such as pens, instead of directing the conversation to a topic on solid mechanics. As described in the following list, method 1 targets general similarity evaluation, methods 2a and 2b focus on evaluating type classifications, and methods 3a, 3b, 3c and 3d target reasoning about various characteristics of an object. We provide examples of suggestions generated using the different strategies:

1. We randomly select two words from the topic and ask the readers to either come up with other related words or identify what the words have in common – "What other things are like *minutes* and *years*?"
2. We generate questions based on hypernym-hyponym relations in two ways:
   (a) Within a topic, we find words that have the same super-class and ask what do the words have in common. For example, for the words "duck" and "swan," by asking "What do ducks and swans have in common?" we would expect and answer similar to "They are both birds."
   (b) If within the topic there is a hypernymy relation between a pair of words, we ask a complementary question:
   "A duck is a bird, what other birds do you know?"
3. We test all possible pairs of words in the topic if they are connected by ConceptNet edges expressing properties or capabilities, and ask a question that tests knowledge about that fact. For example, "Why do birds fly?" for the concepts "bird" and "fly" connected by the edge *Capable Of*. We apply similar patterns for the following edge types:
   (a) *Capable Of* – "Why do *balls roll*?"
   (b) *Made Of* – "Why is a *towel* made of *cotton*?"
   (c) *Part Of* – "Why does a *plant* have a *leaf*?"
   (d) *Has Property* – "Can a *friend* be *important*?"

Note that for the edge-based approaches (methods 2 and 3a,b,c), starting from topics reduces the search for possible pairs from the size of the entire vocabulary used throughout the session (hundreds of words) to a much smaller set (6 words per topic, on average).

In order to transpose the graph representation to a human readable form, we use a number of publicly available language libraries and hand-coded patterns specific to each type of question. These libraries include *pylinkgrammar* [the Python implementation of Link Grammar (Sleator and Temperley 1995)] for checking grammatical correctness of the final result, and *pyinflect* for converting nouns to singular or plural forms. Patterns include fixed translations of edge types to human-readable forms. Finally, we filtered results using a number of hard-coded lists of words, so that no

inappropriate words would be included in the final suggestion list. These included any references to generally offensive words, words about religion, age, and gender.

6.1 Suggestion quality

As with our other methods presented in the previous sections, we conducted a crowdsourced evaluation on the final output of the question generation module. In doing so, our focus was twofold: for the suggestions to be usable in real-world scenarios, the suggestions need to be both grammatically correct and interesting to the readers. Since the quality of applications available for tablets is generally high, our users would quickly start ignoring our suggestions if they were not engaging. Therefore, we used the results of this survey to further filter out unsuitable suggestions.

We conducted the evaluation on the Crowdflower crowdsourcing market. Each worker was presented with a description of the purpose of the task, in which they were informed about the parent–child reading setting and the ultimately educational goals of the project. The survey respondents were restricted to English-speaking countries, but were otherwise not required to have any particular qualification. We made this design choice in order to have the largest pool of respondents and to avoid imposing further restrictions on our method; however, we anticipate that higher-qualified respondents may select more effective suggestions. The tasks presented to respondents showed the following description, which we phrased to be independent of the story: *A parent and her four-year-old child read a story together. Let's suppose that during the reading, the parent thinks of asking the following question to her child. Given this setting and the question, what is your position on the following statements? Focus on the educational and emotional impact the question would have on the child.* We intentionally refrained from providing further details on the story setting for two reasons: (1) to avoid any bias in interpreting the prompts; (2) to increase the degree in which this part of our approach could be automated: having a survey dependent on the story would have required manual task design for each new story.

We asked workers to evaluate individual suggestions using a ten agree/disagree questions. Each suggestion prompt was shown to respondent using the following template: *"The question [SUGGESTION] [SURVEY QUERY]"*, in which *[SUGGESTION]* is replaced by one suggestion generated via our method and *[SURVEY QUERY]* is the list of statements shown on the second column of Table 4. All questions answered in the survey are listed in Table 4. In total 294 suggestions were evaluated via 1470 judgments, where each judgment consisted of answering all ten evaluation questions. The average inter-user agreement per question was high at 70 %, measured as the percentage of votes in favor of the majority choice.

We designed the survey to evaluate two aspects regarding the suggestions generated by our method: (1) whether the respondents considered them appropriate to ask given the setting (items 1-4 in Table 4), and (2) whether their content was appropriate to ask to a child (items 5–10 in Table 4). With respect to the first set of questions, we notice that approximately two thirds of the questions were considered appropriate,

Table 4 Crowdsourced evaluation of the qualitative features of the suggestions generated by our system. For each question, the table shows the proportion of Yes (or Agree) answers and the inter-user agreement

Statement Number	Statement text	"Agree" Answer Proportion	Inter-user agreement
1	Is inappropriate/offensive to anyone	0.34	0.67
2	Is inappropriate only because of the setting, since it is addressed to a child	0.31	0.69
3	Is grammatically correct	0.67	0.7
4	Makes sense logically	0.72	0.71
5	Is worth asking because it is not something necessarily obvious to a four year old child	0.65	0.71
6	Would be considered interesting by most adults	0.38	0.68
7	References only concrete objects	0.51	0.69
8	Requires understanding an abstract concept	0.62	0.71
9	Uses words in their primary meaning	0.7	0.72
10	References concepts outside of the reach of a small child (e.g. theory of relativity)	0.71	0.73

grammatically correct or appropriate to ask a child. Obtaining perfect grammatical correctness was not one of our primary goals, underlined by the 67 % agreement in question 3. For the second set of questions, the study indicates that the prompts generated by our method covered a broader range of topics, given that the questions 7 and 8 respectively received 51 and 62 % participant agreement, showing that our suggestions prompts used both concrete and abstract concepts, which we consider a positive aspect given the educational goals of our application. The survey also revealed some contradictory answers, in particular with respect to the question being appropriate or not to ask. For example, the suggestion "Why does a bird have wing[s]?" received 100 % positive responses from the crowd on questions 5 (i.e. it is worth asking to a child), 80 % positive responses on question 6 (i.e. adults would consider it interesting), 80 % negative responses for question 2 (i.e. it is appropriate to ask), but at the same time 60 % agreement for question 10 (i.e. it is too difficult for a child). This may be due to some survey participants answering the questions literally, without placing them into the context of a shared parent–child setting (question 10 in particular). In the light of this obervation, we used a conservative threshold for the final selection of suggestions for the end-to-end user study, and selected out of the pool of 294 suggestions only those for which the responses had "Agree" answers with high confidence (over .80) for statement 5, resulting in a total of 186 suggestions.

7 User study on suggestion efficacy

For the final evaluation of our work, we conducted a user study to assess the impact our suggestions have on the interaction. The main goal of this study was to verify whether suggestions generated automatically are an effective method of eliciting verbal interaction between parents and children during reading sessions. The study involved 4–8 year old children ($n = 88, M = 5.64, \sigma = 1.33$) and their parents, recruited from the greater Boston area using a combination of email announcements to various family lists and invitations to a volunteer subject list maintained by the Personal Robots Group. We also offered a referral gift if parents recommended subjects that successfully attended the study. All subjects were informed of their rights and parents provided consent as mandated by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES—Committee on the Use of Humans as Experimental Subjects). Two subjects refused consent for video and audio recording and their data was excluded and destroyed, resulting in 86 transcriptions for our corpus. We did not limit the age range of the participating children in order to avoid further reducing the number of participants.

We tested the following hypotheses through the user study:

1. *Vocabulary* We hypothesize that participants exposed to discussion suggestions will talk more with their children, using a greater number of words and with greater variety during the reading session;
2. *Dialog* We hypothesize that participants exposed to discussion suggestions will engage more in conversation, as measured by the number of turn-taking exchanges and number of questions;
3. *Suggestion automation* We hypothesize that semantically-generated topic suggestions will result in effects comparable to expert-generated suggestions, as measured by the aforementioned vocabulary and dialog metrics.

The study took place in controlled lab settings. Figure 4 shows the study setting. Upon arrival to the lab, the parent and child were directed to the study room to sit in two chairs in front of a table (adult on the left, child on the right). Here, a video camera, microphone, and tablet were already present. The researcher started the recording using the application; audio, video, and data about any interaction with the tablet would stream into a computer, also in the room. The researcher would engage in a small talk with the child, for example, asking her name, age, and birth date. The researcher would then explain that the child could touch the words and pictures they see on the screen. Then, the researcher would tell the parent and child that they would be using the tablet to read a story together about a baby duck. The participants were instructed to talk about what they saw on the screen and to treat this like a book they would read together at home. As the two played together, they would have no further interaction with the researcher.

We evaluated two prompting conditions (CROWD and EXPERT) and compared them to a non-prompting baseline (NONE). Participants were assigned randomly to one condition before the study began, and only participated in that condition. We did not include automatic suggestion prompting in our study in order to reduce the influence of a particular delivery strategy. Instead, in order to deliver prompts, we added an

Fig. 4 Annotated video frame from parent–child interaction during user study

element to the TinkRBook application: an animated butterfly character which would appear on-screen which would delivery the prompts through a text-to-speech engine configured for a neutral tone. To avoid potentially causing inconvenience to the users by the tablet producing speech unexpectedly, the application required tapping on the butterfly in order to start speaking the prompts. If the users did not tap on the butterfly character, no suggestions were delivered. In the NONE condition, tapping the butterfly did not have any effect. Our study did not explore the issue of automatic suggestion delivery, and it also did not explore possible alternatives to the butterfly prompting character.

Participants in the CROWD condition viewed suggestions generated by the system described in this paper (186 prompts). When the primer is prompting from crowd-sourced questions, regardless of which scene a prompting character appeared, the prompt was chosen randomly from this entire set and presented to the readers. This is because our data collection method did not account for page numbers in the audio recording. We expect that correlating suggestions generated from vocabulary present on the page will improve the performance of our method. Participants in the EXPERT condition were shown samples from a pool of 28 suggestions manually created by literacy experts. Specifically, this pool of questions was created under guidance from a literacy expert (N. Lasaux, personal communication, April 2013) and from an educator guide to literacy conversations with children (Wasik & Iannone-Campbell, 2012). Each scene had its own limited and specific set of prompts from which the system randomly chose when the prompting character was tapped. The following are examples from the expert-generated set of prompts:

– Describe how Baby D (the protagonist) is feeling right now.
– What color can Baby D be?
– What are you thinking right now?
– What will happen next?
– What can ducks do?

The number of participants for each session was the following: 38 for NONE, 28 for EXPERT, and 20 for CROWD. For six reading sessions, all of which were in the EXPERT condition, users did not tap on the butterfly at all on the course of the reading session and therefore were not exposed to any prompt. In addition the the results presented below, we evaluated our results from the perspective of assigning conditions only based on exposure to the suggestion prompts (i.e. reassigning the six subjects that did not open suggestions to the NONE condition), and found no significant differences. Thus, we only present results corresponding to the initial condition assignments for the sake of brevity. We note that the focus of our study was not to investigate the effectiveness of the prompting method, however, we acknowledge it is essential that the users are likely to read suggestions for the method to be effective overall from the perspective of improving literacy. This behavior indicates it may be worthwhile to investigate delivery methods for prompts.

Furthermore, it is difficult to quantify the parents' understanding and interpretation the suggestion prompts, however, our premise of delivering prompts is that, especially since we are primarily targeting children that are not yet able to read, the parent will make use of them during the conversation. We consider this assumption further implies that (1) the suggestions are delivered in an effective manner that allows the parent to easily notice and interpret them, and that (2) the parent is willing and able to integrate these suggestions into conversation. Our study was not sufficiently instrumented to investigate these aspects, instead focusing on describing the interacting through the relatively broad metrics described below. In our analysis we provide an analysis on the metrics measured corresponding to the initial assignment, since it reflects the performance of our approach in its current implementation.

We transcribed audio recordings obtained during the study using the same method used to transcribe the training corpus and then applied the following set of vocabulary measurements on the transcriptions in order to characterize the verbal interaction:

1. *Full utterances* The number of all complete utterances, including phrases, sentences or distinct words, normalized across session length. For this metric, complete sentences count as one utterance, the same as single words that were not part of a sentence;
2. *Single words* The count of every word spoken during the interaction, normalized across session length. In conjunction with the full utterances metric, this metric can indicate the complexity of the sentences that were used – the greater the *single words* metric is compared to the *full utterances*, the longer the spoken sentences were;
3. *Novel words* Set of unique words not directly included in the story or the suggestions, indicating the richness of the vocabulary introduced by parents, normalized across session length;
4. *Lexical diversity* Measured as the ratio between the number of unique words and total number of spoken words (Dale and Fenson 1996).

One of our goals is for the interaction to incorporate more dialog between the parent and child as they are reading the story. We used a second set of metrics to evaluate the level of dialog present in the interactions using two metrics:

Table 5 Mean and standard deviation for vocabulary and dialog metrics for the three conditions

Metric	NONE	EXPERT	CROWD
Full utterances	0.16 ± 0.07	0.21 ± 0.06	0.20 ± 0.06
Single words	1.05 ± 0.53	1.25 ± 0.35	1.31 ± 0.40
Novel words	0.18 ± 0.08	0.22 ± 0.06	0.21 ± 0.08
Lexical diversity	0.45 ± 0.17	0.37 ± 0.07	0.40 ± 0.11
Turn-taking exchanges	0.07 ± 0.05	0.12 ± 0.06	0.10 ± 0.04
Question utterances	0.07 ± 0.04	0.09 ± 0.04	0.10 ± 0.05

The values were normalized for the session length (measured in seconds) and rounded to two decimals

Table 6 P values corresponding to an omnibus ANOVA (second column) *t* test together with pairwise *t* tests corrected using the Holm-Bonferroni method (columns 3–5)

Metric	ANOVA	NONE – EXPERT	NONE – CROWD	EXPERT – CROWD
Full utterances	**0.007**	**0.006**	**0.036**	0.663
Single words	**0.024**	0.160	0.110	0.610
Novel words	**0.049**	0.067	0.179	0.715
Lexical diversity	0.058	0.027	0.236	0.429
Turn-taking exchanges	**0.009**	**0.001**	0.066	0.078
Question utterances	**0.004**	**0.030**	**0.025**	0.655

Values significant at a 0.05 level are marked in bold

1. *Turn-taking exchanges* A measure of the number of conversational exchanges, normalized across session length, between the parent and the child, indicating a two-way conversation rather than just one participant speaking. The number was approximated from the number of transitions in the text transcripts;
2. *Question utterances* A count of the number of questions the parent asks the child during the interaction, approximated by the number of question marks present in the text transcript, normalized across session length (Blewitt et al. 2009).

Table 5 shows the mean and standard deviation results of tracking these metrics across the three conditions, while Table 6 shows the p values corresponding to an omnibus ANOVA *t* test together with pairwise *t* tests corrected using the Holm-Bonferroni method. We notice improvements in the number of utterances and the number of questions in both the CROWD and EXPERT condition over the non-prompting baseline, which imply that the readers had more verbal interaction while using the application. The number of dialog turns is significantly higher only in the EXPERT condition, however, the CROWD method may also indicate trend of increase (p value of 0.066). The lexical diversity metric was not significant in the omnibus tests. The single words and novel words metrics, which are arguably correlated with lexical diversity, did not show significant improvement between condition although the omnibus test indicated differences. Together, these metrics indicate that breadth of the vocabulary used during reading sessions did not vary significantly between our conditions, despite more verbal interaction occurring for the prompting conditions.

The pairwise test does not indicate significant differences between the EXPERT and CROWD conditions, yet a trend may be present showing a higher number of of turn-takes in the EXPERT condition (p value 0.078). Overall, the study reveals that both prompting conditions increased the amount of verbal interaction between participants, with the automated CROWD condition matching the EXPERT condition on two of the three metrics. We conclude that the CROWD condition showed comparable results to the EXPERT condition in increasing the amount of verbal interaction during reading sessions.

We also performed one-way MANOVA *t* tests between these conditions and found that only the EXPERT condition is globally statistically significant compared to both the NONE and CROWD conditions ($p < 0.01$). While the multivariate test does not indicate that the CROWD condition is significantly different from the NONE condition, we consider this to be partially explained by two factors: (1) the relatively high number of metrics relative to the number of samples per condition; (2) the higher standard deviation in the CROWD condition. With respect to the latter point, five of six metrics show a higher standard deviation compared to the EXPERT condition. This may be a result of the prompts generated using our method being more difficult to interpret by the parents, in which case some parents were able to use them in discussion better than others.

The results of our study reinforce existing results in the literature indicating that the interactivity and richness of joint parent–child reading can be improved through suggestion prompts. We consider that these results indicate that our method could offer a suitable alternative to expert-authored suggestions, and be particularly useful in domains where the input of a literacy expert is difficult to acquire, such as in applications with dynamic content. Our prompt generation strategy focuses on leveraging certain edge types from the semantic network, as described in Sect. 6. The study presented above could be expanded on in the future by introducing different types of automated prompt strategies, since our results indicate that prompts of broader scope, as those authored by literacy experts, are more effective than prompts focusing on facts, such as those produced by our system.

We note a difference between the output of our method and the expert-authored suggestions: latter suggestion set tends to include broader-scope prompts that are less dependent on specific concepts shown in the story. For example, the expert-authored set contained, in addition to the factual question types our system attempted to emulate, general questions about the protagonist such as "What will happen next?" Given our suggestion delivery method, which sampled uniformly from the prompt set instead of attempting to correlate the prompts with the story's content, the expert suggestions may be more effective because some are relevant at any point in the story. While our method could incorporate such general authored prompts to complement the topic-generated suggestions, it is not be capable of producing them algorithmically.

8 Conclusion

The system described and demonstrated in this paper is designed to stimulate dialog between parents and children by providing suggestions to the parent. It comes as

an alternative to guidance offered by child literacy experts, by enabling multimedia applications to support the parent in engaging verbally more with their child. Our contribution supplements a tablet application with the capability to offer intelligent feedback to its users, making the learning experience more effective for pre-literacy children while retaining a game-like approach. Our approach is independent of the application's content, making it suitable for integration with any other story-driven educational application. The methods we introduce are partly-unsupervised and independent of the content of the discussion. Instead of using word associations derived directly from the observed interactions, we use pre-existing semantic networks to infer topics and generate suggestions. By doing so, we are able to model topics from sparse transcriptions of unstructured noisy dialog.

Using a novel method, we generated topics and suggestions using a corpus of 45 parent–child discussions that were recorded in casual contexts during full traversals of the story in a non-prompting condition. We evaluate our system's output and show that both the topics and the suggestions produced by our system are meaningful for people by conducting crowdsourced surveys. We conducted a user study in controlled lab settings to evaluate the relative efficacy our method has compared to not delivering suggestions at all and compared against suggestions authored by child literacy experts. Our results indicate that the suggestions produced by our system have a significant positive impact on the interactions, overall increasing the level of dialog present in the joint reading sessions.

While literacy experts remain preferable if available, our results indicate that automated methods present a viable alternative, particularly in domains where access to expert data is limited or challenging (e.g., dynamically generated content). While our method requires human intervention, a large portion of it is crowdsourced. The annotation process of the audio recordings was outsourced to a professional service, however, a crowdsourcing approach could be expected to complete in near-real-time according to recent work in crowdsourcing which we cover in Sect. 2. While the topic modeling thresholds are determined manually based on survey outcomes, these are set once per story (i.e. training corpus) since they depend on the vocabulary used to generate topics. We also used crowdsourcing for filtering inappropriate suggestions using a threshold on between-user agreement. Because our method also has the potential advantage of generating a larger number of suggestions for each story, a conservative approach of only selecting high-agreement suggestions makes our method less sensitive to setting this threshold. In addition, having a large number of suggestions may lead to a higher interest in using the suggestions on repeated reading sessions. This advantage may prove useful especially if the reading application is used by child care professionals working with numerous children. This work also provides the future potential of customizing suggestions for individual users and for specific literacy goals, since topics and suggestions can be derived from the vocabulary of a subset of users.

One limitation of our method in its current form is its reliance on crowdsourced input. While high-speed crowdsourcing is available, it requires a budget for conducting surveys. With the goal of increasing the level of automation, we consider that our method could benefit from recent work in sentiment analysis. Work in this area has traditionally focused on sentiment analysis at coarser granularity, such as micro-blogs

(Pak and Paroubek 2010; Kouloumpis et al. 2011) or online product reviews (Pang and Lee 2008). Word-level sentiment associations are available in SenticNet (Cambria et al. 2014). Both these methods could be used in conjunction with our method to automatically filter suggestions and thus bypass one stage of crowdsourcing.

While both our method and the expert suggestions improve the speech interaction, we consider that the main reason for which our method is outperformed is that the expert set of suggestions contains a broader variety of prompts. In addition to prompts referring to relations between concepts, which is what our suggestion heuristics focus on, the expert set also contains broader narrative-related questions. The results presented in this study could motivate future research in narrative understanding for suggestion prompt generation. In addition, future research should investigate different prompt delivery strategies and interfaces, such that prompts are highly likely to be read and used, but without intruding on the conversation. One possible solution would be to use the on-board microphone to detect speech (i.e. the users speaking or not (Sohn et al. 1999), which could be arguably solved more reliably speech understanding) and deliver prompts during period of silence.

Acknowledgements This work was supported by National Science Foundation Award Number 1117584.

References

Aharony, N., Gardner, A., Sumter, C., Pentland, A.: Funf: open sensing framework (2011)

Alonso, J.B., Chang, A., Breazeal, C.: Values impacting the design of an adaptive educational storybook. In: International Conference on Interactive Digital Storytelling, pp. 350–353. Springer, Berlin (2011)

Arnold, D.S., Whitehurst, G.J.: Accelerating language development through picture book reading: A summary of dialogic reading and its effect. In: Bridges to literacy: Children, families, and schools, pp. 103–128 (1994)

Bennett, J., Lanning, S.: The netflix prize. In: Proceedings of KDD cup and workshop, vol. 2007, p. 35 (2007)

Bernstein, M.S., Brandt, J., Miller, R.C., Karger, D.R.: Crowds in two seconds: enabling realtime crowd-powered interfaces. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, ACM, pp. 33–42 (2011)

Bessho, F., Harada, T., Kuniyoshi, Y.: Dialog system using real-time crowdsourcing and twitter large-scale corpus. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 227–231. Association for Computational Linguistics (2012)

Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions, pp. 69–72. Association for Computational Linguistics (2006)

Blei, David M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)

Blei, David M, Lafferty, John D: Topic models. Text Min. **10**(71), 34 (2009)

Blewitt, Pamela, Rump, Keiran M., Shealy, Stephanie E., Cook, Samantha A.: Shared book reading: when and how questions affect young children's word learning. J. Educ. Psychol. **101**(2), 294 (2009)

Boteanu, A., Chernova, S.: Modeling discussion topics in interactions with a tablet reading primer. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp. 75–84. ACM (2013a)

Boteanu, A., Chernova, S.: Unsupervised rating prediction based on local and global semantic models. In: 2013 AAAI Fall Symposium Series (2013b)

Bouchard, Bruno, Bouzouane, Abdenour, Giroux, Sylvain: A smart home agent for plan recognition. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents And Multiagent Systems, AAMAS '06, pp. 320–322, New York, NY, USA, 2006. ACM. ISBN 1-59593-303-4. doi:10.1145/1160633.1160687

Boyd-Graber, J.L., Blei, D.M., Zhu, X.: A topic model for word sense disambiguation. In: EMNLP-CoNLL, pp. 1024–1033 (2007)

Burns, M.S., Griffin, P., Snows, C.E.: Starting out right. A guide promoting children's reading success. wdc (1999)
Bus, Adriana G., Ijzendoorn, Marinus H.Van, Pellegrini, Anthony D.: Joint book reading makes for success in learning to read: a meta-analysis on intergenerational transmission of literacy. Rev. Educ. Res. **65**(1), 1–21 (1995)
Cambria, E., Olsher, D., Rajagopal, D.: Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Twenty-eighth AAAI conference on artificial intelligence (2014)
Chang, A., Breazeal, C.: Tinkrbook: shared reading interfaces for storytelling. In: Proceedings of the 10th International Conference on Interaction Design and Children (2011)
Chang, A.: TinkRBooks: tinkerable story elements for emergent literacy. PhD thesis, Massachusetts Institute of Technology (2011)
Chang, A., Breazeal, C., Faridi, F., Roberts, T., Davenport, G., Lieberman, H., and Montfort, N.: Textual tinkerability: encouraging storytelling behaviors to foster emergent literacy. In: CHI'12 Extended Abstracts on Human Factors in Computing Systems, pp. 505–520. ACM (2012)
Dale, Philip S., Fenson, Larry: Lexical development norms for young children. Behav. Res. Methods, Instrum. Comput. **28**(1), 125–127 (1996)
Deerwester, Scott C, Dumais, Susan T, Landauer, Thomas K, Furnas, George W, Harshman, Richard A: Indexing by latent semantic analysis. JASIS **41**(6), 391–407 (1990)
Eisenstein, J., Barzilay, R.: Bayesian unsupervised topic segmentation (2008)
Fellbaum, Christiane: WordNet. Wiley, New York (1998)
Feng, M., Heffernan, N., Koedinger, K.: Addressing the testing challenge with a web-based e-assessment system that tutors as it assesses. In: WWW '06 Proceedings of the 15th international conference on World Wide Web, pp. 307–316 (2006)
Gerlitz, C., Helmond, A.: The like economy: social buttons and the data-intensive web. New Media & Society, page 1461444812472322 (2013)
Haruechaiyasak, C., Damrongrat, C.: Article recommendation based on a topic model for wikipedia selection for schools. In: Digital Libraries: Universal and Ubiquitous Access to Information, pp. 339–342. Springer, Berlin (2008)
Hausendorf, H., Quasthoff, U.: Patterns of adult–child interaction as a machanism of discourse acquisition. J. Pragmat. **17**, 241–259 (1992)
Havasi, C., Speer, R., Alonso, J.: ConceptNet: A Lexical Resource for Common Sense Knowledge, vol. 309, p. 269. John Benjamins Publishing Company, Amsterdam (2009)
Hilbert, Dana D., Eis, Sarah D.: Early intervention for emergent literacy development in a collaborative community pre-kindergarten. Early Child. Educ. J. **42**(2), 105–113 (2014)
Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80–88. ACM (2010)
Korat, O., Shamir, A.: The educational electronic book as a tool for supporting children's emergent literacy in low versus middle ses groups. Comput. Educ. **50**(1), 110–124 (2008)
Kouloumpis, Efthymios, Wilson, Theresa, Moore, Johanna D: Twitter sentiment analysis: The good the bad and the omg!. Icwsm **11**, 538–541 (2011)
Krause, A., Guestrin, C.: Data association for topic intensity tracking. Technical report, In: International Conference on Machine Learning (ICML) (2006)
Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: Proceedings of the Third ACM Conference on Recommender Systems, pp. 61–68. ACM (2009)
Lasecki, W.S., Miller, C.D., Bigham, J.P.: Warping time for more effective real-time crowdsourcing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2033–2036. ACM (2013)
Levelt, Willem J.M.: Models of word production. Trends Cogn. Sci. **3**(6), 223–232 (1999)
Linden, Greg, Smith, Brent, York, Jeremy: Amazon. com recommendations: Item-to-item collaborative filtering. Internet Comput. IEEE **7**(1), 76–80 (2003)
Linell, P.: Approaching Dialogue, vol. 1. John Benjamins Publishing Company, Philadelphia (1998)
Loper, E., Bird, S.: Nltk : The natural language toolkit. Processing **1**(July), 1–4 (2002)
Mashhadi, A.J., Capra, L.: Quality control for real-time ubiquitous crowdsourcing. In: Proceedings of the 2nd international Workshop on Ubiquitous Crowdsouring, pp. 5–8. ACM (2011)

McCallum, A.: Multi-label text classification with a mixture model trained by em. In: AAAI99 Workshop on Text Learning, pp. 1–7 (1999)
McCallum, A., Corrada-Emmanuel, A., Wang, X.: The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email (2005)
McFee, B., Bertin-Mahieux, T., Ellis, D.P.W., Lanckriet, G.R.G.: The million song dataset challenge. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 909–916. ACM (2012)
Miller, George A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
Modi, J., Veloso, M., Smith, F.S., Oh, J.: Cmradar: a personal assistant agent for calendar management. In: Lecture Notes in Computer Science, vol. 3508, pp. 393 (2005)
Nuñez, D.S.: GlobalLit: a platform for collecting, analyzing, and reacting to children's usage data on tablet computers. PhD thesis, Massachusetts Institute of Technology (2015)
Pak, Alexander, Paroubek, Patrick: Twitter as a corpus for sentiment analysis and opinion mining. LREc **10**, 1320–1326 (2010)
Pang, Bo, Lee, Lillian: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
Parish-Morris, Julia, Mahajan, Neha, Hirsh-Pasek, Kathy, Golinkoff, Roberta Michnick, Collins, Molly Fuller: Once upon a time: parent–child dialogue and storybook reading in the electronic era. Mind Brain Educ. **7**(3), 200–211 (2013)
Pazzani, M.J., Billsus, D.: Content-Based Recommendation Systems. The adaptive web, pp. 325–341. Springer, Berlin (2007)
Perkins, J.: Python Text Processing with NLTK 2.0 Cookbook. Packt Publishing Ltd, Birmingham (2010)
Pillinger, C., Wood, C.: Pilot study evaluating the impact of dialogic reading and shared reading at transition to primary school: early literacy skills and parental attitudes. Literacy **48**, 155–163 (2014)
Plisson, J., Lavrac, N., Mladenic, D., et al.: A rule based approach to word lemmatization. In: Proceedings of IS-2004
Ricci, Francesco, Rokach, Lior, Shapira, Bracha: Introduction to Recommender Systems Handbook. Springer, Berlin (2011)
Rich, Charles, Sidner, Candace L: Collagen: A collaboration manager for software interface agents. User Model. User Adapt. Interact. **8**, 315–350 (1998). doi:10.1023/A:1008204020038. ISSN 0924-1868
Schafer, J.B., Konstan, J., Riedl, J.: Recommender systems in e-commerce. In: Proceedings of the 1st ACM Conference on Electronic Commerce, pp. 158–166. ACM (1999)
Segal-Drori, O., Korat, O., Shamir, A., Klein, P.S.: Reading electronic and printed books with and without adult instruction: effects on emergent reading. Read. Writ. **23**(8), 913–930 (2010)
Sleator, D.D.K., Temperley, D.: Parsing english with a link grammar. arXiv preprint cmp-lg/9508004 (1995)
Sohn, Jongseo, Kim, Nam Soo, Sung, Wonyong: A statistical model-based voice activity detection. Signal Process. Lett. IEEE **6**(1), 1–3 (1999)
Sommer, S., Schieber, A., Hilbert, A., Heinrich, K.: Analyzing customer sentiments in microblogs–a topic-model-based approach for twitter datasets. In: Proceedings of the Americas Conference on Information Systems (AMCIS) (2011)
Speer, R., Arnold, K., Havasi, C.: Divisi: Learning from semantic networks and sparse svd. In: Proceedings of 9th Python in Science Conference (SCIPY 2010) (2010)
Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: In KDD Workshop on Text Mining (2000)
van Duursma, E., Augustyn, Marilyn, Zuckerman, Barry: Reading aloud to children: the evidence. Arch. Dis. Child. **93**(7), 554–557 (2008)
Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pp. 977–984, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi:10.1145/1143844.1143967
Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424–433. ACM (2006)
Weld, D., Adar, E., Chilton, L., Hoffmann, R., Horvitz, E., Koch, M., Landay, J., Lin, C., Mausam, M.: Personalized online education a crowdsourcing challenge. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
Wells, G.: Language development in the pre-school years, vol. 2. CUP Archive (1985)

Wenger, E.: Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge. Morgan Kaufmann, Burlington (2014)
Whitehurst, Graver J., Lonigan, Christopher J.: Child development and emergent literacy. Child Dev **69**(3), 848–872 (1998)
Wilbur, John W., Sirotkin, Karl: The automatic identification of stop words. J. Inf. Sci. **18**(1), 45–55 (1992)
Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Advances in Information Retrieval, pp. 338–349. Springer, Berlin (2011)

Dr. Adrian Boteanu is currently a Postdoctoral Associate at Cornell University, working in the Verifiable Robotics group. The work shown in this paper was done as part of his doctoral dissertation while A. B. affiliated with Worcester Polytechnic Institute. He received M.S. and B.S. degrees from the University Politehnica of Bucharest, Romania. His research interests lie in the areas of user modeling and feedback, semantic reasoning, and language grounding.

Dr. Sonia Chernova is the Catherine M. and James E. Allchin Early-Career Assistant Professor in the School of Interactive Computing at Georgia Tech. She received my Ph.D. and B.S. degrees in Computer Science from Carnegie Mellon University, and held positions as a Postdoctoral Associate at the MIT Media Lab and as Assistant Professor at Worcester Polytechnic Institute prior to joining Georgia Tech. S. C. directs the Robot Autonomy and Interactive Learning (RAIL) lab, which develops robots that are able to effectively operate in human environments. Her research interests span robotics and artificial intelligence, including semantic reasoning, adjustable autonomy, human computation and cloud robotics.

David Nunez is Engineering Director at Midnight Commercial, a Brooklyn-based design and innovation agency that creates novel works of art, design, and technology from concept to deployment. David achieved his M.S. in Media Arts and Sciences at the MIT Media Lab while working in the Personal Robots group, and he holds a B.A. degree in Computer Science and Managerial Studies from Rice University.

Dr. Cynthia Breazeal is an Associate Professor of Media Arts and Sciences at the Massachusetts Institute of Technology where she founded and directs the Personal Robots Group at the Media Lab. She is also founder and Chief Scientist of Jibo, Inc. She received her B.S. (1989) in Electrical and Computer Engineering from the University of California, Santa Barbara. She did her graduate work at the MIT Artificial Intelligence Lab, and received her M.S. (1993) and Sc.D. (2000) in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology.