4 Methods Study I

This chapter first details the selection and characteristics of participants in this study[19], followed by a description of the test instruments and the procedure including a detailed explication on how the main variables were quantified, and concludes with a presentation on the strategy governing the data analysis.

4.1 Participants

All participants were recruited through ECEC institutions[20] from districts with high percentages of DLL children in a large German city. The heads of ECEC institutions were contacted by telephone and asked a), if they had more than one DLL Turkish-German child in their institution and b), if they would be interested participating in a research project. If they answered "Yes" to both conditions, brochures were sent out detailing the primary aims of the study as well as its requirements for participation. One to two weeks later, the institutions were contacted again and asked if they were still interested in participating. If so, the institutions were visited to talk through the timeline of the assessments and to clarify any questions.

To be eligible to participate in the study, the children had to meet the following criteria:

- Be between 3 and 6 years of age
- Be DLLs[21] of both Turkish and German, as well as having been systematically exposed to German for at least 10 months
- Are developing at a typical level according to parent(s) and ECEC report(s)

---

[19] Data collection was funded by a research grant by Niedersächsisches Institut für Frühkindliche Bildung und Entwicklung (nifbe) awarded to Ulrike M. Lüdtke and Ulla Licandro, née Grube (nifbe Az. FP 01-12). The author served as the principal investigator and has no financial or nonfinancial relationships relevant to the content of the study.

[20] All ECEC institutions were monolingual German (i.e., all ECEC practitioners exclusively conversed in German).

[21] For the purpose of this study, to be classified as a DLL the children needed to have systematic language contact with German for at least 10 months, attend a German ECEC institution, and produce output in both languages daily (in families who conversed in Turkish exclusively, this criterion was limited to weekdays–when German was spoken in the ECEC institution).

U. Licandro, *Narrative Skills of Dual Language Learners*, Diversität in Kommunikation und Sprache / Diversity in Communication and Language,
DOI 10.1007/978-3-658-14673-3_4

- Are not receiving speech and language services at the time of their participation in the study (according to parental report)

To adhere to current research ethics, written informed consent was obtained from both parents/guardians as well as each child prior to the start of the study, as detailed below.

Informed Consent and Child Assent

To collect parental consent, bilingual (Turkish and German) study information and consent forms were given out to the primary caregivers of potential participants. Questions concerning the study were answered via telephone and/or at a meeting, either in Turkish or in German, depending on the parents' language of choice. Signed consent forms were collected prior to the start of the study. Families did not receive compensation for their participation.

To make allowance for the children's wishes in regard to potential research participation (Dockett & Perry, 2011), child assent was obtained in the ECEC institutions in the following manner: After a familiarization period characterized by engagement in joint play, the examiner sat down with each child individually and walked them through a document that stated the goals and means of the study using child-appropriate language and pictures. Children were then asked if they wanted to participate and, if so, to sign their name or draw a picture on the bottom of the form.

Parental Report on Home Language Use and Proficiency

To profile each participant's language exposure and proficiency (e.g., Guitiérrez-Clellen, Simon-Cereijido, & Wagner, 2008) and to collect further home language data, a parental questionnaire was designed. It consisted of 31 items, which were partly drawn from the bilingual parental questionnaires designed by Asbrock and colleagues (Asbrock, Ferguson, & Hoheiser-Thiel, 2011), Chilla and colleagues (Chilla et al., 2010), and the Alberta Language Development Questionnaire (ALDeQ; Paradis, Emmerzael, & Duncan, 2010). The first part of the questionnaire targeted basic child data (i.e., birthdate, date of joining an ECEC institution, family size). Parental education (here: measured in years) is one factor associated with the family's socio-

economic status which has previously shown direct correlations with children's linguistic and cognitive performance (e.g., Gathercole, Kennedy, & Thomas, 2015; Hair, Hallo, Terry-Humen, Lavelle, & Calkins, 2006) and was therefore included in the questionnaire and used as a demographic control variable (also see Aukrust & Rydland, 2011; Rydland, Grøver, & Lawrence, 2014a, 2014b). To track home language practices, parents were asked to give an hourly breakdown of input in both languages on typical weekdays and weekends. This account included listing all members of the household with whom the child interacted on a regular basis, and a report on each of those person's language abilities (either Turkish only, German only, or mixed)[22]. Furthermore, home literacy practices were addressed by estimating the amount of books in the household and the frequency of shared storybook reading activities. Finally, parents rated their own as well as their child's proficiency in both Turkish and German on a scale from 1 (limited) to 4 (fluent).

A Turkish-German bilingual/bicultural research assistant arranged a meeting with the parent at the child's ECEC institution and administered the questionnaire interview-style in either German or Turkish, depending on the parent's preference. The completion of the questionnaire took around 20 minutes and, except for two cases where the father answered the questions, was conducted with the mother.

Child Demographics

In total, written consent was collected for 56 children, and of those, 5 were excluded from further analyses. Four children did not complete the testing battery (2 children did not participate in any assessment due to repeated absence from the ECEC institution, and 2 children failed to complete the language assessment). Furthermore, one case was eliminated because the child did not have the narrative sample available due to failure of recording equipment.

[22] This detailed breakdown of communication partners was collected because bilinguals may not always consciously notice which of their languages is being spoken (Guiterrez-Clellen & Kreiter, 2003).

The final study sample consisted of 51 Turkish-German DLLs from 15 ECEC institutions in central Germany, all of which enroll large numbers[23] of bilingual children. On average, children were 58 months old (M = 4.83 years, Mdn = 4.75 years, SD = 0.61) and had a mean systematic exposure to German, as assessed by parental report, of 32 months (SD and ranges appear in Table 3). Of the sample, 61% of the children were female (n = 31) and 39% were male (n = 20). Paternal education[24] widely varied across the sample. While six (12%) of the participants' mothers and fathers had no or basic education (i.e., up to four years), the majority had participated in formal schooling for ten years (mothers: n = 24; 47%; fathers: n = 20; 39%). Three mothers (6%) and six fathers (12%) had obtained a university degree. Information on parental education was missing for two children.

While all children were born in Germany and had been living in the country since then, the majority of the children came from successive language backgrounds (i.e., no systematic exposure to German before their third year of life (Chilla, Rothweiler, & Babur, 2010): Forty-five percent of the children learned German and Turkish from age 2 or earlier and 55% of the children started learning German at age 3 or later. The range of language input and output values indicated that the children were spread across the full range that was considered DLL for this study.

Although children were experiencing variation in how much Turkish and German was spoken in their homes, all attended German-only ECEC institutions. While all children had been exposed to Turkish from birth on and were currently exposed to both German and Turkish, at the time of testing, children's language practice spanned the full range from predominant Turkish use to predominant German use.

Based on the children's contact months as well as family and ECEC exposure to each language and the children's patterns of language output, 34% of the children were deemed Turkish dominant (using Turkish over 60% of the time), 32% of the children were balanced bilinguals (using Turkish and German 40 to 60% of the time), and the

[23] At the time of the study, participating institutions enrolled at least 50% DLLs with various language and cultural backgrounds, as reported by the heads of the institutions.

[24] Education in the home country was included in this calculation.

remaining 34% of the children were German dominant (using German over 60% of the time). The average exposure to German was 2 years and 7 months. Forty-five percent of the children were systematically exposed to both Turkish and German before their third year of life, and 55% of the children started learning German at age 3 or later.

Table 3. *Summary Characteristics of Child and Family Demographics*

Variable (*N* = 51)	**Mean (*SD*)**	**Range**
Age in months	57.82 (7.24)	44-72
Mother's education in years [a]	9.82 (3.21)	0-17
Father's education in years [b]	9.96 (3.74)	0-17
Family size (total number of children)	2.22 (1.07)	1-6
Mother's self-rated proficiency in German [c]	2.35 (.64)	0-4
Father's self-rated proficiency in German [c]	2.43 (.65)	0-4
Mother's frequency of language mixing [a]	1.53 (1.24)	0-4
Father's frequency of language mixing [c]	1.40 (1.33)	0-4
Number of persons addressing the child in Turkish [c]	6.25 (2.01)	1-10
Frequency of shared storybook reading	1.92 (0.94)	1-3
Months of systematic exposure to German	32.04 (14.89)	10-68
ECEC participation in months	17.85 (11.87)	1-49
Parental rating of child language skills [b]		
- **Turkish**	2.70 (1.13)	1-4
- **German**	2.90 (1.02)	1-4
Average language input patterns		
- **mainly Turkish**	*n* = 18 (35.3 %)	
- **approximately balanced**	*n* = 16 (31.4 %)	
- **mainly German**	*n* = 17 (33.3 %)	

Note. Systematic exposure to German was determined by exposure rates of at least 20 % per week-day. All children were exposed to Turkish from birth. Language input patterns were derived from parental questionnaires as specified in section 4.2.
[a] *n* = 49, [b] *n* = 50, [c] *n* = 48.

4.2 General Procedure and Test Instruments

All assessments were administered in two separate sessions in their ECEC institutions. Children were tested individually while sitting at a table with a female examiner. Prior to the assessment sessions, the examiners had visited the children in their ECEC institutions on one or more occasions to establish familiarity. The entire session was audiotaped. As part of the test battery, children completed a standardized German receptive and productive language assessment, a nonverbal intelligence screen (means and *SD*s for the standardized assessments are reported in Table 4), and produced a narrative sample based on a wordless picture book (Frog Story, see section 4.3). Furthermore, the children's parents completed a questionnaire about family background data, including the child's level of exposure to Turkish and German.[25] The contents, administration, and scoring for all standardized assessments and the parental report are specified in the following sections.

German Language Assessment

Children's language abilities in German were measured via the standardized test 'Linguistische Sprachstandserhebung—Deutsch als Zweitsprache' (Lise-DaZ) *[Linguistic language assessment for children with German as a second language]* (Schulz & Tracy, 2011). LiSe-DaZ was chosen for the current study, as it contains culturally and linguistically appropriate items and was normed on a DLL population (norm data exist for successive bilingual children aged 3;0 to 7;11 years and children aged 3;0 – 6;11 years growing up with German as their first language), both of which are central aspects to consider in the assessment of DLLs (e.g., Paradis et al., 2010). Using a picture-with-question-design, LiSe-DaZ assesses receptive language skills via three subscales: *Verb meaning*, *wh-questions*, and *negation*. Productive language abilities are elicited via an elicited production task using a picture sequence and assessed on further sub-scales: *Word classes* (conjunctions, prepositions, focus particles, main verbs, auxiliary and modal verbs), *case marking*, *sentence structure*, and *subject-verb agreement*.

[25] Also, ECEC practitioners filled out a questionnaire on children's language and literacy behaviors in the ECEC institutions (Sismik; Ulich & Mayr, 2003), which was not included in the current investigation.

To establish the children's level of language ability in German, all participants[26] of the current study completed LiSe-DaZ. However, for several important reasons, the current study used raw scores instead of operating with T-scores. As previously discussed, children growing up with more than one language, even if coming from the same linguistic and cultural backgrounds, constitute a very heterogeneous population, which was reflected in the study's participants. Therefore, the study sample was not entirely representative of the standardization sample, which includes only two main exposure groups (German as a first language and exposure to German after the second year of life). Furthermore, the goal was not to compare the participants to a statistically determined norm, but rather to compare them within the study population. Therefore, the current study applied raw scores in consideration of children's age, the exact number of contact months as well as language input patterns in further analyses.

To reduce item dimensionality, raw score sums were calculated for both expressive[27] as well receptive language subtests. Both individual composite scores yielded Cronbach's α values higher than 0.7 and stayed above this level when applying an 'alpha if item deleted' analysis. In accordance with Kline (1999), it was determined that an alpha value of at least 0.7 indicates good reliability and both composite scores were applied in further analyses.

Table 4. *Summary of Standardized Child Assessments*

Variable (*N* = 51)	**Mean (*SD*)**	**Range**
Expressive language German[a]	30.24 (14.52)	3-59
Receptive language German	21.86 (6.01)	8-34
Raven CPM	16.10 (3.96)	8-28

Note. Scores reported for expressive and receptive language German are sums based on LiSe-DaZ subtests; CPM, Coloured Progressive Matrices; provided data are raw scores.
[a]*n* = 50.

[26] One child did not complete the expressive subtests of the LiSe-DaZ.

[27] The subtests *sentence structure* and *subject-verb agreement* were not included in the expressive language composite score, because they yield group assignments instead of raw scores.

Nonverbal Intelligence

To assess the children's nonverbal intelligence potential, the book form of the Raven Coloured Progressive Matrices (CPM) (Raven, 1995) was administered. Because verbal instruction is kept to a minimum, the test can be considered a culturally fair measure of intellectual function and was previously used in studies with preschool-age DLLs (e.g., Scheele, Leseman, & Mayo, 2010). The CPM consists of 36 perceptual and conceptual matching exercises in which the child is required to complete a pattern by pointing to the correct picture out of six pictures. The German version includes norm data for children aged 3;9 to 11;8 years of life (Raven, Raven, & Court, 2010). The child was given a score for each correct answer and testing ended when children failed five consecutive items. The raw score sum (maximum score: 28) for each participant was further analyzed in this study.[28]

4.3 Narrative Sample Collection, Transcription, Coding, and Scoring

The following sections serve to substantiate the choice of the narrative prompt, the procedure for collecting narrative samples, as well as to provide detailed information on transcription, coding, and narrative analysis procedures.

Narrative Prompt

When it comes to selecting the type of stimulus for the assessment of fictional narratives of preschool-age children, pictures are commonly the prompt of choice. While single-pictures might elicit short and unelaborated stories and yield inconsistent output across children (Kaderavek & Sulzby, 2000; Shapiro & Hudson, 1991), the highly structured stimuli of a sequence of pictures are supportive of narrative organization (Eisenberg et al., 2008; Hedberg & Westby, 1993). Indeed, clearly sequenced illustrations with high episodic complexity will likely elicit elaborate and complex narratives from young children (e.g., Curenton & Justice, 2004; Fiestas & Peña, 2004). These

[28] Three children from the sample were 3;8 years of age, which is one month younger than the starting range of standardization. However, as the children were able to complete the test and raw scores rather than standardized ranks were applied for further analyses, the application of the assessment for all children was deemed acceptable in this study.

types of picture sequences typically can be found in picture books. Referential and communicative context information offered by a picture story is rather clear and can provide a developmentally appropriate stimulus for the generation and structuration of rather rich fictional narratives (Bamberg, 1987). Furthermore, the given (temporal) flow of events encourages the production of substantial and connected output allowing for further multifaceted analyses (Reese, Sparks, & Suggate, 2012). For these reasons, the current study utilized a wordless picture book to elicit narrative productions.

Specifically, to examine the participants' narrative competence, the children were presented with the wordless picture book "Frog, Where Are You?" (Mayer, 1969), which has also been commonly referred to as the 'Frog Story.' The book depicts the story of a boy and his dog whose pet frog escapes at night. On their search for the frog, the boy and the dog enter a forest where they encounter different animals that in some way interfere with the search. Eventually, they find the frog surrounded by his family and walk away with a baby frog as their new pet. Besides including the global search theme and a series of temporally sequenced and causally linked events, the plot line offers plenty of opportunities to make inferences about the characters' relationships, thoughts, feelings, and motivations. Therefore, while being cognitively challenging, the prompt is suitable for child narrators. For this reason, the Frog Story has been applied extensively as a narrative stimulus across typically and atypically developing monolingual and DLL populations[29] (e.g., Colle, Baron-Cohen, Wheelwright, & van der Lely, 2008; Curenton & Justice, 2004; Greenhalgh & Strong, 2001; Justice et al., 2010; Mills, 2015; Reilly, Losh, Bellugi, & Wulfeck, 2004; Montanari, 2004; Peets & Bialystok, 2015; Tager-Flusberg & Sullivan, 1995) as well as in cross-linguistic work (e.g., Berman & Slobin, 1994; Fiestas & Peña, 2004; Montanari, 2004; Verhoeven & Strömqvist, 2001), including preschool-age children acquiring Turkish (Aksu-Koç, 1994) and German (Bamberg, 1987; 1994) and Turkish children in Germany (Pfaff, 2001). Importantly, previous work on child narrative skills using the Frog Story yielded high productivity rates in young DLLs as needed for productivity and complexity

[29] In fact, De Fina and Georgakopoulou argue that the Frog Story is the best known prompt for elicitation of narratives, used in „at least 150 studies in fifty languages" (2011, p. 13).

measures (e.g., Bedore, Fiestas, Peña, & Nagy, 2006; Lofranco, Peña, & Bedore, 2006).

A critical issue in any narrative investigation is that the examiner has "obtained a valid representation of the subject's generative processes in narrative production" (Liles, 1993, p. 877). In general, while eliciting the narrative probe without a model ensured that the collection of child stories occurred without any influence or imposition of a certain style, telling stories from wordless picture books can pose specific challenges to children, especially those who may not be familiar with the demands of such a task. However, as preschool-age children attending ECEC institutions are familiar with shared picture book reading and storytelling (e.g., van Kleeck, Stahl, & Bauer, 2008; Wasik & Bond, 2001) and all participants had attended ECEC for at least 10 months, the task was deemed appropriate for participating children.

Procedure

All narratives were collected in a quiet room of the children's ECEC institution, and they were seen individually by an examiner they were familiar with through previous warm-up and assessment sessions. The picture book was new to all participants, and they were not told about the story beforehand. Fictional narratives were elicited following the protocol developed by Berman and Slobin (1994). Children were given time to first view the whole book in silence to get a sense of the plot, before telling the story in their own words based on the illustrations, going page by page. In eliciting the spoken narrative, the examiner instructed the child, „Ich habe dir ein Buch mitgebracht. Es erzählt die Geschichte von einem Jungen, einem Hund und einem Frosch. Als Erstes möchte ich, dass du dir alle Bilder anschaust. Schau dir jedes Bild genau an. Danach sollst du mir die Geschichte erzählen." *[I brought you a book. It tells the story of a boy, a dog, and a frog. First, I would like you to look at all the pictures. When you are finished looking at all the pictures, I would like you to tell me the story.]* When the child indicated that she or he was ready to tell the story, the book was flipped back to page one. At this point, the examiner remained silent except to demonstrate interest using a selected array of minimal prompts and backchannel responses such as nodding,

"yes," "mhm," "anything else?", and "continue." No time limit was given to the narration. When the child arrived at the end of the book, the examiner asked if he or she wanted to add anything, or if they were finished telling the story. When children indicated that they were finished, the recording was stopped. All samples were audiotaped using a digital voice recorder (Olympus DM-650) for later transcription.

Transcription, Coding and Narrative Analysis Procedures

All digital sound files were transferred to a computer and were transcribed while using headphones. While the transcription of oral narratives is not standardized (Pavlenko, 2008), in keeping with common practice in child language research (e.g., Peets & Bialystok, 2015), the entirety of each narrative was transcribed using the Codes for the Human Analysis of Transcripts (CHAT) system developed as part of the Children's Data Exchange System (CHILDES) (MacWhinney, 2000). Mainly following Justice et al.'s (2010) transcription rules, the transcription process started with the examiner's prompt and ended after the child had indicated she or he was finished telling the story. While incomplete and uninterpretable verbal utterances were also transcribed following the conventional use of the CHAT symbols, only complete and intelligible child utterances were included in later analysis. Discourse by the examiner and all child utterances unrelated to telling the story (e.g., questions about other books and comments about the room) were transcribed, but excluded from the analysis reported here, similar to child repetitions of examiner recasts. If a child self-corrected, the corrected form was scored. Also, as preschoolers do not yet reliably produce conventional features of stories such as formal endings, e.g. 'the end' (Cain, 2003), they were not included in further analysis.

In accordance with Alamillo and colleagues, sentences, which might be a suitable descriptive unit for written texts, or "utterances," which are frequently used for transcribing very young children's speech, were considered too imprecise a definition to be able to undertake corpus annotation and quantitative analyses (Alamillo, Colletta, & Guidetti, 2013). Also, when assessing syntactic complexity in utterances longer than three words, the traditional measure of mean length of utterance (MLU) does not deliver an accurate estimate of syntactic skills in children (Scarborough, Rescorla, Tager-

Flusberg, Fowler, & Sudhalter, 1991). Therefore, utterances were segmented into communication units (C-units; Loban, 1976), a conventional procedure designed to organize and analyze children's narrative productions in meaningful and grammatical utterances (Hughes et al., 1997; Retherford, 2000). Based on these authors, C-units were defined as syntactic units consisting of one main clause and any dependent constituents, including subordinated clauses and phrases, to achieve a better estimate of children's syntactic skills. Accordingly, dependent clauses were transcribed in one C-unit, while series of successive main clauses as well as clauses connected by a coordinating conjunction were segmented in different C-units. Because single-word utterances and/or utterances lacking clausal structure are quite common in the narratives of younger children and those with limited previous second language exposure (e.g., Bedore et al., 2006, Strömqvist & Verhoeven, 2004), they were included in the analysis. A narrative had to consist of at least two C-units, following Labov's (1972a) definition of a minimal narrative.

Furthermore, in accordance with Gagarina et al. (2012), all filled pauses, repetitions, reformulations, and disfluencies were considered mazes. They were transcribed accordingly, but excluded from further analysis (except for the measures on percentage of maze use). This resulted in the elimination of 8.16% word tokens (*SD* = 6.48) from the language samples.

By reducing inflectional forms and derivationally related forms of a word to their word roots, a process referred to as lemmatization, it was ensured that measures of word use were not inflated by the presence of multiple forms of single words. Accordingly, verb forms were linked to their word roots. For example, kommt [*comes*] and kam [*came*] were both linked to kommen [*to come*]. This process was deemed especially important working with language samples of young DLLs who regularly produce "creative but wrongly inflected verb forms or plural forms" (Bedore, Peña, Gillam, & Ho, 2010, p. 504), which could lead to inflated lexical diversity measures. To adequately account for compound words, which commonly occur in the German language, credit was given for the two stem words. For example, Babyfrosch *[baby frog]* was linked to Baby *[baby]* and Frosch *[frog]*.

4.4 Analytical Framework for Narrative Measures

When investigating a child's oral fictional narrative performance, consideration needs to be given to the type of measures that are included. To be able to objectively compare the participants' narrative productions with respect to one another, as well as to derive measures of narrative skill for the examination of relationships between narrative and other indices of child development and family environment, a wide-scoped and integrative narrative scoring system was developed on the basis of current approaches to micro- and macrostructural narrative analysis, as presented in the following sections.

4.4.1 Microstructural Measures of Narrative Performance

As presented in Table 5, for the current profile of oral narrative ability, five transparent, frequently used measures of narrative microstructure known to be sensitive to language ability in young DLLs (e.g., Hipfner-Boucher, 2011; Uccelli & Páez, 2007) were selected from established guidelines on child microstructure analysis (Gagarina et al., 2012, 2015; Justice et al., 2006, 2010). Measures were derived from children's stories based on all complete and intelligible utterances to targeted general productivity, lexical diversity, as well as syntactic complexity and features.

Table 5. *Applied Measures of Narrative Microstructure*

Abbreviation	Narrative Measure	Indicator of
TNW	total number of word tokens without mazes	general productivity
TNCU	total number of utterances (in C-units)	narrative length / verbal productivity
NDW	number of different words (in lemmas)	lexical diversity based on lemmas
VOCD	vocabulary diversity	lexical diversity accounting for sample length
MLCU	mean length of C-units in words	syntactic complexity / grammatical ability

Productivity

Verbal narrative productivity was calculated on the token as well as on the C-unit level. TNW was a sum score of all produced tokens excluding mazes (as specified in section 3.3), while TNCU was a count of all C-units. Both measures were computed using the *freq* command of the Child Language Analysis software (CLAN; MacWhinney, 2000).

Lexical Diversity

Measures of lexical diversity, that is, indicators of how many different words are used in a language sample, are a key feature of the language structure of children's narratives and can be seen as a measure of expressive vocabulary size (Curenton & Lucas, 2007). Two different measures were computed representing lexical diversity: Number of different words (NDW), and the D statistic. NDW[30] is a traditional approach to measuring the range of vocabulary in a language sample; it was calculated by summing up all lemmas produced for one narrative. When comparing samples of different lengths, however, an obvious limitation of this approach is that it does not account for productivity, despite the relation of number of word types and tokens, i.e., the longer the sample, the more tokens it likely contains (Malvern, Richards, Chipere, & Durán, 2004). A simple solution to this problem is to calculate the ratio between the types and the tokens, for example by calculating the historically widely used TTR (division of the number of different word types (here: lemma types) by all the words (here: lemmas) produced). Again, however, this approach bears the inherent flaw of disregarding the overall sample length (e.g., Pavlenko, 2008). This is problematic when comparing multiple samples, as the introduction of new types is substantially affected by sample length and gradually decreases over the sample length. Therefore, the D statistic (henceforth termed VOCD, calculated via the *vocd* command in the CLAN program) was used to compute an additional measure of lexical diversity. Other than the traditional measure of TTR, this newer approach corrects for typical variation in type-token

[30] To avoid inflation of rates, the current study measured NDW in lemmas. The lemmatization process is presented in section 4.3.

ratio over a range of text lengths and is proposed to more robustly measure children's lexical diversity (Malvern et al., 2004). However, as the VOCD computation relies on a certain sample length (Koizumi & In'nami, 2012) and narratives of participating children were likely to greatly vary in terms of productivity, the traditional measure of lexical diversity, NDW (in lemmas), was also computed.

Syntactic Complexity

Mean length of C-units in words was chosen as a well-established measure of syntactic complexity and overall grammatical ability. Because of the previous segmentation of utterances into C-units (i.e., syntactic units consisting of one main clause and any dependent constituents, including subordinated clauses and phrases), the mean length of C-units across a narrative production serves as a good indicator of a child's spontaneous syntactical construction skills in a narrative context. This measure was computed using the *mlu* command of the CLAN program.

4.4.2 Composite Measures of Narrative Complexity

A variety of analyses have been proposed for examining mono and dual language learning children's expression of narrative macrostructure, focusing on story grammar/episodic complexity and organization (Fiestas & Peña, 2004; Liles, Duffy, Merritt, & Purcell, 1995; Petersen, Gillam, & Gillam, 2008; Peterson & McCabe, 1983; Stein & Glenn, 1979), expressive elaboration (Ukrainetz et al., 2005; Ukrainetz & Gillam, 2009), and high-point analysis (McCabe, Bliss, Barra, & Bennett, 2008).

Analytic approaches to child narrative in the story grammar tradition have been criticized for putting a too limited focus on specific episodes and not enough emphasis on higher-level narrative skills (Heilmann, Miller, Nockerts, & Dunaway, 2010). The expressive elaboration dimension (e.g., as expressed by evaluative language use) is especially valuable, though, because when telling a complete story, it is not only important to convey the mere facts on what happened, but also the meaning behind the narrated events. Therefore, to capture higher level narrative skills, inspired by Hipfner-Boucher (2011), the current study employed a scoring rubric based on three different parts: an adapted version of the Index of Narrative Complexity (adapted from Petersen, Gillam,

& Gillam, 2008), a binary decision tree for scoring the overall level of narrative elaboration and complexity (based on Westby, 2005), and a categorical scheme for evaluative language use. The final scoring rubric was termed Extended Index of Narrative Complexity (EINC) (see Appendix B for the complete instrument and Appendix C for a scoring example).

Adaptation of the Index of Narrative Complexity

The Index of Narrative Complexity (INC; Petersen et al., 2008) was developed as a criterion-referenced assessment protocol for the clinical evaluation and investigation of school-aged children's oral fictional narrative productions. Foundational to the INC are the traditional high point analysis of Labov (1972a), the well-known story grammar analysis put forth by Stein and Glenn (1979, 1982), and refinements of Peterson and McCabe (1983). The instrument uses a rubric to assign scores on a scale from 0 to 2 or 0 to 3 to a range of categories related to episodic complexity and narrative cohesion in oral narratives. The derived total score reflects the overall complexity of a narrative referring to central features: *characters, setting, initiating events, internal responses, plans, action/attempts, complications, consequences, narrator evaluations, formulaic markers, temporal markers, dialogue*, and *causal adverbial clauses*. Preliminary analyses of reliability and validity conducted by the creators of the INC yielded high interscorer agreement (90% to 96%), good test–retest correlations with 1 month between testing (.60 to .90), and strong concurrent criterion evidence for validity (.60 to .83) with the standardized assessment Test of Narrative Language (Gillam & Pearson, 2004).

To ensure reliable scoring of preschool-age DLLs' narratives, several modifications and clarifications had to be made to the instrument. Most significantly, the 'narrator evaluations' as well as the 'internal response' categories were eliminated, as they could not be found in the preschool-age DLLs' narrative productions. For the same reason, the highest point category (3) for initiating event, consequence, and knowledge of dialogue was not applied. Instead of formulaic markers, additive markers/conjunctions were included in the analysis, as they are much more common in this

age group's narrations (e.g., Bedore et al., 2010). As character introduction and information on setting were rarely elaborated, only 1 point was granted per category. Furthermore, because none of the children produced more than one causal marker, a maximum of 1 point was granted in this category. Other minor modifications included adding own examples to ease the scoring procedure. Also, following Spencer and Slocum (2010) it was specified that the problem and action/attempt had to be in reference to the main character and not a secondary character. As for temporal markers, only 1 point was assigned for repeated production of dann *[then]*, as it was used excessively by some children and otherwise could have inflated the measures. Redundant mentions of story grammar elements, such as the repeated notion of the consequences that the boy and the dog had found the frog, or recasts of previously mentioned story grammar elements using different words, were net coded twice; only the first instance was coded.

In sum, the instrument yielded an aggregated score of a child's fictional narrative performance on a macrostructure level including aspects of narrative microstructure (i.e., narrative cohesion).

Story Structure Level

For further analysis of narrative macrostructure, a story structure decision tree—a graphic tool for guiding the narrative analysis of children's stories—was chosen: In accordance with recommendations for the assessment of narratives in mono- and dual language learning children (e.g., Gagarina et al., 2012, 2015; Paul, 2007), Westby's (2005) Story Grammar Decision Tree based on Stein and Glenn's (1979) classic description of story grammar (also see Hughes et al., 1997, p. 120) was used for holistically assessing the overall maturity of narrative organization (from descriptive sequence to complete episode) from a goal-directed viewpoint. The decision tree consists of a flow chart containing a series of yes or no questions. Each "yes" answer moves the user to the next question/level, while a "no" response prompts the user to exit the flow chart whereby the narrative sequence level is indicated (see Figure 5). Optionally, scores can be assigned for each level reached.

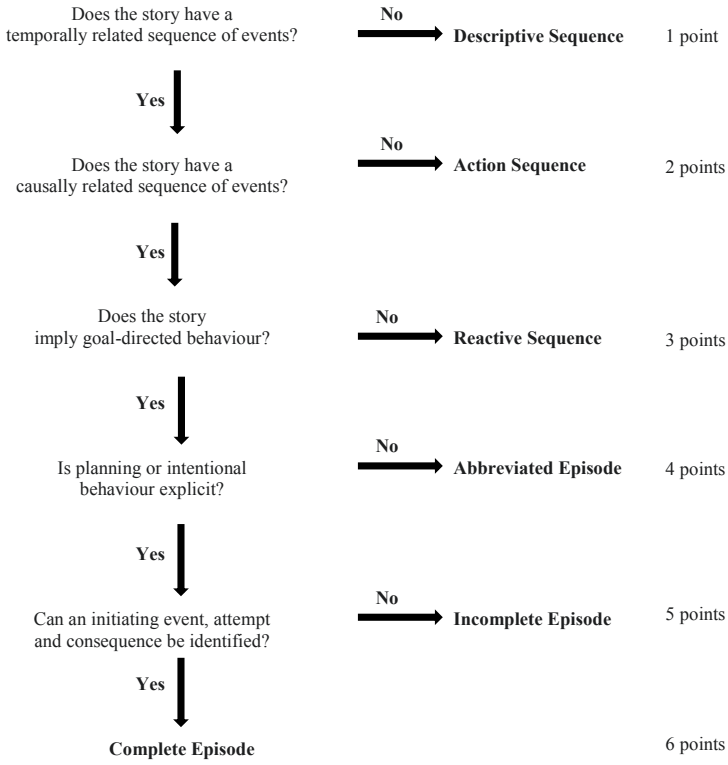

Figure 5. Binary Story Structure Decision Tree. As displayed in Paul (2007, p. 497), also see Westby (2005).

Evaluative Language Features

Furthermore, a categorical matrix for evaluative and literate language use was adopted from Hipfner-Boucher (2011, p. 64), which tallied the absence or presence of the following categories: modifiers (adjectives/adverbs), expression of intent, metacognitive verbs, emotional state terms, physical state terms, and knowledge of dialogue (see Table 6).

Table 6. *Coding and Scoring Procedures for Evaluative Language (Categorical)*

Variable	Description
Dialogue	A score of 1 indicated the presence of character dialogue (both direct and free, see examples). A score of 0 indicated the absence of dialogue. Indirect reports of speech (e.g., *he called for the frog*) were not coded. Examples: - „Jetzt endlich hab ich dich gefunden, Frosch." *["Now I finally found you frog."]* - Und dann rufen er mit dem Hund: „Wo bist du, Frosch?" *[And then he call with the dog, "where are you frog?"]*
Modifiers	A score of 1 indicated the presence of at least one modifier (adjective or adverb). A score of 0 indicated the absence of a modifier. Examples: - Hat er aber ein(en) **richtig schlechten** Tag. *[He had a **really bad** day.]* - Und da riecht ekelig. *[And there smells **disgusting**.]* - Und der Frosch war **immer noch** nicht da. *[And the frog **still** was not there.]*
Expressions of intent	A score of 1 indicated the presence of at least one expression of intent. A score of 0 indicated the absence of an expression of intent. Examples: - Fund die Frosch **muss**. *[Find the frog **must**.]* - Und sie **will** auch in die Baum. *[And she also **wants** to go in the tree.]* - Er **versucht** das zu holen. *[He **tries** to get it.]*
Metacognitive verbs	A score of 1 indicated the presence of at least one metacognitive verb. A score of 0 indicated the absence of a metacognitive verb. Examples: - Er **dachte**, der Hund hat ihn freigelassen. *[He **thought** the dog set him free.]* - Den Frosch **weiß** nicht wo der Wauwau. *[The frog does not **know** where the doggy.]*
Emotional state terms	A score of 1 indicated the presence of at least one emotional state term. A score of 0 indicated the absence an emotional state term. Examples: - Dann wird er **böse**, weil da Hund da ist. *[Then he gets **angry**, because there dog is there.]* - Kriegt er **Angst**. *[He gets **scared**.]* - Dann war der **froh**, weil er ein Babyfrosch bekommen hat. *[Then he was **happy**, because he got a baby frog.]*
Physical state terms	A score of 1 indicated the presence of at least one physical state term. A score of 0 indicated the absence of a physical state term. Examples: - Das **tut** ihm **weh**. *[That **hurts** him.]* - Der war **müde**. *[He was **tired**.]*

Note. Categories and scoring system from Hipfner-Boucher, 2011, p. 64. Own examples were added from narratives produced in this study. The INC category *dialogue* was included here for a more comprehensive picture of evaluative language use.

Combined Instrument – Extended Index of Narrative Complexity (EINC)

The individual parts were each scored[31] manually on a scoring sheet (see Appendix C). A composite score was calculated for all three parts described above, yielding a maximum of 26 points.

Speech Production Process

The speech production process/verbal fluency in narrative production was targeted via maze use (i.e., disfluencies such as false starts, filled pauses, repetitions, and revisions). More specifically, by dividing the number of maze tokens over the number of word tokens without mazes, the proportion of maze tokens was obtained.

4.4.3 Reliability for Transcription and Narrative Measures

A consensus procedure was used for transcription and segmentation in C-units. Following initial transcription by a trained research assistant, a second research assistant examined[32] the transcript in its entirety for errors in the area of spelling, to ensure accurate word counts, and in the area of utterance segmentation, to ensure accurate TNCU and MLCU calculations. Finally, language transcripts (100%) were reviewed by the author. Three remaining cases of disagreement with respect to C-unit segmentation and two cases with respect to maze use were resolved by listening to the audio recording and by discussion.

Lemmatization, i.e., reducing inflectional forms and derivationally related forms of a word to their word roots, was performed by transferring the CLAN list of words computed by the *freq* command into an Excel-worksheet and manually sorting the tokens into the following lexical categories: nouns, verbs, adjectives, adverbs, articles (definite and indefinite), pronouns, prepositions, conjunctions, and numerals. Twenty per-

[31] Bearing in mind concerns expressed by Muñoz and colleagues (2003), O'Neill, Pearce, and Pick (2004), and in accordance with Hipfner-Boucher (2011) a child-based approach rather than a text-based approach (e.g., Berman & Slobin, 1994) was adopted in scoring macrostructural aspects. As such, children were given credit for the inclusion of story grammar elements that were particular to the story they chose to tell; in this way, children were not evaluated on the basis of their ability to match the story intended by the examiner, but on their ability to generate a well-structured story.

[32] If in doubt, the research assistant listened to the audio recording while simultaneously checking the transcript.

cent of the stories were then randomly selected by a second research assistant for reliability purposes. Interrater reliability was very good overall, as measures for each word category (tokens) and NDW measure based on lemmas exceeded 90%.

To determine EINC rating consistency, a research assistant who was blinded from participant information was trained on the coding system and independently re-coded 41% of the narrative samples according to the procedures outlined above. Cohen's κ revealed high interrater agreement, $\kappa = .84$.

Finally, maze use was calculated by removing all transcription conventions for mazes from the transcripts and rerunning the CHAT *freq* count. This procedure was repeated by a research assistant for a random sample of 20% of the transcripts. To obtain an interrater agreement score, the total number of agreements was divided by the total number of item comparisons and multiplied by 100. The mean reliability score was 98.5% (ranging from 90.9% to 100%). Any disagreements were resolved through discussion before data analysis. The reliability was not counted for TNW and MLCU because the CLAN software automatically calculated these values.

4.5 Analytic Strategy

In response to the study's main research aim—to examine the fictional narrative skills of preschool-age DLLs—three research questions were derived. Analyses specific to each research question are detailed below.

After preliminary analyses for sex differences, as an initial step, descriptive analyses were run on all study measures to determine mean performance on the various narrative microstructure and macrostructure measures for the entire sample of children ($N = 51$). Then, to determine associations between narrative indices and other measures, correlations were run and analyzed for significance, directionality, and strength. Spearman rank correlation tests were computed, because this procedure does not require assumption of normality and is less sensitive to bias due to the effect of outliers, which were likely to occur in the current sample. Two-tailed correlations were run because they are more conservative than one-tailed tests, thus accounting for the limited sample size in the present study. Cohen's (1988) standard was followed to evaluate the correlation coefficient to determine the magnitude of the effect size, or the

strength of the relationship. Coefficients between .10 and .29 represented a small association, coefficients between .30 and .49 represented a medium association, and coefficients above .50 represented a large association (Cohen, 1988, pp. 77-81). Furthermore, following Rosenthal's (1996) suggestions, an effect size equivalent to or greater than r_s = .70 was considered very large (also see Ellis, 2010; Grissom & Kim, 2005). Finally, to identify factors contributing to complexity in narrative generations, univariate and multiple regression analyses were computed through the generalized linear model options in SPSS. For all analyses, potential impacts on study findings and interpretations are discussed in the results section (see section 6.6 specifically for statistical considerations).