Lexical diversity in bilingual speakers of Croatian and Italian[1]

Gordana Hržica
University of Zagreb
Faculty of Education and Rehabilitation Sciences
Department of Speech and Language Pathology

Maja Roch
University of Padua
Department of Developmental Psychology and Socialisation

Abstract

The aim of this study was to determine whether measures of lexical diversity (LD) can be used to assess general lexical knowledge. This study focused on 30 sequential Croatian/Italian bilinguals aged between 5 and 7 years, who were compared with 30 Croatian and 30 Italian monolinguals matched in age. LD was assessed from narrative language samples using standard measures and compared with Peabody Picture Vocabulary Test (PPVT) results. Two LD measures, the type-token ratio (TTR) and the measure proposed by Maas (1972), did not differentiate between bilinguals and monolinguals, whereas several other did: total number of words (NTW), number of different words (NDW), measure D, moving average type-token ratio (MATTR) and hypergeometric diversity of D (HD-D). NTW, NDW, D, MATTR and HD-D were predictive of PPVT results for monolinguals, whereas NDW, TNW and D (with replacements, Dr) were predictive of PPVT results for bilinguals. Our results suggest that LD measures can be reliable measures of bilinguals' vocabulary knowledge and can adequately reflect PPVT results in languages other than English.

Key words: bilingualism, lexical diversity, cross-language associations

[1] This work has been supported in part by Croatian Science Foundation under the project *Multilevel approach to spoken discourse in language development* (UIP-2017-05-6603)

1. Introduction

Bilingual children usually have a smaller vocabulary in each of the developing languages than their monolingual peers do. Nevertheless, total lexicon size across both languages in bilingual children appears to be comparable to the vocabulary of monolingual peers (overview: Hamers & Blanc, 2000). This insight highlights the need to examine bilingual children *across* both their languages when studying early lexical development, yet this can be challenging, especially when early development of the languages involved has been sparsely studied in the monolingual context and when few standardised tests are available. Language samples, i.e. transcribed recordings of spoken language, provide an alternative to standardised language assessment. Diverse language sampling methods have been developed. For example, a stimulus can be used to elicit narrative discourse, which can provide rich information on several language levels, including the lexical level. Such narratives are generated in non-testing conditions, which can be particularly useful in studies with children. Not surprisingly, the narrative methodology has been used extensively in bilingual research (among others Damico et al., 1983; Gutiérrez-Clellen, 2002; Bedore et al., 2010). Language samples, however, can lack reference measures needed for determining group and individual differences. There are also challenges in establishing reliability of such measures, especially when they have not been rigorously explored cross-linguistically.

Observing how results obtained from the analysis of language samples correspond to those obtained from tests of language knowledge might help to build a more detailed picture of bilingual speakers and can benefit research in languages (or language pairs) for which standardised language assessment materials are lacking. This would serve a growing need for reliable procedures for discriminating between bilinguals with typical development and bilinguals with language disorder in cases where no tests or only monolingually normed tests are available.

In an effort to provide insights to guide cross-language analysis in bilingual children, the present study applied two methodologies, language testing and several lexical diversity (LD) measures, to assess vocabulary knowledge of bilingual children speaking Italian and Croatian and to compare them to monolingual peers.

2. Background

2.1. LD measures

Longstanding, widely used measures of lexical development based on language samples are the number of different words (NDW), number of total words (NTW) and number of lemmas (NLW). These measures served as the basis for the first LD index, the type-token ratio (TTR; Templin, 1957). All these measures are highly sensitive to text length (e.g. Malveren et al., 2004), so they are considered to be comparable only across language samples of similar length. Such homogeneity is usually ensured by eliciting responses using stimuli such as pictures or narratives, or by selecting a subset of utterances for detailed analysis (e.g.

Tweedie & Baayen, 1998 & McCarthy, 2005). This sample truncation may lead to bias, such as when it leads to comparisons between complete and partial narratives. In addition, it is unclear to what extent NTW, NDW, NLW or TTR apply across languages. Although such cross-language comparability has been proposed (e.g. Simon-Cereijido & Gutiérrez-Clellen, 2009), it seems likely that these measures are sensitive to language-specific features, such as morphological richness, diversity of functional words and word segmentation principles.

Researchers have attempted to correct for differences in sample length. For example, several authors (e.g. Guiraud, 1960; Carroll, 1964; Maas, 1972) have proposed different TTR transformations to compensate for the variability in sample length, but none of these is sufficiently reliable (Malveren et al. 2004). Several authors have proposed mathematical or statistical treatments of LD measures that are highly sensitive to text length. The measure D, proposed by Malvern & Richards (1997) and further developed by Malvern et al. (2004), is calculated in VOCD software (McKee et al., 2000) using a curve-fitting approach. To avoid the problem of sample length, 100 random samples of 35–50 tokens are used to generate mean TTR values, which are plotted as a TTR curve and then compared to a series of ideal curves that the programme generates based on the highest and lowest possible LD for a given text. The value of D is derived from this comparison. Since the samples vary with each iteration of the calculation, the VOCD procedure is run three times, and the result is averaged. While D is used extensively in the literature and appears to resolve problems associated with other measures (Malvern et al., 2004; Jarvis, 2002; McCarthy, 2005), it may still be sensitive to text length (McCarthy & Jarvis, 2007, 2010). Those authors reported that D performs reasonably well for samples with 100–400 tokens (correlation coefficient with tokens, r = .22), consistent with results reported by Koizumi & In'nami (2012). The measure known as "hypergeometric diversity of D" (HD-D) has been proposed as more accurate and more stable than D (McCarthy & Jarvis, 2007; deBoer, 2014), but other researchers suggest that HD-D and D itself may be more sensitive than other LD measures to language sample length (Fergadiotis et al., 2015).

Several additional metrics to assess LD in text samples have been developed. The measure known as textual lexical diversity (MTLD) divides text into sequences with the same TTR (by default, 0.72) and calculates the mean length of sequences with this TTR (McCarthy, 2005). Several researchers claim that the measure is independent of text size in the range of 100–2000 words (McCarthy, 2005; McCarthy & Jarvis, 2007; McCarthy & Jarvis, 2010). More recent work supports the validity of MTLD, especially for samples longer than 100 tokens (Koizumi & In'nami, 2012; Fergadiotis et al., 2015). The moving-average TTR (MATTR; Covington, 2007; Covington & McFall, 2010) calculates TTR for a moving window of fixed length. The window moves through the text, and the individual TTRs are averaged to obtain the final score. Covington & McFall (2010) describe MATTR as a measure independent of text length that can detect changes *within* a text as well as *between* texts. Fergadiotis et al. (2015) found MATTR and MTLD to have greater validity than D and the index proposed by Maas (1972), usually referred to as the Maas index.

2.2. LD measures in bilingual research

It is still not clear to what extent LD measures are applicable to languages other than English, especially since they usually involve different theoretical assumptions (reviewed in Fergadiotis et al., 2015). It is also unclear whether the measures can apply to cross-language comparisons, which is essential for accurately assessing bilingual language knowledge (Simon-Cerijido & Gutiérrez-Clellen, 2009). It seems reasonable to expect that such comparisons cannot always be made, particularly between typologically different languages, which differ in the number of formally distinct inflected word-forms per lemma (paradigm richness; e.g. Xanthos, Laaha, Gillis et al., 2011), as well as in the functions and frequencies of free-standing morphological markers (copulas, articles, prepositions). This issue of cross-language comparisons is often neglected in bilingual research. In fact, several studies have used NTW and NDW to compare lexical production across languages in bilingual speakers without acknowledging the potential pitfalls of this approach (Ucelli & Paez, 2007; Bedore et al., 2010; Simon-Cerijido & Gutiérrez-Clellen, 2009).

Some studies have applied LD measures to languages other than English; most of these studies deal with first language acquisition (reviewed in Malvern et al., 2004; see Johansson, 2008 for the case of Swedish). Additionally, there are studies that report some lexical measures in languages other than English or/and for bilingual speakers (e.g. Reuterskiöld Wagner et al 1999 for Swedish, Fiestas & Peña 2004 for Spanish/English, Acarlar & Johnston 2010 for Turkish, Gagarina 2016 for Russian/German, Kunnari et al 2016 for Finnish/Swedish, Tsimpli et al 2016 for Greek). Reported measures include NTW, NDW and sometimes TTR, but the focus in those studies is not on LD, so the authors do not reflect on these measures or discuss their relevance to language acquisition or bilingual studies.

Only a handful of studies have applied more elaborate LD measures to both languages of bilingual speakers, and they have focused on written language, typically essays of bilinguals. Such studies involved typologically different languages and tried to tackle the problem of language differences. Three approaches were used to control for language effects: standardisation of the results by using Z-scores (indicates how many standard deviations an element lies from the mean), lemmatisation of language samples (reducing inflectional forms of a word to a common base form or lemma), and sample truncation (shortening transcripts according to pre-established criteria, usually a certain number of words or utterances). In their comparison of LD in two groups of Turkish-German bilinguals, one German-dominant and the other Turkish-dominant, Daller et al. (2003) used Z-scores. They were able to demonstrate opposite, mirror-like performance in the two bilingual groups. In their study of English-Polish late bilinguals, Treffers-Daller & Korybski (2016) tried to determine language dominance using two LD measures, which they compared to extralinguistic factors such as language exposure. They lemmatised the language samples to control for differences between English and inflectionally richer Polish. This led to surprising results because the mean number of tokens was larger for English than for Polish, even though English was the participants' second language. Those authors attributed this result to typological differences between languages such as the absence of articles in Polish, pro-drop in Polish, differences in formation of the past tense (suffix vs. auxiliary), indication of case using suffixes (Polish) vs.

prepositions (English), as well as differences in usage and frequencies of forms. These and other arguments suggest the impossibility of eliminating cross-language differences through sample normalisation. The third approach to control for language effects is sample truncation. In their study of two groups of bilinguals differing in language dominance, Daller et al. (2011) tried to establish measures of language dominance for bilinguals who speak structurally different languages, in their case, German and Turkish. They applied several measures to narrative language samples, including the number of tokens. The agglutinative nature of Turkish meant that it contained significantly fewer tokens than German. To correct for this, the authors took nine pairs of texts, one German and one Turkish, in which the two texts were direct translations of each other. They calculated the difference in size between original texts and their translations and determined that Turkish texts were approximately 10% shorter, so they used this length difference to adjust the number of tokens measured in German narratives. While this approach shows promise, it has some pitfalls. For one thing, it may be quite sensitive to how texts in one language are translated directly into another; numerous studies indicate that translations can differ from the original text in several respects, including the number of tokens (reviewed in Cvrček & Chlumská 2015). In addition, cross-language differences measured using this translation approach are likely to be valid only within the same genre and sample length.

The studies of Daller et al. (2003), Treffers-Daller & Korybski (2016), and Daller et al. (2011) managed to identify differences in language dominance that have matched other measures (e.g. self-evaluation). The external validity of these differences may be a concern, however, since all three studies examined groups of early and late bilinguals. This may be appropriate for those authors' research goals of comparing language dominance with pre-established membership in one or the other group. But it may not be appropriate for understanding bilingual speakers without preestablished language preference or with similar competence in both languages.

Since LD indices aim to measure the same construct, it should come as no surprise that previous studies have revealed correlations among them. In a study by Treffers Daller & Treffers-Daller (2014), very strong significant correlation above .9 was found between HD-D and D. Other studies also revealed significant, albeit less strong, correlations. In another study, Treffers-Daller et al. (2016) measured correlations of TTR with the Guiraud measure (Guiraud, 1954; coefficient, .99), HD-D (.84) and MTLD (< .8). Similar results were obtained for correlations of the number of different words with Guiraud, HD-D and MTLD. McCarthy and Jarvis (2010) investigated the Maas index, D, and Yule's K (a probabilistic index). They found that MTLD was moderately to strongly correlated with all three indices (with respective coefficients of .84, .69, and .85), but not with TTR. Such strong correlations raise questions about the need for computing complex measures if simple ones might provide comparable results. Fergadiotis et al. (2015) compared four types of discourse and concluded that techniques such as MTLD and MATTR might perform differently in different types of discourse that have not been studied yet. The studies mentioned above explored LD indices (TTR, Guiraud, HD-D, MTLD, Maas index, D and Yule's K) on written language samples; much less is known about spoken language.

2.3. Croatian and Italian language features

Croatian belongs to the family of South-Slavic languages. It is a fusional language in which bound grammatical morphemes (endings) usually express three grammatical categories (case, number and gender for nouns and adjectives and person, number and tense for verbs). The canonical word order is SVO, but, due to morphological features, word order is relatively free. Declensional classes are based on three genders, but there are also different declensional classes across genders. Verbs are subdivided in nine conjugational classes (Jelaska, 2005) according to differences in stem alterations. Both nominal and verbal morphology are influenced by a number of morphologically conditioned phoneme alterations, leading to many cases of allomorphy. All this makes for a rather rich morphological system.

Italian is a Romance language. It is a fusional language in which bound grammatical morphemes (endings) express two or three grammatical categories (number and gender for nouns; person, number and tense for verbs; number and gender for articles). Italian nouns do not have cases but do have plural forms. A classification by D’Achille & Thornton (2003) distinguishes six classes based on plural formation. Additionally, there are both four indefinite and seven definite articles. There are three regular conjugational classes for verbs. The basic word order is SVO, but it is more flexible than in less fusional languages (Iacobini, 2006).

The two languages show several typological differences that might influence measures of vocabulary diversity. First, Italian has obligatory articles, which might influence the number of total words produced (NTW), along with all measures at least partially based on this measure. Second, Italian does not have cases and it has only three conjugational classes compared to the nine in Croatian. This suggests that the two languages differ in morphological richness or, more precisely, paradigmatic richness. Paradigmatic richness refers to the number of morphological forms per lemma in a certain language (Dressler, 2004). Additionally, when dealing with language acquisition or language processing, as Xanthos et al. (2011) point out, it is important to understand the difference between the grammatical structure of a language system as a whole and as the parts of the system available in the input. In other words, a language may have a very rich morphology measured in the number of forms, but this richness does not materialize in language usage. Although both Croatian and Italian are inflected languages, Italian is only weakly so, while Croatian is strongly so (e.g. Dressler, 2007). As already pointed out, differences in morphological richness might lead to differences in TTR and in measures based on TTR. Several studies have shown that morphologically rich languages have higher TTR (e.g. Bisazza & Monz, 2014; Kann & Schutze, 2017). Such languages are also expected to have a higher number of different words (NDW). It is difficult to predict how other LD measures will vary in different languages, since the measures are so complex.

2.4. Aims of the study

How to measure the different LD dimensions remains controversial, and the literature is biased toward studies of English. Studies involving other languages, particularly those comparing across languages in bilinguals, have generally used quantitative measures such as the number of total words (NTW), the number of different words (NDW) or TTR (e.g. Malvern et al., 2004). Nevertheless, several studies have explored alternative measures (Daller, van Hout & Treffers-Daller, 2003; Laufer, 2003; Dewaele & Pavlenko, 2003; Šiškova, 2012; Daller & Treffers-Daller, 2014, Treffers-Daller & Korybski 2016), but they have focused on written language. This literature, while providing many useful insights into bilingualism, has not provided a clear and consistent framework for measuring LD across languages, particularly across typologically different languages, and in spoken language. Identifying measures useful in this regard may substantially improve research and assessment of languages for which adequate testing materials are unavailable in the mono- or bilingual context. The present study examined bilingual speakers of Croatian and Italian to find out whether LD measures developed for English can be used to study lexical competence in bilingual speakers of other languages. The following questions were addressed:

1. How well can measures developed for English be used for other languages?

It is expected that LD measures, when language differences are kept in mind, will outline differences between participants already established by other assessment material and/or language status.

2. To what extent do LD measures reflect general lexical knowledge?

It is expected that LD measures that have proven more reliable in previous research (D, HD-D, MATTR) will better predict results of general lexical knowledge than those developed without taking into consideration the problem of transcript size (NTW, NDW, TTR, Maas index.

3. Method

3.1. Participants

Participants in this study were mono- and bilingual speakers of Croatian and Italian. Bilingual speakers were members of the Italian-Croatian bilingual community in areas near the Italian border. Bilingualism in the areas of Istria and Rijeka is maintained and encouraged: bilingualism is official policy, kindergartens and schools offer programmes in Italian, economic ties to Italy are strong, and Italian culture is readily accessible. Bilingual participants were preschool children who obtained similar scores on the Test for Reception of Grammar (TROG-2) (Bishop, 2003) adapted for Italian (Bishop, Suraniti et al., 2009) and adapted for Croatian (Bishop, Kuvač Kraljević et al., 2013). Each participant was age-matched with one monolingual speaker of Italian and one monolingual speaker of Croatian. All bilingual children had been attending kindergarten in Italian for at least 18 months prior to the study. We obtained written permission from the parents or legal guardians of 48 children aged 5 and 6. Two children self-excluded from the study. Therefore, we tested 46 children on the Test for Reception of Grammar (TROG-2) (Bishop, 2003) adapted for Italian (Bishop, Suraniti et al., 2009) and for Croatian (Bishop, Kuvač Kraljević et al., 2013). As TROG has been adapted for two languages with rigorous standardisation, the two test versions are expected to involve the same level of difficulty. Although the constructions differ between the two versions, they were carefully selected and adjusted to each language in order to adequately measure language acquisition. TROG-2 was used to determine whether participants had a certain command of both languages.

Although communities with higher ethnolinguistic vitality are typically considered to contain balanced bilingual speakers, community members, especially young ones, can vary in their bilingualism because of extralinguistic factors. Therefore, we imposed inclusion criteria to select participants with appropriate and relatively balanced knowledge of both languages: (a) their performance on TROG-2 in both languages had to be > 70 (corresponding to –2 SD); and (b) the difference between the two languages had to be < 20 points (corresponding to 1.5 SD). Application of these criteria led to the exclusion of 16 children. The remaining group of 30 children showed a mean difference of 5.33 (SD = 12.36) between their scores on the Croatian and Italian TROG-2. The parents of these children indicated that they had been exposed to both languages for more than 2 consecutive years. The mean time of exposure was 32 months. All participants were consecutive bilinguals who were exposed daily to both languages: they estimated that in a typical day, children were exposed to Italian for 38% of the time. According to parental report, 50% of the families belonged to medium socioeconomic status (SES). The remaining families ranged from medium-low (22%) to medium-high SES (28%). Monolingual participants were recruited from institutions in monolingual areas of Padua (Italy) and Zagreb (Croatia). They were chosen to be within ±2 months of the age of the bilingual participants. They were tested with the corresponding Italian or Croatian version of the TROG-2, and their scores fell within the range of average standard results. According to parental report, both groups of monolinguals ranged from medium-low to medium-high SES.

The final group of participants consisted of three subgroups: 30 bilingual Croatian-Italian speakers, 30 monolingual Croatian speakers and 30 monolingual Italian speakers. The three groups did not differ in SES (p>.05). Table 1 illustrates the main characteristics of the participants.

Table 1. Characteristics of the three groups of participants

	Bilinguals (N=30)	Croatian monolinguals (N=30)	Italian monolinguals (N=30)
Mean age, months (SD)	73 (5)	75 (6)	73 (7)
Range, months	61–84	61–83	61–82

TROG-2, Croatian, mean standard score (SD)	91 (11.8)	97.3 (11.3)	
Range	72–117	76–115	
TROG-2, Italian, mean standard score (SD)	86 (18.1)		113.3 (12.7)
Range	63–131		89–136

3.2. Materials

3.2.1. Language samples

LD was calculated on oral language samples obtained using elicitation narrative material in the Multilingual Assessment Instrument for Narratives (MAIN) (Gagarina, Klop, Kunnari, Tantele, Valimaa, Balčiūnienė, Bohnacker, Walters, 2012) adapted for Croatian (Hržica & Kuvač Kraljević, 2012) and for Italian (Roch & Levorato, 2012). MAIN was designed to assess narrative skills in children who acquire one or more languages from an early age: available in 26 languages, it is suitable for children from 3 to 10 years and evaluates both comprehension and production of narratives. It allows for the assessment of two languages of the participant since it contains four parallel stories, two for telling and two for retelling, each with a six-picture sequence. The stories are controlled for cognitive and linguistic complexity, parallelism in macro- and microstructure, and cultural appropriateness and robustness. This elicitation material has not been norm-referenced, but it was developed in research involving approximately 270 participants and the procedures have been standardised.

Two MAIN forms for telling were used, 'Baby goats' and 'Baby birds'. Picture material was presented according to the computer presentation model of MAIN. To control for the effects of shared knowledge and joint attention (Van der Lely, 1996; Serratrice, 2007), each child was asked to select a story from the first screen, which showed four differently coloured squares. The child clicked on his or her choice of square (with assistance from the examiner, if necessary), which launched a PowerPoint® presentation. In fact, the selection process was simulated: the examiner had preselected the story. First, the child viewed the entire set of six pictures in the middle of the screen, after which he or she pressed a key on the keyboard to reveal each pair of pictures. During the initial viewing and telling of the story, only the participant could see the computer screen.

Half the monolingual participants were presented with the 'Baby Birds' story, while the other half were presented with the 'Baby Goats' story. Bilingual participants were divided into two groups, one of which told the ‘Baby Goats’ story in Croatian and ‘Baby Birds’ in Italian, while the other told ‘Baby Birds’ in Croatian and ‘Baby Goats’ in Italian.

Table 2. Children (N) receiving each type of elicitation material

Group	‘Baby goats’		‘Baby birds’	
	Croatian	Italian	Croatian	Italian
Bilinguals	8	7	7	8
Croatian monolinguals	15		15	
Italian monolinguals		15		15

Storytelling was recorded, transcribed using the Codes for Human Analysis of Transcripts (CHAT) system, and further coded using the Computerised Language Analysis (CLAN) programme. CHAT and CLAN are parts of the Child Language Data Exchange System (CHILDES; MacWhinney, 2000). Transcription and coding were done by monolingual speakers of Croatian and Italian (for monolingual samples), and by bilingual speakers of Croatian and Italian. All had received training in language research methodology and coding. Transcripts were also checked by two experienced researchers. All transcripts successfully passed the CHECK programme in CLAN. In each group of language samples, six samples (20%) were randomly selected and re-transcribed. Word-by-word transcription and coding for each of the samples showed at least 86% identity with the corresponding original.

Non-words, hesitations, revisions, repetitions, onomatopoeia, codeswitching and fillers were coded using predefined transcription codes and excluded from further analysis. Each form that corresponded to a known word in either of the two languages (including dialectal words) was treated as a word. In both Italian and Croatian, clitics are usually written as separate words, and this orthographic rule was respected in the transcription. More than 90% of the lexical material on each bilingual transcript belonged to the target language, and this percentage was 100% on most bilingual transcripts.

3.2.2. LD indices

Several measures were used to measure lexical production. The following quantitative measures were calculated in CLAN: NDW, NTW, TTR, D, and MATTR. Gramulator software (McCarthy, Watanabe & Lamkin, 2012) was used to calculate the Maas index (1972) and HD-D. Language samples were converted to plain text format and cleared of CHAT codes prior to Gramulator analysis.

Table 3. LD indices

Category	Name	Acronym	Software
Quantitative measures	Number of total words	NTW	CLAN
	Number of different words	NDW	CLAN
TTR and transformations	Type-token ratio	TTR	CLAN

	Maas index	MAAS	Gramulator
Model-based indices	VOCD-D	D	CLAN
	Hypergeometric diversity of D	HD-D	Gramulator
	Moving-average type-token ratio	MATTR	CLAN

3.2.3. Vocabulary test

Participants were tested using the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997) adapted and standardised for Croatian (Dun & Dunn, Kovačević et al., 2010) and for Italian (Dun & Dunn, Stella, Pizzoli & Tressoldi, 2000). The adapted versions of the PPVT maintain the same procedure as the original but introduce changes in the lexical material (order of words, exclusion/inclusion of words) to ensure similar level of difficulty based on standardisation data. PPVT is a receptive vocabulary test in which the participant is asked to choose one of four pictures upon hearing the target word. Testing finishes when the participant makes 8 errors in 12 consecutive words.

Recent literature has shown that children growing up in a bilingual environment show a discrepancy between receptive and expressive language that is greater than the gap usually shown by monolinguals. This is in part a consequence of very poor expressive vocabulary (Gibson, Pena & Bedore, 2014). Such results, obtained from a number of studies (overviewed in ibid.), suggest that classical tests measuring expressive vocabulary through naming might not reflect realistic lexical knowledge of bilinguals, which might be better represented by measures of receptive vocabulary. LD measures elicited by narratives tap into children's ability to use their vocabulary knowledge. This provides the rationale for examining the relationship between receptive vocabulary and expressive vocabulary measured through LD. A receptive vocabulary test was also selected because it is the only type of vocabulary test standardised in both languages.

Although the Croatian PPVT (Dun & Dunn, Kovačević et al., 2010) has been standardised, a recent validation study showed slightly skewed results in a sample of 521 preschool children aged from 3;0 to 6;6, who were chosen to be representative of the national preschool population in terms of gender and geographical location (Kuvač Kraljević, Hržica, Padovan, Kologranić Belić, Olujić, Matić, Srebačić & Peretić, 2015). That sample showed an average standard PPVT score of 112.19, which contrasts with the score of the population used to validate the Croatian version of the PPVT. The ages of participants in the present study fell within the distribution of ages in the recent validation study (Kuvac Kraljevic et al., 2015) (Figure 1, Table 4). Given the skewed age distribution discovered in Kuvac Kraljević et al. (2015), we adjusted the standard PPVT scores of the children in the present study by subtracting 12 points.

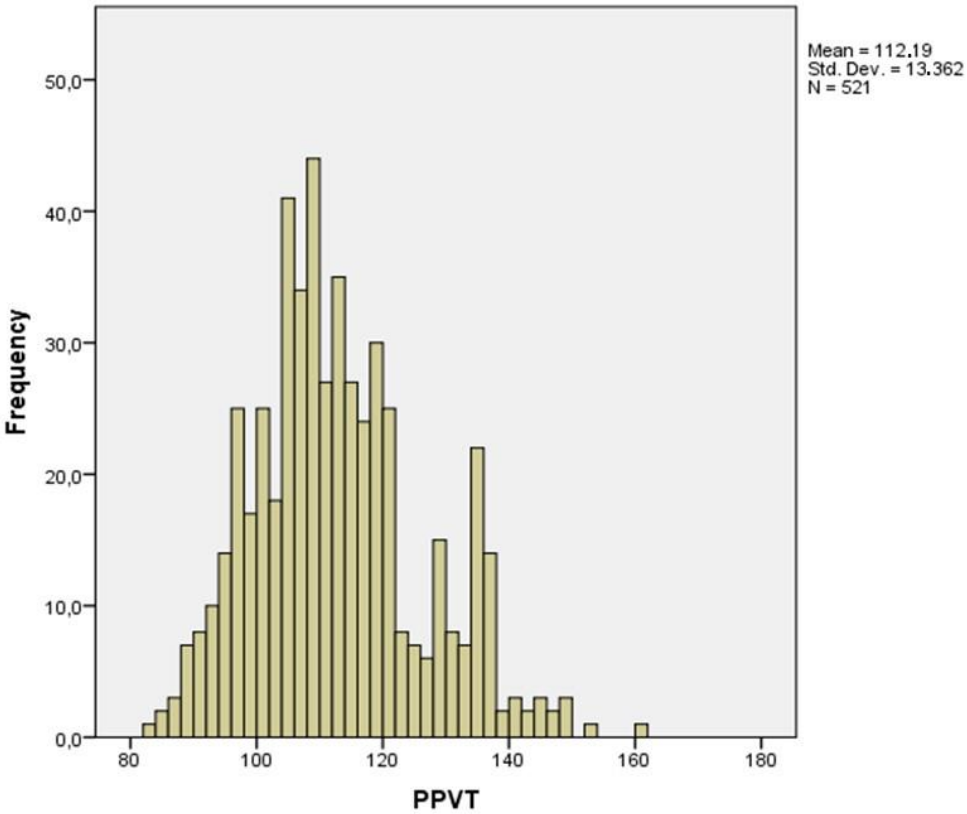


Figure 1. Distribution of mean scores on the Croatian PPVT in the validation study by Kuvač Kraljević et al. (2015) for children from 5;0 to 6;6. The age range of children in the present study was from 5;1 to 7;0.

*Table 4. Statistical analysis of scores on the Croatian PPVT in the validation study by Kuvač Kraljević et al. (2015), stratified by age group.**

Age group	N	min	max	mean	SD
5;0–5;5	85	91	148	115	14.59
5;6–5;11	58	86	140	112	14.07
6;0–6;6	78	83	160	114	15.14
All	222	83	160	114	14.67

*The age range of children in the present study was from 5;1 to 7;0.

4. Results

In the first section descriptive analyses are reported. For each LD measure, a comparison between bilinguals and the two groups of monolinguals is reported first, and then a comparison is made between the two languages based only on monolinguals.

4.1. PPVT

Bilingual children in our study scored significantly lower on the PPVT in both languages than their monolingual peers (Croatian, 83.7 vs. 101.1; Italian, 70.5 vs. 102.5), and these differences were significant for both languages [Croatian, t(58)=4.931, p <.001, d=1.29; Italian, t(58)=9.737, p <.001, d=2.56].

Table 5. Croatian and Italian PPVT mean standard scores, SD and ranges across the three groups of participants

Test		Bilinguals	Croatian monolinguals	Italian monolinguals
PPVT Croatian	mean	83.7 (16.3)	101.1 (11.9)	–
	range	48–115	75–126	–
PPVT Italian	mean	70.5 (14.6)	–	102.5 (10.5)
	range	60–109	–	80–116

When PPVT scores were compared across languages, bilinguals showed significantly higher scores in Croatian than Italian [$t(58)=6.324$, $p<.001$, $d=1.66$], while the two groups of monolinguals showed similar scores in their respective languages [$t(58)=.208$, $p=.836$, $d=.05$].

4.2. Quantitative LD measures

Bilinguals showed lower NDW and NTW than their monolingual peers (Table 6), and the differences between bi- and monolinguals were significant in Italian [NDW, $t(57)=5.217$, $p<.001$, $d=1.38$; NTW, $t(57)=4.440$, $p<.001$, d=1.18] but not in Croatian [NDW, $t(57)=1.962$, $p=.055$, $d=.52$; NTW, $t(57)=1.784$, $p=.08$, $d=.47$].

Table 6. Quantitative LD measures across the three groups of participants.

Measure	Bilinguals	Croatian monolinguals	Italian monolinguals
	mean (SD), range	mean (SD), range	mean (SD), range
NDW Croatian	33.6 (11.6), 15–66	40 (13.8), 26–69	–
NDW Italian	32.4 (13.4), 11–68	–	53 (16.7), 30–93
NTW Croatian	55.2 (22.6), 19–122	67 (28.1), 36–128	–

NTW Italian	57 (28.7), 13–144	–	103.3 (48.5), 40–208

Comparisons of NDW and NTW between the two languages showed no significant differences in bilingual speakers [NDW, *t*(56)=.359, *p*=.721; NTW, *t*(56)= –.264, *p*=.793]. In contrast, both measures were significantly higher for Italian than Croatian among monolinguals [NDW, *t*(58)=3.270, *p*=.002, *d*=.86; NTW, *t*(58)=3.539, *p*=.001, *d*=.93].

4.3. TTR and its transformations

Analysis of TTR showed similar results between bi- and monolingual speakers in Croatian [*t*(57)=.667, *p*=.507, *d*=.17] and in Italian [*t*(57)=1.403, *p*=.166, *d*=.37] (Table 7). Similar results were obtained using the Maas index, for which lower scores indicate greater diversity. Scores were similar between bi- and monolinguals in Croatian [*t*(57)=.660, *p*=.512, d=.16] and Italian [*t*(57)=1.649, *p*=.105, *d*=.37].

Table 7. TTR and its transformations across the three groups of participants

	Bilinguals	Croatian monolinguals	Italian monolinguals
Measure	mean (SD), range	mean (SD), range	mean (SD), range
TTR Croatian	0.63 (0.09), 0.4–0.8	0.61 (0.07), 0.5–0.8	–
TTR Italian	0.59 (0.09), 0.5–0.9	–	0.55 (0.1), 0.4–0.7
Maas Croatian	100.7 (25.4), 53.8–152.5	97.1 (16), 54.5–127.9	–

Maas Italian	113.2 (30.7), 69.6–195.1	–	101.55 (23.4), 51.8–148.1

Among bilinguals, TTR was similar between Croatian and Italian [*t*(56)=1.762, *p*=.084, *d*=.47], as was MAAS [*t*(56)=1.688, *p*=.097, *d*=.45].

4.4. Model-based indices

Measure D was similar between bi- and monolingual speakers in Croatian [*t*(33)=1.426, *p*=.163, *d*=.50], whereas it was significantly better for monolinguals than bilinguals in Italian [*t*(44)=2.840, *p*=.007, *d*=.86] (Table 8). Similar results were obtained for HD-D, which was similar between mono- and bilinguals in Croatian [*t*(57)=1.662, *p*=.102, *d*=.44], but significantly better for monolinguals in Italian [*t*(57)=2.924, *p*=.005, *d*=.77]. The same results were also obtained when comparing MATTR between mono- and bilinguals in Croatian [*t*(57)= 1.274, *p*=.208, *d*=.33] and Italian [*t*(57)= 4.298, *p*<.001, *d*=1.14].

Table 8. Model-based indices across the three groups of participants

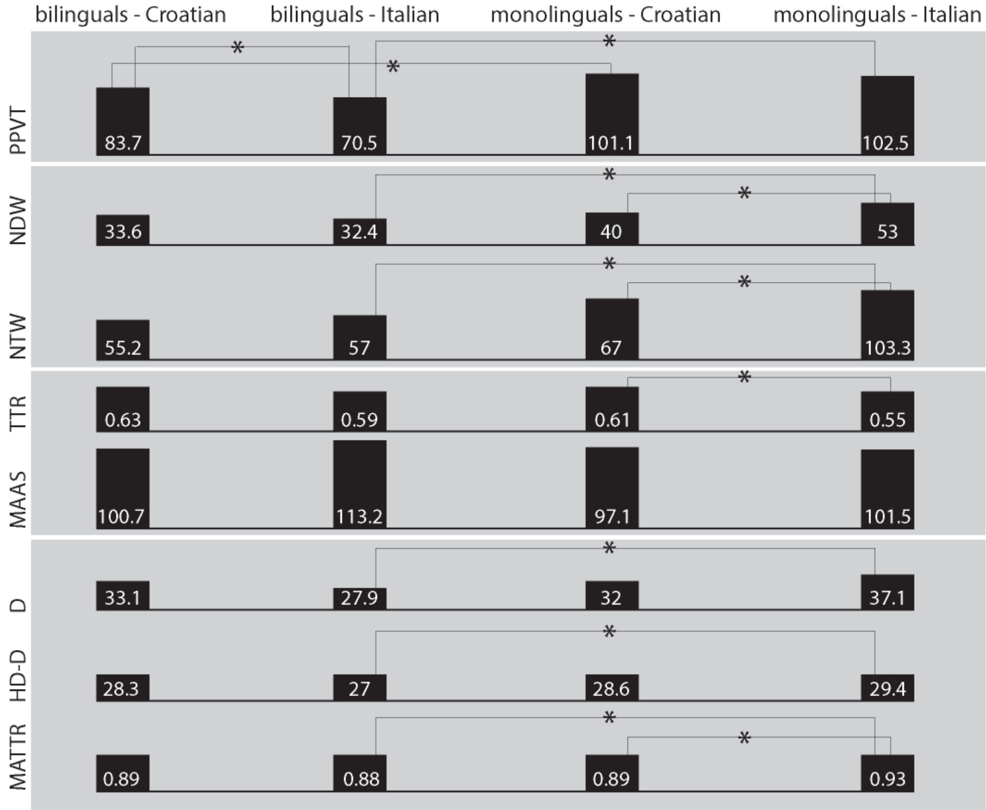
Measure	Bilinguals	Croatian monolinguals	Italian monolinguals
	mean (SD), range	mean (SD), range	mean (SD), range
D Croatian	33.06 (13.48), 13.59–62.57	32.04 (11.71), 16.32–64.71	–
D Italian	27.89 (7.18), 13.73–40.72	–	37.09 (12.45), 17.7–79.64
HD-D Croatian	28.32 (2.68), 23.56–33.25	28.56 (2.13), 23.56–32.56	–
HD-D Italian	27.05 (2.67), 19.69–30.88	–	29.38 (2.24), 24.44–34.55
MATTR Croatian	0.89 (0.04), 0.81–0.96	0.89 (0.03), 0.82–0.94	–
MATTR Italian	0.88 (0.05), 0.74–0.94	–	0.93 (0.02), 0.86–0.96

Cross-language comparison among bilinguals showed no significant differences in D [*t*(33)=1.426, *p*=.163, *d*=.49], HD-D [*t*(56)=1.540, *p*=.129, *d*=.41] or MATTR [*t*(56)=.265, *p*=.792, *d*=.07]. In contrast, the two monolingual groups differed significantly in MATTR

[t(58)=5.291, p<.001, d=1.38], but not in D [t(46)=1.419, p=.163, d=.43] or HD-D [t(58)=1.714, p=.092, d=.45].

Figure 2 summarises the differences in lexical measures across the three groups of participants.

Figure 2. Pairwise differences in lexical measures across the three participant groups



4.5. Correlations among measures

Possible correlations were explored among all lexical measures, within each of the three groups of participants (see Figure 3). Most LD measures showed medium or strong correlations, but not necessarily in all four sets of language samples. While significant, most of the correlations were associated with coefficients < .80. This provides evidence that the different LD measures indeed measure different features. NDW correlated positively with all model-based indices (D, Dr, HD-D and MATTR), but only the correlation with Dr was significant across all sets of language samples. NDW correlated with D and HD-D among bilinguals and monolinguals in Croatian, while NDW correlated with MATTR only among Croatian monolinguals. NTW correlated with Dr among bi- and monolinguals in both languages, while it correlated with D, HD-D and MATTR only for Italian samples from bilinguals.

TTR correlated strongly with D in Croatian samples from bilinguals and in samples of both languages from monolinguals. It showed either a tendency to correlate or a significant correlation with HD-D in all sets of language samples. This correlation was positive except for Italian samples from bilinguals, where it was negative. TTR showed either a tendency to correlate or a significant correlation with MATTR in all sets of language samples except Italian samples from bilinguals. TTR showed a correlation with Dr only in the case of Italian samples from bilinguals, where the correlation was negative.

The Maas index correlated strongly and negatively with D, HD-D and MATTR in all four sets of language samples. Maas index also correlated negatively with Dr, but only in Croatian samples from bilinguals. Lower Maas values indicate greater LD.

In all sets of language samples, MATTR correlated with the other model-based indices D, HD-D and Dr. The index HD-D correlated strongly with D and less strongly with Dr, while D itself correlated with Dr.

Figure 3. Correlations among LD measures

D	BI-CRO	BI-IT	MO-CRO	MO-IT
NDW	.607** p=.010	.818** p<.001	.499* p=.025	.130 p=.511
NTW	.123 p=.639	.757** p<.001	.271 p=.248	-.144 p=.466
MATTR	.845** p<.001	.760** p<.001	.604** p=.005	.470* p=.012
HD-D	.905** p<.001	.927** p<.001	.867** p=.005	.929** p<.001
Dr	.880** p<.001	.926** p<.001	.786** p<.001	.538** p=.003
TTR	.847** p<.001	.342 p=.165	.652** p=.002	.636** p<.001
MAAS	-.851** p<.001	-.700** p=.001	-.744** p<.001	-.717** p<.001

Dr	BI-CRO	BI-IT	MO-CRO	MO-IT
NDW	.941** p<.001	.955** p<.001	.905** p<.001	.873** p<.001
NTW	.670** p<.001	.922** p<.001	.822** p<.001	.682** p<.001
MATTR	.690** p<.001	.608** p<.001	.399* p=.029	.259 p=.167
HD-D	.789** p<.001	.561** p<.001	.582** p=.001	.529 p=.003
TTR	.091 p=.637	-.437* p=.018	-.254 p=.176	.323 p=.082
MAAS	-.499* p=.006	-.349 p=.063	-.109 p=.566	-.090 p=.637

HD-D	BI-CRO	BI-IT	MO-CRO	MO-IT
NDW	.677** p<.001	.610** p<.001	.452* p=.012	.134 p=.481
NTW	.365 p=.051	.574** p<.001	.278 p=.136	.130 p=.511
MATTR	.799** p<.001	.519** p<.001	.604** p<.001	.517** p=.003
TTR	.359 p=.056	-.563** p=.001	.468* p=.001	.578** p<.001
MAAS	-.847** p<.001	-729** p<.001	-.844** p<.001	-.765** p<.001

MATTR	BI-CRO	BI-IT	MO-CRO	MO-IT
NDW	.501** p<.001	.517** p=.004	.243 p=.196	.049 p=.796
NTW	.143 p=.459	.453** p=.014	.129 p=.495	-.082 p=.665
TTR	.511** p=.005	.019 p=.920	.375* p=.041	.361 p=.050
MAAS	-.752** p<.001	-.687** p<.001	-.488* p=.006	-.374* p=.041

4.6. LD measures as predictors of PPVT performance

Quantitative LD measures in this study point to the conclusion that the bilinguals are more proficient in Croatian than Italian. At the same time, the PPVT results indicate a significant difference between the bilinguals and each monolingual group. This raises the question of whether the measures are associated with the same language phenomena as the PPVT. In other words, to what extent can the quantitative measures predict performance on the PPVT, and can the same prediction be made across the three groups of participants? Regression analysis was used to examine whether and to what extent LD measures predicted

PPVT for each of the three participant groups (Table 9). Separate analyses were run for each LD measure because of the high collinearity among the measures.

NDW and D predicted the PPVT results for all three groups of speakers in both languages. In contrast, TTR and Maas index did not predict these results for any group or language. HD-D predicted PPVT results in both groups of monolingual speakers, whereas NTW and MATTR predicted PPVT results only in Croatian monolinguals. NTW predicted PPVT performance based on Italian samples from bilinguals. Overall, fewer indices predicted PPVT results in bilinguals than in monolinguals, and all predictive measures explained a greater proportion of variance in monolingual performance than bilingual performance.

Table 9. Regression analysis to identify predictors of PPVT scores in the three groups

			R^2	β
Bilinguals Croatian	Quantitative measures	NTW	.130	.360
		NDW	.165*	.406
	TTR and transformations	TTR	.002	–.048
		MAAS	.003	–.053
	Model-based measures	D	.173	.416
		HD-D	.036	.191
		MATTR	.062	.249
Bilinguals Italian	Quantitative measures	NTW	.145**	.381
		NDW	.181***	.426
	TTR and transformations	TTR	.044	–.210
		MAAS	.044	–.210
	Model-based measures	D	.050	.224
		HD-D	.081	.284
		MATTR		
Monolinguals Croatian	Quantitative measures	NTW	.210°	.458
		NDW	.290°°	.539
	TTR and transformations	TTR	.001	–.003
		MAAS	.070	–.265
	Model-based measures	D	.309°°°	.556
		HD-D	.242°°°°	.492
		MATTR	.183°°°°°	.428

	Quantitative measures	NTW	.086	.294
		NDW	.219^	.468
Monolinguals-Italian	TTR and transformations	TTR	.001	–.003
		MAAS	.070	–.256
	Model-based measures	D	.252^^	.502
		HD-D	.254^^^	.504
		MATTR	.074	.271

Regression analyses were performed on difference scores for each measure.
*$F(1,27)=5.337$, $p<.05$, ** $F(1,27)=4.574$, $p<.05$, *** $F(1,27)=5.986$, $p<.05$,
° $F(1,28)=7.430$, $p<.01$, °° $F(1,28)=11.449$, $p<.01$, °°° $F(1,28)=8.050$, $p<.01$, °°°° $F(1,28)=8.922$, $p<.01$, °°°°° $F(1,28)=6.276$, $p<.05$
^ $F(1,28)=7.846$, $p<.01$, ^^ $F(1,28)=8.776$, $p<.01$, $p<.001$, ^^^ $F(1,28)=9.552$, $p<.01$,

The finding that several lexical indices predicted PPVT results across the three participant groups confirms that the measures are interrelated, as previously suggested from the correlational analyses. Nevertheless, the measures capture different skills to some extent, given that standardised measures explained less variance than non-standardised measures, and they explained different proportions of variance among bi- or monolinguals.

4.7. Analyses of differences across languages in bilinguals

To complement our comparisons of bi- and monolinguals using direct comparisons of test norms, we also compared the two languages *within* the bilingual group. To standardise results and thereby avoid monolingual bias or the influence of language, we calculated Z-scores for the measures previously shown to be discriminatory or predictive of PPVT: NTW, NDW, D, HD-D and MATTR. Mean scores on all measures fell between 0 and 1. For all measures except MATTR, the proportion of bilinguals showing above-average performance in Italian was higher than the proportion showing below-average performance. It is important to keep in mind that these results apply only within the bilingual group. Subtracting the Z-score in one language from the Z-score in the other language provides a measure of differential language performance. In our study, Italian scores were subtracted from Croatian ones, so a negative difference indicates better performance in Italian, while a positive difference indicates above-average performance in Croatian (see Table 10).

Table 10. Z-scores on measures of LD and differences in Z-scores between languages in bilingual participants

Measure	Language	Mean (SD)	Mean difference (SD)
NTW	Croatian	.67 (.94)	–.07 (.83)
	Italian	.80 (.82)	
NDW			

	Croatian	.71 (.72)	.02 (.87)
	Italian	.79 (79)	
D	Croatian	.12 (1.04)	–.21(1.21)
	Italian	.42 (.82)	
HD-D	Croatian	.19 (.93)	–.16 (1.81)
	Italian	.42 (.66)	
MATTR	Croatian	.49 (1.07)	.13 (1.70)
	Italian	.34 (.71)	

Then PPVT scores for each language were compared within the group of bilinguals to assess dominance. For this analysis, Italian PPVT scores were subtracted from Croatian scores. Differences in NTW and NDW predicted the observed differences in PPVT results (Table 11).

Table 11. Regression analysis to identify predictors of differences in PPVT score between languages in bilingual participants

Type of measure	Measure	R^2	β
Quantitative	NTW	.228*	.447
measures	NDW	.274**	.524

Regression analyses were performed on difference scores for each measure
*F(1,27)= 7.954, p<.01, **F(1,27)= 10.212, p<.01

5. Discussion

The aim of the study was to analyse to which extent LD measures are appropriate for measuring lexical skills of sequential bilingual children speaking typologically different languages (Croatian and Italian). LD measures were calculated from narrative language samples and compared with PPVT results. Some of the LD measures revealed differences between participants already established by other assessment material. Furthermore, the findings indicated that LD measures (D, HD-D, MATTR) better predicted results of receptive vocabulary test than those developed without taking into consideration the problem of transcript size (NTW, NDW, TTR, MAAS), as previously suggested in the literature (reviewed in Fergadiotis et al., 2015). Finally, correlations among measures were moderate or high, suggesting that in short narrative production, LD measures might only partially capture the same construct. The results of the present work have theoretical implications for research on bilingual language development as well as practical implications for achieving desired educational outcomes.

Debate continues in the literature about which measures are more appropriate and reliable for assessing language in children who develop in a multilingual environment. In particular, the debate concerns measures of lexical knowledge since they are more influenced by exposure and linguistic dominance than other measures of linguistic knowledge. While it is well recognised that lexical knowledge is one of the main prerequisites for bilingual academic achievement, how to measure it remains controversial. This is mainly because evaluating the language abilities of bilingual children requires taking into account their distributed language knowledge. The present study employed several measures that have been proposed as more reliable for measuring bilingual vocabulary knowledge. The study population is particularly under-represented in the literature: bilingual children aged 5–7 who speak typologically different languages other than English, French or German. In addition, LD was assessed with ecological language procedures, i.e. narrative language sampling. These measures are thought to be particularly effective for assessing the ability to use vocabulary in real-life situations. The present work advances the debate on which ecological LD measures are reliable in showing differences between bi- and monolinguals, and which measures can quantify cross-language differences in bilinguals. This is particularly relevant when other parallel measures are not available in all the languages spoken by the child. The Croatian-Italian bilinguals in this study live in an environment where they can use both languages in different contexts of their everyday life. Since Croatian is an official language in Croatia, it is expected that they will show good command of Italian, but overall better achievements in Croatian. These bilinguals were compared with age-matched Croatian and Italian monolinguals.

The PPVT test in our study, which has been standardised for monolingual Italian and Croatian populations, revealed, as expected, no significant difference between the two monolingual groups. In contrast, the test uncovered a significant difference between bilinguals and both monolingual groups: bilinguals scored lower on the receptive vocabulary test than their monolingual peers, as reported in previous studies (reviewed in Bialystok, 2009), even when scores on vocabulary tests are calculated in different ways (de Villiers, 2015).

Quantitative and model-based LD measures turned out to be reliable measures of lexical diversity, which captures children’s ability to use their lexical knowledge. In particular, bilingual speakers scored lower on quantitative measures of Italian language than their Italian monolingual peers, but they scored similarly on quantitative measures of Croatian as the corresponding monolingual group. The same pattern emerged with model-based measures. On all measures except TTR and the Maas index, bilinguals performed lower than Italian monolinguals but similarly as Croatian monolinguals. TTR and Maas index were similar among the three groups of participants. The results suggest that the quantitative measures and model-based measures applied in our study adequately reflect the abilities of bilingual speakers.

Different patterns from those with inter-group comparisons emerged when we compared results between languages within the group of bilinguals. Bilingual speakers scored lower on the PPVT than both monolingual groups, and also lower on many LD measures than the Italian, but not Croatian, monolingual group. In contrast to the PPVT results, neither quantitative nor model-based LD measures showed differences between the two languages within the bilingual group. These results indicate that LD indices do not measure the same phenomena as the PPVT test. The test measures receptive vocabulary, while LD measures are used to assess usage of vocabulary. In addition, standardised measures of receptive vocabulary might underestimate bilingual speakers' vocabulary: most have been standardised using a normative sample of monolingual speakers, since bilingual norms are unavailable. Many standardised tests for certain languages were not developed in parallel, nor have they been validated for comparisons across languages (e.g. see Gutiérrez-Clellen & Simon-Cereijido, 2007 for tests of English and Spanish). One good example is provided in recent work by Lindgren (2018) on Swedish, Swedish-German and Swedish-Turkish children 4–6 years old, where vocabulary in narrative production (MAIN) was compared with vocabulary production scores on the Cross-linguistic Lexical Task (Haman et al. 2015). We relied on the PPVT for lack of a parallel task that measures vocabulary production in Italian and Croatian. By relying on ecological assessment of language samples rather than standardised tests, the present study answers the first research question posed at the outset: LD measures are reliable and appropriate measures of lexical diversity for languages other than English. More importantly, these measures detect subtle differences between bi- and monolingual performance as well as similar performance between languages in bilingual speakers.

The second research question was whether and to what extent LD measures are related to each other and to receptive vocabulary measured by PPVT. This research question was addressed by running correlations among all lexical measures, within each of the three groups of participants and through regression analyses with lexical measures predicting PPVT. Our results show that some LD measures can predict PPVT results, and therefore that LD indeed reflects other aspects of lexical knowledge.

Regression analysis between LD measures and PPVT scores was able to explain 40% of the variance in PPVT scores of monolinguals. This is a substantial achievement given individual differences in vocabulary knowledge. As far as we know, the present work is the strongest demonstration of this relationship so far. Our observations that fewer measures (NTW, NDW) significantly predicted PPVT scores among bilinguals and that these measures explained only 18% of variance may reflect greater individual differences in lexical knowledge among bilinguals. Of all the LD measures employed in this study, the ones that significantly predicted PPVT are also those shown to differ more between bi- and monolinguals: NTW, NDW, D, HD-D and MATTR. NDW reliably predicted PPVT scores of bilinguals in both languages. This result may be highly relevant to researchers, clinicians and educators since it identifies the most reliable measures of vocabulary knowledge in bilinguals that also reflect general lexical knowledge.

Measures that predicted receptive vocabulary scores, and the amount of variance they explained, differed between the mono- and bilinguals in our study. This may indicate that standardised tests such as PPVT and ecological assessment based on narrative samples are not equally effective in bi- and monolinguals for assessing vocabulary comprehension. Our results suggest that ecological assessment based on language usage is more effective for bilinguals, consistent with the literature and with Grosjean's claim (1992) that language competencies are structured and organised differently in bilinguals than in monolinguals. Consistent with this idea, the same measures (NTW, NDW) were able to predict PPVT performance for each language within the bilingual group in our study. Our observation of correlations between Dr and quantitative measures leads us to conclude that only quantitative measures predict receptive vocabulary scores in bilingual speakers.

Potential overlap between quantitative measures and model-based LD measures was assessed by calculating correlations across all three participant groups and within each group separately. The two types of indices showed moderate or strong correlations across and within groups. Exceptionally strong correlations, with coefficients greater than .9, were identified between D and HD-D and between Dr and NDW. The first correlation corroborates the claim that D is equivalent to HD-D (McCharty & Jarvis, 2007, deBoer, 2014). The second correlation shows the influence of short language samples on the calculation of Dr. Even though some authors opt to randomise with replacements since they believe it represents the subject's vocabulary stream more faithfully, most researchers believe that random sampling without replacements is more appropriate in the case of fixed language samples (Malvern et al., 2004). The VOCD programme in CLAN performs analyses of subsamples of 35–50 tokens in steps of one token. Since many narrative samples in the present study contained around 50 tokens or fewer, it was not possible to follow the rule that the final number of tokens in a subsample cannot be as large as the number in the entire transcript (MacWhinney, 2000). As a result, TTR was calculated for the entire set of language samples. Our results suggest that when the VOCD programme in CLAN is executed with replacements to analyse short language samples, other LD measures, namely NTW and NDW, should be used. In this way, the data answer the second research question posed at the outset: LD measures are connected and are indeed related to receptive vocabulary. Taken together, our findings support the legitimacy and importance of assessing bilingual language knowledge using ecological measures, alone or as a complement to standardised tests. Our results identify practical, flexible and reliable measures for assessing the quality of lexical production. At the same time, our study has important limitations and it leaves open questions that need to be addressed in the future.

One such open question is how to obtain reliable comparisons of two languages of a bilingual speaker without having to test two control groups of monolingual speakers. Future work is needed to isolate factors that contribute to differences in performance in different languages. More work is also needed to assess the ability of different methodologies to compare performance in different languages, including sample truncation, sample lemmatisation and standardisation using Z-scores. A second open question concerns the cross-linguistic comparison. There are obvious limitations for adopting PPVT as a comparison measure. There should be an effort to build and adapt cross-linguistic tasks to more languages, in order to have reliable measures of lexical comprehension and production. As mentioned earlier, the Cross-linguistic Lexical Task (Haman, et al., 2015) is a good example in that direction, and future work should consider the correlation between this Task and LD measures. Another open question is the validity of the measures in the present study for the relatively short texts generated by our participants. The number of tokens in the texts ranged from 19 to 208, with most texts having 50–150; overall, the samples were short and variable in length. Most of the LD measures in this study were developed and validated for longer texts. More recent studies have examined the reliability of these measures for shorter texts, but no clear conclusions have emerged. For example, Koizumi & In’nami (2012) examined the dependence of Maas index, TTR, D and HD-D on sample length. They found TTR to be most sensitive to length for shorter texts, and they found D, HD-D and Maas index to be more sensitive to length than previously reported in studies involving longer texts (McCarthy & Jarvis, 2010). These differences may reflect, in part, the fact that Koizumi & In’nami examined spoken narratives, while McCarthy & Jarvis examined written texts: McCarthy & Jarvis (2007) reported that assessment of spoken narratives is more likely to be sensitive to length. These issues should be examined systematically in future work.

Another open question is why Italian and Croatian monolinguals differed in quantitative measures (NTW and NDW) and TTR, although they showed similar PPVT. The Italian group showed higher NTW and NDW but lower TTR. There are two possible explanations for this. One is based on typological differences between the two languages: the articles in Italian may contribute to higher NTW, while TTR should be higher for morphologically richer Croatian. The second explanation considers the sample-size effect: since Italian monolingual files are longer, lower TTR should be expected. In support of the second explanation, MATTR was the only other measure that differentiated between the two monolingual groups. Stills (2016) showed that of the measures they tested, MATTR depended the most strongly on sample size (it increased with sample size). Of course, both explanations might be interconnected: samples are longer because of language features such as the existence of articles. We failed to observe differences between the two groups of monolinguals on the PPVT test or on other LD measures that are less sensitive to sample-size effects. This supports the idea that the two groups showed similar receptive and productive lexical skills. Future work is needed to isolate factors that contribute to differences in performance in different languages.

6. Conclusion

The current study has shown that LD measures are very useful tools in analyses of children’s vocabulary in storytelling, indicating differences between monolinguals and bilinguals similar to those indicated by standardised tests of vocabulary. LD measures in this study largely accounted for variance in general vocabulary knowledge. Therefore, when for any reason it is impossible or inappropriate to measure a child's vocabulary through standardised tests, LD may be reliable for estimating vocabulary knowledge.

There is an ongoing discussion in the literature about the role of LD assessment as more than just a measure of lexical development or lexical skills, but as a proxy for general language ability. Several studies report correlations of LD measures and overall language proficiency. Crossley et al. (2013) reported that the Maas index was the second strongest predictor of overall language proficiency in essays written by L2 learners. Follow-up study (Crossley et al., 2014) reported correlations with coefficients from 0.8 to 0.9 between holistic ratings of essays and LD measures. While these studies were conducted on written work by L2 learners, they reveal an interesting pattern that may also be relevant for other target groups and other types of assessments. If further studies corroborate the link between LD and overall language abilities, this will have significant consequences on language assessment, especially of languages for which assessment materials are lacking.

References

Acarlar, F. & Johnston, J. (2010). Turkish SALT: Computer-assisted language sample analysis. In S. Topbaş & M. Yavaş (Eds.), *Communication disorders in Turkish* (pp. 119–139). Bristol: Multilingual Matters.

Bedore L. M., Pena, E. D., Gillamb, R. B. & Hoa, T. H. (2010). Language sample measures and language ability in Spanish English bilingual kindergarteners. *Journal of Communication Disorders,* 43(6), 498–510. doi: 10.1016/j.jcomdis.2010.05.002

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition,* 12(1), 3–11. doi:10.1017/S1366728908003477

Bisazza, A. & Monz, C. (2014). Class-based language modelling for translating into morphologically rich languages. In J. Tsujii & J. Hajic (Eds.) *Proceedings of COLING 2014: 25th International Conference on Computational Linguistics* (1918–1927). Dublin: Dublin City University and Association for Computational Linguistics.

Bishop D. (2003). *Test for reception of grammar, Version 2*. London: The Psychological Corporation.

Bishop, D., Kuvač Kraljević, J., Hržica, G., Kovačević, M. & Kologranić Belić, L. 2013. *TROG-HR Test razumijevanja gramatike*. Zagreb: Naklada Slap.

Bishop, D., Suraniti, S., Ferri, R. & Neri, V. (2009). *Test for reception of grammar, version 2 TROG-2, Edizioni Italy*. Florence: Giunti OS.

Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Covington, M. A. (2007). *CASPR research report 2007–05. MATTR user manual*. Athens: The University of Georgia.

Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio MATTR. *Journal of Quantitative Linguistics,* 17(2), 94–100. doi: 10.1080/09296171003643098

Crossley, S. A., Salsbury, T., Macnamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics,* 5(1), 570–590. doi: 10.1093/applin/amt056

Crossley, S. A., Salsbury, T., Macnamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. H. Daller (Eds.), *Human ratings and automated measures* (pp. 105–134). Amsterdam: John Benjamins.

Cvrček, V. & Chlumská, L. (2015). Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics*, 39(3), 309–325. doi: 10.1007/s11185-015-9151-8

D'Achille, P. & Thornton, A. M. (2003). La flessione del nome dall'italiano antico all'italiano contemporaneo. In N. Maraschio & T. Poggi Salani (Eds.), *Italia linguistica anno Mille – Italia linguistica anno Duemila. Atti del XXXIV congresso internazionale di studi della Società di Linguistica Italiana* (pp. 211–230). Roma: Bulzoni.

Daller, H., van Hout, R. & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics,* 24(2), 197–222. doi: 10.1093/applin/24.2.197

Daller, M. H. & Treffers-Daller, J. (2014). Moving between languages: Turkish returnees from Germany. In B. Menzel & C. Engel (Eds.), *Rückkehr in die Fremde? Ethnische Remigration russlanddeutscher Spätaussiedler* (Ost-West-Express. Kultur und Übersetzung 21, pp. 185-212). Berlin: Frank & Timme.

Daller, M., Yildiz, C., De Jong, N., Kan, S. & Basbagi, R. (2011). Language dominance in Turkish German bilinguals: Methodological aspects of measurements in structurally different languages. *The International Journal of Bilingualism,* 15(2), 215–236. doi: 10.1177/1367006910381197

Damico, J. S., Oller, J. W. & Storey, M. E. (1983). The diagnosis of language disorders in bilingual children: Pragmatic and surface-oriented criteria. *Journal of Speech and Hearing Disorders* 48(4), 385–394.

de Villiers, J. (2015). Taking account of both languages in the assessment of dual language learners. *Seminars in Speech and Language,* 36(2), 120–132. doi: 10.1055/s-0035-1549107

deBoer, F. (2014). Evaluating the comparability of two measures of lexical diversity. *System,* 47, 139–145. doi: 10.1016/j.system.2014.10.008

Dewaele, J. M. & Pavlenko, A. (2003). Productivity and lexical diversity in native and non-native speech. A study of cross-cultural effects. In V. Cook (Ed.), *Effects of the second language on the first* (120–141). Clevedon: Multilingual Matters.
Dressler, W. (2004). Degrees of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics,* 15(1), 31–62.
Dressler, W. U. (2007). Introduction to: Early development of nominal and verbal morphology from a typological perspective. In S. Laaha & S. Gillis (Eds.), *Typological perspectives on the acquisition of noun and verb morphology* (Antwerp Papers in Linguistics, vol. 112, pp. 3–9). Antwerpen: University of Antwerpen.
Dunn L. M. & Dunn L. M. (1997). *Peabody Picture Vocabulary Test-PPVT – 3rd Edition*. Minneapolis: American Guidance Service.
Dunn, L. M. & Dunn, L. M., Kovačević, M., Padovan, N., Hržica, G., Kuvač Kraljević, J., Mustapić, M., Dobravac, G. & Palmović, M. (2010). *PPVT-III-HR Peabody slikovni test rječnika*. Zagreb: Naklada Slap.
Dunn, L. M. & Dunn, L. M., Stella, G., Pizzoli, C. & Tressoldi, P. (2000). *PPVT-revised*. Torino: Omega Edition.
Fergadiotis, G., Wright, H. & Green, S. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research* 58(3), 840-852. doi: 10.1044/2015_JSLHR-L-14-0280
Fiestas, C. E. & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. *Language, Speech, and Hearing Services in Schools,* 35(2), 155–168. doi: 10.1044/0161-1461(2004/016)
Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Valimaa, T. Balčiūnienė, I. Bohnacker U. & Walters, J. (2012). *MAIN: Multilingual assessment instrument for narratives* (ZAS Papers in linguistics, vol. 56). Berlin: ZAS.
Gagarina, Natalia (2016). Narratives of Russian-German preschool and primary school bilinguals: Rasskaz and Erzaehlung. *Applied Psycholinguistics,* 37(1), 91–122. doi: 10.1017/S0142716415000430
Gibson, T., Peña, E.D. & Bedore, L.M. (2014). The receptive-expressive gap in bilingual children with and without primary language impairment. *American Journal of Speech-Language Pathology*. 23(4), 655–667. doi: 10.1044/2014_AJSLP-12-0119
Grosjean, F. (1992). Another view of bilingualism. In R. Harris (Ed.), *Cognitive processing in bilinguals*. Amsterdam: North-Holland.
Guiraud, P. (1954). *Les caractéristiques statistiques du vocabulaire*. Paris: Presses. Universitaires de France.
Guiraud, P. (1960). *Problemes et methodes de la statistique linguistique*. Paris: Presses Universitaires de France.
Gutiérrez-Clellen V. F. (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education,* 13(2), 175–197. doi: 10.1016/S0898-5898(01)00061-4
Gutiérrez-Clellen, V. F. & Simón-Cereijido, G. (2007). The discriminant accuracy of a grammatical measure with Latino English-speaking children. *Journal of Speech, Language, and Hearing Research,* 50(4), 968–981. doi: 10.1044/1092-4388(2007/068)
Haman, E., Łuniewska, M. & Pomiechowska, B. (2015). Designing cross-linguistic lexical tasks (CLTs) for bilingual preschool children. In S. Armon-Lotem, J. de Jong & N. Meir (Eds.) *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 196–240). Bristol: Multilingual Matters.
Hamers, J. & Blanc M. (2000). *Bilinguality and Bilingualism* (2nd ed.). Cambridge: Cambridge University Press.
Hržica, G. & Kuvač Kraljević, J. (2012). MAIN Croatian adaptation. In N. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Valimaa, I. Balčiūnienė, U. Bohnacker & J. Walters, J., *MAIN: Multilingual assessment instrument for narratives* (ZAS Papers in linguistics, vol 56). Berlin: ZAS.
Iacobini, C. (2006). Italian as fusional language. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd Edition, vol 6, pp. 64–69). Oxford: Elsevier.

Jarvis, S. (2002). Short texts, best fitting curves, and new measures of lexical diversity. *Language Testing*, 19(1), 57–84. doi: 10.1191/0265532202lt220oa
Jelaska, Z. (2005). Glagolske vrste ‘Conjugational classes’. In Z. Jelaska (Ed.), *Hrvatski kao drugi i strani jezik* ‘Croatian as second and foreign language’ (pp. 153–170). Zagreb: Hrvatska sveučilišna naklada.
Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund University Department of Linguistics and Phonetics Working Papers*, 53, 61–79.
Kann, K. & Schütze, H. (2017). Unlabelled data for morphological generation with character-based sequence-to-sequence models. In M. Faruqui, H. Schuetze, I. Trancoso & Y. Yaghoobzadeh (Eds.), *Proceedings of the 1st Workshop on Subword and Character Level Models in NLP* (pp. 76–81). Copenhagen: Association for Computational Linguistics.
Koizumi, R. & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System,* 40(4), 522–532. doi: 10.1016/j.system.2012.10.012
Kunnari, S., Välimaa, T. & Laukkanen-Nevala, P. (2016). Macrostructure in the narratives of monolingual Finnish and bilingual Finnish-Swedish children. *Applied Psycholinguistics,* 37(1), 123–144.
Kuvač Kraljević, J., Hržica, G., Padovan, N., Kologranić Belić, L., Olujić, M., Matić, A., Srebačić, I. & Peretić, M. (2015). *Probirni test jezičnih sposobnosti e-POTJEH* 'Language Screening Test'. Zagreb: ERF.
Laufer, B. (2003). The influence of L2 on L1 collocational knowledge and on L1 lexical diversity in free written expression. In V. Cook (Ed.) *Effects of the second language on the first* (pp. 19-31). Clevedon: Multilingual Matters.
Lindgren, J. (2018). *Developing narrative competence: Swedish, Swedish-German and Swedish-Turkish children aged 4–6* (Studia Linguistica Upsaliensia, vol. 19). Uppsala: Uppsala University.
Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Lange eines Textes. *Zeitschrift fur Literaturwissenschaft und Linguistik,* 8, 73–79.
MacWhinney, B. (2000). *The CHILDES Project. Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
Malvern, D. D. & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language*. *Papers from the Annual Meeting of the British Association of Applied Linguists* (pp. 58–71). Clevedon: Multilingual Matters.
Malvern, D., Richards, B., Chipere, N. & Durán, P. (2004). *Lexical diversity and language development*. New York: Palgrave Macmillan.
McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity MTLD* (unpublished doctoral dissertation). University of Memphis.
McCarthy, P. M. & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing,* 24(4), 459–488. doi: 10.1177/0265532207080767
McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods,* 42(2), 381–392.
McCarthy, P. M., Watanabe, S. & Lamkin, T. A. (2012). The Gramulator: A tool to identify differential linguistic features of correlative text types. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 311-332). Hershey, PA: IGI Global.
McKee, G., Malvern, D. & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing,* 15(3), 323–337. doi: 10.1093/llc/15.3.323
Reuterskiöld Wagner, C., Sahlén, B. & Nettelbladt, U. (1999). What’s the story? Narration and comprehension in Swedish preschool children with language impairment. *Child Language Teaching and Therapy,* 15(2), 83–93. doi: 10.1177/026565909901500202
Roch, M. & Levorato, C. 2012. MAIN Italian adaptation. In N. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Valimaa, I. Balčiūnienė, U. Bohnacker & J. Walters, J., *MAIN: Multilingual assessment instrument for narratives* (ZAS Papers in linguistics, vol 56). Berlin: ZAS.

Serratrice, L. (2007). Cross-linguistic influence in the interpretation of anaphoric and cataphoric pronouns in English–Italian bilingual children. *Bilingualism: Language and Cognition,* 10(3), 225–238. doi:10.1017/S1366728907003045

Simon-Cereijido, G. & Gutiérrez-Clellen, V. (2009). A cross-linguistic and bilingual evaluation of the interdependence between lexical and grammatical domains. *Applied Psycholinguistics,* 30(2), 315–337. doi: 10.1017/S0142716409090134

Siskova, Z. 2012. Lexical richness in EFL students' narratives. In C. Ciarlo & D. S. Giannoni (Eds.) *University of Reading language studies working papers* (Vol. 4, pp 26–36), Reading: University of Reading.

Stills, M. (2016). *Language sample length effects on various lexical diversity measures: An analysis of Spanish language samples from children* (undergraduate honors thesis, Paper 233). Portland State University: University Honors Theses. doi: 10.15760/honors.250

Templin M. C. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.

Treffers-Daller, J. & Korybski, T. (2016). Using lexical diversity measures to operationalise language dominance in bilinguals. In C. Silva-Corvalan & J. Treffers-Daller, J. (Eds.) *Language dominance in bilinguals: Issues of measurement and operationalization* (pp. 106–123). Cambridge: Cambridge University Press.

Treffers-Daller, J., Parslow, P. & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, 39(3), 302–327.

Tsimpli, I. M., Peristeri, E. & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics*. 37(1), 195–216.

Tweedie, F. J. & Baayen, R. H. (1998). How variable may a constant be? Measures in lexical richness in perspective. *Computers and the Humanities*, 32(5), 323–352.

Uccelli, P. & Paez, M. M. (2007). Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among English and Spanish skills. *Language, Speech, and Hearing Services in Schools*, 38(3), 225–236. doi: 10.1044/0161-1461(2007/024)

Van der Lely, H. K. J. (1996). Grammatical specific language impaired children: Evidence for modularity. In C. Koster & F. Wijnen (Eds.), *The Groningen assembly on language acquisition* (pp. 283–292). Groningen: University of Groningen Press.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., Gagarina, N., Hržica, G., Ketrez, N. F., Kilani-Schoch, M., Korecky-Kröll, K., Kovačević, M., Laalo, K., Palmović, M., Pfeiler, B., Voeikova, M. D. & Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language,* 31(4), 461–479.

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary-management profile. *Language,* 67(4), 763–778.