Profiling complex word usage in the speech of English preschool children: frequency patterns and transparency characteristics

Article

Accepted Version

Profiling complex word usage in the speech of English preschool children: frequency patterns and transparency characteristics

Abstract

This corpus-based study provides a baseline of complex word usage patterns in the spontaneous speech of English preschool children to ascertain the characteristics of their derivative vocabulary before literacy development affects language skills. Frequencies of suffixed derivatives produced by (N=243) children aged 2-5 and their caregivers were extracted for 58 suffix variants, yielding 558 types from the former and 1,364 from the latter. Between the youngest and oldest groups, 11 suffix categories increased significantly in type frequency, compared with 22 in the caregiver dataset. All derivative types were classified for transparency of meaning and simplicity of form on a 5-point analysability scale. Around 59% of both the child and caregiver derivative vocabulary sets were classified as transparent regardless of age, suggesting that the potential analysability of the preschool child's input remains surprisingly invariant over time. The study provides baseline data for future studies on the development of morphological awareness in English-speaking schoolchildren.

Keywords

derivational morphology, corpus linguistics, preschool children, spontaneous speech, transparency

Profiling complex word usage in pre-schoolers

Introduction

Complex words are formed through the attachment of derivational affixes to a base, e.g., help-*ful*. The first derivatives that very young children come across are treated as whole lexical items (Clark, 2014); it is only after repeated exposure to a base such as *help* and other complex words such as *hopeful* and *wonderful*, in addition to an appreciation of the form-meaning relationship inherent in the suffix *-ful*, that by the age of two the child is in a position to recognise word constituents and identify combinatorial processes of this type as a word-formation device (Derwing & Baker, 1986).

The occurrence of established derivatives such as *helpful* in spontaneous child speech does not indicate whether these forms have been analysed by the child. A large corpus of child speech that spans a number of years can, however, tell us a great deal about the kinds of linguistic features that make up the child's output at different stages of development (Demuth, 2008). A substantial body of research has provided ample evidence for the supportive function of the child-directed speech register and verbal interaction in early lexical development (Hoff & Naigles, 2002) and the role of type frequency in the linguistic input in the acquisition of morphological forms (Ambridge, Kidd, Rowland, & Theakston, 2015; Clark, 1993). Therefore, a corpus of child-caregiver speech can reveal the range of affixes that are integral to the complex words used in interactions and their relative frequencies, and provide an age-related profile of derivative usage of child output and caregiver input. One of the questions the current research raises is to what extent does the former profile reflect the latter? To this end, the study reported here focuses exclusively on usage patterns of suffixes occurring in English-speaking child-caregiver interactions; follow-on work could usefully extend the approach to other language types, but such an endeavour was outside the scope of the current project which acts as a starting point for future cross-linguistic comparisons.

The development of morphological awareness in English-speaking children has been explored through a variety of elicitation tasks and the recording of spontaneous new coinages (Berko, 1958; Clark & Cohen, 1984; Clark & Hecht, 1982). That body of research has focused on children ranging in age from kindergarten to secondary school and, although the acquisition of derivatives occurs predominantly during the school years, children as young as 2 have been observed to create new lexemes using, for example, the agentive suffix *-er* (Clark, 1981), thus demonstrating that they have acquired that morpheme and the associated word-formation rule. Clark (1993, 2014) proposes that in addition to the frequency of the affix-meaning pairing in the child's linguistic environment, the factors which facilitate the development of morphological awareness include transparency in meaning of both the base and affix, and simplicity of form in the construction of the complex word. Another question that the current research raises is to what extent are the derivatives in caregiver input and child output transparent in meaning and simple in form?

The aim of this research was threefold. Firstly, to identify the profiles of derivatives in the spontaneous speech of monolingual English children with normally-developing language, from their earliest utterances at age two, to the age of five when reading ability starts to affect language skills. The corresponding derivative profiles of caregivers in the same interactions were also identified. The derivatives in question contained the most commonly used suffix categories in the child data: 58 suffix variants were analysed.

The second aim was to chart the rate of expansion of suffix categories in child output over age groups, compared with the corresponding profiles of caregiver input, and evaluate these usage patterns against an independent baseline derived from a more generic English-speaking sample of everyday adult speech, the British National Corpus (BNC).

The third aim was to examine the degree to which derivatives in child and caregiver speech display transparency in meaning and simplicity of form, as defined by Clark (1993), and whether age-related derivative profiles change over the preschool years with respect to these characteristics.

The output of this naturalistic study is not intended as a yardstick for assessing the child's ability to analyse complex words, nor does it propose a measure of morphological awareness; rather, it provides age-related profiles of derivatives produced spontaneously by preschool children that can be compared with concurrent input from caregiver speech and an independent source of adult speech from the wider community. In addition, the evaluation of transparency in meaning and simplicity of form of child-caregiver derivative profiles provides an index of the analysability of complex word datasets that relate to the preschool years.

Characteristics of derivational morphemes and the nature of productivity

The distinction between inflectional and derivational morphology has been extensively documented in the morphological literature (see Booij, 2006 for a comprehensive review). In broad terms, inflectional morphology expresses grammatical relations between a word and its context, e.g., plurality or past tense, whereas derivational morphemes are used to construct new words from a stem by the application of prefixes and/or suffixes, e.g., *un*-shock-*able*. These affixes change the meaning of the base and, primarily in the case of suffixes, may also change its word class. In English, inflectional morphology is characterised by 8 inflections and involves suffixation exclusively, whereas derivational suffixes are considerably more numerous: Stein (2007) lists 164 distinct forms.

Derivational affixes fall into two classes (Siegel, 1974): Class I (non-neutral) tend to be derived from Latinate forms, e.g., *-ity*, *-ic*, *-ion* and *-ous* and, due to their etymology, mostly attach to bound stems, a process which often results in a sound and/or stress shift in the base, as in the transition from *impress* to *impression* and *átom* to *atómic*. Class II affixes (neutral) are generally derived from Germanic origins, e.g., *-ful*, *-less*, *-ness* and *-ly*, and tend to attach to free bases without incurring any sound or stress change to the base, as in *seriousness* and *wonderful*. However, membership to one affix class or the other is not completely predictable. For example, the attachment of non-neutral *-ous* produces two sound changes in the formation of *vicious* from *vice*, a stress change in the generation of *ridiculous* from *ridicule* and no changes when it is attached to *humour* to produce *humorous*. Therefore, each affix category, e.g., *-ic* or *-ous*, must be evaluated in relation to each derivative it produces, since such phonological and stress-related changes are key to the notions of transparency in meaning and simplicity of form.

The term 'productivity' refers to different phenomena depending on the nature of the linguistic approach (see Baayen, 2009 for a review of frequency measures and types of productivity). Bauer (1983, p.18) defines the productivity of a morphological category as the degree to which "it can be used synchronically in the production of new forms". Such a characteristic is relevant when evaluating children's level of morphological awareness, either through their spontaneous production of new words, or through their ability to coin new lexemes in elicitation tasks (Clark, 1993).

The number of derivative types bearing a particular affix (affix category size) is dependent on a number of factors including etymological and linguistic constraints, such as phonological, syntactic and semantic features (Bauer, Lieber, & Plag, 2013). Affixes vary greatly in their type frequency; for example, agentive *-er*, has produced several hundred

derivatives, whereas *-ern* is predominantly restricted to the four compass directions. Baayen (2009) refers to affix category size as 'realised productivity', i.e., type frequency reflects the relative success with which the affix has produced new derivatives over time. The larger the category size, the more productive the morpheme is considered to have been. However, contextual factors constrain the extent to which the realised productivity of an affix is expressed in any sample. For example, the more formal the register, the larger the category size of non-neutral derivatives; the more informal the context, the greater the predominance of neutral derivatives (Laws & Ryder, 2018; Schmid, 2011). Therefore, affix category size is dependent on the context in which it is measured.

Since the current study is concerned with the range of suffixes used by children over the preschool years in relation to corresponding usage patterns occurring in caregiver input and an independent sample of adult speech, the measure used to perform these comparisons was affix type frequency, or category size, as a function of age and speaker source.

Derivational morphology and child language development

Research on grammatical development in children (Brown, 1973; de Villiers & de Villiers, 1973) has successfully documented the order of acquisition of inflectional morphology. Furthermore, given the rule-based nature of inflectional morphology, its acquisition is amenable to assessment through the productive application of the limited rule set to novel word stems, as demonstrated by the wug test (Berko, 1958). By contrast, far less is known about the order of acquisition of derivational morphology in the preschool years and this endeavour is considerably more challenging given the large number of affix categories in English and the wider range of meanings that derivational affixes represent (Booij, 2006).

The productive use of derivational processes appears later in language development than compounding, owing to its greater complexity (Berko, 1958; Derwing & Baker, 1986). There is, however, evidence that neutral suffixes are employed creatively at early stages of language development, owing to their greater transparency. The principal suffixes include: diminutive *–ie*/*y*, adjectival *-y* and agentive *-er* at age 2, instrumental *-er* by age 2-3, *-ness* at age 3 and *-ist*, *-ment* and verb-forming *-en* at age 4; whereas innovations involving potentially opaque non-neutral suffixes, such as *–ity*, -(*a*)*tion* and *-ical*, only appear after initial reading skills have been acquired (Clark, 1993).

It is during the early school years that the mutually reinforcing relationship between vocabulary knowledge and morphological awareness occurs (Anglin, 1993; Berninger, Abbott, Nagy, & Carlisle, 2010; Freyd & Baron, 1982; McBride-Chang, Wagner, Muse, Chow, & Shu, 2005). Studies on school children between 5 and 11 years have shown that increased morphological awareness can enhance spelling performance (Carlisle, 1988; Deacon & Bryant, 2005), reading comprehension (Cunningham & Carroll, 2015) and the efficacy of general language instruction (Moats & Smith, 1992). Indeed, one of the very few studies on pre-schoolers (Lyster, Lervåg, & Hulme, 2016) demonstrates that training children in morphological awareness 8 months before starting school can produce benefits in literacy skills that are still observable when those children reach Grade 6. Therefore, it would be useful to identify the profile of derivational usage patterns of preschool children before they learn to read and write, in order to maximise opportunities for building effectively on their existing knowledge of complex words.

Inflectional processes are almost exclusively semantically regular, whereas derivational processes frequently express a range of polysemous interpretations (Booij, 2006). For example, *-er* may denote an agent (*writer*), the most frequent interpretation, but also an

instrument (*printer*), experiencer (*hearer*), stimulus (*pleaser*), patient (*looker*), denominal noun (*Londoner*), measure (*fiver*) and location (*diner*) (Lieber, 2004). Thus, not only is the child faced with the need to acquire a vast array of many-to-many form-function combinations in the same syntactic context, a prerequisite for the extraction of word-formation rules involves protracted exposure to a wealth of complex words in order that analogies can be drawn.

As mentioned in the Introduction, Clark (1993) suggests that the child's ability to analyse complex words is enhanced by input frequency, transparency of meaning and simplicity of form. Children are sensitive to the type frequency of morphological forms in the linguistic input (Ambridge et al. 2015) and acquire more productive forms earlier, as demonstrated by Clark and Cohen (1984) who compared memory performance of 4-5 year-old children on novel words constructed with the semantically related suffixes *-er*, *-ist* and *-ian*. Recall measures were significantly greater for items containing the most productive suffix *-er*. Thus, affixes belonging to large category sizes should be acquired earlier than those from small category sets.

Meaning transparency refers to the ability to identify the word stem and affix easily: *hopeful* is transparent because it has a free base and is formed with a neutral suffix, whereas *station* is not because it contains a bound stem and is formed with a non-neutral suffix. Children should therefore acquire more transparent, neutral categories earlier than more opaque, non-neutral categories, as has been observed in studies involving children of school age (Gordon, 1989; Moats & Smith, 1992; Tyler & Nagy, 1989).

Simplicity of form refers to the case where the attachment of an affix does not invoke morpho-phonological changes to the base: again, *hopeful* is simple in form as the base *hope* undergoes no sound or stress change when neutral *-ful* is attached; the base *impress*, however, undergoes a final consonant sound change when non-neutral *-ion* is added (*impression*); similarly, *féstive* undergoes a stress shift when non-neutral *-ity* is added to form *festívity*. Therefore, children should acquire neutral affixes earlier than non-neutral ones, in fact Clark (1993) reports that most Latinate vocabulary is acquired from the age of 7. Furthermore, it is worth noting that there are important educational implications relating to children's knowledge of non-neutral morphemes: Latinate forms feature predominantly in academic register (Schleppegrell, 2004), both in the classroom and in school texts, and early exposure to these low-frequency word types has been shown to facilitate academic success (Dickinson & Tabors, 2001).

To the author's knowledge, researchers have not to date examined the degree of transparency of meaning and simplicity of form inherent in the derivative repertoire of language users in various contexts; this study, therefore, contributes a methodology for devising such measures in order to evaluate these characteristics in relation to age-related vocabulary sets of preschool children and their caregivers. If, as Clark (1993) suggests, transparency of meaning and simplicity of form support the child's development of morphological awareness, to what extent are these facilitating characteristics actually present in the derivatives they use and in the input they receive from caregivers at different ages?

The current study

The motivation for the current study was to use a corpus-based approach to examine variation in derivational profiles of children and caregivers in the same interactions over the preschool years, to compare those profiles against an independent corpus of adult speech, and to evaluate the extent to which the age-related complex words used by these children and

caregivers are transparent in meaning and simple in form with respect to morphological category size. Given the predominance in English of suffixation over prefixation in child language (Slobin, 1973) and functional and semantic distinctions between these morphological phenomena (Clark, 2014; Lieber, 2005), the former was selected as the focus of this study which posed the following research questions (RQs):

1. What is the nature of suffix usage patterns in child-caregiver speech between the ages of 2 and 5?

a) Which suffixes increase in use most markedly in the child data across the age groups?

b) What changes occur in caregiver suffix usage across the age groups?

With respect to 1a, it was predicted that neutral suffix categories would expand most rapidly across the age range. Furthermore, derivatives bearing non-neutral suffixes would appear more predominantly in the older age groups. Regarding 1b, it was expected that non-neutral suffix categories would be observed in caregiver speech at all ages but that related type frequencies would increase across age groups.

2. To what extent do type frequency patterns in child speech reflect those in caregiver speech in the same interactions and an independent corpus of adult speech?

Following expectations 1a and b, it was predicted that child profiles would contain a greater proportion of neutral to non-neutral suffixes compared with the caregiver data, and that caregiver suffix profiles would closely reflect those extracted from the BNC, with the exception of diminutives as they are associated with child-directed speech.

3. What are the characteristics of derivatives occurring in the child-caregiver datasets with respect to transparency of meaning and simplicity of form?

It was predicted that measures of transparency of meaning and simplicity of form of derivatives in the age-related caregiver datasets would be lower than those in the corresponding child sets on the basis that adults would be more likely to use opaque forms more frequently. In addition, it was predicted that the new forms entering the children's vocabulary sets would include a larger number of opaque forms as age increased, thus showing greater conformity to caregiver speech patterns.

Method

Data sources: The age-based child-caregiver sub-corpora and the BNC

The age groups were: 2;0-2;6, 3;0-3;6, 4;0-4;6 and 5;0-5;6. In determining an adequate corpus size for each age group, reference was made to Raban (1988) and Swan (2000). The former compiled a "spoken vocabulary of five-year old children" based on a corpus of 69,166 words (Raban, 1988, p.12) from 96 children aged 5;0 to 5;6. To investigate the spontaneous production of derivational innovations and inflectional over-regularisations, Swan compiled a corpus of almost 110,000 tokens from the speech of one child from the age of 2;6.5 to 5;11.13. On the basis of these two studies, it was deemed that a corpus size of between 70,000 and 110,000 tokens would provide a representative sample, so a target of 100,000 for all tokens in each age group was set.

The four age-related corpora were compiled from the CHILDES database (MacWhinney, 2000). CHAT files coded with the %mor tier were downloaded until the target of 100,000 tokens was reached for each group. Care was taken to ensure that total token counts within each group were balanced for gender. Since it was not possible to reach the target token count

of 100,000 tokens from UK transcripts alone, it was necessary to construct the dataset with an equal number of tokens from UK and US transcripts; data file sources are reported in Appendix A. The number of children contributing to the study was 243, but the number included in each age by country subgroup varies; this was an inevitable consequence of the fact that studies include different numbers of children and transcriptions vary in length. The goal here was to ensure that the number of tokens in each age by gender by country subgroup was as close as possible to 50,000.

Raw CHAT files contain 'unusable' types, e.g., punctuation characters ('!', '.', '?') and uninterpretable, or unknown (unk) types ('www', 'xxx', 'yyy', etc.). These were removed to obtain sub-corpora containing only 'usable' types. The final gender and age group contributions are presented in Table 1.

Table 1. Child tokens by age group

Age Range	Gender	Mean Age	Tokens	Gender %
2;0–2;6	**Boys**	2;03.08	50,107	50.08
	Girls	2;02.30	49,953	49.92
			100,060	
3;0–3;6	**Boys**	3;02.17	50,034	50.00
	Girls	3;02.10	50,026	50.00
			100,060	
4;0–4;6	**Boys**	4;02.01	49,988	49.96
	Girls	4;01.27	50,070	50.04
			100,058	
5;0–5;6	**Boys**	5;00.21	50,131	50.09
	Girls	5;00.25	49,952	49.91
			100,083	
Total			**400,261**	

Caregiver speech from the interactions that contributed to the child data in Table 1 was extracted from the same transcriptions. Token counts necessarily varied across groups because this corpus was based on the fixed set of child speech files in each age band. Table 2 presents the number of usable caregiver tokens and their distribution across age and gender.

Table 2. Caregiver tokens by age group

Age Range	Gender	Mean Age	Tokens	Gender %
2;0–2;6	**Boys**	2;03.08	146,292	53.17
	Girls	2;02.30	128,858	46.83
			275,150	
3;0–3;6	**Boys**	3;02.17	87,905	40.79
	Girls	3;02.10	127,608	59.21
			215,513	
4;0–4;6	**Boys**	4;02.01	165,363	52.57
	Girls	4;01.27	149,207	47.43
			314,570	
5;0–5;6	**Boys**	5;00.21	90,602	50.37
	Girls	5;00.25	89,265	49.63
			179,867	
Total			**985,100**	

The total corpus contained 1,385,361 tokens, from which the contributions from children and adults were 28.89% and 71.11%, respectively.

The British National Corpus (BNC) was compiled in 1994 and thus constitutes a representative baseline of language usage contemporaneous with the child-caregiver data from CHILDES. The demographically-sampled (DS) sub-corpus of the BNC was used as the

independent sample of adult everyday spoken English; it contains 4,233,938 tokens from 1,405 native speakers of British English balanced for age, gender and socio-economic class. The DS therefore provides a means of estimating the potential size of suffix categories used in everyday speech, thus indicating the number of types the child will eventually employ. Laws and Ryder (2014) compiled a database of derivative type and token frequencies from the spoken BNC. The DS-related category sizes of the 58 suffix variants used here were extracted from this database.

Suffixes selected for analysis

Laws and Ryder (2014) identified a comprehensive set of 141 prefixes and 131 suffixes in English as part of a project on derivative usage patterns and register variation. A subset of the 41 most frequently occurring suffixes were selected for the current study. Suffixes had been drawn from Stein (2007), a dictionary that sets out the primary function and meaning of each affix variant according to the word class of derivatives created by the attachment of that affix. For example, *-ful* generates both adjectives (*hopeful*) and nouns (*handful*); Stein distinguishes these polysemous variants by means of superscripts: *-ful*^{1} and *-ful*^{2}, respectively. This convention also applies to homonymous variants of orthographically identical suffixes. For example, the two homophonous but distinct forms of *-ly* (Bauer et al. 2013) create adjectives (*homely*) and adverbs (*quickly*), which are coded *-ly*^{1} and *-ly*^{2}, respectively. Phonological allomorphs, such as *-ation*, *-ion* and *-ition* (Bauer et al. 2013) constitute separate entries in Stein's dictionary and were thus treated as three unique suffixes in the current study. Within each word-class based entry, Stein then lists finer-grained senses conveyed by each variant. Stein's classification system formed the basis for coding the 41 target suffixes, yielding 58 variants. These are organised alphabetically in Table 3 according to the word class of the derivatives they produce.

A few minor adjustments were applied to Stein's convention. When type numbers were small, counts for Adjective/Noun-forming suffixes were combined: *-an*$^{1/2}$, *-ic*$^{1/2}$, *-ese*$^{1/2}$ and *-ist*$^{1/2}$. Stein includes the orthographic allomorphs *-able* and *-ible* in one entry (Bauer et al., 2013), but given the phonological difference between these suffixes it was deemed appropriate to profile them separately.

In the context of the current study, which set out to analyse suffix usage patterns in child speech, the following four suffixes were identified as requiring a more differentiated coding system than that provided by Stein: *-ish*, *-er*, *-or* and *-ie/y*. Stein (2007) lists four senses of the polysemous suffix *-ish*: having characteristics of the base (*foolish*), belonging to the base nationality (*Spanish*), an approximation to the base (*fortyish*) and tending to carry out the verb base (*peckish*). Only the first two of these occurred in the child dataset and since their senses were considered sufficiently diverse to warrant profiling independently, they were labelled *-ish*^{1} and *-ish*^{2}, respectively.

The derivational suffix *-er* is assigned one entry in Stein (2007) with four senses that correspond to the four polysemous variants listed in Table 3 as *-er*$^{1\text{-}4}$. These include: *-er*^{1} (animate and agentive), *-er*^{2} (animate and non-agentive), *-er*^{3} (inanimate and agentive) and *-er*^{4} (inanimate and non-agentive). Entities denoted by *-er*^{2} and *-er*^{4} are not agents or instruments but may be experiencers, patient/themes, or may have other non-agentive interpretations (Lieber, 2004, p.17). Similarly, Stein's single entry for *-or* lists two senses which correspond to the polysemous variants *-or*$^{1\text{-}2}$ in Table 3: *-or*^{1} (animate) and *-or*^{2} (inanimate). Where the animacy of the item was ambiguous (*cleaner* and *monitor*), the context was checked in the transcripts.

Finally, three variants of the diminutive *-ie/y* suffix were identified that occur frequently in child speech. These relate to different conditions of the base: free base, *-ie/y*[1] (*horsey*), unknown origin of the base according to the *Oxford English Dictionary*, *-ie/y*[2] (*bunny*) and truncated base *-ie/y*[3] (*hankie*). Only the first of these is listed in Stein (2007).

Table 3. Target suffix variants

	Adjective-forming	Noun-forming	Adjectives/Nouns	Verb-forming	Adverb-forming
	-able (*adaptable*)	*-age* (*storage*)	*-al*[1] (*critical*)	*-ate*[3] (*appreciate*)	*-ly*[2] (*quickly*)
	-ate[1] (*fortunate*)	*-ance* (*entrance*)	*-al*[2] (*hospital*)	*-en*[1] (*tighten*)	
	-en[2] (*wooden*)	*-ancy* (*pregnancy*)	*-an*[1/2] (*Italian*)	*-ify* (*pacify*)	
	-ful[1] (*painful*)	*-ate*[2] (*candidate*)	*-ant*[1] (*pleasant*)	*-ize* (*recognize*)	
	-ible (*sensible*)	*-ation* (*combination*)	*-ant*[2] (*consultant*)		
	-ish[1] (*foolish*)	*-cy* (*privacy*)	*-ary*[1] (*necessary*)		
	-ish[2] (*Spanish*)	*-ence* (*difference*)	*-ary*[2] (*secretary*)		
	-ive (*disruptive*)	*-ency* (*tendency*)	*-ent*[1] (*confident*)		
	-less (*useless*)	*-er*[1] (*farmer*)	*-ent*[2] (*resident*)		
	-ly[1] (*homely*)	*-er*[2] (*officer*)	*-ese*[1/2] (*Siamese*)		
	-ous (*nervous*)	*-er*[3] (*sharpener*)	*-ian*[1] (*Canadian*)		
	-y (*fussy*)	*-er*[4] (*rubber*)	*-ian*[2] (*magician*)		
		-ery/ry (*crockery*)	*-ic*[1/2] (*plastic*)		
		-ess (*lioness*)	*-ist*[1/2] (*elitist*)		
		-et (*packet*)			
		-ette (*rosette*)			
		-ful[2] (*spoonful*)			
		-ie/y[1] (*horsey*)			
		-ie/y[2] (*bunny*)			
		-ie/y[3] (*hankie*)			
		-ion (*occasion*)			
		-ition (*competition*)			
		-ity (*reality*)			
		-ment (*movement*)			
		-ness (*sweetness*)			
		-or[1] (*decorator*)			
		-or[2] (*radiator*)			
Totals	**12**	**27**	**14**	**4**	**1**

The examples in Table 3 illustrate that the derivatives selected for this study included forms with both free and bound bases. Since the objective was to identify a representative set of derivatives occurring in the spontaneous speech of preschool children and their caregivers, the inclusion of both transparent and opaque forms was essential.

Coding of speech data by suffix category

CLAN was used to extract the speech data from CHAT files. The %mor tier was used to compute word frequency. To calculate the frequency of target morphemes, all word types in the combined child-caregiver corpus of 1,385,361 tokens were checked and derivatives were extracted manually by source and age group. The etymology of each potential complex word was checked in the *Oxford English Dictionary*. A small proportion of the resultant derivatives contained a multimorphemic stem (*activity, carefully*): 1.08% in the child and 2.93% in the caregiver datasets. These forms were included in the analysis and classified according to the outer-most suffix (*-ity*, *-ly*); this decision was motivated by the fact that derivational suffixes determine the word class of the complex word, and that the focus of current study was suffix usage rather than stem composition patterns. The final derivative set was checked for variant spellings (*horsey / horsie*) and misspellings (*Dalmation*); all erroneous items were rectified.

To ensure that derivatives were drawn from the child's vocabulary and had not resulted from the child repeating an item that had been produced by the caregiver, the context of each complex word was checked using 'grep' (Global regular expression print). By searching the context, it as possible to ascertain whether the item was produced spontaneously by the child, rather than in response to hearing the word previously in that exchange. Only 'first-use' instances of complex words contributed as new items for any age group and 'mirrored' items were discarded.

Type and token frequencies of the 58 suffix variants were calculated for the child and caregiver corpora. Equivalent data for the independent sample of adult speech was extracted from the DS-related database (Laws & Ryder, 2014). Type frequencies for each suffix variant provided the vocabulary size measures for the three data sources.

Coding for transparency and simplicity

To identify the transparency of meaning and simplicity of form characteristics of derivatives occurring in the child-caregiver datasets, a Transparency-Simplicity measure was devised based on Clark's (1993, p.115-122) definitions of these constructs.

Transparency refers to the availability of meaning of the components of a complex word, i.e., the base and affixes. Simplicity of form refers to the absence of sound and/or stress changes to a base as the result of suffix attachment. The Transparency-Simplicity measure involved a 0-1 scale for Sound/Stress features and 0-3 for Base characteristics, as shown in Table 4.

Table 4. Sound+Stress+Base scoring system

Score	Sound	Stress	Base
0	- phonological change (*sharp-en*)	- stress change (*length-en*)	Free base (*pack-age*)
1	+ phonological change (*exhaust-ion*)	+ stress change (*major-ity*)	Truncated base (*apolog-ize*) or lengthened base (*talk-ative*)
2			Base+affix meaning no longer interpretable (*poor-ly*)
3			Base meaning no longer accessible (*stat-ion*)

The Sound+Stress+Base measure for each derivative was calculated by adding the three scores. One category of derivative did not conform easily to the scoring system: types with an opaque base such as *station* could not be scored for Sound or Stress, since this information is no longer available without etymological knowledge; therefore, a score of '1' was assigned to each in addition to the score of '3' already assigned to the base for opaqueness, thus producing an overall score of '5'. An inter-rater reliability check was conducted on all the derivatives analysed: the author and colleague independently assigned the Sound+Stress+Base scores to each of the combined list of child and adult derivative types. Initial agreement was 99%; all discrepant assignments occurred in the Intermediate band and were readily resolved through discussion to produce 100% agreement.

Table 5. Examples of Sound+Stress+Base Scores

Examples	Sound change	Stress change	Base change/access	Total score: S+S+B	Transparency Classification
active, gardener	0	0	0	**0**	Transparent
exclusive, studious	1	0	0	**1**	
excellent, photographer	0	1	0	**1**	
marriage, stimulate	0	0	1	**1**	
resident, stability	1	1	0	**2**	
teddy, tummy	1	0	1	**2**	
allergic, historic	0	1	1	**2**	Intermediate
recital, organize	0	0	2	**2**	
Canadian, optician	1	1	1	**3**	
cavity, fiddly	1	0	2	**3**	
restaurant, university	0	1	2	**3**	
confident, position	1	1	2	**4**	
station, decent	1	1	3	**5**	Opaque

Table 5 provides some examples of Sound+Stress+Base (S+S+B) scores grouped as Transparent (0), Intermediate (1-4) and Opaque (5).

Results and discussion

Table 6 presents the age-based type and token counts of the 58 suffix variants from the child-caregiver datasets and equivalent totals from the DS.

From the total of 400,261 tokens, children produced 558 derivative types and 6,809 tokens (1.70%). The baseline 2;0-2;6 group produced 41% and 32% respectively of all the types and tokens observed. The relatively uniform increase of roughly 20% in the accumulation of additional derivative types year-on-year between the ages of 3;0 and 5;6 indicates a constant expansion in derivative vocabulary size. This suggests that potential literacy skills of the children in the 5;0-5;6 age group do not affect the rate at which their repertoire of derivatives expands, compared with children in the 3:0-3;6 and 4;0-4;6 age groups.

Table 6. Derivative type and token counts by age group

	2;0-2;6	3;0-3;6	4;0-4;6	5;0-5;6	Totals	DS
Child types	228 (41%)	111 (20%)	115 (21%)	103 (18%)	**558**	
Child tokens	2,193 (32%)	1,366 (20%)	1,620 (24%)	1,630 (24%)	**6,809**	
Caregiver types	645 (47%)	334 (24%)	306 (22%)	79 (6%)	**1,364**	5,483
Caregiver tokens	5,312 (27%)	4,452 (22%)	6,648 (33%)	3,459 (17%)	**19,871**	102,507

Caregivers produced 1,364 derivative types from a corpus of 985,100 tokens; 2.02% of the total token count is composed of the target derivative forms. The equivalent proportion of derivatives in the DS was 2.42%, reflecting, unsurprisingly, greater diversity in adult-directed everyday speech. As mentioned in the Method section, the relative proportion of child-to-caregiver speech necessarily varied across age groups, because the number of child tokens was constant (Table 2). Therefore, caregiver speech is slightly over-represented in the 4;0-4;6 compared with the 2;0-2;6 and 3;0-3;6 bands, but markedly under-represented in the 5;0-5;6 group.

The baseline for caregiver types in the youngest group is 645; this represents 47% of the total number of types observed across all the age groups (1,364), which is larger than that reported for the child baseline (41%), indicating, as expected, a higher degree of diversity in

caregiver speech. Token contributions of caregivers in this group, however, are lower (27%) than in the child dataset (32%), indicating less repetition of derivatives in the former sub-corpus. Type counts increase by around 23% between age groups, with the obvious exception of the difference between the 4;0-4;6 and 5;0-5;6 datasets.

In the two following sections, profiles of suffix usage as a function of age are identified by examining the relative distribution of suffix categories within the child, caregiver and DS datasets. The χ^2 test was used to evaluate differences in suffix category sizes between age groups. Since in all cases N=2 and df=1, these values have not been reported for each result. For the χ^2 test to provide reliable results, values compared must be 5 or above; therefore, numbers in *italics* in Tables 8 and 9 indicate where values are below 5; the results have nevertheless been provided for completeness. The following statistical conventions are used: *p<0.05; **p<0.025; ***p<0.01; ****p<0.001; *****p<0.0001. Effect sizes are indicated as: L=large (φ=>0.5), M=medium (φ=>0.3); S=small (φ=>0.1).

Child data: suffix profiles across age groups

Thirty-eight of the 58 suffix categories occurred in the 2;0-2;6 child dataset, but across all age groups 55 were observed. Table 7 presents the age of first occurrence of the 17 categories that did not appear in the younger age groups and the 3 that failed to appear at all.

Table 7. Order of appearance for suffix categories across age groups - child data

Suffix	Example	2;0-2;6	3;0-3;6	4;0-4;6	5;0-5;6	Caregiver Rank	DS Rank
-able	*comfortable*	-	✓	✓	✓	22	9
-ary[1]	*ordinary*	-	✓	✓	✓	41	40
-en[2]	*wooden*	-	✓	✓	✓	55	57
-ful[2]	*handful*	-	✓	✓	✓	48	48
-ish[1]	*greenish*	-	✓	✓	✓	40	18
-ish[2]	*Spanish*	-	✓	✓	✓	45	47
-ity	*activity*	-	✓	✓	✓	14	11
-ary[2]	*library*	-	-	✓	✓	44	39
-ate[3]	*motivate*	-	-	✓	✓	9	8
-ese	*Chinese*	-	-	✓	✓	50	51
-ian	*magician*	-	-	✓	✓	43	44
-ive	*expensive*	-	-	✓	✓	17	15
-ize	*hypnotize*	-	-	✓	✓	35	23
-ate[2]	*certificate*	-	-	-	✓	46	45
-ency	*emergency*	-	-	-	✓	52	49
-ify	*pacify*	-	-	-	✓	51	42
-ition	*competition*	-	-	-	✓	49	50
-ancy	*redundancy*	-	-	-	-	54	54
-cy	*privacy*	-	-	-	-	57	55
-less	*careless*	-	-	-	-	32	27

Key: - Absence of suffix; ✓ Occurrence of the suffix

The greater the frequency and range of suffix category members available in the linguistic environment, the more likely that suffix category will be reflected in child speech. Therefore, to ascertain the input frequency of the late/absent suffix categories in Table 7, the rank order of categories from 1 to 58 (1 is the highest) in the caregiver and DS data have been included; e.g., the category size of *-able* ranks 22nd in the caregiver dataset but 9th in the DS, indicating a marked under-representation of this suffix in the former compared with the latter dataset.

Unsurprisingly, most suffix categories appearing only in the older groups, or not at all, ranked lower than 30 in caregiver speech, reflecting relatively smaller category sizes for these

suffixes. Based on high rankings in the DS, the suffixes that would be expected to occur, but do not appear in the younger datasets, are: *-able*, *-ish*[1], *-ity*, *-ate*[3], *-ive*, *-ize* and *-less*. As predicted in relation to RQ1a, few abstract Latinate suffixes occurred in early child speech, but the absence of the relatively transparent categories *-ish*[1] and *-less* is unexpected, perhaps indicating that concepts of 'approximation' and 'negative quality' are less useful in communication at that early stage.

Table 8 presents the eleven suffix categories that expanded significantly between 2;0 and 5;6. Type frequencies from the 2;0-2;6 suffix baselines are compared with corresponding cumulative frequencies from the four age groups (2;0-5;6). Appendix B presents the full set of derived forms used by the children by age group. Table 8 includes the type rank (1-58) and χ^2 statistics for suffix categories where significant growth was observed between the youngest (2;0-2;6) and the cumulative (2;0-5;6) datasets. The magnitude of χ^2 indicates the increase in suffix category size across groups, therefore, results are presented in descending order of χ^2 to highlight those suffix categories that underwent the greatest expansion. For comparative purposes, the ranking for the caregiver and DS data are also included.

Table 8. Suffix categories that increased significantly in the child data

		2;0-2;6		2;0-5;6					Diff in	Cgiver	DS
Suffix	**Example**	**Rank**	**Types**	**Rank**	**Types**	χ^2	φ	***p***	**rank**	**Rank**	**Rank**
-y	*noisy*	**1**	47	**1**	110	25.28	M	*****	**0**	1	2
***-er*[1]**	*shopper*	**4**	14	**4**	43	14.75	L	****	**0**	4	3
***-ly*[2]**	*quickly*	**6.5**	9	**5**	33	13.71	L	****	**+1.5**	2	1
***-ie/y*[1]**	*horsie/y*	**2**	33	**2**	64	9.91	M	***	**0**	6	21
***-er*[3]**	*buzzer*	**3**	29	**3**	52	6.53	M	**	**0**	5	6
-ous	*nervous*	**16.5**	*3*	**10**	13	6.25	L	**	**+6.5**	11	10
-ion	*vacation*	**9**	6	**7**	18	6.00	L	**	**+2**	3	4
***-er*[4]**	*folder*	**8**	8	**6**	20	5.14	M	**	**+2**	13	22
***-or*[1]**	*visitor*	**32**	*1*	**18**	7	4.50	L	(*)	**+14**	34	20
-ity	*activity*	**47.5**	*0*	**26.5**	*4*	4.00	L	(*)	**+21**	14	11
-ive	*massive*	**47.5**	*0*	**26.5**	*4*	4.00	L	(*)	**+21**	17	15

Key: *p<0.05; **p<0.025; ***p<0.01; ****p<0.001; *****p<0.0001. Effect sizes: L=large (φ=>0.5), M=medium (φ=>0.3); S=small (φ=>0.1)

The most statistically robust category size increases occurred with the following four neutral suffixes: the adjective-forming *-y*, animate / agentive nominaliser *-er*[1], adverb-forming *-ly*[2] and the nominal diminutive *-ie/y*[1]. The top three of these are among the largest in the child's linguistic environment, as indicated by the rank ordering of caregiver and DS datasets; this is an expected consequence of child speech reflecting input frequency. As predicted, owing to the nature of the child-caregiver interaction, the diminutive form *-ie/y*[1] ranks considerably higher in the caregiver dataset than it does in the DS.

Two suffixes that increase in type count to a lesser degree are the inanimate nominalisers *-er*[3], and *-er*[4]. The cumulative category size (52) of *-er*[3] (*buzzer*) exceeds that (43) of the animate agentive form *-er*[1] (*farmer*), indicating that, in this sample, the highly frequent but polysemous suffix *-er* is used here more widely when it denotes inanimate than animate entities. Furthermore, inanimate non-agentive *-er*[4] (*folder*) also increases significantly, indicating that the children's growing vocabulary accommodates non-central functions of this suffix as well as more predictable meanings, and that the ranking of this category (6) is relatively higher than it is for the caregiver (13) and DS (22) datasets.

The two non-neutral suffixes *-ous* and *-ion* also display a reliable increase in category size, which, together with the remaining Latinate suffixes listed (agentive *-or*[1], *-ity* and *-ive*),

indicate a marked increase in the less transparent suffix categories, although the increase in these last three is less reliable (*) owing to low type values. Latinate derivatives tend to represent abstract concepts (*infinity*, *negative*) and increase more markedly in the school years (Schleppegrell, 2004), but the increase in rank of +21 for *-ity* and *-ive* is noteworthy for spontaneous speech in preschool children. By contrast, as predicted, all Latinate suffixes produce large category sizes in both the caregiver and DS datasets, particularly *-ion*.

Only two adjacent age comparisons just reached significance; both occurred between 2;0-2;6 and 3;0-3;6 and involved *-er*[1] (from 14 to 27 types), where χ^2=4.19, p<0.05, φ=M, and *-ly*[2] (from 9 to 22 types), where χ^2=5.45, p<0.025, φ=M. These findings indicate that growth in the use of the animate agentive nominaliser and the most common adverb-forming suffix is most pronounced in the second youngest age group examined here.

Caregiver data: suffix profiles across age groups

All 58 suffix variants occurred in the caregiver dataset for all age groups, except for noun-forming *-cy* (*privacy*), which appeared for the first time in the 4;0-4;6 sample.

Lexical choice reflects greater complexity as adults engage in conversation with older children; therefore, as predicted in relation to RQ1b, the proportion of suffix categories that demonstrate a significant increase was greater for the caregiver than child dataset when the baseline (2;0-2;6) and cumulative (2;0-5;6) datasets are compared. Table 9 presents the 22 suffixes where diversity in caregiver speech increased significantly.

Table 9. Suffix categories that increased significantly in the caregiver data

		2;0-2;6		2;0-5;6					Diff in	DS
Suffix	**Example**	**Rank**	**Types**	**Rank**	**Types**	χ^2	φ	p	**rank**	**Rank**
-y	*noisy*	**1**	109	**1**	207	30.39	M	*****	**0**	2
-ion	*vacation*	**6**	30	**3**	89	29.25	L	*****	**+3**	4
***-er*[1]**	*shopper*	**5**	34	**4**	83	20.52	M	*****	**+1**	3
***-al*[1]**	*normal*	**8**	19	**7**	52	15.34	M	*****	**+1**	5
***-ly*[2]**	*quickly*	**2**	72	**2**	127	15.20	S	*****	**0**	1
-ity	*activity*	**30**	5	**14**	23	11.57	L	****	**+16**	11
***-ate*[3]**	*operate*	**10**	15	**9**	40	11.36	M	****	**+1**	8
***-an*[1/2]**	*Italian*	**40.5**	*3*	**20.5**	17	9.80	L	***	**+20**	14
-ment	*equipment*	**10**	15	**10**	37	9.31	M	***	**0**	12
-ous	*nervous*	**14.5**	12	**11**	32	9.09	M	***	**+3.5**	10
***-ie/y*[1]**	*horsie/y*	**4**	35	**6**	64	8.49	S	***	**-2**	21
***-ic*[1/2]**	*plastic*	**7**	22	**8**	46	8.47	M	***	**-1**	7
***-al*[2]**	*recital*	**18.5**	9	**12.5**	24	6.82	M	***	**+6**	16
-ist	*therapist*	**35**	*4*	**24.5**	15	6.37	L	(**)	**+10.5**	19
-ize	*realize*	**47**	*2*	**35**	11	6.23	L	(**)	**+12**	23
***-ent*[2]**	*resident*	**40.5**	*3*	**32**	12	5.40	L	(**)	**+8.5**	46
***-ant*[1]**	*important*	**40.5**	*3*	**32**	12	5.40	L	(**)	**+8.5**	37
-ation	*conversation*	**26**	6	**20.5**	17	5.26	M	**	**+5.5**	17
***-er*[3]**	*buzzer*	**3**	52	**5**	78	5.20	S	*	**-2**	6
-ness	*illness*	**30**	5	**24.5**	15	5.00	L	*	**+5.5**	13
***-en*[1]**	*tighten*	**18.5**	9	**15**	22	5.45	M	*	**+3.5**	34
-ive	*massive*	**18.5**	9	**16.5**	21	4.80	M	*	**+2**	15

Key: *p<0.05; **p<0.025; ***p<0.01; ****p<0.001; *****p<0.0001. Effect sizes: L=large (φ=>0.5), M=medium (φ=>0.3); S=small (φ=>0.1)

Five suffixes showed the greatest increase across age groups: *-y*, *-ion*, *-er*[1], adjective-forming *-al*[1] and adverb-forming *-ly*[2]; all, except for *-al*[1], were identified as showing significant growth in the child dataset. These are also the five highest ranking category sizes in the DS, indicating that with respect to these suffixes, caregiver usage patterns are representative of the more general adult speech community, as predicted, despite the child-directed nature of this sample. Furthermore, only four suffixes have a lower ranking than 22 in the DS: *-ize*, *-ent*[2], *-ant*[1] and *-en*[1]. This observation conforms to the prediction that, as the caregiver dataset increases across age groups, the richest suffix categories expand most markedly, in line with DS category size.

Comparison of suffix profiles across the child, caregiver and DS datasets

To address RQ2, suffix profiles across the three corpora were compared; the relative size of each suffix category was calculated for the child, caregiver and DS datasets. Since the corpora differ considerably in size (400,261, 985,100 and 4,233,938, respectively), normalisation procedures were inappropriate for cross-corpus comparisons. Therefore, the relative contribution of each suffix category was calculated as a percentage of the derivatives in that dataset. For example, in Table 8, the child's cumulative type count for *-y* is 110. Since the total derivative type count from the child dataset is 558 (Table 6), the proportion of derivatives bearing the *-y* suffix is 19.71%; for the adult dataset, this proportion is 15.18% (207/1,364, Tables 6 and 9); for the DS it is 9.92% across the 58 suffix variants (544/5,483, Table 6 and Laws & Ryder, 2014). These percentages represent usage patterns of each suffix for each dataset.

The DS reflects the language characteristics that children will eventually adopt when they become adults. Therefore, the rank-ordered profile for the DS was used as the baseline for examining the child-caregiver profiles. Figure 1 shows the relative over/under-representation of suffix categories in relation to this DS baseline.

Figure 1 reveals that around six suffix categories in the child data are over-represented compared with the caregiver and DS datasets; these include: *-y* (*noisy*), *-er*[3] (*buzzer*), *-ie/y*[1] (*horsie*/y), *-er*[4] (*folder*), *-et* (*packet*) and *-ie/y*[3] (*hankie*). All except for *-et* were reported as showing significant growth across age ranges in Table 8. As noted in that discussion, reference to inanimate agentive entities (*-er*[3] and *-er*[4]) is more frequent than it is to animate forms (*-er*[1]). This finding is unexpected, given that animate agentive *-er* is one of the earliest suffixes children use and tends to outnumber the occurrence of instrument references (Clark & Hecht, 1982); however, the predominance of inanimate objects observed here might reflect the use of toys in the interactions. With respect to *-y*, some of the adjectives are characteristic of child speech, e.g., *stinky*, *yucky* and *yummy*, but the majority (Appendix B) are conventional uses typical of adult speech, e.g., *crazy*, *rainy* and *wobbly*. Surprisingly few innovative uses were observed, three notable examples being *glue-y*, *beary* and *poisony*. Clark (1981) observes that children readily employ *-y*; here, its category size increase is the largest over the preschool years, even though its use appears to be mainly restricted to conventional forms.

Diminutives (as predicted) and inanimate agentive forms occur more frequently in the caregiver than DS dataset. In line with the child data, *-y* is also over-represented in the caregiver sample which contains a number of 'invented' adjectives: *apple-y* (hands covered in apple), *burn-y* ('hot'), *ginger-y* (hair), *injection-y* (items in a doctor's kit), *squirrel-y* (behaviour), *wedding-y* (a dress that looks like a wedding dress). Thus, caregivers in the child

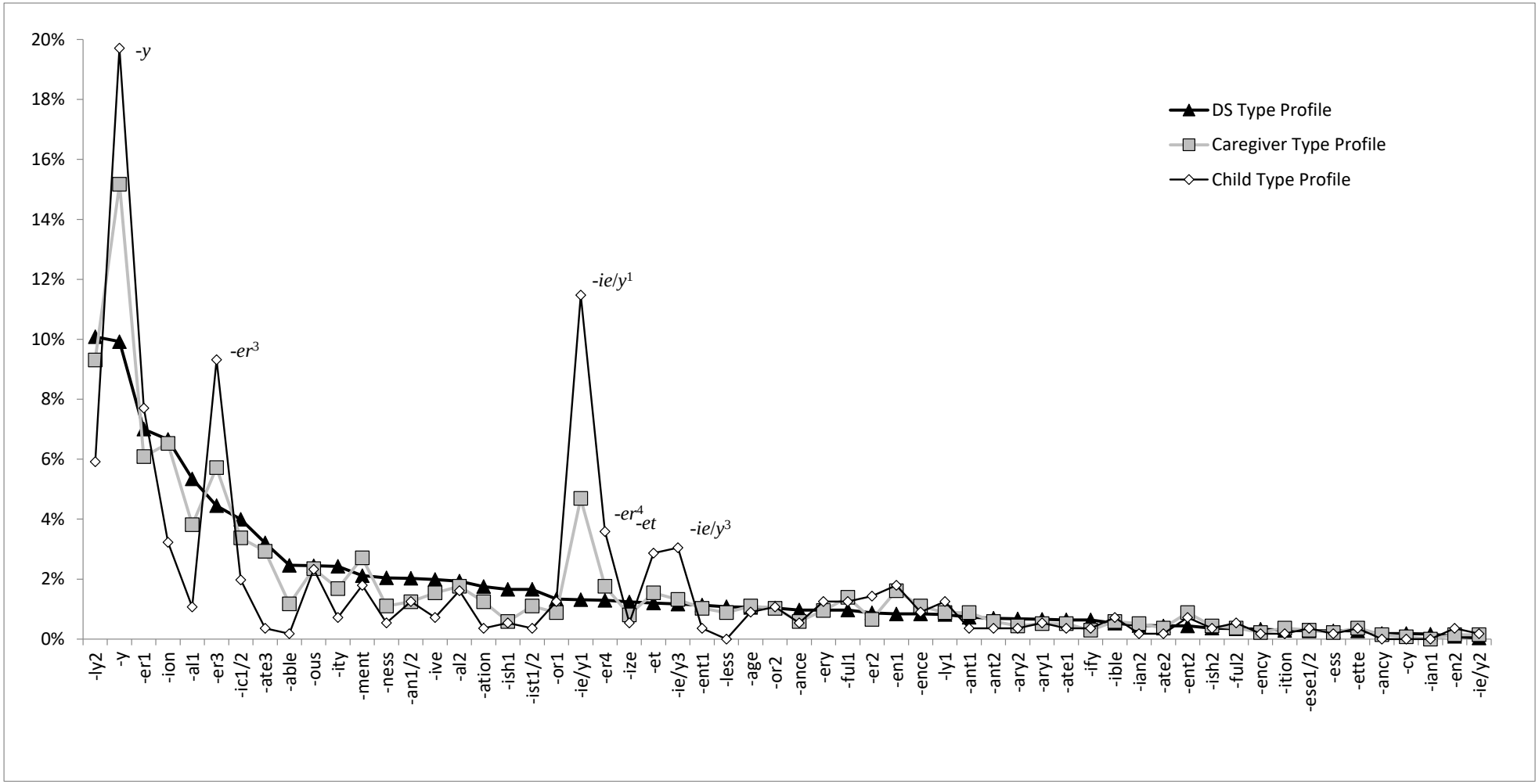

Figure 1. Child, Caregiver and DS Suffix Profiles

interactions use this adjective-forming suffix to an extent that exceeds the expectation in everyday adult speech.

Of particular interest are the suffixes that have low representation in the child dataset. Apart from *-ly*[2], *-able*, *-ness*, *-ish*[1], and *-less*, all suffix categories that fall at least 1% below the DS baseline in Figure 1 are non-neutral: *-al*[1], *-ion*, *-ate*[3], *-ic*[1/2], *-ity*, *-ation*, *-ist*[1/2] and *-ive*. Of these, *-ion*, *-ity* and *-ive* showed significant growth across age bands (Table 8), but nevertheless remain under-represented compared with the caregiver and DS corpora. This finding was predicted, as discussed in relation to RQ1a, given that Latinate derivatives are associated with the academic register and appear more frequently in child language in the school years (Schleppegrell, 2004).

Despite the observation that adverb-forming *-ly*[2] showed a robust increase in category size (Table 8), this suffix is very poorly represented in the child data compared with caregiver and DS sources. By contrast, Figure 1 reveals that *-ly*[2] is only marginally under-represented in the caregiver data compared with DS (0.85%), where it is the second largest suffix category; therefore, unexpectedly, the rich set of *-ly*[2] forms provided by caregivers in the child's immediate linguistic environment is not reflected in the current child data up to the age of 5;6. In terms of transparency of meaning and simplicity of form, *-ly*[2] scores highly, since its meaning is available (it converts adjectives to adverbs) and it predominantly attaches to free stems with no sound or stress changes, similar to agentive *-er* and diminutive *-ie/y*[1], which are pervasive in child speech. It appears, then, that *-ly*[2] adverbs are comparatively less useful to the child, despite the significant increase in category size between 2;0 and 5;6.

Characteristics of derivatives in relation to transparency and simplicity

In response to RQ3, Transparency and Simplicity scores were calculated for each derivative in the child and caregiver datasets. Examples of the 0-5 Sound+Stress+Base scores were provided in Table 5. The percentage of derivatives falling into Transparent-Intermediate-Opaque classifications for each age group are presented in Figure 2 for the child-caregiver datasets. Graphs 1(a) and 1(b) plot the percentage of 'new' derivative types entering vocabulary sets in each age sample; Graphs 2(a) and 2(b) represent the percentage of all derivative types, and Graphs 3(a) and 3(b) the percentage of all derivative tokens.

A few common observations relate to all six graphs. Firstly, regardless of whether percentages refer to 'new' derivative types, total derivative types or tokens, Transparent complex words were statistically more numerous than Intermediate or Opaque derivatives in all cases, and there was no significant difference between Intermediate and Opaque comparisons. Furthermore, all inter-age group comparisons failed to reach significance, therefore, for clarity, the statistical results below relate to the mean percentage of Sound+Stress+Base values for each line on the graphs, i.e., Transparent, Intermediate and Opaque means.

Graphs 1(a) and 1(b) indicate that 'new' derivatives entering the vocabulary set for each age group are predominantly Transparent (Child: 65%; Caregiver: 54%) compared with those classified as Intermediate (Child: 19%; Caregiver: 26%) or Opaque (Child: 17%; Caregiver: 19%). The difference between Transparent and Intermediate means was significant for the child (χ^2=25.46, p<0.0001, φ=L) and caregiver datasets (χ^2=9.74, p<0.01, φ=M), and in the latter case, as predicted, this gap is narrower.

Graphs 2(a) and (b) relate to all derivative types in each age group. In both datasets, the percentage of Transparent derivatives (Child: 64%; Caregiver: 56%) outweighed the proportion of Intermediate (Child: 18%; Caregiver: 24%) and Opaque (Child: 19%;

Caregiver: 20%) complex words. Again, Transparent derivatives reliably exceeded the Intermediate percentages in both datasets (Child: χ^2=26.09, p<0.0001, φ =L; Caregiver: χ^2=13.31, p<0.01, φ=M).

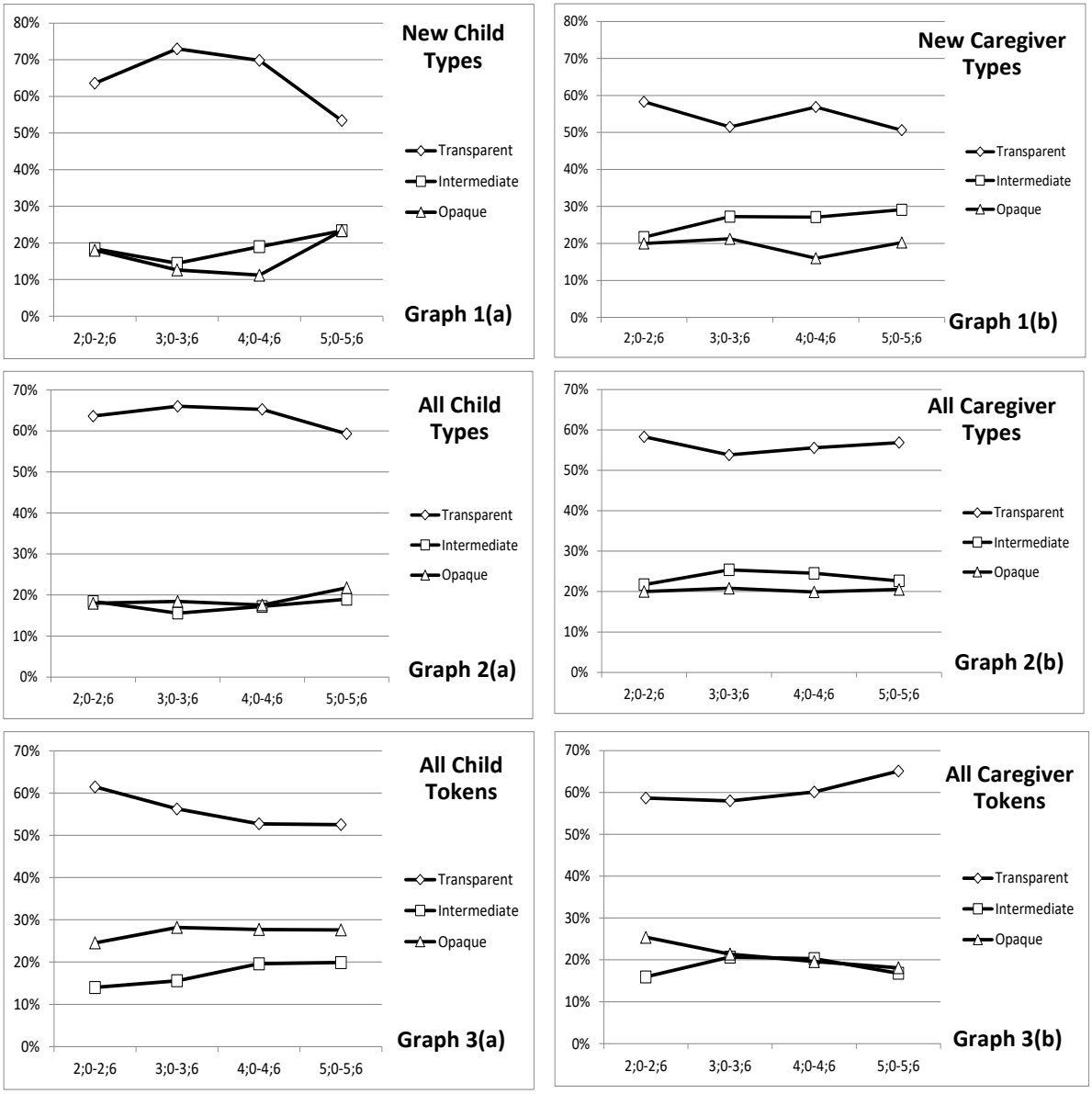

Figure 2. Percentage of transparent, medium and opaque types and tokens in the child and caregiver datasets

Graphs 1(a) to 2(b) represent type frequency, i.e., diversity in lexical choice. By contrast, Graphs 3(a) and 3(b) plot token frequencies which reflect repeated use of derivative types. The overall type and token patterns are very similar, except that Intermediate tokens are least frequent: Transparent derivatives (Child: 56%; Caregiver: 60%) outnumber the Opaque (Child: 27%; Caregiver: 21%) and Intermediate (Child: 17%; Caregiver: 18%) complex words. The proportion of Transparent derivatives exceeded the Opaque in both datasets (Child: χ^2=9.97, p<0.01, φ=M; Caregiver: χ^2=18.97, p<0.001, φ=M).

Therefore, contrary to predictions, the proportion of Transparent derivatives in caregivers' speech is comparable across age groups. Thus, when addressing older children, the early

acquired diminutive forms used by caregivers give way to other derivative types with similar Transparency ratings.

Taken together, these findings demonstrate that the largest proportion of derivatives in both child and caregiver speech is Transparent, regardless of whether the measure relates to diversity (type count) or density (token count). On average this proportion is around 61% for the child and 57% for the caregiver speech samples. The implications are that in the linguistic environment of pre-schoolers, 50-60% of derivative forms are directly analysable for decomposition into base-suffix components, and that, contrary to expectations, this proportion tends to persist in a uniform fashion over the preschool years.

Conclusions

The results reported here indicate an expected correspondence between suffix diversity in caregiver input and growth in category size in child output. However, certain suffix categories were employed more widely than anticipated, when compared with everyday adult usage patterns (DS), and others were markedly underused, despite their high degree of transparency in meaning and simplicity of form.

In accordance with early studies on school children (Carlisle, 1988; Gordon, 1989), the five suffix categories that increased most markedly across age groups in the child data were all neutral; furthermore, these suffixes also ranked among the six largest categories in the caregiver and DS speech samples. Contrary to expectations, both children and caregivers employed a greater range of instrumental than animate agentive *-er* derivatives in the younger groups, although the effect was reversed in older age bands. The significant increase in children's use of non-agentive *-er* forms (*folder*) indicates that as the child's repertoire of the more prevalent and conventional use of this suffix expands, their vocabulary set of derivatives with polysemous, non-central, less transparent meanings of this suffix (Lieber, 2004) also increases rapidly. Although the ability of preschool children to produce novel agentive and instrumental *-er* derivatives has been successfully demonstrated (Clark & Hecht, 1982), research has not to date explored whether non-central meanings of this ubiquitous suffix may also be used by children to produce neologisms.

As predicted, large suffix categories in the input are correspondingly reflected as sizable categories in the child's output, but the reverse was also observed: the transparent adverb-forming *-ly*[2] suffix was markedly under-represented in the child data compared with the caregiver and DS datasets. Although significant growth in *-ly* adverbs was observed across the age groups, diversity, commensurate with the adult data, was not mirrored in the child data. Further research is required to explore the types of *-ly* adverbs that are absent from the spontaneous speech of preschool children.

The study also revealed that differences in child and caregiver speech are negligible with respect to transparency in meaning and simplicity of form for both type or token data over the preschool years. As expected, the child's vocabulary set reflects the linguistic environment, but the results here indicate a close correspondence in derivative analysability between the input and output, despite the greater richness of the input. The current study serves as a benchmark to explore the degree to which this phenomenon persists over the early school years and beyond.

The analytical framework employed gave equal weighting to the Sound+Stress+Base components of the transparency measure; it is for follow-on work to ascertain the relative weightings of these factors in preschool child-caregiver derivative profiles. Furthermore, given that accessibility of the base has been shown to promote word reading accuracy in

school children between Grades 4 and 6 (Deacon, Whalen, & Kirby, 2011), an assessment of base frequency (e.g., the frequency of *hope* and its morphologically-related forms, such as *hopeful* and *hopeless*) as well as the surface frequency measure employed here, would throw additional light on the nature of the expanding repertoire of derivative types observed in preschool children.

This corpus-based study contributes to our understanding of derivational usage patterns in the spontaneous speech of English-speaking preschool children before they learn to read and write, both in relation to immediate caregiver input and adult-directed speech. Although the current results are necessarily limited to production, and thus do not include those derivatives that are comprehended but not uttered by the children, they provide a baseline of derivative types and their transparency characteristics from which to explore the factors that facilitate the development of morphological awareness in the early school years. Derivational profiles of pre-schoolers can inform the development of primary school literacy programmes that directly build on children's existing repertoire of complex words. Furthermore, such age-related normative datasets can provide a baseline for evaluating derivative vocabulary development in language-delayed children who, not only encounter difficulties with inflectional morphology (Bishop 1997), but also show production errors with complex words that are not observed with children whose language is developing normally (Marshall & van de Lely 2007). Therefore, the derivative profiles reported here provide valuable information concerning the relative frequency of English suffixes associated with normal language development and can be employed as a baseline in future naturalistic studies and as an informed source for the design of experimental materials.

Acknowledgements

The author would like to thank Dr Chris Ryder for performing the inter-rater reliability check of the transparency measures reported.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42, 239–273.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*. *58* (Series No. 238). Chicago: University of Chicago Press.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp.899–919). Berlin: Mouton de Gruyter.

Bauer, L. (1983). *English Word-Formation*. Cambridge: Cambridge University Press.
Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.
Berko, J. (1958). The child's learning of English morphology. *Word 14*, 150-177.
Berninger, V.W., Abbott, R.D., Nagy, W. & Carlisle, J. (2010). Growth in phonological, orthographic and morphological awareness in Grades 1 to 6. *Journal of Psycholinguistic Research 39*, 141–163.
Bishop, D. V. M. (1997) *Uncommon understanding: development and disorders of language comprehension in children*. Hove: Psychology Press.
BNC*web* (CQP-Edition) Version 4.3, November 2013. Retrieved from https://bncweb.lancs.ac.uk/
Booij, G. (2006). Inflection and Derivation. In: Brown, K. (Ed.) *Encyclopedia of Language & Linguistics*, Second Edition, Volume 5, pp. 654-661. Oxford: Elsevier.
Brown, K. (1973). *A first language: the early stages.* Cambridge, M.A: Harvard University Press.
Carlisle, F. (1988). Knowledge of derivational morphology and spelling ability in fourth, sixth and eighth graders. *Applied Psycholinguistics 9*, 247-266.
Clark, E.V. (1981). Lexical innovations: How children learn to create new words. In W. Deutsch (Ed.), *The child's construction of language*, (pp.299-328). London: Academic Press.
Clark, E.V. (1993). *The lexicon in acquisition.* Cambridge: Cambridge University Press.
Clark, E.V. (2014). Acquisition of derivational morphology. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology*, (pp.424-439). Oxford: Oxford University Press.
Clark, E. V., & Cohen, S. (1984). Productivity and memory for newly formed words. *Journal of Child Language 11*, 611-625.
Clark, E. V., & Hecht, B. F. (1982). Learning to coin agent and instrument nouns. *Cognition 12*, 1-24.
Cunningham, A. J. & Carroll, J. M. (2015). Early predictors of phonological and morphological awareness and the link with reading: Evidence from children with different patterns of early deficit. *Applied Psycholinguistics 36*, 509-531.
Deacon, S. H. & Bryant, P. (2005). What young children do and do not know about the spelling of inflections and derivations. *Developmental Science 8*, 583-594.
Deacon, S. H., Whalen, R. & Kirby, J. R. (2011). Do children see the *danger* in *dangerous*? Grade 4, 6 and 8 children's reading of morphologically complex words. *Applied Psycholinguistics 32*, 467-481.
Demuth, K. (2008). Exploiting corpora for language acquisition research. In Behrens, H. (Ed.), *Corpora in language acquisition research: Finding structure in data*, (pp.199–205). Amsterdam: John Benjamins.
Derwing, B. L., & W. J. Baker. (1986). Assessing morphological development. In P. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development*, 2nd edition, (pp.326-338). Cambridge: Cambridge University Press.
de Villiers, J. G. & de Villiers, P. (1973). A cross-sectional study of the acquisition of grammatical morphemes. *Journal of Psycholinguistic Research 2*, 267-278.
Dickinson, D. K., & Tabors, P. O. (2001). *Beginning Literacy and Language: Young Children Learning at Home and at School*. Baltimore, MD: Brookes Publishing.
Freyd, P. & J. Baron. (1982). Individual differences in acquisition of derivational morphology. *Journal of Verbal Learning and Verbal Behavior 21*, 310-332.
Gordon, P. (1989). Levels of affixation in the acquisition of English morphology. *Journal of Memory and Language 28*, 519-530.

Hoff, E. & Naigles, L. (2002). How children use input in acquiring a lexicon. *Child Development, 73*, 418–433.
Laws, J. & Ryder, C. (2014). *MorphoQuantics*: http://morphoquantics.co.uk
Laws, J. & Ryder, C. (2018). Register variation in spoken British English: The case of verb-forming suffixation. *International Journal of Corpus Linguistics*, 23 (1): 1-27.
Lieber, R. (2004). *Morphology and lexical semantics*. Cambridge: Cambridge University Press.
Lieber, R. (2005). English Word-Formation Processes. Observations, Issues, and Thoughts on Future Research. In: Štekauer, P. & Lieber, R. (Eds.) *Handbook of Word-Formation*, 375-427, Dordrecht: Springer.
Lyster, S. A. H., Lervåg, A. O. & Hulme, C. (2016). Preschool morphological training produces long-term improvements in reading comprehension. *Reading and writing 29*, 1269-1288.
MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.
Marshall, C. R., & van der Lely, H. K. J. (2007). Derivational morphology in children with Grammatical-Specific Language Impairment. *Clinical Linguistics and Phonetics 42*, 71–91.
McBride-Chang, C., Wagner, R. K., Muse, A., Chow, B. W.-Y. & Shu, H. (2005). The role of morphological awareness of children's vocabulary acquisition in English. *Applied Psycholinguistics 26*, 415-435.
Moats, L. C. & Smith, C. (1992). Derivational morphology: why it should be included in language assessment and instruction. *Language, Speech, and Hearing Services in Schools 23*, 312-319.
Raban, B. (1988). *The spoken vocabulary of five-year old children*. University of Reading: Reading and Language Information Centre.
Schleppegrell, M.J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah, NJ: Lawrence Erlbaum.
Schmid, H-J. (2011). *English morphology and word-formation. An introduction*. Berlin: Erich Schmidt Verlag.
Seigal, D. I. (1974). *Topics in English morphology*. New York: Garland.
Slobin, D. (1973). Cognitive prerequisites for the acquisition of grammar. In C. A. Ferguson, and D. I. Slobin (Eds). *Studies of child language development*, (pp.173-208). New York: Holt, Reinhart and Winston.
Stein, G. (2007) *A Dictionary of English Affixes: Their function and meaning*. LINCOM Studies in English Linguistics, LINCOM EUROPA academic publications.
Swan, D. W. (2000). How to Build a Lexicon: A Case Study of Lexical Errors and Innovations, *First Language 20*, 187–204.
Tyler, A., & Nagy, W. (1989). The acquisition of derivational morphology. *Journal of Memory and Language 28*, 649-667.