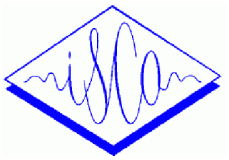
ISCA Archive
http://www.isca-speech.org/archive



Third Workshop on Child, Computer and Interaction (WOCCI 2012)
Portland, OR, USA
September 14, 2012

Coherence in Child Language Narratives: A Case Study of Annotation and Automatic Prediction of Coherence

Khairun-nisa Hassanali[1], *Yang Liu*[1], *Thamar Solorio*[2]

[1]Computer Science Department, The University of Texas at Dallas, Richardson, TX, USA
[2]Department of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, AL, USA

nisa@hlt.utdallas.edu, yangl@hlt.utdallas.edu, solorio@uab.edu

Abstract

Coherence is an important aspect of language ability. In this study, we analyze and annotate child language samples of story retell sessions for coherence and presence of narrative structure and narrative quality constructs. We use these constructs as features and use existing Natural Language Processing (NLP) techniques to build models that automatically predict coherence and language impairment in narratives. Our feature analysis results give us an insight into some of the important narrative quality features such as the use of cognitive inferences and social engagement devices. Our study shows that modeling of coherence in the context of language development in children is promising.

Index Terms: Natural language processing, child language, machine learning, coherence, narrative

1. Introduction

Speech or language interface is an important feature to consider when developing child computer interaction systems. A system with language understanding abilities will have a dramatic impact on many applications, such as language learning or measurement of language development, tutoring systems, companion for children, and intervention technology for autistic children.

Language sample analysis has been a common technique used by speech and language pathologists to measure language development. Some of the language development tests are based on grammatical and syntactical abilities. Such tests include the Developmental Sentence Scoring (DSS) [1] and Index of Productive Syntax (IPSyn) [2] which measure a child's proficiency in using certain syntactic constructs. Other measures such as the Peabody Picture Vocabulary Test (PPVT) and Total Percentage Phonemes Correctly repeated (TPPC) measure other aspects of language development.

The ability to coherently express oneself is also a mark of language development. As a child masters the syntax and semantics of the language, the child gathers the abilities to express himself more coherently. Coherence in language is especially more apparent in narratives where the narrative structure requires the narrator to be coherent in order to communicate effectively with the audience.

In this paper, we use story retells to analyze coherence in child language. To our knowledge, this has not been studied earlier. In story retells, all the children narrate the same story. This allows us to compare skills across children, which would be difficult with spontaneous speech. We had the story retells annotated by native English speakers for coherence, narrative structure and narrative quality features. We describe the annotation process and the models we built for automatic prediction of coherence. We perform feature analysis to see what are the top most features that account for the coherence of story retells. The top most features include the use of cognitive inferences, social engagement devices, instantiation of the story and the resolution of the story.

In the context of Language Impairment (LI), children with LI face more difficulty with language as opposed to Typically Developing (TD) children. It is therefore expected that children with LI will have different language skills compared to that of TD children. The question we ask ourselves is: Are TD children more coherent than children with LI? If so, what differentiates the TD children from children with LI? We use general coherence and narrative related features to build models to automatically predict LI from child story retells. Our analysis reveals that more than 80% of TD children produce coherent narratives as opposed to 35% of LI children. Our experiments show that while coherence and narrative structure related features are by themselves not sufficient to predict LI effectively, using these features in addition to existing features used in the prediction of LI significantly improves performance of the model.

2. Related Work

Narratives have been studied extensively in the context of language development. The Bus Story is a widely used measure of narrative ability in the UK and is a predictor

of persistent language impairment [3]. The "Frog, Where are You?" picture book has been used extensively in researching narratives produced by TD children [4, 5] and children with LI [6].

Language impairment has been studied in the communication disorder field. Children that perform according to the expected norm are called Typically Developing (TD) children whereas children who lack in some aspect of language development are called LI children. It is important to identify the LI children at an early age so that they can get help. Identifying the particlar aspects of language that LI children need help on will assist in developing technologies that focus on needs of children with LI. For example, softwares that teach a child language can focus more on the syntactic constructs that children with LI perform poorly on.

Traditional methods of detecting language impairment include cutoff methods on standard tests. Gabani et al. [7, 8, 9] explored the use of automated methods for analyzing transcripts of monolingual English speaking children to predict the presence or absence of language impairment. They exploit corpus-based approaches inspired by the fields of natural language processing and machine learning. They use features that focus on different aspects of language such as language productivity, morphosyntactic skills, vocabulary knowledge, probabilities from language models and sentence complexity. They compare results against a cut off baseline and find their methods are superior, reaching F-measures of above 73.7%.

Coherence has been used in the evaluation of linguistic quality. Pitler et al. [10] considered the use of coherence measures in the automatic evaluation of linguistic quality in multi-document summarization. Some of the coherence measures they considered were the use of local cohesive devices, adjascent sentence similarity, coreference chains and word co-occurence patterns.

Our work is unique since it deals with predicting and modeling coherence on child language transcripts. Coherence has so far been explored in the context of written text, for evaluating the quality of essays or automatically generated summaries. To our knowledge there has been no work done on automatically predicting coherence for child language transcripts. These child language transcripts pose a challenge since they contain disfluencies and other characteristics of spoken language. Our work differs from Pitler et al.'s in the features we use and our ultimate goal of predicting the level of coherence of a narrative. Further, we also look at coherence in the context of language development.

3. Data

The dataset we use for the experiments contains transcripts for a story telling task that is based on Mayers 24 page wordless picture storybook "Frog, Where Are You?" [11]. The story is about a boy and his two pets, a dog and a frog. In the night, the frog escapes from his jar and runs away. In the morning, the boy and the dog discover the frog is missing and set out searching for him. During the search they meet many different animals and experience a number of mishaps. Finally, they find their frog with a family of his own and take one of the baby frogs home with them [4].

This dataset contained 118 transcripts, of which 99 belonged to the TD group and 19 belonged to the LI group. The TD group consisted of 99 adolescents (61 female, 38 male), aged 14.5 years on average. English was their first and only language and they had no history of speech or language therapy. The LI group consisted of 19 adolescents (with the average age being 14.3) who had LI at one time point during the duration of the study [12].

4. Annotation Scheme

The annotators were six computer science undergraduate students who were native English speakers. The annotators were instructed to annotate the transcripts for coherence, narrative structures, and the usage of certain narrative quality constructs such as cognitive inferences and social engagement devices. For the annotation of narrative structures and narrative quality constructs, we follow the scheme proposed by Reily et al. [6]. We describe in detail the annotation process below:

4.1. Coherence

The narratives were annotated for coherence on the story level. While there are several specific types of coherence such as referential coherence, our intent was to measure coherence on a more general scale. The annotators were instructed to annotate the story as coherent if there was no block in understanding the story at any point in time. Furthermore, in order to validate their annotations and understand their concept of coherence, we also asked them to give us reasons why they judged a story as coherent or incoherent.

The narratives were annotated for coherence on 2 scales. The first scale was a 2-level scale: 0 if the story was coherent and 1 if the story was incoherent.

The second scale was a 3-level scale (coherent, somewhat coherent, and incoherent). Burstein et al. [13] discovered that annotating on a 3-level scale led to low inter-annotator agreement. If the annotator's response changed from coherent or incoherent to somewhat coherent, we asked them why they changed their answer. Our motivation behind getting a 3-level annotation was to verify two things: was there less inter-annotator agreement on a 3 level scale? and what caused the annotators to change their decision from coherent or incoherent to somewhat coherent?

We asked them to disregard spelling mistakes and grammatical errors during coherence annotation unless they felt it severely impacted their ability to understand the story. Furthermore, we asked the annotators to rate on a scale of 1 to 5 the extent to which they felt spelling mistakes and grammatical errors were hard to ignore, with 1 being that there was no difficulty in understanding the story, and 5 being that the annotators were really impacted by the spelling and grammatical mistakes.

4.2. Maintaining the Search Theme

As described earlier, the dominant theme in this narrative is the search for the frog. We felt that reiteration of the search theme would make for a more coherent story. It was also essential that the child mention the main basis for the story: the frog is missing in order for the story to be coherent.

Here we asked the annotators to annotate for the search theme as follows:

- 0: The child did not mention that the frog was missing and the boy was searching for the frog.
- 1: Only one of the following was mentioned: the frog is missing or the boy is searching for the frog.
- 2: The child mentioned both of the following: the frog was missing and the child was searching for the frog.
- 3: There was one additional mention of the boy searching for the frog.
- 4: There were two or more additional mentions of the boy searching for the frog.

4.3. Narrative Structure

One of the crucial factors of a coherent narrative is the presence of all the critical components of the narrative. We asked the annotators to go through the narratives and annotate the story for the presence or absence of the narrative structure components. A score of zero was given if the component was not present and a score of one was given if the component was present. The narrative structure components are:

- Instantiation of the story:
  This includes the frog is missing.
- Search episodes:
  The story consists of five search episodes. We asked the annotators to annotate for the presence or absence of each search episode.
- Resolution of the story:
  This includes the boy finding the missing frog and taking a baby frog home.

4.4. Narrative Quality

We examined the story retells for certain narrative quality features that we felt were crucial to an overall understanding and coherence of the narratives. The annotators were asked to mark the specific utterances in which the narrative quality features present. We describe these constructs below:

- Use of Cognitive Inferences :
  Cognitive inferences include inferences of character motivation, mental states and causality. An example of character motivation is “The boy turned around and pushed the dog off”. An example of mental states is “He thinks that the frog might be in the hole”.
- Use of Social Engagement Devices:
  Social engagement devices include using phrases or exclamations that are used to capture the audience attention. Examples of social engagement devices would be using sound effects such as “Woof!” to illustrate the dog barking. Another example would be the use of character speech such as “The boy said “Shhhhhh!”. These constructs make the narrative much more interesting and easier to understand. We asked the annotators to also look for audience hookers that they felt kept them engaged. An example of a audience hooker was “Look at the cute little doggie!”.
- Usage of Hedges:
  Hedges indicate a level of certainity that is expressed by the characters in the story. From a cognitive perspective, the use of hedges in the story tells us about the narrators thought process. An example of an utterance with a hedge would be “The boy probably thinks that the frog is in the hole”.
- References to Affective States:
  A reference to an affective state allows the audience to relate to the character in the story. It also makes the story much more interesting. An example of such an utterance would be “He was crying” or “The boy was suspicious about the deer”.
- Use of Intensifiers:
  The use of intensifiers in narratives provides emphasis on a certain action. Examples of the usage of intensifiers would be “The boy searched very hard for the frog” or “The boy searched and searched for the frog.” Here repetition gives us an insight into the usage of intensifiers.

4.5. Analysis

Table 1 gives the proportion of transcripts that were annotated on the 2-level coherence scale. While more than

Coherence Scale	TD	LI	Total
Coherent	81	6	87
Incoherent	18	13	31
Total	**99**	**19**	**118**

Table 1: TD and LI distribution on a 2-scale coherence level

Coherence Scale	TD	LI	Total
Coherent	45	6	51
Somewhat coherent	43	9	52
Incoherent	11	4	15
Total	**99**	**19**	**118**

Table 2: TD and LI distribution on a 3-scale coherence level

80% of the TD transcripts were judged as coherent, in comparision only 32% of the LI transcripts were judged as coherent.

Table 2 shows the results for the 3-level coherence annotation. As we can observe, there was a significant prortion of transcripts that moved from the coherent/incoherent category to the somewhat coherent category. Our analysis of the annotators' comments on the change of classification revealed that certain disfluencies such as repetitions and overall flow of story resulted in a change of classification from coherent to somewhat coherent. In narratives that were earlier classified as incoherent, if there was a reasonable flow the label changed from incoherent to somewhat coherent. Regarding the impact of spelling mistakes and disfluencies, the annotators gave a score of 2 and 3 to most of the transcripts on how hard they found it to ignore spelling mistakes, indicating that spelling and grammatical mistakes were moderately difficult to ignore.

We had 2 annotators annotate 37 transcripts and calculated interannotator agreement for coherence label. The overall interannotator agreement was 78.38% on the 2-scale coherence annotation and 42.34% on the 3-scale coherence annotations. As we can observe, there was greater disagreement on the 3-level scale compared to the 2-level scale annotation of coherence, which is as expected.

5. Automatic Prediction of Coherence

We treat the task of prediction of coherence in story telling transcripts as a binary classfication task: a transcript was classified as being coherent or incoherent. We explore different features for the prediction of coherence including narrative and CohMetrix features.

5.1. Narrative Features

The narrative features we used in the automatic prediction of coherence were as follows:

1. Search theme: This feature takes a value from 0 to 4 (0 being the search theme and frog is missing was not mentioned at all, and 4 being the search theme was mentioned at least thrice along with the fact that frog is missing was mentioned).

2. Narrative Structure: This feature set consists of seven features, each of which could take a value of 0 or 1. They denote the presence or absence of the instantiation, five search episodes, and the resolution of the story.

3. Cognitive inference: Number of occurences of cognitive inference constructs in the narrative.

4. Social engagement devices: Number of occurences of social enagement devices in the narrative.

5. Intensifiers: Number of intensifiers in the narrative.

6. Affective states: Number of references to a mental state in the narrative.

7. Hedges: Number of hedges present in the narrative.

The narrative features were based on manual annotation. We explored with the usage of binary features (presence or absence) for the narrative quality constructs (features 3-7 above) and found that using actual counts of narrative quality constructs as features worked better.

5.2. Coh-Metrix Coherence Features

Coh-Metrix [1] is a tool that provides an implementation of 54 features in the psycholinguistic literature that is known to correlate with coherence of human written texts. We therefore included the features in this study. Some of the features that are generated by Coh-Metrix tool are described below:

1. Readability metrics: These are the Flesh-Kincaid grade level and the Flesh reading ease score.

2. Situational model features: These are based on the micro-world that a text is about. Some of these features are: repetition score for tense and aspect.

3. General word and text features: These are based on surface text properties such as basic count and frequency features. Some of these features are: number of words, number of utterances and mean frequency of content words.

[1] For more information, refer to the document at http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm

Feature Set	Coherent			Incoherent			Accuracy (%)
	Precision	Recall	F-1	Precision	Recall	F-1	
Narrative Structure	0.869	0.839	0.854	0.588	0.645	0.615	78.814
Coh-Metrix	**0.950**	**0.970**	**0.960**	**0.824**	**0.737**	**0.778**	**93.220**

Table 3: Automatic classification of coherence on a 2-scale coherence level

4. Syntactic features: These assess the syntactic complexity of the text. Some of these features are: number of noun phrases, ratio of pronouns to noun phrases and number of connectives.

5. Referential and semantic features: These look at argument overlap and use Latent Semantic Analysis (LSA) to calculate conceptual similarity. Some of these features are: number of anaphora references between utterances, and proportion of adjascent utterances that share one or more arguments.

5.3. Results and Analysis

We evaluated several classifiers including the naive Bayes classifier, support vector machines, Bayesian network and logistic regression classifiers using the WEKA [14] toolkit. All the experiments were performed using leave one out cross validation. Of all these classifiers, the Bayesian network classifier performed the best. We only report the Bayesian network classifier results in Table 3. We can see the usage of narrative structure and narrative quality constructs as features yielded an accuracy of 78.814%. This is comparable to the inter-annotator agreement on classification of coherence. Using Coh-Metrix features yields an accuracy of 93.22% which is much higher than that using just the narrative structures. The fact that they perform well on oral narratives too makes the Coh-Metrix an important tool that can be used to model coherence of speech as well as text.

We performed feature selection using the narrative structure and narrative quality features to gain an insight into the features that contributed the most to the automatic prediction of coherence. Additionally, we also looked at the logistic regression coefficients to look at the features that have the largest positive coefficients. Based on our analysis the following features contributed the most to the automatic prediction of coherence:

1. Presence or absence of instantiation of the story
2. Number of social engagement devices
3. Presence or absence of resolution of the story
4. Number of cognitive inferences
5. Presence or absence of search episode 1

The presence or absence of instantiation and resolution of a story as top scoring features make sense since the introduction and conclusion of the story would contribute to the understanding of the audience. Furthermore, the first search episode follows the introduction and a well narrated episode also seems to contribute to the coherence of narratives. The use of social engagement devices makes a story more interesting. It is an interesting finding to see that social engagement devices contribute to the coherence of a story. The use of cognitive inferences on the other hand makes a story more clear since the narrator gives some insight to the audience.

6. Automatic Prediction of Language Impairment

As mentioned earlier, the corpus we used was annotated for LI status. 99 of the transcripts were produced by TD children and 19 transcripts by children with LI. Since more than 80% of TD children produced coherent stories compared to 32% of children with LI, we decided to use the coherence features in the prediction of language impairment.

Here we explored the use of the narrative structure, narrative quality, and the coherence scores as features in the automatic prediction of language impairment. In addition to the features used in the automatic detection of coherence, we used the 2-scale coherence score as a binary feature (0 for coherent and 1 for incoherent). We did not use the 3-scale coherence score as a feature due to the low inter-annotator agreement. Preliminary experiments also revealed that the 3-scale coherence score was not an appropriate feature. We also combine these features with those used in existing work, and examine whether these features add any improvement to existing work. We built several models using the naive Bayes, support Vector Machine, Bayesian network, and logitBoost classifiers using the WEKA toolkit. The best performance was obtained using the logitBoost classifier.

We report the results that were obtained using logitboost classifier in Table 4. These results were obtain using leave one out cross validation. We report the Precision (P), Recall (R), and F-1 measure (F-1). We consider LI as the class of interest and report those results.

We use the features used by Gabani et al. [9] as a baseline. The features they used are language productivity features, morphosyntactic skills, vocabulary knowledge features, speech fluency features, probabilities from language models, standard scores, sentence complexity,

Feature	P	R	F-1
Gabani [9]	0.737	0.737	0.737
Coherence	0.285	0.263	0.313
Coherence + Gabani	**0.889**	**0.842**	**0.865**

Table 4: Automatic classification of language impairment

and error patterns features. We augmented the feature set used by them with the coherence and narrative structure quality features. As we can see from Table 4, this results in an improvement of F-1 measure from 0.737 to 0.865 with an increase in both precision and recall. This results indicates that while the coherence score and narrative structures are not enough by themselves to effectively predict language impairment, they are definitely useful in predicting language impairment when combined with other features. Note, the coherence score and narrative structure features are based on human annotation. We plan to explore the automatic extraction of these features in future work.

7. Conclusion and Future Work

In this paper, we described a scheme for annotating coherence, narrative structure, and narrative quality features in story retells. We used the narrative structure and narrative quality annotations as features in the automatic prediction of coherence. We found the coherence related features generated by the Coh-Metrix tool to be quite effective in the automatic prediction of coherence.

Our analysis showed that more TD children produced coherent narratives as opposed to children with LI. This reinforces the fact that narratives are an important tool in measuring language development. We expect that findings from this study will be useful when designing child computer interaction systems that need to understand children's language or speech. In future, we plan to further explore coherence in child language transcripts and look at various aspects of coherence including exploiting cohesion related features in the prediction of coherence.

8. Acknowledgements

This research is supported by an NSF award IIS-1017190 and 1018124.

9. References

[1] L. L. Lee and S. M. Canter, "Developmental sentence scoring: A clinical procedure for estimating syntactic development in children's spontaneous speech," *Journal of Speech and Hearing Disorders*, vol. 36, no. 3, p. 315, 1971.

[2] H. S. Scarborough, "Index of productive syntax," *Applied Psycholinguistics*, vol. 11, no. 01, pp. 1–22, 1990.

[3] D. Bishop and A. Edmundson, "Language-impaired 4-year-olds: Distinguishing transient from persistent impairment," *Journal of Speech and Hearing Disorders*, vol. 52, no. 2, pp. 156–173, 1987.

[4] M. Bamberg and R. Damrad-Frye, "On the ability to provide evaluative comments: Further explorations of childrens narrative competencies," *Journal of Child Language*, vol. 18, no. 3, pp. 689–710, 1991.

[5] R. A. Berman, "On the ability to relate events in narrative," *Discourse Processes*, vol. 11, no. 4, pp. 469–497, 1988.

[6] J. Reilly, M. Losh, U. Bellugi, and B. Wulfeck, "frog, where are you? narratives in children with specific language impairment, early focal brain injury, and williams syndrome," *Brain and Language*, vol. 88, no. 2, pp. 229–247, 2004.

[7] K. Gabani, M. Sherman, T. Solorio, Y. Liu, L. M. Bedore, and E. D. Peña, "A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 46–55.

[8] K. Gabani, "Automatic identification of language impairment in monolingual English-speaking children," Master's thesis, The University Of Texas At Dallas, 2009.

[9] K. Gabani, T. Solorio, Y. Liu, K. Hassanali, and C. A. Dollaghan, "Exploring a corpus-based approach for detecting language impairment in monolingual english-speaking children," *Artificial Intelligence in Medicine*, vol. 53, no. 3, pp. 161–170, 2011.

[10] E. Pitler, A. Louis, and A. Nenkova, "Automatic evaluation of linguistic quality in multi-document summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 544–554.

[11] M. Mayer, *Frog, where are You?* Dial Press New York, 1969.

[12] G. Conti-Ramsden, N. Botting, and B. Faragher, "Psycholinguistic markers for specific language impairment (SLI)," *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 42, no. 06, pp. 741–748, 2001.

[13] J. Burstein, J. Tetreault, and S. Andreyev, "Using entity-based features to model coherence in student essays," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 681–684.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.