Corresponding author:
Madlen Mangold;
Email: madlen.mangold@uni-wuerzburg.de

CAMBRIDGE
UNIVERSITY PRESS

Assessing vocabulary of bilingual German-Turkish preschool children

Madlen Mangold [iD], Wolfgang Lenhard [iD], Julia Schindler [iD], Daniel Schulz [iD] and Tobias Richter [iD]

Department of Psychology IV, University of Würzburg, Wurzburg, Germany

Abstract

The vocabulary of bilingual children is determined by various linguistic factors that develop depending on the language input and individual factors of these children. To understand vocabulary development and to be able to support these children accordingly, the assessment instruments essentially need to be adapted to this individual process. The current study examined factors influencing productive vocabulary performance by means of explanatory item-response models on a sample of 126 German monolingual and Turkish—German bilingual preschool children aged between 3 and 6 years. We analyzed item features with respect to lexical category, length, frequency, age of acquisition, and complexity in dependence of the language background of the children. Apart from a general delay in the development and some language specific characteristics, the results indicate a parallel shifted acquisition of vocabulary in bilingual children, which is nonetheless shaped by the same factors as in monolinguals.

Introduction

In our increasingly globalized world, it is becoming more common for children to grow up with two or more languages. Extensive research has established that children's linguistic abilities play a crucial role in their academic achievements (e.g., Cunningham & Stanovich, 1997; Walker et al., 1994; Whitehurst & Lonigan, 1998). Therefore, the lexical development, which refers to the acquisition and expansion of vocabulary, holds significant importance in overall language development for both monolingual and bilingual individuals. Understanding the factors influencing lexical development in bilingual children is essential for supporting their language skills and academic success. Vocabulary knowledge is typically distinguished into two categories: productive and receptive knowledge (Nation & Meara, 2002). The focus is on comprehending the components of vocabulary, organizing lexical resources, and attaining proficiency and fluency in both receptive and productive language usage. Productive knowledge specifically pertains to the capacity to autonomously produce words during speaking or writing (Schmitt & Schmitt, 2020). The process extends beyond mere word familiarity, encompassing a comprehensive grasp of the multifaceted aspects of productive vocabulary acquisition and its associated constructs (Nation, 2013).

Evidence has shown that monolinguals and bilinguals use one common system of conceptual lemma representations (Costa et al., 1999; Kroll & Stewart, 1994; Lee & Williams, 2001), although the lexemes of the two languages seem to be stored in separate lexical storage systems (Kroll & Stewart, 1994; Potter, 1979; Snodgrass, 1984). This divergence in conceptual representations leads to differences between monolinguals and bilinguals in the lexical processing – as measured, for instance, by vocabulary tasks. One such assessment methods are naming tasks, which are commonly used as a diagnostic tool for evaluating vocabulary, especially in bilingual children. These tasks involve participants naming photographs or line drawings. In contrast to the assessment of the receptive vocabulary, this active process of production is more demanding and requires subjects to demonstrate more aspects of knowledge. While the word comprehension process is primarily characterized by the recognition of the phonological representation and activation of the related semantic concept stored in long term memory, the production process involves not only the selection of the appropriate lemma but also the retrieval and maintenance of the phonological representation throughout the planning and execution of the phonetic plan (Dapretto & Bjork, 2000). Consequently, we get more information about the extent to which the children can actively use the words. For that reason, assessing productive vocabulary through picture naming tasks is a widespread and effective method to evaluate the vocabulary of bilingual children (Webb, 2008).

Even high-proficient bilinguals usually perform more poorly in these tasks than monolinguals and show not only less accuracy but also a time delay in picture naming in both the non-

This article has earned badges for transparent research practices: Open Data. For details see the Data Availability Statement.

dominant and the dominant language (e.g., Gollan et al., 2005, 2008; Ivanova & Costa, 2008). Furthermore, bilinguals appear to have more difficulties in activating language specific representations (Gollan & Acenas, 2004), and they show a naming disadvantage in comparison to monolingual children, which increases with age (Klassert et al., 2014). Because of these differences between monolingual and bilingual children, the use of common naming tasks might be problematic, especially when the instruments to assess the vocabulary have not been specifically designed for and normed on bilingual children.

The aim of our study was to illustrate the construction of a language assessment inventory that balances well established factors from language acquisition research in order to generate parallel and fair versions for L1 and L2. Here we will focus on the effects of these factors and analyze how they interact with the linguistic background of children in the societal language, by comparing the structure and item features of the naming task for monolingual German and bilingual Turkish—German children. We demonstrate how to validate a test by using explanatory item-response models (De Boeck & Wilson, 2004) and to examine the construct validity, if the test was to be used not only for monolingual but also for bilingual children based on a productive vocabulary task. In the following, we first discuss features that are known to influence the access to words, especially for bilingual children – namely, lexical category, frequency, age of acquisition, and complexity. We address the limitations of existing diagnostic approaches in the assessment of vocabulary. We then turn to present the principles of constructions used in our productive vocabulary task.

Influences on lexical access and phonological encoding

Transforming an idea into spoken words requires first to activate a lexical concept and to retrieve a lemma from the mental lexicon (Levelt et al., 1999). Levelt (1989) defined a lemma as a lexical item that contains semantic and syntactic specification. In the end of this conceptual processing, an activation of semantically related concepts is linked to specific lexemes. The language production process and its conceptual processing is generally agreed to consist of the two major components of lexical access and phonological encoding (e.g., Dell, 1986; Levelt, 1992; Levelt et al., 1999). Lexical access involves the search and selection of appropriate lemma within the mental lexicon (Levelt et al., 1999), and phonological encoding provides acoustic and articulatory properties of the words (Wheeldon & Levelt, 1995). To say a word like *frog*, the connected lexical nodes and its syntactic features must be activated in the semantic system and the phonological information must be encoded (Caramazza, 1997; Levelt, 1989). When acquiring vocabulary, the corresponding information must be integrated into the language system, making it susceptible to various influencing factors. The following section will discuss significant factors that influence the word retrieval process, with a focus on those that have been reported as significant and for which there are available findings. These factors that influence the word retrieval process operate at various levels, as illustrated in Figure 1a, and can be broadly considered as language specific or non-language specific. At the lexeme level, for instance, language-specific information is involved, such as word length and complexity (Heuven & Dijkstra, 2010).

Age of Acquisition (AoA)

Age of Acquisition (AoA) in the context of language learning refers to the age at which a concept or a lexical item is acquired

(Hernandez & Li, 2007). Early-learned words are processed differently from late-learned words (Gilhooly & Watson, 1981), but the reasons for this difference are not yet fully understood. One hypothesis, known as the semantic locus hypothesis, proposes that early learned words have a semantic advantage over those learned later because they enter the representational network first and influence the semantic representations of subsequently learned words (Brysbaert et al., 2000). Specifically, lexical nodes that are acquired earlier have more connections within the lexical system than later established nodes, establishing a basic semantic structure that allows later word learning to accelerate (Steyvers & Tenenbaum, 2005). As a result, early-learned words lead to a more frequent activation, better accuracy and faster response latencies in picture-naming tasks (e.g., Barry et al., 1997; Carol & White, 1973; Cuetos et al., 1999; Meschyan & Hernandez, 2002; Snodgrass & Yuditsky, 1996). Accordingly, we consider AoA to be a significant predictor of naming tasks.

Frequency

In language learning, words that have been known for a longer time have been encountered more frequently, and higher word frequency is associated with greater exposure. A hypothesis known as the cumulative frequency hypothesis suggests that the total number of times a word has been encountered can account for both AoA and frequency effects (Lewis et al., 2001). However, there are also suggestions to treat the frequency effect as an independent factor that influences lexical access (Barry et al., 1997; Carol & White, 1973; Snodgrass & Yuditsky, 1996). This robust effect, first described by Oldfield and Wingfield (1965), refers to the fact that the production of less frequent words is substantially slower than the production of frequent words. Jescheniak and Levelt (1994) described a modular two-stage model with frequency-coded lexemes' activation. According to this model, an initial phase with the activation and the selection of lemmas and their semantic-syntactic information is followed in a second stage by word form retrieval. Unlike lemmas, lexemes appear to be affected by word frequency. Familiarity with the words seem to make the access of the word forms easier. The activation threshold for word forms is low for words with a high frequency and high for words with a low frequency (Jescheniak & Levelt, 1994).

Lexical category

Another aspect of language production processes is the lexical category. For example, nouns are acquired before verbs in most languages because of their referential bindings (Gentner, 1982). Children form stable object concepts during their first year of life, and the reference to objects allows a transparent semantic mapping to the perceptual world (Spelke, 1990). Empirical studies showed a processing advantage of concrete words in contrast to abstract words (De Groot et al., 1994; Schwanenflugel et al., 1988) because of their perceptual referents (Bolognesi & Steen, 2018). First acquired nouns are concrete concepts and thus children may retrieve them easier than verbs (Chiarello et al., 1999). The conceptual components of verbs seem to be more difficult to detect, as well as to combine and to organize them. Thus, children learn to access them later compared to nouns (Gentner, 2006). Accordingly, toddlers' vocabulary consists mainly of nouns and few verbs (Gentner, 1982; Nelson, 1973). This propensity to learn nouns compared to verbs is supported by findings that show a strong effect of lexical category. Haman et al. (2017) analyzed data from monolingual preschool children across 17 languages in lexical tasks and found accuracy for nouns to be higher than for verbs.
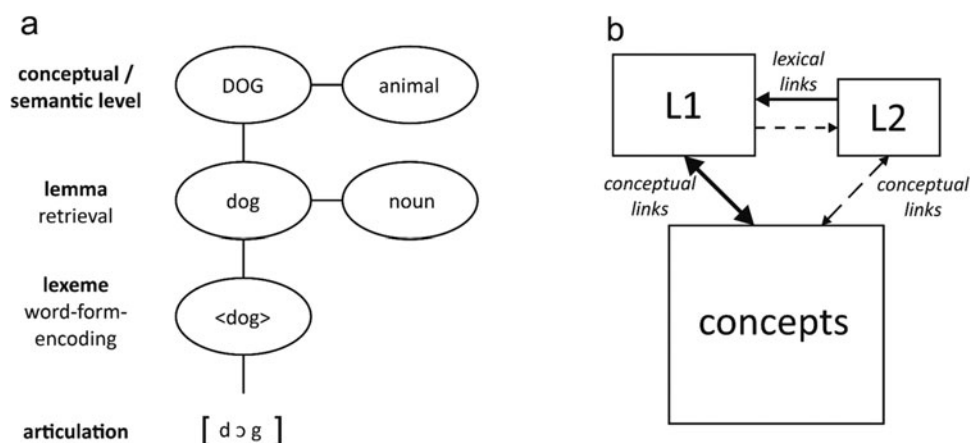
Figure 1. Model of lexem and lemma representation (a) and the relation of lexems and concepts in L1 and L2 (b), based on the revised hierarchical model of lexical and conceptual representation in bilingual memory (Kroll & Stewart, 1994)

Complexity of words

At some point in the language acquisition process, children start to use compounds to differentiate their semantic knowledge, to build finer-grained categories, and to thus increase their vocabulary (Clark et al., 1985). Since compounds in the German language are common, they are acquired very early, from an age of about 2 years (Clark, 1998). The production of compound words, lexemes that consist of more than one stem, requires linguistic and conceptual knowledge about the constituents (Gagné & Spalding, 2009). To date, no general consensus exists about how compounds are represented in the mental lexicon (Li et al., 2017). Depending on how they are represented, however, it should affect the retrieval. Stored as a whole word should result in frequency effects (Janssen et al., 2008) and the retrieval should not differ from the simple words. However, representation of the constituents has found strong support (Marelli et al., 2014) and understanding this representation requires knowing about the processing and how the constituents are organized and combined. Compounds are an important component of vocabulary, and they contribute to understanding its organization and development.

Word length

The effects mentioned above refer primarily to the lexical access, whereas word length belongs more to the phonological encoding (Jescheniak & Levelt, 1994). Speakers retrieve lexeme information from a syllable based articulatory program. The phonological encoder first selects segments and then combines them into the word form (Levelt et al., 1999). Accordingly, effects of word length should be found not only on response latency but also on accuracy in speech production (Vance et al., 2005). Speech onset latencies for long words are longer than for short words and this happens because speakers first must phonologically encode the entire word and activate all articulatory programs before speaking (Meyer et al., 2003). All these findings on lexical access and phonological encoding are considered independent of the language, yet important differences across languages must still be addressed when constructing testing material and should thus also be included in the analysis.

The bilingual speech production

Strong evidence has shown that bilingual children comparably organize semantic content across their dual lexicons (Holowka et al., 2002). Based on a hierarchically related representation of the children's memory proposed by Kroll and Stewart (1994), Potter (1979) and Snodgrass (1984), we assume bilingual children to have a common semantic concept for both languages (Figure 1b). However, concrete material has a processing advantage and the processing might depend on lexical category, which means that concrete translation pairs tend to share a conceptual representation more often than abstract translation pairs (Van Hell & De Groot, 1998a). Possible reasons for this difference might be that concrete words have an imaginal referent (e.g., De Groot, 1989; Plaut & Shallice, 1993; Van Hell & De Groot, 1998a), whereas abstract words depend more on a linguistic context and are thus more language-specific (Van Hell & De Groot, 1998b). At the lexical level, researchers widely agree that units correspond to words and mediation occurs between the conceptual-semantic level and the level of individual phonemes (Rapp & Goldrick, 2000). In contrast to monolinguals, bilinguals are assumed to have a separate lexical store for each of the two languages and these lexicons contain language-specific information (Dijkstra & Van Heuven, 2002). The content of these stores depends largely on the input the children receive, and the input differences result in different acquisition rates for each of the languages (Rinker et al., 2017). Accordingly, we assume AoA to be affected by the input the children receive, with strong influences on the lexicon from the environment – for example, from grandparents, friends, or caregivers. This assumption is supported by findings about a strong Turkish dominance in children raised in Germany who acquire both Turkish and German (Rinker et al., 2017). The authors examined very young children at the age of 2 and 3 years and showed a dominance of lexical items in Turkish. They attributed this dominance to a more frequent exposure and a higher quality of the input (Rinker et al., 2017). Cultural aspects are assumed to play an important role on the variation of lexical content as well (Caselli et al., 1999) – for example, specific foods and drinks that are less common in the respective languages (Rinker et al., 2017).

Assessing vocabulary in bilingual children

Studies comparing the vocabulary of bilingual children across languages are relatively scarce (Bornstein et al., 2004). Moreover, only a small number of studies have investigated the development of vocabulary in bilingual children from the same perspective

(Cote & Bornstein, 2014). Often, assessments are merely translated without being adapted, which leads to inadequacies such as linguistically uncontrolled test items and culturally inappropriate targets, pictures or procedures (Schaefer et al., 2019). Given the lack of appropriate testing material, practitioners and researchers often need to rely on instruments that were developed for a specific population – usually monolinguals or unselected population representative samples – to assess lexical abilities in children with bilingual background. In some cases, rating scales completed by educators are used to assess vocabulary. One of these instruments is the MacArthur-Bates Communicative Development Inventories (MB-CDI; Fenson et al., 2007) that examine productive vocabulary of young children by using parental vocabulary checklists. The original English version has been adapted for several other languages. According to the authors, the instrument provides a reliable and valid assessment of productive vocabulary in theses languages (Fenson et al., 2007). For the Turkish language, Acarlar et al. (2009) adapted and standardized the MB-CDI on monolingual Turkish children (Türkçe İletişim Gelişimi Envanterleri; TIGE) and for a subgroup of bilingual Turkish (— German) children raised in Germany, the Türkce Ifade ve Lisan Dizelges Araştırması (TILDA; Sachse et al., 2016) was developed. Parental checklists and reports are an economic assessment tool that can certainly contribute additional information about children's vocabulary, but they suffer from drawbacks of subjective rating scales – for example, positively biased parental ratings. Consequently, an objective view on the status of the child's language development should be substantiated at least by one reliable and valid performance assessment. Our aim was to show by way of example that the construct validity of a vocabulary test must also be checked for non-native speakers to be able to use this test for multilingual children. Despite the large number of bilingual Turkish—German children, hardly any assessment tools are available to date.

Rationale of the present study

In the current study, we demonstrate how a language assessment inventory can be constructed in a balanced way to adequately capture productive vocabulary in mono- and bilingual children. We ensured the comparability of tests across multiple languages to facilitate cross-linguistic comparisons. To achieve this, we implemented a parallel construction of linguistic twin elements, characterized by consistent semantic and syntactic features. Utilizing a standardized language corpus (Corpora Collection Leipzig, 2011), we verified that both the original items and their twin counterparts exhibited comparable frequencies within their respective languages. To this end, we describe the construction process and test the construct validity of German word items by jointly modelling item parameters and person features by means of explanatory item-response models. Specifically, we examined whether observable item difficulties can be predicted by item features and the language background of 3 to 6 year-old mono- and bilingual German- and Turkish-speaking children in a German picture-naming task.

We expected the theoretically based item difficulties to be reflected in the children's performance in the productive vocabulary tests and relied primarily on the item characteristics known to affect vocabulary production. Since the AoA is an important predictor of productive vocabulary skills, we expected an influence on the item difficulty in our picture-naming task, with early-acquired words being easier in terms of accuracy than words that were acquired later (Hypothesis 1a). Further, the described word-forms activation threshold for low versus high-frequency words

(Jescheniak & Levelt, 1994) should result in a better performance of the children for higher-frequency words. Accordingly, word frequency should determine picture-naming accuracy as well, with highly frequent words resulting in a lower difficulty and thus higher naming accuracy (Hypothesis 1b). The performance in the naming task furthermore depends on the lexical category (Haman et al., 2017). We expected the production of nouns to be easier in terms of accuracy than the production of verbs (Hypothesis 1c). Finally, we expected word length to affect item difficulty such that longer words should be associated with lower picture-naming accuracy (Hypothesis 1d). Finally, compounds should be associated with lower picture-naming accuracy than simple word forms (Hypothesis 1e). If these item features significantly predict the empirical item difficulties, this finding would not only confirm findings on language development, but it would also be an indicator of the test's construct validity (Hartig & Frey, 2012).

We analyzed the effect of the person characteristics of age and language background. As children improve their vocabulary with age, this effect should be reflected in the data as well (Hypothesis 2). Although environmental factors play an important role in acquiring a second language, the language development process for very young children that are raised with two languages is also subject to biologically controlled mechanisms (Petitto et al., 2001). Accordingly, we expected an identical pattern in the acquisition of vocabulary regardless of language background, reflected by the same factors exhibiting significance in both populations as outlined in Hypotheses 1a to 1d. To be more precise, the test should show construct validity when being modelled for both mono- and bilingual children. However, bilingual children might start accessing the society language later and use it less often than monolingual children. Thus, we expected a bilingual disadvantage because of the lack of content in the lexicon of these children. Consequently, bilinguals should perform worse compared to the monolinguals, and language background should negatively affect the performance in the naming task (Hypothesis 3).

Method

Participants

Participants were 126 preschool children (62 girls) from Switzerland and Southern Germany. The 75 German monolingual and 51 Turkish—German bilingual children aged between 33 and 78 months ($M = 51.35$, $SD = 7.88$) were recruited through daycare centers and through contacts in the Turkish speaking community. The highest educational attainment of the parents was coded to determine their socioeconomic status, using the International Standard Classification of Education (ISCED; UNESCO Institute for Statistics, 2012) as a reference. Educational level obtained in Germany or Turkey was indicated on a 7-step scale (level 1 = primary education, level 7 = university/tertiary education). The mothers' education ranged from level 1 to level 6 ($M = 4.28$, $SD = 1.99$). Similarly, the fathers' education ranged from level 1 to level 7 ($M = 4.89$, $SD = 2.36$). Characteristics of the study sample are provided in Table 1. Sociodemographic data, such as age and sex of the children and information about the parents' graduation and academic achievement, were collected via a parental questionnaire. Only children with written parental consent participated in the study. Responses of eight children were excluded from analyses because of technical problems with the audio recordings or children terminating the test session early, amounting to a loss of 6.3% of the children's data (Table 1).

Table 1. Characteristics of the Study Sample

Characteristics	SLL	DLL	Full sample
Participants	75	51	126
Gender			
Female	39	23	62
Male	36	28	64
Age in months			
Mean	50.4 (6.42)	52.9 (9.62)	51.4 (7.88)
Min	35	33	33
Max	69	78	78
Educational level			
Mother	5.19 (1.39)	2.90 (1.98)	4.28 (1.99)
Father	6.00 (1.67)	3.18 (2.26)	4.89 (2.36)
Highest	5.51 (1.26)	3.06 (2.14)	4.54 (2.05)

Note. The table provides information about the study sample. Standard deviations are given in brackets. SLL = Single language learners (monolinguals); DLL = Dual language learners (bilinguals).

Procedure

The productive vocabulary test was part of a cross-national study assessing language, metacognition, and socio-emotional competences in 3–6 year-old mono- and bilingual children. The tests were completed by the children in their homes, childcare facilities, or in an examination room of the university. Children were tested on two separate days with individual assessment sessions of 40–60 min per day by research assistants. All the bilingual children were tested by a native speaker in the dominant language. The testing was embedded in a story of crocodile 'Sammy' who wanted to find a treasure and asked the children to help her. One part of the testing, including the productive vocabulary task, was computerized and conducted on a 14" Windows 10 convertible. The other part of the test consisted of offline tasks that also were accompanied by Sammy as a hand puppet.

The instruction in the productive vocabulary task was: "I am going to show you a picture and you tell me what you see in the picture. Are you ready?" The children first received two practice items. If the child had trouble understanding the instructions, they were repeated. During the test application, children were asked to specify their answer in case of underspecified responses, overspecification, part-whole problems (e.g., "bird" instead of "wings") or noun-verb problems (e.g., "Lion" instead of "to roar") by a standardized prompt such as "Do you know another word for it?" or "And what does she do?" If the child named the first or the second exercise object incorrectly, Sammy said: "Oops that was unfortunately not quite right. Try it again." The task was still performed, even if the child had solved both exercise items incorrectly.

Instruments

Productive vocabulary was evaluated using a computer-based task that comprised 32 items, each varying in difficulty level based on factors such as awareness, frequency, and complexity. The task aimed to assess vocabulary knowledge across domains including activities (e.g., drink, draw), animals (e.g., cow, badger), everyday objects (e.g., sponge, dice), and other terms relevant to the children's immediate or broader life experiences. Children were presented with pictures designed specifically for this study, which they had to name (Figure 2). The scale featured an adaptive testing approach with a routing set of 16 items across different difficulty levels. The categorization into routing, basic, and advanced sets was based on information on the words' AoA, length, class, and frequency. All children were required to complete the routing set. Depending on the number of items solved correctly in the routing set, the children proceeded with a set of advanced items (8 items), or they were assigned to the basic item set (8 items). In total, each child completed 24 items. The scale displayed a good expected posterior reliability/plausible value (EAP/PV) in the one-parameter logistic model (1PL) item response theory (IRT; De Boeck & Wilson, 2004; Hartig et al., 2012; Schindler et al., 2018) scaling of $r = .84$. An overview of the items and detailed information regarding item features can be found in the online supplementary materials.

Items were selected based on the effects of AoA, frequency, lexical category, word length, and complexity of words. The items should reliably represent as wide a range as possible within this selected item features. The item characteristics were operationalized as follows. For the operationalization of the frequency, we applied the normed lemma frequencies from the *Childlex Corpus* (Schroeder et al., 2015. Missing values were implemented from the *Core Corpus for the German Language of the 20th Century* (DWDS), a large database for German word frequency standards. The frequency in one million words in the corpus was used as the basis for statistical evaluations (Geyken et al., 2006). For the AoA, we used estimations from Birchenough et al. (2017). German-speaking people over 18 years evaluated these AoA standards on a 7-point-Likert scale for 3,259 German words. We completed missing values on the basis of the Kuperman et al. (2012) AoA ratings, which are ratings for 30,000 English content words from the SUBTLEX corpus. For the length of words, we used the number of phonemes and adjusted these values for phonological processes – that is, counting the sounds that were actually spoken. In the following section, we report the procedure and results of the explanatory item-response models for response accuracy as dependent variable for mono- and bilingual children separately.

Data analysis strategy

The objective of our study was to develop a vocabulary test that incorporates established factors from language acquisition research effectively, aiming to create equitable and comparable versions for both L1 and L2. Our focus will be on investigating the impact of these factors and examining their interaction with the linguistic background of the children in the societal language. We will accomplish this by comparing the structure and item characteristics of the naming task between monolingual German children and bilingual Turkish—German children. We investigated whether the item characteristics can explain a substantial proportion of the variance in item difficulties with a three-step procedure (see Hartig & Frey, 2012; Hartig et al., 2012; Schindler et al., 2018). In Step 1, empirical item difficulties were estimated by using a 1PL IRT model for the accuracy data, implemented as a Generalized Linear Mixed Model (GLMM; Dixon, 2008). The items were dummy-coded with one item that was not represented by a dummy-coded fixed effect, as the reference item. The items were included as predictors (fixed effects) for logit-transformed response accuracy as dependent variables. The intercept was allowed to vary randomly between participants.
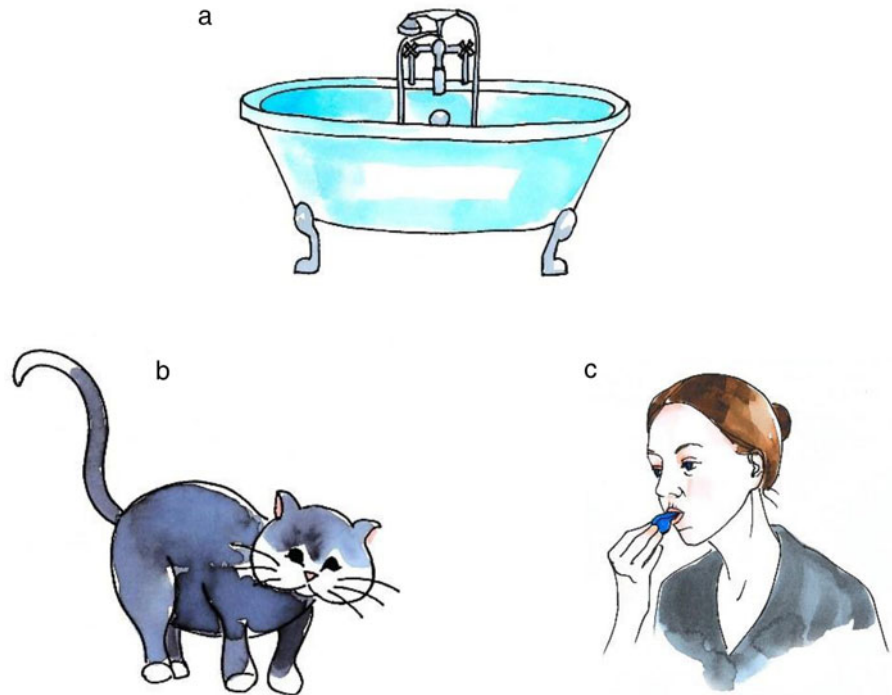
Figure 2. Examples of test items for a compound noun (Badewanne 'bathtube' (a)), a simple noun (Katze 'cat' (b)) and a verb (pfeifen 'whistle' (c)).

The 1PL IRT model estimated in Step 1 for accuracy was as follows:

$$\text{logit (accuracy}_{ij}) = b_{0j} + b_{1j}X_{1ij} + b_{2j}X_{2ij} + \ldots + b_{(k-1)j}X_{(k-1)ij} + r_{ij}$$

$$b_{0j} = g_{00} + u_{0j} \quad \text{Random coefficient, } u_{0j}: \text{person parameter}$$

$$\left. \begin{array}{l} b_{1j} = g_{10} \\ b_{1j} = g_{10} \ldots b_{(k-1)j} = g_{(k-1)} \end{array} \right\} \quad \text{Fixed coefficients: item parameters}$$

$$(1)$$

In Step 2, response accuracy was modeled in a Generalized Linear Mixed Model (GLMM; Dixon, 2008) as a function of the item features (lexical category, AoA, frequency, word length, complexity). Equation 2 shows the estimation of the theoretically derived item features predictors in a maximal model (Barr et al., 2013).

$$\text{logit}(accuracy_{ij}) = \beta_0 + b_{0j} + (\beta_1 + b_{1j}) * (feature_1)$$
$$+ (\beta_2 + b_{2j}) * (feature_2) + \ldots + (\beta_q + b_{qj}) * (feature_q) + r_{ij}.$$

$\beta_0 + b_{0j}$ Fixed intercept + Random intercept for person j

$(\beta_n + b_{nj}) * (feature_n)$ Fixed slope feature n and Random slope for feature n for person j

r_{ij} Error term

$$(2)$$

In Step 3, we assessed model fit by correlating predicted and empirical item difficulties. A significant and substantial R^2 would indicate that empirical item difficulties can be predicted from our selected item features. The reflection of individual differences in response accuracy can be considered as evidence of construct validity.

For analyzing the effect of age (Hypothesis 2), we included centered age in months as a predictor into the model. To test whether the vocabulary item features have the same effect on item difficulty in both populations, in mono- and bilingual children, we used the three-step-procedure of explanatory-item-response-modeling for response accuracy as dependent variable for the total sample (Model 1) and for the two subgroups – monolinguals (Model 2) and bilinguals (Model 3). We included AoA, lexical category, word length, complexity, and frequency as fixed effects. Intercept and slopes for item features were allowed to vary randomly between participants (random intercept and slopes). If development had occurred in parallel with both subgroups, then both models should show significant effects. Assuming that bilingual children should perform more poorly compared to the monolingual subgroup, we relied on interaction analysis to explore whether a significant interaction existed between the individual item features and whether the child spoke one or two languages (Hypothesis 3). As item features, we included AoA, lexical category, word length, complexity, and frequency into the model.

All models were estimated with the software packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) for the statistical software R (R Core Team, 2019; Version 3.6.1). Parameters were estimated with Maximum Likelihood (ML). All significance tests were based on a Type I error probability of .05.

Results

The estimation of the empirical item difficulties by using a 1PL IRT model for the accuracy data, implemented as a Generalized Linear Mixed Model (GLMM, Dixon, 2008; step 1), yielded a distribution of item difficulties ranging from -1.84 ("Katze"; "cat") to 2.08 ("Korken"; "cork"). As expected and intended, the set of routing items covered a wider span of difficulty levels, whereas the basic items set was easier on average, and the advanced set was more difficult. The parameter estimates for the fixed and

random effects of the GLMM with logit-transformed response accuracy as dependent variable (Step 2) are provided in Table 2.

AoA (Hypothesis 1a), frequency (Hypothesis 1b), lexical category (Hypothesis 1c) and the complexity of the vocabulary (Hypothesis 1e) as main effects reached significance. As expected, the earlier vocabulary was acquired the easier it was accessed as indicated by higher naming accuracy ($β = −0.65$; $z = −8.36$; $p < .001$). Children were less accurate in naming verbs compared to nouns ($β = −0.79$; $z = −4.41$; $p < .001$). The higher the word frequency, the more accurately children named the word ($β = 0.48$; $z = 6.11$; $p < .001$). In addition, compounds (Hypothesis 1e) were named less accurately than simple words ($β = −0.73$; $z = −2.31$; $p = .021$). Word length (Hypothesis 1d), however, seemed to not affect naming accuracy. Overall, the regression model explained 47% of the variance in the empirical data, (Step 3). To put it in different words, item characteristics were able to successfully predict empirical item difficulties, which is an indicator of the scale's construct validity.

We also hypothesized the person characteristic age to be an important predictor of picture-naming accuracy (Hypothesis 2). As expected, we found a significant age effect ($β = 0.05$; $z = 3.92$; $p < .001$) indicating that children's performance in the picture-naming task increased with increasing age. However, no further interaction of age and the item features was found. Thus, the effects of the item features remained constant across age with the items being generally easier for older children.

We expected similar influences of item features on item difficulties for both populations – for the monolingual German and bilingual German-Turkish children. Accordingly, we computed the GLMM with response accuracy as dependent variable for the total sample (Model 1) and for the two subgroups – monolinguals (Model 2) and bilinguals (Model 3). Table 2 shows the coefficients and variance components for response accuracy as dependent variable. AoA, lexical category, word length, complexity, and frequency were modeled as fixed effects. In both populations, we found an effect of AoA on the performance of the picture naming, which means that vocabulary acquired earlier was easier to access for both the monolingual ($β = −0.61$; $z = −7.60$; $p < .001$) and the bilingual children ($β = −0.88$; $z = −5.87$; $p < .001$). Naming nouns was easier than naming verbs for monolinguals ($β = -0.72$; $z = −3.60$; $p < .001$) and bilinguals ($β = −0.96$; $z = −3.01$; $p = .003$). Somewhat unexpected but in accordance with the results from Model 1, no significant effect was found for word length in both subsamples. Compounds were more difficult to name than simple word forms for the bilinguals ($β = −1.26$; $z = −2.07$; $p = .038$), whereas no significant effect was found for the monolinguals. As expected, we found a significant effect of frequency in both subsamples, indicating that frequency was positively associated with naming accuracy for the monolinguals ($β = 0.37$; $z = 4.37$; $p < .001$) and for the bilinguals ($β = 0.71$; $z = 6.27$; $p < .001$).

To test differences between the mono- and bilingual children, we additionally modeled the interaction terms between language background and the main effects. The model revealed no significant differences. We found only main effects of language background and of the item features lexical category ($β = −0.79$; $z = −3.91$; $p < .001$), AoA ($β = −0.65$; $z = −8.10$; $p < .001$), frequency ($β = 0.39$; $z = 4.51$; $p < .001$), and complexity ($β = −0.64$; $z = −2.07$; $p = .039$). Word length approached significance ($β = 0.21$; $z = 1.77$; $p = 0.08$) as well as frequency ($β = 0.25$; $z = 1.84$; $p = .066$). Overall, bilingual children performed slightly more poorly than monolingual children. These findings indicate that development between both groups is similar, albeit time-delayed.

Discussion

The objective of this study was to investigate the factors that influence the performance of productive vocabulary in order to contribute to our understanding of vocabulary development in monolingual and bilingual individuals. We developed a productive vocabulary task and used explanatory item-response models to demonstrate the validity of construct, making it suitable not only for monolingual but also for bilingual children. The items were selected based on various factors, including age of acquisition,

Table 2. Fixed Effects and Variance Components in the LLTMs for Response Accuracy for the Subsamples of Monolinguals and Bilinguals

	Response Accuracy		
	Model 1 (Total sample)	Model 2 (Monolingual)	Model 3 (Bilingual)
Parameter	β (SE)	β (SE)	β (SE)
Fixed Effects			
Intercept	0.61 (0.15)***	1.05 (0.13)***	−0.39 (0.30)
Age of acquisition	−0.65 (0.08)***	−0.61 (0.08)***	−0.88 (0.15)***
Lexical category	−0.79 (0.18)***	−0.72 (0.20)***	−0.96 (0.32)**
Word length	0.19 (0.11)	0.19 (0.11)	0.24 (0.21)
Simplex vs. Compound	−0.73 (0.31)*	−0.57 (0.31)	−1.26 (0.61)*
Frequency	0.48 (0.08)***	0.37 (0.09)***	0.71 (0.11)***
Variance Components			
Participant	0.98 (0.99)	0.42 (0.65)	2.80 (1.67)

Note. The table displays the fixed effects and variance components in the LLTM for the response accuracy in the picture-naming task. Age of acquisition in month was centered and lexical category dummy coded (noun = 0; verb = 1). Word length specifies the number of phonemes (phonological processes- adjusted) and complexity was again dummy coded (simplex = 0; compound = 1). Frequency per millions tokens.*$p < .05$, **$p < .01$, ***$p < .001$.

frequency, lexical category, length, and complexity. Our findings suggest that both mono- and bilingual children's lexicon are influenced by the same factors. Furthermore, the results indicate that our measurement of vocabulary is valid for bilingual children raised in a bilingual context.

As expected for the total sample, vocabulary acquired early and words that are more frequent were easier to access as well as nouns in comparison to verbs. Complexity of words negatively affected the performance in the naming task, and compounds were more difficult to name for bilingual children than for monolingual children. Furthermore, we assumed that children improve their vocabulary with age, which was confirmed by our analyses. Although the nature of the data collection is cross sectional and thus the results simply represent an age effect, they nonetheless at least indicate parallel development in monolingual and bilingual children. Influences of the item features were expected to be similar for the subgroups, which means that effects of frequency, AoA, lexical category, complexity, and word length should be reflected in the results of both groups. Indeed, our analysis showed this similarity for frequency, AoA, lexical category, and complexity of words.

We expected the results to depend on the language background of the children, with a bilingual disadvantage in the number of items solved (e.g., Gollan et al., 2008; Rinker et al., 2017), but further analyses between the language background and item features revealed no interactions. This finding suggests that the two language groups have no structural differences in vocabulary. The construction of this picture-naming task and its theoretically based test items with systematically varied item features allows for an interpretation in terms of individual differences in the vocabulary of mono- and bilingual children. In the remaining discussion, we discuss the results of our analysis and underline its importance for the construction of vocabulary tests assessing the language abilities of single and dual language learners.

Dual language learners are not necessarily confronted with the same words as monolingual children, or the frequency of words can interact in a complex way depending on the topic and life circumstances (e.g., Caselli et al., 1999; Rinker et al., 2017). For example, dual language learners might be confronted with words that depict household items or body parts in the foreign language, but they might learn academic word material in institutional settings. The question remains whether words acquired early are generally more frequent. No general consensus exists on the relationship of these influential factors. In a sample of 161 students that performed different tasks on word processing in Dutch language, Ghyselinck et al. (2004) showed AoA and frequency to be highly correlated. Given that the conditions of language learning in a bilingual context obviously differ from those of the monolinguals, we should be careful in interpreting the relationship between AoA and frequency. The reference for the AoA in the second language is the age at which the bilingual child begins to learn this language and makes use of it in interaction with other speakers. This may differ considerably from the chronological age, which leads to some implications in the language development process. When bilingual children start to access a second language, they can draw on existing structures of the first language, which adds new words to their lexicon that are not necessarily high-frequency words but are important for everyday life. Nonetheless, we hypothesized that words acquired early and those with high frequency in everyday usage have a semantic advantage within the lexical system (Brysbaert et al., 2000; Steyvers & Tenenbaum, 2005). Consequently, the

retrieval of these words should be easier and more accurate compared to words with later AoA or low-frequency words (Hernandez & Li, 2007; Jescheniak & Levelt, 1994), as demonstrated in our study. This effect is expected to be even more pronounced in terms of response latency, as these words are expected to have a positive impact on retrieval speed due to their more stable connections. However, when conducting studies with very young children around the age of 3, practical limitations arise, particularly in measuring retrieval speed, which becomes challenging.

In principle, a difference between age and the period of acquisition of the second language can also be assumed with the influence of the lexical category (e.g., Hernandez et al., 2005; Montrul & Foote, 2014). As argued before, we expected an advantage in naming nouns over verbs and this aligns with our findings. We found a significant effect of lexical category for monolinguals and bilinguals, albeit stronger for the bilingual group. The development seems to be similar, but the influence of the lexical category differs slightly in the bilingual context. The acquisition of verbs probably starts at a point when the children already have some syntactic structures at their disposal and their communication intentions are more complex. Accordingly, children need to access verbs: which has implications about the order of the acquisition of verbs in relation to nouns. Investigations of Naigles and Hoff-Ginsberg (1998) on the syntactic environments of verbs showed that the variety of sentence structures affected the acquisition of verbs. Verbs that appeared in many different syntactic frames were acquired earlier in comparison to verbs that children heard in fewer forms of sentence structures. Assuming that bilingual children develop in a more complex language environment, the acquisition of verbs should occur earlier than for monolingual children, which could explain our results. All the bilingual children in our study were raised with Turkish as their home language. When comparing Turkish and German, clear structural differences can be observed, posing an additional challenge for bilingual children. Turkish, being an agglutinating language, has fewer but longer words compared to German (Daller et al., 2011). The structural disparities between these two languages specifically draw the bilingual child's attention to the word material during the language acquisition process, resulting in a positive impact on metalinguistic knowledge. Bilingual children can make use of this metalinguistic knowledge, which might facilitate the acquisition of the language, including the acquisition of verbs (Bialystok, 2001).

Interestingly, we found an effect of word complexity only for the bilingual group. These children seemed to have more problems with accessing compounds compared to the monolingual group. We suppose that monolingual children in our study were in an age range when the influences of this feature are less important. Data from Birchenough et al. (2017) and Kuperman et al. (2012) show for this group that the AoA for the compound *spiderweb* ("Spinnennetz") was 4.90, and *bathtub* ("Badewanne") was 3.84. The question of how compounds are stored in this context also seems to be important. Compounds are multimorphemic words that are composed of at least two free morphemes. Assuming that first compounds are stored as a whole word representation, children should have little difficulty formulating these complex words. Access to the whole word meaning eliminates the need to combine the meaning of the word's constituents and activate structural knowledge, which allows for more economic processing, especially for frequent word material. To our knowledge, research on the access of compounds in the

second language (L2) is scarce. In principle, we expect some differences in the formation and structure of compounds in contrast to the first Language (L1), which should be an additional hurdle for the bilingual child. Given the assumption of a whole-word representation for well-known and frequent compounds, the worse performance of the bilingual group might also substantiate the time-delayed development course for the bilingual child. Thus, the effect for bilinguals (and not monolinguals) in our study might indicate a lag in their linguistic development.

Notably, our analyses showed no significant effects of word length in addition to the other vocabulary features for monolingual and bilingual children. The length of the items related to the phonemes, which were adjusted to the phonological processes. A consistent strong effect of word length on the encoding process can be found in the literature (Vance et al., 2005). Our null finding of this effect may be due to the age range of our sample. This effect is particularly significant at the very beginning of language acquisition and weakens as language acquisition progresses. The children we tested were between 3 and 6 years-old. At this age, we assume that these effects are less important for the access of vocabulary. Moreover, we used relatively few compounds compared to simple words, which might explain the lack of effect. Nevertheless, age was an important factor that influenced the performance in the naming task. As expected, we found that children performed better with age, which means that the test items are sensitive to developmental differences.

The findings of the present study indicate that the structure of language acquisition develops in parallel for L1 and L2 learners. In addition to a general delay in the development of L2, further important factors influence the acquisition of vocabulary. Considering the benefits of first language competencies in the language learning process, a delayed but parallel acquisition of the second language should not necessarily be disadvantageous for the bilingual child. The lack of significant interaction effects in our findings strongly indicates structural validity and that the development process is comparable. Since the children can draw on existing concepts, it makes sense, for example, when learning or supporting L2, to offer appropriate vocabulary material to augment the lexicon. The acquisition of L1 cannot be transferred absolutely to the acquisition of L2. Thus, it is not a question of imitating the acquisition of L1 but rather of supplementing it accordingly in the acquisition of L2.

Limitations

A cross-sectional research design is limited in providing information about the development process in the vocabulary of bilingual children that a longitudinal study could provide. In addition to methodological challenges associated with comparing test results across languages with different structures, there is a broader issue concerning the sample selection. The use of a between-subject design to compare the performance of bilingual individuals with demographically matched monolingual counterparts is a common approach in bilingualism research. While such designs can reveal fundamental performance differences between groups, they often lack sensitivity to variations within each group. Finding a homogeneous group of participants who share all demographic variables except for language experience is particularly challenging, as language experience itself correlates with other factors (Luk & Bialystok, 2013). Further data collection is required not only for accuracy but also for latency analyses to allow for conclusions on

the automation and stability in the word retrieval process. Our results might have also been influenced by the circumstances stemming from the Corona pandemic. During the testing period, the living conditions of many of the participating children had become more difficult. For example, they were unable to attend childhood facilities for a long period, or the conditions of the survey changed because of governmental regulations. These external variables could have adversely affected the comparability of the data collected, although we assume the effect was the same for mono- and bilingual children. Lastly, given the collection of many parameters, the test session was very long for the children. Consequently, the number of items for the productive vocabulary was also small and therefore the data should be interpreted with caution.

Conclusion

The present study corroborates the body of research on the main determinants of vocabulary acquisition. The findings further underscore that these factors seem to influence vocabulary acquisition comparably for mono- and bilingual children. The study also demonstrated that valid assessment tools can be developed to identify individual differences in both groups of children. These findings have direct practical relevance, given the increasing number of bilingual or multilingual children and the need to examine these children with appropriate tests – for example, for school aptitude diagnostics. An important point is that these tests should not only be validated for monolingual children but also that they can map developmental differences in bilingual children, as was shown with our analysis. The results suggest that the structure of vocabulary acquisition in L2 is similar to the acquisition in L1, and children can draw on previous knowledge from L1. Our findings are directly and practically relevant because they contribute to understanding and providing support for bilingual children in their language learning process.

References

Acarlar, F., Aksu-Koç, A., Küntay, A. C., Maviş, İ., Sofu, H., Topbaş, S., & Turan, F. (2009). Adapting MB-CDI to Turkish: The first phase. In S. Ay, Ö. Aydın., İ. Ergenç, S. Gökmen, S. İşsever S. & D. Peçenel (Eds.), *Essays on Turkish linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6th–8th, 2008* (pp. 313–320). Harrassowitz Verlag.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 560–585. https://doi.org/10.1080/783663595

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bialystok, E. (2001). Metalinguistic aspects of bilingual processing. *Annual Review of Applied Linguistics*, *21*(1), 169–181. https://doi.org/10.1017/S0267190501000101

Birchenough, J. M. H., Davies, R., & Connelly, V. (2017). Rated age-of-acquisition norms for over 3,200 German words. *Behaviour Research Methods 49*, 484–501. https://doi.org/10.3758/s13428-016-0718-0

Bolognesi, M., & Steen, G. (2018). Editors' introduction: Abstract concepts: Structure, processing, and modeling. *Topics in Cognitive Science*, *10*(3), 490 -500. https://doi.org/10.1111/tops.12354

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S. Y., Pascual, L., Pêcheux, M. G., Ruel, J., Venuti, P., & Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, *75*(4), 1115–1139. https://doi.org/10.1111/j.1467-8624.2004.00729

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta psychologica*, *104*(2), 215–226. https://doi.org/10.1016/s0001-6918(00)00021-4

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive neuropsychology*, *14*(1), 177–208. https://doi.org/10.1080/026432997381664

Carol, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture naming latency. *Quarterly Journal of Experimental Psychology*, *25*, 85–95. https://doi.org/10.1080/14640747308400325

Caselli, C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language*, *26*(1), 69–111. https://psycnet.apa.org/doi/10.1017/S0305000998003687

Chiarello, C., Shears, C., & Lund, K. (1999). Imageability and distributional typicality measures of nouns and verbs in contemporary English. *Behavior Research Methods, Instruments & Computers*, *31*(4), 603–637. https://doi.org/10.3758/BF03200739

Clark, E. (1998). Morphology in language acquisition. In A. Spencer A and A. M. Zwicky (eds.), *The handbook of morphology* (pp. 374–389). Blackwell.

Clark, E., Gelman, S. A., & Lane, N. M. (1985). Compound nouns and category structure in young children. *Child Development*, *56*(1), 84–94. https://doi.org/10.2307/1130176

Corpora Collection Leipzig. (2011). *Leipzig Corpora Collection-German*. https://corpora.uni-leipzig.de/en?corpusId = deu_mixed_2011_dan.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection?. *Journal of Memory and Language*, *41*(3), 365–397. https://doi.org/10.1006/jmla.1999.2651

Cote, L. R., & Bornstein, M. H. (2014). Productive vocabulary among three groups of bilingual American children: Comparison and prediction. *First Language*, *34*(6), 467–485. https://doi.org/10.1177/0142723714560178

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments & Computers*, *31*(4), 650–658. https://doi.org/10.3758/BF03200741

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Daller, M. H., Yıldız, C., de Jong, N. H., Kan, S., & Başbaği, R. (2011). Language dominance in Turkish-German bilinguals: Methodological aspects of measurements in structurally different languages. *International Journal of Bilingualism*, *15*(2), 215–236. https://doi.org/10.1177/1367006910381197

Dapretto, M., & Bjork, E. L. (2000). The development of word retrieval abilities in the second year and its relation to early vocabulary growth. *Child Development*, *71*(3), 635–648. https://doi.org/10.1111/1467-8624.00172

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 824–845. https://doi.org/10.1037/0278-7393.15.5.824

De Groot, A. M., Dannenburg, L., & Van Hell, J. G. (1994). Forward and backward word translation by bilinguals. *Journal of Memory and Language*, *33*(5), 600–629. https://doi.org/10.1006/jmla.1994.1029

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dijkstra, T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. https://doi.org/10.1017/S1366728902003012

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories* (2nd ed.). Baltimore, Brookes Publishing.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*(1), 20–35. https://doi.org/10.1016/j.jml.2008.07.003

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading*. Technical Report; No. 257.

Gentner, D. (2006). Why Verbs Are Hard to Learn. In Hirsh-Pasek, K., & Golinkoff RM (eds.), *Action meets word: How children learn verbs* (pp. 544–564). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195170009.003.0022

Geyken, A., Habermas, J., Klemperer, V., Kraus, K., & Lenz, S. (2006). *The DWDS corpus: A reference corpus for the German language of the 20 century.*

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, *115*(1), 43–67. https://doi.org/10.1016/j.actpsy.2003.11.002

Gilhooly, K. J., & Watson, F. L. (1981). Word age-of-acquisition effects: A review. *Current Psychological Reviews*, *1*(3), 269–286. https://doi.org/10.1007/BF02684489

Gollan, T. H., & Acenas, L. A. R. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and Tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 246-269. https://doi.org/10.1037/0278-7393.30.1.246

Gollan, T. H., Montoya, R. I., & Bonanni, M. P. (2005). Proper names get stuck on bilingual and monolingual speakers' tip of the tongue equally often. *Neuropsychology*, *19*(3), 278–287. https://doi.org/10.1037/0894-4105.19.3.278

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always a means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of memory and language*, *58*(3), 787–814. https://doi.org/10.1016/j.jml.2007.07.001

Haman, E., Łuniewska, M., Hansen, P., Simonsen, H. G., Chiat, S., Bjekić, J., Blažienė, A., Chyl, K., Dabašinskienė, I., Engel de Abreu, P., Gagarina, N., Gavarró, A., Håkansson, G., Harel, E., Holm, I. E., Kapalková, S., Kunnari, S., Levorato, C., Lindgren, J., … Armon-Lotem, S. (2017). Noun and verb knowledge in monolingual preschool children across 17 languages: Data from cross-linguistic lexical tasks (LITMUS-CLT). *Clinical Linguistics & Phonetics*, *31*(11–12), 818–843. https://doi.org/10.1080/02699206.2017.1308553

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten [Construct validity and scale interpretation in competence diagnostic by predicting task difficulty]. *Psychologische Rundschau*, *63*(1), 43–49. https://doi.org/10.1026/0033-3042/a000109

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665–686. https://doi.org/10.1177/0013164411430707

Hernandez, A., & Li, P. (2007). Age of acquisition: Its neural and computational mechanisms. *Psychological Bulletin*, *133*(4), 638–650. https://doi.org/10.1037/0033-2909.133.4.638

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, *9*(5), 220–225. https://doi.org/10.1016/j.tics.2005.03.003

Heuven, W. V., & Dijkstra, T. (2010). Language comprehension in the bilingual brain: fMRI and ERP support for psycholinguistic models. *Brain Research Reviews*, *64*, 104–122. https://doi.org/10.1016/j.brainresrev.2010.03.002

Holowka, S., Brosseau-Lapré, F., & Petitto, L. A. (2002). Semantic and conceptual knowledge underlying bilingual babies' first signs and words. *Language Learning*, 52(2), 205–262. https://doi.org/10.1111/0023-8333.00184

Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta psychologica*, 127(2), 277–288. https://doi.org/10.1016/j.actpsy.2007.06.003

Janssen, N., Bi, Y., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes*, 23(7–8), 1191–1223. https://doi.org/10.1080/01690960802250900

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 824–843. https://doi.org/10.1037/0278-7393.20.4.824

Klassert, A., Gagarina, N., & Kauschke, C. (2014). Object and action naming in Russian-and German-speaking monolingual and bilingual children. *Bilingualism: Language and Cognition*, 17(1), 73–88. http://doi.org/10.1017/S1366728913000096

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connection between bilingual memory representations. *Journal of Memory and Language*, 33(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lee, M. W., & Williams, J. N. (2001). Lexical access in spoken word production by bilinguals: Evidence from the semantic competitor priming paradigm. *Bilingualism: Language and Cognition*, 4(3), 233–248.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MIT Press

Levelt, W. J. M. (1992). Accessing words in speech production: stages, processes and representations. *Cognition*, 42(1–3), 1–22. https://doi.org/10.1016/0010-0277(92)90038-j

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and brain sciences*, 22(1), 1–75. https://doi.org/10.1017/s0140525×99001776

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: Are they simply cumulative-frequency effects? *Cognition*, 78(2), 189–205. https://doi.org/10.1016/S0010-0277(00)00117-7

Li, M., Jiang, N., & Gor, K. (2017). L1 and L2 processing of compound words: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition*, 20(2), 384–402. https://doi.org/10.1017/S1366728915000681

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of cognitive Psychology*, 25(5), 605–621. doi: 10.1080/20445911.2013.795574. PMID: 24073327; PMCID: PMC3780436.

Marelli, M., Zonca, G., Contardi, A., & Luzzatti, C. (2014). The representation of compound headedness in the mental lexicon: a picture naming study in aphasia. *Cognitive neuropsychology*, 31(1–2), 26–39. https://doi.org/10.1080/02643294.2013.860024

Meschyan, G., & Hernandez, A. (2002). Age of acquisition and word frequency: determinants of object-naming speed and accuracy. *Memory & cognition*, 30(2), 262–269. https://doi.org/10.3758/bf03195287

Meyer, A. S., Roelofs, A., & Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, 48(1), 131–147. https://doi.org/10.1016/S0749-596X(02)00509-0

Montrul, S., & Foote, R. (2014). Age of acquisition interactions in bilingual lexical access: A study of the weaker language of L2 learners and heritage speakers. *International Journal of bilingualism*, 18(3), 274-303. https://doi.org/10.1177/1367006912443431

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, 25(1), 95–120. https://doi.org/10.1017/s0305000997003358

Nation, I. S. P. (2013). *Learning vocabulary in another language*, 2nd edn. Cambridge University Press.

Nation, P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 35–54). London, England: Arnold

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, 38(1–2, Serial No 149), 136. https://doi.org/10.2307/1165788

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology*, 17(4), 273–281. https://doi.org/10.1080/17470216508416445

Petitto, L. A., Katerelos, M., Levy, B. G., Gauna, K., Tétreault, K., & Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. *Journal of Child Language*, 28(2), 453–496. https://doi.org/10.1017/S0305000901004718

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10(5), 377–500. https://doi.org/10.1080/02643299308253469

Potter, M. C. (1979). Mundane symbolism: The relations among objects, names, and ideas. *Symbolic functioning in childhood*, 41–65. Lawrence Erlbaum Associates Inc.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, 107(3), 460–499. https://doi.org/10.1037/0033-295X.107.3.460

R Core Team (2019) *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

Rinker, T., Budde-Spengler, N., & Sachse, S. (2017). The relationship between first language (L1) and second language (L2) lexical development in young Turkish-German children. *International Journal of Bilingual Education and Bilingualism*, 20(2), 218–233. DOI: 10.1080/13670050.2016.1179260

Sachse, S., Budde-Spengler, N., & Rinker, T. (2016). Türkischsprachiges Instrument zur frühen Wortschatzerfassung–Türkçe Ifade ve Lisan Dizelges Araştırması TILDA [Turkish language instrument for the early assessment of vocabulary skills]. *Unpublished assessment*. University of Education Heidelberg/University of Ulm/University of Konstanz.

Schaefer, B., Ehlert, H., Kemp, L., Hoesl, K., Schrader, V., Warnecke, C., & Herrmann, F. (2019). Stern, gwiazda or star: Screening receptive vocabulary skills across languages in monolingual and bilingual German–Polish or German–Turkish children using a tablet application. *Child Language Teaching and Therapy*, 35(1), 25–38. https://doi.org/10.1177/0265659018810334

Schindler, J., Richter, T., Isberner, M. B., Naumann, J., & Neeb, Y. (2018). Construct validity of a process-oriented test assessing syntactic skills in German primary schoolchildren. *Language Assessment Quarterly*, 15(2), 183–203. https://doi.org/10.1080/15434303.2018.1446142

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge university press.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex—Eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen [childLex—A Lexical Database for Print Language for Children in German]. *Psychologische Rundschau*, 66(3), 155–165. https://doi.org/10.1026/0033-3042/a000275

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, 27(5), 499–520. https://doi.org/10.1016/0749-596X(88)90022-8

Snodgrass, J. G. (1984). Concepts and their surface representations. *Journal of Verbal Learning & Verbal Behavior*, 23(1), 3–22. https://doi.org/10.1016/S0022-5371(84)90479-1

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments & Computers*, 28(4), 516–536. https://doi.org/10.3758/BF03200540

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14(1), 29–56. https://doi.org/10.1207/s15516709cog1401_3

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

UNESCO Institute for Statistics. (2012). International Standard Classification of Education (ISCED) 2011. doi:10.15220/978-92-9189-123-8-en

Vance, M., Stackhouse, J., & Wells, B. (2005). Speech-production skills in children aged 3–7 years. *International Journal of Language & Communication Disorders*, *40*(1), 29–48. https://doi.org/10.1080/13682820410001716172

Van Hell, J. G., & De Groot, A. M. (1998a). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*, *1*(3), 193–211. https://doi.org/10.1017/S1366728998000352

Van Hell, J. G., & De Groot, A. M. (1998b). Disentangling context availability and concreteness in lexical decision and word translation. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *51*(1), 41–63. https://doi.org/10.1080/713755752

Walker, D., Greenwood, C. R., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development*, *65*(2), 606–621. https://doi.org/10.2307/1131404

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. https://doi.org/10.1017/S0272263108080042

Wheeldon, L. R., & Levelt, W. J. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language*, *34*(3), 311–334.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, *69*(3), 848–872. https://doi.org/10.2307/1132208