BERT-based Annotation of Oral Texts Elicited via Multilingual Assessment Instrument for Narratives

Timo Baumann and Korbinian Eller
Faculty for Informatics and Mathematics
OTH Regensburg, Germany
timo.baumann@oth-regensburg.de

Natalia Gagarina
Leibniz-Centre General Linguistics
Berlin, Germany
gagarina@leibniz-zas.de

Abstract

We investigate how NLP can help annotate the structure and complexity of oral narrative texts elicited via the Multilingual Assessment Instrument for Narratives (MAIN). MAIN is a theory-based tool designed to evaluate the narrative abilities of children who are learning one or more languages from birth or early in their development. It provides a standardized way to measure how well children can comprehend and produce stories across different languages and referential norms for children between 3 and 12 years old. MAIN has been adapted to over ninety languages and is used in over 65 countries. The MAIN analysis focuses on story structure and story complexity which are typically evaluated manually based on scoring sheets. We here investigate the automation of this process using BERT-based classification which already yields promising results.

1 Introduction

The ability to produce comprehensible oral narratives is a fundamental skill for functioning in society, and influences well-being and health (Bliss et al., 1998; McCabe, 1996). Narrative competence is therefore a key component of early childhood development, bridging the gap between spoken and written language (Hadley, 1998). A strong link between children's oral narrative abilities and early literacy, particularly reading (e. g. Catts et al., 1999; Sénéchal and LeFevre, 2002; Tabors et al., 2001; Charity et al., 2004; Reese et al., 2010), as well as broader academic and life success (Bishop and Edmundson, 1987; Gutiérrez-Clellen, 2002; McCabe, 1996; McCabe and Rollins, 1994; Norris and Bruning, 1988; Swanson et al., 2005; Torrance and Olson, 1984; Wallach, 2008) makes their understanding indispensable. Given the critical role of narrative skills in overall child development, they are increasingly used to diagnose early language disorders in both monolingual (Ringmann and Siegmüller, 2013; Schneider et al., 2006; Skerra et al., 2013) and bilingual children (Iluz-Cohen and Walters, 2012; Tsimpli et al., 2016), as well as to identify children at risk for delayed reading development (Reese et al., 2010; Suggate et al., 2011).

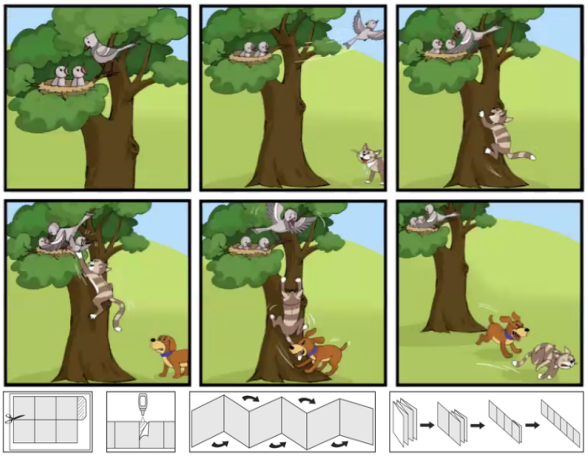

Figure 1: Example of the Baby Birds cartoon with multiple, partially overlapping story elements (bird feeds chicks, cat stalks chick, dog chases cat; reproduced with permission from Gagarina et al., 2012).

While there is a growing body of research on narrative acquisition, much of it is not grounded in theory-based materials. Instead, it often relies on existing wordless picture books and culturally specific materials, such as Frog, Where Are You? (Mayer, 1969; Berman and Slobin, 1994), Bus Story Test (Cowley and Glasgow, 1994), or Test of Narrative Language (Gillam and Pearson, 2004).

A group of researchers from the COST Action IS0804 Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment (www.bi-sli.org), closed the gap and created a theory-driven picture-based narrative elicitation tool featuring multiple parallel stories, the Multilingual Assessment Instrument for Narratives (Gagarina et al., 2012, 2019)[1], known as LITMUS MAIN, part of the LITMUS *Language Impairment Testing*

[1] https://main.leibniz-zas.de

in Multilingual Settings network. MAIN includes standardized pictorial stimuli, elicitation protocols, background questionnaires, and scoring methods for four stories: Baby Birds (shown in Figure 1), Baby Goats, Cat, and Dog.

In this paper we describe further the background and structure of the MAIN approach to assessing narrative capabilities, and the required annotations. We describe our corpus of annotated narrations in German and discuss our prototype system for automated annotation and its performance.

2 Theoretical Background of MAIN: Story Structure and Story Complexity

MAIN is grounded in a multidimensional model of high-order story organization or macrostructure, which suggests an alternative to the classical story grammar (Stein and Glenn, 1979), postulating that a comprehensive narrative includes seven components. The macrostructure represents the overarching structure of texts and exhibits a cross-linguistic nature (Heilmann et al., 2010). One of its key features is the correct representation of causal and temporal sequences. Smaller units within the macrostructure, known as episodes, are composed of individual components which are: an internal state as initiating event, a goal, an attempt, an outcome, and a resulting internal state. This model assesses episodes by means of both: story structure and complexity, providing a comprehensive framework for evaluating children's narrative skills.

Story structure components offer a quantitative measure of a narrative's macrostructure, while story complexity examines the combination of these components and evaluates narrative on a higher-order level. Essentially, the quantitative score reflects how many story structure components a child includes in their narrative, whereas the qualitative complexity score considers the interplay of goals, attempts, and outcomes within an episode.

This approach provides a comprehensive evaluation of narrative macrostructure by considering both quantity (the total number of episode components) and quality (the complexity level based on how these components are combined). In this paper, we focus on narrative structure rather than complexity.

3 Elicitation and Annotation of MAIN

Child language researchers all over the world use the MAIN elicitation schema to transcribe and annotate data manually.[2] They use the annotation described in the scoring sheets, e. g. to assess the need for interventions based on the total of episode components in the narrative (0–17).

MAIN narrative elicitation is conducted according to detailed guidelines[3] by trained native speakers. For bilingual children, MAIN is conducted several times so that different stories, e. g. Cat and Dog are collected in either language (but note that stories are structurally similar). Elicitation usually begins with warm-up questions, followed by the presentation of two or three colored envelopes. The child takes one envelope, opens it and takes a folded cartoon as shown on Figure 1. The child then tells or retells the story and answers comprehension questions. The child's production is audio-recorded and transcribed both verbatim and orthographically normalized in the CLAN format (MacWhinney, 2000).

Once the oral text is transcribed, the annotator manually identifies the presence or absence of story components as described in a scoring sheet (see Example 1 and Table 1).

```
@G: 1
*CHI: Ok, eines Tages war hm war die Vogelmutter bei
      ihren Kindern.
*CHI: Und hat auf Vogelsprache [x2] gesagt, sie
     solln (sollen) hier kurz warten, weil sie Es
     [//] Fressen holen will.
@G: 2
*CHI: Und dann flog sie weg.
*CHI: Aber eine Katze hat gesehen, dass die Kueken
     ganz allein sind, also die Entenkinder ganz
     allein sind.
*CHI: Und deswegen dachte sie, sie kenn [//] sie
     haette gutes Frass gefunden.
@G: 3
*CHI: Dann kletterte die Katze auf den Baum und
     wollte sich ein Vogel schnappen.
@G: 4
*CHI: Aber ein Hund bemerkte das und wollte nicht
     zulassen, dass die Katze die ho [//] die Voegel
      frisst.
@G: 5
*CHI: Also biss der Hund ihr in den Schweif.
*CHI: Und dann [//] und damit hat er sie abgehalten
     &hm und damit hat er sie abgehalten, ein Vogel
     zu essen.
*CHI: <Die Vogelmutter hat es bemerkt> [x2] und
     deswegen hat sie sich erschrocken.
@G: 6
*CHI: Der Hund hat sie runtergeholt und sie gejagt.
*CHI: Und die Voegel [//] und die Voeg [//]
     Vogelmutter mit ihren Kueken, also Vogelbabys,
     waren ziemlich froh.
*CHI: Und die Geschichte jetzt zu Ende.
*EX1: Ok.
```

Example 1: Baby Birds narrative of a child, 9 years 10 months. Each utterance is segmented as a sentence and starts with the sign `*`. `@G` markers indicate progression through the pictures of the cartoon. `[x2]` indicates repetition, `[//]` indicates pausing.

[2] https://main.leibniz-zas.de/en/worldwide-network

[3] http://www.zas.gwz-berlin.de/zaspil.html

Table 1: Scoring sheet for the cartoon depicted in Figure 1 and narrated in Example 1.

		Examples of correct responses	Score
A1.	Setting	Time and/ or place reference, e.g. once upon a time/ one day/ long ago... in a forest/ in a meadow/ in a garden/ in a field/ in a bird's nest/ up a tree	0 1 2
Episode 1: Mother/ Bird (Episode characters: mother bird and baby birds)			
A2.	IST as initiating event	**Baby birds** were hungry/ wanted food/ cried for food/ asked for food **<Mother/ Bird/ Parent, etc.>** saw that baby birds were hungry/ wanted food	0 1
A3.	Goal	**Mother bird** wanted to feed baby birds/ to catch/ bring/ get/ find food/ worms (In order) to + VERB (get food)	0 1
A4.	Attempt	**Mother bird** flew away/ went away/ looked for food/ was fetching food **Mother bird** tried to + VERB (get food)	0 1
A5.	Outcome	**Mother bird** got/ caught/ brought/ came back with food/ a worm/ fed the babies **Baby birds** got food/ a worm	0 1
A6.	IST as reaction	**Mother bird** was happy/ satisfied/ pleased **Baby birds** were happy/ satisfied/ pleased/ not hungry any more	0 1
Episode 2: Cat (Episode characters: cat and baby bird(s))			
A7.	IST as initiating event	**Cat** saw mother flying away/ saw that baby birds were all alone/ saw that there was food **Cat** was hungry/ thought "yummy"	0 1
A8.	Goal	**Cat** wanted to eat/ catch/ kill baby bird/-s (In order) to + VERB (eat, catch, kill, get)	0 1
A9.	Attempt	**Cat** was/ is climbing up the tree **Cat** tried to reach/ get baby bird **Cat** climbed/ jumped up (the tree)	0 1
A10.	Outcome	**Cat** grabbed/ got baby bird **Cat** nearly/almost + VERB (caught, got)	0 1
A11.	IST as reaction	**Cat** was happy **Bird/-s** was/ were scared/ crying/ screaming with pain	0 1
Episode 3: Dog (episode characters: dog, cat and baby bird(s))			
A12.	IST as initiating event	**Dog** saw that the bird was in danger/ saw that cat caught/ got the bird **Bird/-s** was/were in danger	0 1
A13.	Goal	**Dog** decided/ wanted to stop the cat **Dog** decided/ wanted to help/ protect/ save/ rescue the bird(-s) (In order) to + VERB (stop, rescue, help)	0 1
A14.	Attempt	**Dog** was/is pulling/ dragging the cat down/ biting/ attacking the cat/ grabbing the cat's tail **Dog** tried to + VERB (pull, drag, get down) **Dog** pulled/ dragged the cat down/ bit/ attacked the cat/ grabbed the cat's tail	0 1
A15.	Outcome	**Dog** chased the cat (away)/ scared the cat off/ away **Cat** let go of the baby bird/ ran away **Bird/-s** was/ were saved/ rescued	0 1
A16.	IST as reaction	**Dog** was relieved/ happy/ proud (to have saved/ rescued the baby bird) **Cat** was angry/ disappointed/ feeling bad/ mad/ scared/ in pain/ cat's tail hurt **Bird/-s** was/ were relieved/ happy/ safe **Mother bird** was relieved/ happy	0 1
A17.		**Total score out of 17:**	

The components consist of terms describing the setting and then each of the three episodes of the story can consist of opening *internal state terms* (IST), a description of the attempted action, its goal and the outcome of the action, again followed by closing IST.

We focus our study below on the binary criteria A2–16 (A1 is ternary), which can be grouped as 3 groups of quintuples, one for each of the three episodes, and the sum of A2–16.

4 Dataset

We work with 927 narrations (roughly equally distributed among the four cartoons) in German, collected mostly from children aged 5–9 years most of which are bilingual. They contain a total of 20,894 utterances with 122,104 words for an average of 23 utterances per narration and 5.8 words per utterance.

Table 2 reports the average scores achieved by the subjects in each criterion as well as averaged over and across episodes. Overall, we find that criteria differ (with ISTs being most difficult to achieve) but that averaged scores are similar across the three episodes.

Table 2: Average scores for binary criteria (A2–16) in the corpus and their averages.

	IST	goal	attempt	outcome	IST	mean
Episode 1	.33	.22	.51	.54	.03	.32
Episode 2	.21	.37	.52	.57	.15	.36
Episode 3	.25	.12	.52	.61	.18	.34
mean	.26	.23	.52	.57	.12	.34

The sum of A2–16 for each subject has a broad, fairly normal distribution (min/max: 0/13) and a mean/stddev/median of 5.1/2.7/5.

5 Classifier Implementation

We have implemented a prototype of an automated annotator that classifies texts wrt. the fifteen binary features. Following the discussion by Johannßen et al. (2020) and to leverage the power of pre-trained models, we build classifiers based on BERT-extracted features as has been done for psychometric scoring (Schäfer et al., 2020) which arguably is roughly similar to our task.

We tokenize, parameterize and aggregate the (orthographic) textual representation of the narration with the transformers library (Wolf et al., 2019) using a German cased BERT model[4]. We did not yet experiment with other models or fine-tune the base model. In the rare cases that the text exceeds the token limit of the transformer, we truncate it.

The BERT aggregation is followed by one inner layer followed by the classification layer. We implement three approaches for the classification:

Single implements 15 individual binary classifiers for each of the 15 features, which are trained in isolation.

Multi shares the inner layer among the 15 binary classifiers, which may help to overcome sparsity and overfitting.

Multi-G receives four BERT aggregations, one for each episode of the story in addition to the full text (as above) and then shares the inner layer.

We use a 512-dimensional inner layer with dropout before and after, a decision that we did not fine-tune. We train each model for 2000 epochs using SGD and a learning rate of .01. In preliminary experiments with the Single setup, we found

[4] http://huggingface.co/dbmdz/bert-base-german-cased

Table 3: F-measure for individual classification decisions (and their aggregations) for the three models.

	IST	goal	attempt	outcome	IST	
Single						
Episode 1	.39	.23	.69	.79	0	
Episode 2	0	.70	.75	.74	0	
Episode 3	.22	0	.74	.82	.20	
overall						.42
Multi						
Episode 1	.38	0	.71	.73	0	
Episode 2	0	.69	.79	.81	0	
Episode 3	0	0	.79	.85	.48	
overall						.42
Multi-G						
Episode 1	.35	.19	.58	.70	0	
Episode 2	.27	.64	.75	.71	0	
Episode 3	.53	.19	.73	.89	.22	
overall						**.45**

the models overfitting for some classes early while only yielding meaningful classifications after very many epochs for others. This is why we chose a large number of epochs. We randomly split our data into 90 % training and 10 % test data.

Each automated annotator also computes the sum of the positive classifications which is similar to the total score on the scoring sheet (except that the score for the three-valued A1 is missing).

6 Results and Discussion

We evaluate all classifiers by the individual and average F-measures for the binary classifications which we report in Table 3. We furthermore compute the *root mean squared error* (RMSE) of the estimated score vs. the sum of human annotations.

We find that classification performance differs radically across categories while it is more stable across episodes. Specifically, the presence of internal state term components seems to be most difficult to estimate and there is a tendency of

$$outcome > attempt > goal > IST.$$

While the overall performance in F-measure is not very high, the performance for some categories, specifically outcome and attempt, appear usable.

It is interesting to note that an outcome is the most concretely observable and an attempt a slightly more abstract (and a goal even more abstract) property of a story. It may be that the linguistic variation for describing more abstract properties is higher and that therefore models perform worse. We cannot exclude that class imbalance also weakens the performance (see Table 2).

The performance of the classification approaches is quite similar and we are surprised that apparently features that are relevant to describe categories in different positions of the story (early, mid, end) are properly retrieved from the 768 BERT features. Overall, Multi-G yields slightly higher performance which is also more equalled out across the different categories. Single is much slower to train without providing any benefit.

With respect to RMSE of the aggregated scores, we find Multi-G (2.20) to be inferior to Multi (1.89) and both much better than Single (4.04). We believe that the individual decisions of the Single classifier are much more correlated than in Multi-G and Multi (as they take decisions individually) and hence that errors, when they happen, are also more clustered for instances. In cases where the overall aggregate is used for narration assessments (e. g. via thresholds for interventions), a lower RMSE may be more relevant than a higher F-measure.

7 Conclusions, Limitations and Future Work

We find that some of the annotation categories can already be automatically inferred from the transcribed texts alone. However, we intend to analyze further the influence of age, bilinguality, and other factors known about the subjects on their narrative performance. Beyond our current prototype, we believe that the classification performance of our models can still be boosted significantly, for example by fine-tuning the underlying BERT parameters.

Automatic speech recognition transcripts of developmental language use are often riddled with further difficulties, which is why we focused on human transcripts in the present study. In future work, we intend to study the interrelations of narrative capabilities with lexical and phonetic development. While we believe that such interrelations could be useful to inform the narration annotation with additional information from the speech signal, we are also interested in studying the more general developmental implications.

We believe that final judgements about interventions on subjects, especially children, should always be made by qualified human experts. However, this resource is limited and a gradation of simple cases can help free this resource to actually help in interventions rather than over-focusing on the assessment.

Acknowledgements

We thank Nathalie Topaj and Alyona Sternharz for help in compiling and proofing the dataset annotations as well as Felix Wensky for help with programming and analytics.

References

R.A. Berman and D.I. Slobin. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Lawrence Erlbaum, Hillsdale, NJ.

D. V. M. Bishop and A. Edmundson. 1987. Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders*, 52(2):156–173.

Lynn S. Bliss, Allyssa McCabe, and A. Elisabeth Miranda. 1998. Narrative assessment profile: Discourse analysis for school-age children. *Journal of Communication Disorders*, 31(4):347–363.

Hugh W. Catts, Marc E. Fey, Xuyang Zhang, and J. Bruce Tomblin. 1999. Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, 3(4):331–361.

Anne H. Charity, Hollis S. Scarborough, and Darion M. Griffin. 2004. Familiarity with school english in african american children and its relation to early reading achievement. *Child Development*, 75(5):1340–1356.

J Cowley and C Glasgow. 1994. The renfrew bus story–north american edition.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ingrida Balčiūnienė, Ute Bohnacker, and Joel Walters. 2012. MAIN: Multilingual assessment instrument for narratives. *ZAS Papers in Linguistics*, 56.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: Multilingual assessment instrument for narratives–Revised. *ZAS Papers in Linguistics*, 63.

Ronald Bradley Gillam and Nils A Pearson. 2004. *Test of narrative language*. Pro-ed, Austin, TX.

Vera F Gutiérrez-Clellen. 2002. Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education*, 13(2):175–197.

P. A. Hadley. 1998. Early verb-related vulnerability among children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 41:1384–1397.

John Heilmann, Jon F. Miller, Ann Nockerts, and Claudia Dunaway. 2010. Properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology*, 19(2):154–166.

Peri Iluz-Cohen and Joel Walters. 2012. Telling stories in two languages: Narratives of bilingual preschool children with typical and impaired language. *Bilingualism: Language and Cognition*, 15(1):58–74.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer. 2020. Germeval 2020 task 1 on the classification and regression of cognitive and motivational style from text. In *Proceedings of the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 1–10, Zurich, Switzerland.

Brian MacWhinney. 2000. The childes project. *Computational Linguistics*, 26(4):657–657.

Mercer Mayer. 1969. *Frog, Where are you?* Dial Press, New York.

Allyssa McCabe. 1996. Relating events in narrative: a crosslinguistic developmental study. *Journal of Child Language*, 23(3):715–723.

Allyssa McCabe and Pamela Rosenthal Rollins. 1994. Assessment of preschool narrative skills. *American Journal of Speech-Language Pathology*, 3(1):45–56.

Janet A. Norris and Roger H. Bruning. 1988. Cohesion in the narratives of good and poor readers. *Journal of Speech and Hearing Disorders*, 53(4):416–424.

Elaine Reese, Alison Sparks, and Diana Leyva. 2010. A review of parent interventions for preschool children's language and emergent literacy. *Journal of Early Childhood Literacy*, 10(1):97–117.

Svenja Ringmann and Julia Siegmüller. 2013. Die Beziehung zwischen Satzgrammatik und Erzählfähigkeit im unauffälligen und auffälligen Spracherwerb. *Forschung Sprache*, 1(1):36–50.

Henning Schäfer, Ahmad Idrissi-Yaghir, Andreas Schimanowski, Michael R Bujotzek, Hendrik Damm, Jannis Nagel, and Christoph M Friedrich. 2020. Predicting cognitive and motivational style from german text using multilingual transformer architectures. *Proceedings of the GermEval 2020 Task*, 1:17–22.

Phyllis Schneider, Denyse Hayward, and Rita Vis Dubé. 2006. Storytelling from pictures using the Edmonton narrative norms instrument. *Journal of speech language pathology and audiology*, 30(4):224.

Monique Sénéchal and Jo-Anne LeFevre. 2002. Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, 73(2):445–460.

Antje Skerra, Flavia Adani, Natalia Gagarina, T Fritzsche, CB Meyer, A Adelt, and J Roß. 2013. Diskurskohäsive Mittel in Erzählungen als diagnostischer Marker für Sprachentwicklungsstörungen. *Spektrum Patholinguistik*, 6:127–158.

NL Stein and CG Glenn. 1979. An analysis of story comprehension in elementary school children. *New directions in discourse processing/Ablex*.

Sebastian P. Suggate, Elizabeth A. Schaughency, and Elaine Reese. 2011. The contribution of age and reading instruction to oral narrative and pre-reading skills. *First Language*, 31(4):379–403.

Lori A. Swanson, Marc E. Fey, Carrie E. Mills, and Lynn S. Hood. 2005. Use of narrative-based language intervention with children who have specific language impairment. *American Journal of Speech-Language Pathology*, 14(2):131–141.

Patton O Tabors, Catherine E Snow, and David K Dickinson. 2001. *Homes and schools together: Supporting language and literacy development.* Paul H. Brookes Publishing Co.

N Torrance and D Olson. 1984. Oral language competence and the development of literacy. *The development of oral and written language in social contexts rpo*, pages 167–182.

Ianthi Maria Tsimpli, Eleni Peristeri, and Maria Andreou. 2016. Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics*, 37(1):195–216.

Hanna Megan Wallach. 2008. *Structured topic models for language*. Ph.D. thesis, University of Cambridge Cambridge, UK.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.