Scientific Studies of Reading

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hssr20

# Effects of Target Age and Genre on Morphological Complexity in Children's Reading Material

Nicola Dawson, Yaling Hsiao, Alvin Wei Ming Tan, Nilanjana Banerji & Kate Nation

**To cite this article:** Nicola Dawson, Yaling Hsiao, Alvin Wei Ming Tan, Nilanjana Banerji & Kate Nation (2023): Effects of Target Age and Genre on Morphological Complexity in Children's Reading Material, Scientific Studies of Reading, DOI: 10.1080/10888438.2023.2206574

**To link to this article:** https://doi.org/10.1080/10888438.2023.2206574

SCIENTIFIC STUDIES OF READING
https://doi.org/10.1080/10888438.2023.2206574

Routledge
Taylor & Francis Group



# Effects of Target Age and Genre on Morphological Complexity in Children's Reading Material

Nicola Dawson[a], Yaling Hsiao[a], Alvin Wei Ming Tan[a], Nilanjana Banerji[b], and Kate Nation[a]

[a]Department of Experimental Psychology, University of Oxford, Oxford, UK; [b]Oxford University Press, University of Oxford, Oxford, UK

**ABSTRACT**

**Purpose:** Morphological regularities are an important feature of the English writing system, and exposure to written morphology may be key in the development of skilled word recognition. Our aim was to investigate children's experiences of written morphology by analyzing a large-scale corpus of children's reading materials spanning a target age range from 5 to 14 years.

**Method:** Analysis was based on the Oxford Children's Reading Corpus. We examined frequency distributions of derived and compound words by target age and genre, as well as type and token frequencies for individual derivational suffixes.

**Results:** We found that the proportion of morphologically complex words – and derived words particularly – increased in line with target age, and that nonfiction contained more complex words than fiction. Frequencies of individual suffixes also varied by target age and genre, with Germanic forms more common in fiction and texts for younger children, and Latinate forms more common in nonfiction and texts for older children.

**Conclusion:** These findings provide a comprehensive picture of how children's experience with written morphology changes over the course of reading development. We discuss these findings in the context of developmental changes in morphological processing, and the benefits and limitations of using large-scale natural language datasets.

The English writing system is morphophonemic, meaning that in addition to systematic mappings between letters and sounds, letters and letter combinations also reflect underlying morphological structure. Morphemes are the smallest meaningful unit in a language (Carstairs McCarthy, 2002). A morphologically complex word such as *player* combines a base (*play*) and an affix (*er*), and both base and affix operate in systematic ways across multiple words (e.g., *replay*, *playful, playground*; *teacher*, *writer*, *reader*). Skilled visual word recognition is characterized by rapid and implicit analysis of morphological structure (Amenta & Crepaldi, 2012; Rastle & Davis, 2008), but this behavior emerges relatively late in reading development (Beyersmann et al., 2012; Dawson et al., 2018). The mechanisms that drive this developmental change are not well understood, but one proposal is that acquiring insight into morphological regularities in the writing system may be an important factor in the transition from novice to expert reader, and

**CONTACT** Nicola Dawson nicola.dawson@psy.ox.ac.uk Department of Experimental Psychology, University of Oxford, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, United Kingdom

The Oxford Children's Corpus is a growing database of writing for and by children developed and maintained by Oxford University Press for the purpose of children's language research. Our data and analysis scripts are available on the Open Science Framework at https://osf.io/4xurw/?view_only=acd838535db344d3a78911f53b1b9e9c.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10888438.2023.2206574.

specifically in building direct pathways from orthography to meaning (Rastle, 2019b). Our aim was to characterize developmental variation in written morphology by analyzing a large corpus of reading material targeted at 5–14 year-olds. From this, we derived lexical statistics relating to morphological features, which we discuss in the context of morphological development in children's reading.

Morphological structure within a writing system provides "islands of regularity" that support direct connections between spelling and meaning (Rastle, 2019a). This is nicely illustrated by the English past tense. While monomorphemic words such as *act*, *bake*, and *play* follow regular spelling-sound mappings, pronunciation of *–ed* in their past tense forms (/*ækt***ɪd**/ ,/*be***ɪ***k***t**/and/*ple***ɪ***d***/) varies according to the word-final phoneme of the base, even though spelling of the suffix does not change. In this way, the written forms of the suffixed words provide a consistent cue to meaning (in this case, past tense), at the expense of consistency in the relationship between spelling and sound (Rastle, 2019a). Corpus analyses examining spelling patterns of morphologically complex words have indicated that the optimization of semantic information over phonological consistency is a characteristic of English orthography more broadly, and this is observed across both the inflectional and derivational systems. Berg and Aronoff (2017) showed that affix spellings typically diverge from spellings representing the same sequence of sounds in non-morphological units. For example, they found that the phoneme combination/ə*s*/took the orthographic form *-ous* wherever it appeared as an adjectival suffix (e.g., *nervous*, *advantageous*), but not when it appeared in word-final position in alternate parts of speech (e.g., *bonus*, *genius*). Ulicheva et al. (2020) extended this work to 154 English suffixes, revealing a high degree of consistency in the relationship between suffix spellings and word class. In this way, morphological regularities provide a direct link from spelling to meaning, and overlaps in meaning between morphologically-related words may be particularly salient in written, compared to spoken, language. This systematicity substantially reduces the scale of unique orthography-semantics mappings that need to be learned (Brysbaert et al., 2016). In addition, adults are sensitive to morphological structure as they read words and there is strong evidence that skilled visual word recognition draws on rapid and implicit analysis of morphological structure (Amenta & Crepaldi, 2012).

Children do not seem to process complex written words in the same way as skilled adult readers. Evidence from masked priming studies with adults suggests that in English at least, the initial stages of complex word recognition is based on recognition of morphological structure at the orthographic level, meaning that words such as *corner* comprising a pseudo-stem and suffix structure are parsed in the same way as "true" complex words, such as *reader* (Rastle & Davis, 2008; Rastle et al., 2004). This pattern of morpho-orthographic processing is not observed in developing readers until mid-to-late adolescence (Dawson, Rastle, et al., 2021). Similar conclusions can be drawn from studies probing the morpheme interference effect. While adults are slower to correctly reject a pseudoword with apparent morphological structure (e.g., *earist*) compared to a matched nonmorphological control (e.g., *earilt*; Crepaldi et al., 2010) in lexical decision, children do not show this sublexical effect in reaction times until late adolescence, although they are less accurate in rejecting morphological pseudowords (Dawson et al., 2018, but see Burani et al., 2002; Casalis et al., 2015; Hasenäcker et al., 2017 for diverging findings in other European languages).

One possible account for these developmental differences – in English at least – is that an individual must accumulate sufficient exposure to morphological regularities in writing for this information to become embedded in the reading system. In particular, experience of bound morphemes (e.g., affixes) may be key. Compared to stems, affix meanings are often less transparent, and by definition they are never encountered in isolation (Grainger & Beyersmann, 2017; Schreuder & Baayen, 1995). Children must therefore build associations between these letter patterns and their meanings by encountering them repeatedly across different contexts (Tamminen et al., 2015). Although children demonstrate some knowledge of morphological relationships in spoken language well before they learn to read (Berko, 1958), it is

clear from work by Ulicheva et al. (2020) and Berg and Aronoff (2017) that access to written English permits unique insights into consistencies between form and meaning. Indeed, the strength of these consistencies predicts people's sensitivity to these cues in reading and spelling tasks (Ulicheva et al., 2020). In turn, this suggests that accumulated experience of morphology in written language is an important factor in shaping skilled reading behavior, although this has yet to be directly tested.

Consistent with the view that experience with written language shapes morphological processing are experimental data showing that corpus-derived properties of complex words and their component morphemes, such as frequency and productivity, influence visual word recognition in skilled readers (Bertram et al., 2000; De Jong et al., 2000; Ford et al., 2010; Lázaro et al., 2013; Schreuder & Baayen, 1997; Taft & Ardasinski, 2006; Taft, 1979; Xu & Taft, 2015). For example, complex words containing higher frequency stems are recognized faster in lexical decision compared to words with lower frequency stems, even when they are matched on surface frequency (Taft & Ardasinski, 2006; Xu & Taft, 2015). Processing is also influenced by morphological family size, such that a base word with a large number of morphological relatives is recognized more rapidly and accurately (Ford et al., 2010; Schreuder & Baayen, 1997) and some studies report effects of affix type and token frequency under certain conditions (Burani & Thornton, 2003; Sánchez-Gutiérrez et al., 2018). On the basis that morphological processing in skilled reading reflects the distribution of base words and affixes in written language, documenting morphological complexity in the materials children read is a critical first step toward understanding how this input might shape the development of morphological processing through childhood and adolescence and into adulthood.

Segbers and Schroeder (2017) detailed developmental changes in morphology based on a set of 500 children's books targeted at German readers aged 6–12 years. They took repeated samples of the corpus at different lexicon sizes and used a vocabulary "test" to determine the probability of any given test item also appearing in the lexicon as a function of the size of the lexicon. These parameters were then used to estimate the words known by participants from Grade 1 through to adulthood who took the same vocabulary test. Within these sample lexicons, the authors categorized words based on the number of morphemes they contained and their morphological category (derived, compound, or both), and found that the proportion of all words accounted for by morphologically complex items increased significantly as lexicon size grew. They further reported that while derivation formed the largest morphological category in Grades 1 and 2, compounds showed the biggest increase in line with lexicon size, and became the most dominant morphological category for all other age groups.

The morphological category patterns reported by Segbers and Schroeder (2017) contrast somewhat with observations from studies of spoken language acquisition in English, where it is clear that derivational knowledge develops over a long period (Anglin, 1993; Berko, 1958; Tyler & Nagy, 1989), and well into adolescence (Nippold & Sun, 2008). This protracted development may in part reflect the polysemous and quasi-regular nature of derivational affixation in English, as well as its breadth. As such, mastery of derivational morphology in English may be particularly dependent on rich exposure to derived words and their constituents across a broad range of contexts. It is important to consider how the linguistic environment experienced by children changes over this period too. Learning to read brings radical changes to a child's linguistic experience (Nation et al., 2022; Nation, 2017). As books written for young children contain more morphologically complex words than child-directed speech (Dawson, Hsiao, et al., 2021), the onset of literacy provides children with access to more sophisticated vocabulary (Nagy & Anderson, 1984). More relevant to our investigation, however, is how learning to read provides children with new insights into morphemes and form-meaning regularities, especially those that are less salient in spoken language (Berg & Aronoff, 2017; Rastle, 2019a; Ulicheva et al., 2020). By taking developmental slices through a large corpus of language written for

children to read, our study aimed to capture and quantify changes in morphological complexity as reading experience builds.

In addition to developmental variation, contextual features such as genre and register may also influence morphological characteristics of written texts. Academic language is information-dense and concise, and quite distinct from informal spoken discourse or fiction (Nation et al., 2022). In part, this effect is driven by grammatical processes such as nominalization, whereby ideas that could otherwise be communicated via an entire phrase are condensed into a single (often abstract) noun (Snow, 2010). Suffixes such as *–ion* and *–ity* are characteristic of this process, meaning that they are likely to occur more frequently in written material adopting a formal, academic style, such as expository texts, compared to fiction or conversational language (Biber et al., 1998). Aside from some exceptions, these suffixes are often non-neutral: they attach to bound stems (e.g., *quantify*), cause shifts in the pronunciation of the base word (e.g., *decide – decision*), and are typically low in productivity (i.e. they attach to a restricted set of base words – e.g., *warmth*). By contrast, neutral suffixes attach to freestanding stems (e.g., *hopeful*), do not change the pronunciation or stress pattern of the stem (e.g., *teacher*), are more productive (e.g., *-ness* attaches to many adjectives to form a noun), and are usually mastered at an earlier stage of development (Tyler & Nagy, 1989). Analysis by genre (fiction vs. nonfiction) provides an opportunity to track these patterns as reading experience builds. Of additional interest is to detail how variation in morphological complexity by genre may overlap with variation by developmental stage. Children in the primary school years encounter more narrative fiction than nonfiction texts (Duke, 2000; Yopp & Yopp, 2006), while older readers are expected to engage increasingly with expository and nonfiction texts as the emphasis in education shifts to reading as a vehicle for learning (Graesser et al., 2011; Nippold, 2016).

## Current study

Our aim in this study was to examine morphological complexity in a large and varied corpus of reading materials targeted at children and adolescents aged 5–14 years. Specifically, we asked a) how the prevalence of morphologically complex words in texts differs according to the age of the intended audience and the genre of the text (fiction vs. nonfiction); b) whether distributions of complex word types (derived, compound, compound derived) vary with target age and genre; c) how frequencies of individual suffixes vary with target age and genre.

In these analyses, we report both token and type frequency data. Token counts of complex words and suffixes indicate how frequently children may encounter those units as a function of their reading experience, while type counts represent the diversity of complex words a child may encounter, or the number of unique contexts in which they experience a given suffix. This distinction is relevant when considering the link between reading experience and processing of complex words, as both computational and empirical evidence point to a differing contribution of token and type frequencies to retrieval and generalization of derived morphological forms (Reichle & Perfetti, 2003; Tamminen et al., 2015).

Our investigation follows in two parts. We first examined how the proportion of derived and compound words in these texts varied in accordance with target age and genre. We anticipated that the overall proportion of morphologically complex words would increase in line with target age (Segbers & Schroeder, 2017), and that derived forms would show the biggest expansion, paralleling protracted growth in children's knowledge of derivational relationships in English (Anglin, 1993; Nippold & Sun, 2008; Tyler & Nagy, 1989). We also expected that the proportion of complex words would be greater in nonfiction compared to fiction, given that academic texts are characterized by dense informational content and a formal register (Biber et al., 1998; Snow, 2010).

Secondly, we investigated in more detail the frequency distributions of 80 derivational suffixes, again comparing across target age and genre. We focused in particular on derivational suffixation as this aspect of morphology is most commonly the focus of developmental (and adult) studies of morphological processing in the context of visual word recognition (Beyersmann et al., 2012; Burani et al., 2002; Casalis et al., 2015; Dawson et al., 2018; Dawson, Rastle, et al., 2021; Hasenäcker et al., 2016; Rastle et al., 2004). Additionally, exposure to derivational suffixes in texts targeted at different age groups is of particular interest given the close association between suffix spellings and meaning (Ulicheva et al., 2020), and the protracted development of children's derivational knowledge (Anglin, 1993; Nippold & Sun, 2008). In line with previous work (Baayen, 2008; Biber et al., 1998; Laws, 2019; Plag et al., 1999; Tyler & Nagy, 1989), we predicted that suffixes most characteristic of texts targeted at younger children or fiction would tend to be neutral (attaching to free bases and resulting in no change to the pronunciation of the base word), and semantically transparent. By contrast, we expected that texts targeted at older readers or nonfiction would be more strongly associated with nominalizing suffixes with Latinate origins, such as *–ion* and *–ity*, typical of a more formal register.

## Method

### Corpus

Our data were taken from the reading section of the Oxford Children's Corpus, a dynamic and growing corpus created by Oxford University Press in 2006 to inform the development of children's dictionaries (Wild et al., 2013). The full version of the corpus comprises around 21,000 documents targeted at children aged 5–16 years, for a total word count of around 47 million words, although our analyses were based on a subsection of the full corpus (see below). These documents were sampled from a broad range of contexts spanning fiction and nonfiction texts, curriculum materials, and text extracted from children's websites. Metadata associated with the corpus include document information (e.g., title, author, and publisher), genre (fiction vs. nonfiction), and target Key Stage. Key Stages refer to education levels in England and Wales, determined by a child's age. Key Stages 1–4 equate to age groupings of 5–7 years, 7–11 years, 11–14 years, and 14–16 years, respectively.

### Procedure

#### Corpus processing

The full corpus was made available by Oxford University Press as vertical text files. These files had been pre-processed using the Oxford English part-of-speech tagset to generate lemmatized forms of each token (i.e. removing inflections) and add part of speech tags. We converted these vertical text files into.csv files in which each row corresponded to a single token along with its lemmatized form, part of speech tag, and document identifier allowing us to link each document to its associated metadata. These files formed the basis of all subsequent analyses.

### Coding of morphological structure

We created a reference database for the purposes of analyzing the morphological structure of words in the Oxford Children's Corpus. We first generated a list comprising all words in the corpus occurring 50 times or more, along with their frequencies. Our analysis was based on lemmas rather than word forms, such that all regular inflected words (e.g., *plays*, *played*, *playing*) were included under the head word (e.g., *play*), and pseudoderived words such as *corner* were

counted as monomorphemic. However, where inflectional suffixes appeared in contexts other than verbs (for example, adjectives ending in *–ing* [*exciting*] or *–ed* [*tired*]), these words were treated as separate lemmas because of their shift in word class, although note that this may have inflated estimates of complex word-type frequency relative to approaches that exclude all inflected forms.

Our primary source of reference for coding of morphological structure was the MorphoLex database (Sánchez-Gutiérrez et al., 2018). MorphoLex comprises morphological information for each complex word listed in the English Lexicon Project (Balota et al., 2007), including number of morphemes and segmentation of complex words into bases and affixes, as well as statistics on morphological family size, affix productivity and length, and summed token frequencies. We cross-referenced each word in the Oxford Children's Corpus with entries in the MorphoLex database. Following this procedure, data were available for 74% of total lemma types in the Oxford Children's Corpus. The remaining items were hand-coded following segmentation protocols outlined in Sánchez-Gutiérrez et al. (2018) by the first and third authors of this paper and a research assistant with training in linguistics. Thus, our final database contained information on a) number of morphemes; b) segmentation structure; and c) occurrence of individual base words and derivational affixes for each item in the Oxford Children's Corpus. For example, the word *uncertainly* returned a morpheme count of three, a segmentation structure of <*un*<{(*certain*)}>*ly*> with associated prefix-root-suffix (PRS) tag of 1,1,1, and contributed to frequency counts of the prefix *un–*, the base *certain*, and the suffix *–ly*.

We additionally tagged each lemma as monomorphemic, morphologically complex or other. The category "other" comprised a range of additional codes for items that are generally considered non-lexical, such as multiword phrases (e.g., *every so often*), abbreviations (e.g., *UK*), and proper nouns (e.g., names of people; Brysbaert et al., 2016), and comprised around 4% of tokens in the version of the corpus used in our analyses. These items were included in calculations of total corpus size, but were never coded as morphologically complex. Finally, we coded each morphologically complex word as either derived (e.g., *teacher*), compound (e.g., *football*) or compound-derived (e.g., *footballer*). For example, if a segmented form in MorphoLex contained two or more base forms, e.g., {(*any*)(*body*)}, this was tagged as compound; if it contained a base form and a prefix or suffix, e.g., {(*able*)}>*ity*>, it was tagged as derived, and if it contained two or more base forms and an affix, e.g., {(*dish*)}{(*wash*)}>*er*>, it was tagged as compound-derived.

## Analysis

*(i) Morphological Complexity by Key Stage and Genre*

We first examined differences across Key Stage and genre in the percentage of word tokens and types classed as morphologically complex. Not all documents available in the Oxford Children's Reading Corpus contained Key Stage metadata, so our target corpus was a subsample of the whole corpus (approximately 44% of the total number of documents, but still a comprehensive sample of 22 million words). For comparison, we include statistics on the number of documents, words, and complex

**Table 1.** A breakdown of document and word counts for each subsection of the corpus.

| Key Stage | Genre | Number of documents | Number of words | Number of unique word types |
|---|---|---|---|---|
| KS1 | Fiction | 192 | 794,495 | 9,991 |
| | Non-fiction | 14 | 18,756 | 1,864 |
| KS2 | Fiction | 392 | 8,989,480 | 18,487 |
| | Non-fiction | 3,014 | 2,802,046 | 17,200 |
| KS3 | Fiction | 294 | 5,566,314 | 18,033 |
| | Non-fiction | 5,676 | 1,965,653 | 16,272 |

words in fiction vs. nonfiction texts with and without Key Stage data in Appendix A. These values indicate a similar split across genre in these key characteristics between the subcorpus that was used in our analyses (with Key Stage data) and the subcorpus that was discarded (without Key Stage data).

Documents targeted at Key Stages 1, 2 and 3 contained examples of both fiction and nonfiction. However, Key Stage 4 texts were all fiction and comprised only 17 documents, the majority of which were 19th century novels. Because of this confound between Key Stage and genre, our reported analyses are based on texts targeted at children in Key Stages 1–3. This comprised 9,582 documents (approximately 20 million words), all of which were tagged with Key Stage (1–3) and genre (fiction vs. nonfiction) information.

Table 1 provides a summary of document and word counts for each subsection of this sample. We calculated the percentage of morphologically complex words in each document, excluding lemma types occurring with a frequency of less than 50 across the full corpus to reduce the impact of tokenization or spelling errors on type frequency counts, and to align with information in our morphology database.

We examined the types of complex words occurring most frequently in texts targeted at different age groups, and across fiction and nonfiction materials. Here, we examined the distribution of complex word types and tokens classed as derived, compound or compound-derived across Key Stage and genre separately.

*(ii) Suffix Frequencies by Key Stage and Genre*

Our coding of morphological structure (see above) allowed us to identify the morphological constituents of each complex word appearing in the corpus. To examine variation in suffix frequencies, we created a list of 80 derivational suffixes that comprised a subset of the suffixes identified by our analysis that also appeared in other relevant work (Laws, 2019; Ulicheva et al., 2020). We then recorded the summed token frequency (i.e. the total number of occurrences of the suffix) and type frequency (the number of unique words the suffix appears in) for each suffix, separately for Key Stage and genre (see Appendix B for a list of target suffixes). For words with multi-morphemic stems, frequency counts were based on word-final suffixes, such that a word like *thoughtful* counted toward the token and type frequencies for the suffix *–ful*, while *thoughtfulness* contributed to frequency counts for *–ness*, but not *–ful*. This approach was taken because the outermost suffix determines part of speech of the complex word, and little is known about how children process suffixes when they are embedded in complex words with more than two constituents (Kuperman et al., 2009). Across the Key Stage 1–3 corpus, 7.71% of all derived words contained multiple suffixes. Note that our analysis was based on orthographic form, such that we did not differentiate between polysemous variants of the same suffix (e.g., the animate form of *–er* [*teacher*] vs. the inanimate form [*toaster*]). However, because our analysis was based on lemmas, frequency counts for *–er* were not confused with the inflectional suffix bearing the same orthographic form (e.g., comparative form [*higher*]).

Given the variation in size of the subcorpora, token frequency counts were normalized in two ways: first, by dividing the count of an individual suffix by the total number of words in that subcorpus, and multiplying by 1 million to give frequency per million words, and second, by taking number of suffixed words as the denominator to account for variation in the overall number of complex words across the subcorpora. To obtain comparable raw type frequency counts, we randomly sampled whole documents from each subcorpus to equate to a total word count of ±10% of the size of the smallest subcorpus (Key Stage 1). This was repeated 100 times for each subcorpus, and suffix type frequencies were calculated as the mean type frequency across all 100 iterations. These frequencies were also divided by the total number of unique suffixed word types in a given subcorpus and multiplied by 1 thousand to give type frequency per thousand suffixed word types.

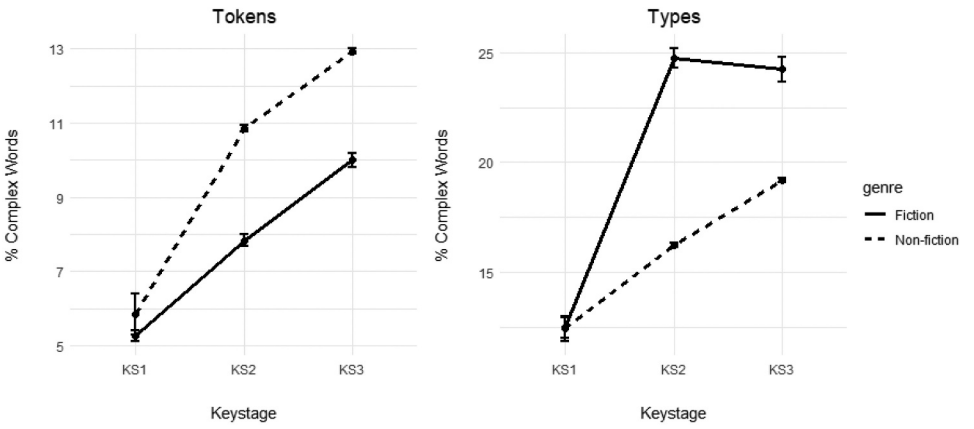


**Figure 1.** Mean percentage morphologically complex lemma tokens (left panel) and lemma types (right panel) in the oxford children's corpus, split by key stage and genre.

## Results

### *(i) How Does the Prevalence of Morphologically Complex Words in Texts Differ According to Target Age and Genre?*

#### *Token frequencies*

Figure 1 shows the mean percentage morphologically complex lemma tokens, split by Key Stage and genre. These means indicate a clear increase in the percentage of words classed as morphologically complex between Key Stages 1–3, and a higher proportion of complex words in nonfiction vs. fiction texts. These trends were confirmed using a linear regression model with percentage complex words as the outcome variable (each observation representing a single text; $n = 9{,}582$), Key Stage and genre as predictors, and document length in words as a covariate to account for the presence of a few very short texts containing a disproportionately high percentage of complex words. Predictor variables were coded using successive differences contrasts so that we could test stepwise changes in morphological complexity across Key Stage.

Our analyses revealed that the percentage of complex word tokens increased by approximately 3 percentage points between Key Stages 1 and 2 ($\beta = 3.02$, $SE = 0.37$, $t = 8.16$, $p < .001$) and by around 2 percentage points between Key Stages 2 and 3 ($\beta = 2.09$, $SE = 0.10$, $t = 20.87$, $p < .001$). There was also a significant effect of genre, with a greater proportion of complex word tokens in nonfiction compared to fiction; an effect size of around 2.6 percentage points ($\beta = 2.64$, $SE = 0.20$, $t = 13.12$, $p < .001$). Document word count showed a significant inverse relationship with percentage of complex words ($\beta = -0.00$, $SE = 0.00$, $t = -3.34$, $p < .001$): for each increase of 1 word in document length, the percentage of complex words decreased by an average of 0.00001 percentage points. Finally, we examined whether effects of genre were consistent across Key Stages by adding the interaction term to the above model, but this did not significantly improve model fit (LRT: $\chi^2$ [2] = 1.39, $p = .249$).

#### *Type frequencies*

Figure 1 shows a similar increase in the percentage of complex word types across Key Stage, indicating that complex words account for an increasing proportion of all unique words that individuals encounter in texts as they become more proficient readers. However, in contrast to the trend observed for word tokens, the proportion of unique word types classed as morphologically complex is greater in fiction compared to nonfiction. This indicates that while nonfiction contains a higher proportion of complex words overall (as demonstrated by the token frequency analysis), fiction draws on a wider range of complex word types as a proportion of total word types.

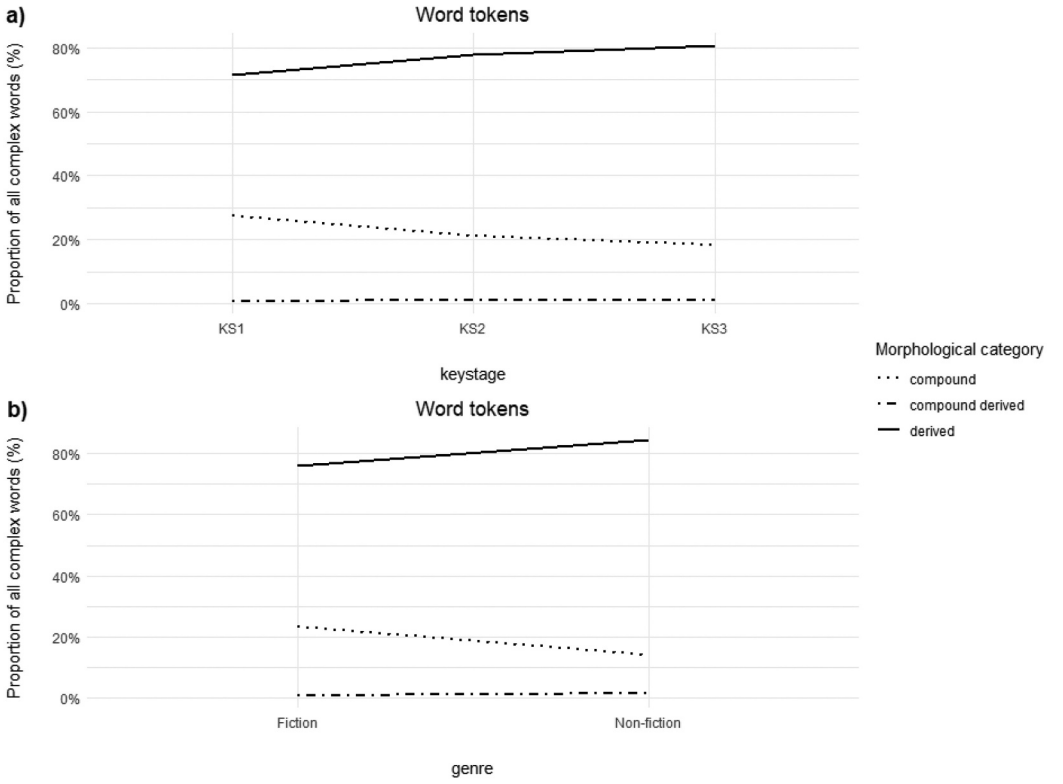


**Figure 2.** Proportion of complex word tokens by morphological category, plotted by key stage (panel a) and genre (panel b).

**Table 2.** Standardized chi-square residuals comparing observed and expected word token frequencies for different morphological categories by key stage and genre.

| | | Morphological Category | | |
|---|---|---|---|---|
| | | Compound | Derived | Compound derived |
| Key Stage | Key Stage 1 | 36.56 | −17.74 | −6.64 |
| | Key Stage 2 | 21.66 | −10.19 | −6.77 |
| | Key Stage 3 | −35.41 | 16.80 | 9.79 |
| Genre | Fiction | 73.94 | −34.08 | −29.20 |
| | Non-fiction | −104.08 | 47.97 | 41.10 |

Chi-square residuals represent the difference between the observed and expected values for a given cell.

We again used linear mixed effects models to test these effects. With document length in words included in the models as a covariate, we observed a significant increase in percentage of complex word types between Key Stages 1 and 2 by around 6.6 percentage points ($\beta = 6.64$, $SE = 0.85$, $t = 7.83$, $p < .001$), and a smaller increase between Key Stages 2 and 3 of 1.6 percentage points ($\beta = 1.57$, $SE = 0.24$, $t = 6.51$, $p < .001$). There was also a significant effect of genre, this time with a higher proportion of complex word types in fiction compared to nonfiction ($\beta = -2.37$, $SE = 0.58$, $t = -4.07$, $p < .001$); a difference of 2.4 percentage points. Finally, there was again a significant effect of document length, with each additional word in length associated with a marginally higher proportion of complex word types ($\beta = 0.00$, $SE = 0.00$, $t = 27.17$, $p < .001$).

The interaction between Key Stage and genre was also significant (LRT: $\chi^2$ [2] = 21.91, $p < .001$). Examination of the coefficients showed that the difference in percentage of complex word types between fiction and nonfiction was significantly smaller in Key Stage 1 texts relative to Key Stage 2 (β

**Table 3.** Standardized chi-square residuals comparing observed and expected word type frequencies for different morphological categories by key stage and genre.

| | | Morphological Category | | |
|---|---|---|---|---|
| | | Compound | Derived | Compound derived |
| Key Stage | Key Stage 1 | 1.18 | 0.21 | −3.87 |
| | Key Stage 2 | 0.14 | −0.24 | 1.08 |
| | Key Stage 3 | −0.96 | 0.09 | 1.64 |
| Genre | Fiction | 0.15 | 0.06 | −0.68 |
| | Non-fiction | −0.16 | −0.07 | 0.70 |

Chi-square residuals represent the difference between the observed and expected values for a given cell.

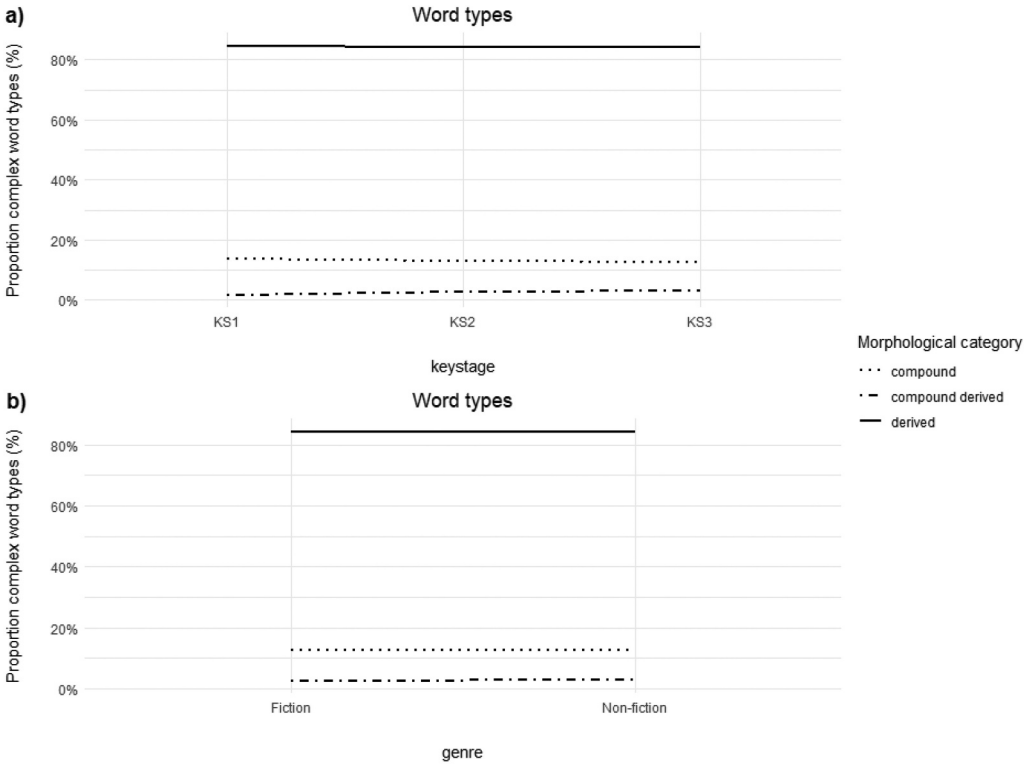


**Figure 3.** Proportion of complex word types by morphological category, plotted by key stage (panel a) and genre (panel b).

= −5.60, *SE* = 1.70, *t* = −3.30, *p* < .001), and larger for Key Stage 2 relative to Key Stage 3 texts (β = 2.96, *SE* = 0.48, *t* = 6.12, *p* < .001).

### *(ii) Do distributions of complex word types (derived, compound, compound derived) vary with target age and genre?*

#### *Token frequencies*

Turning to the distribution of complex word categories, we used chi-square statistics to examine whether the frequency of derived, compound and compound-derived words was associated with Key Stage or genre. For token frequencies, analysis revealed that the distribution of complex word categories was associated with Key Stage ($\chi^2$ [4] = 3945.2, *p* < .001) and with genre ($\chi^2$ [2] = 22304, *p* < .001). Figure 2

**Table 4.** Suffix token frequencies by key stage, normalized per million tokens and per million suffixed tokens.

| Suffix | Suffix frequency per million tokens | | | Suffix frequency per million suffixed tokens | | |
|---|---|---|---|---|---|---|
| | KS1 | KS2 | KS3 | KS1 | KS2 | KS3 |
| able | 480 | 825 | 1078 | 13804 | 16400 | 18897 |
| ac | 1 | 3 | 2 | 35 | 56 | 27 |
| acy | 11 | 20 | 45 | 313 | 401 | 782 |
| ade | 13 | 12 | 8 | 382 | 244 | 141 |
| age | 197 | 239 | 290 | 5668 | 4752 | 5080 |
| aire | 69 | 5 | 6 | 1982 | 107 | 100 |
| al | 462 | 2170 | 3109 | 13282 | 43129 | 54508 |
| an | 12 | 301 | 321 | 348 | 5983 | 5629 |
| ance | 395 | 777 | 1248 | 11370 | 15447 | 21883 |
| ant | 785 | 1609 | 1940 | 22566 | 31984 | 34003 |
| ar | 272 | 511 | 568 | 7823 | 10157 | 9954 |
| ard | 1 | 15 | 7 | 35 | 299 | 122 |
| arian | 7 | 5 | 12 | 209 | 107 | 211 |
| ate | 300 | 827 | 1172 | 8623 | 16438 | 20554 |
| dom | 40 | 129 | 164 | 1147 | 2560 | 2871 |
| ee | 6 | 58 | 67 | 174 | 1161 | 1173 |
| eer | 13 | 56 | 55 | 382 | 1109 | 973 |
| en | 794 | 1036 | 1004 | 22844 | 20597 | 17602 |
| er | 2858 | 3726 | 3581 | 82197 | 74049 | 62788 |
| erie | 28 | 4 | 1 | 800 | 76 | 14 |
| ern | 17 | 218 | 215 | 487 | 4337 | 3767 |
| esque | 2 | 3 | 3 | 70 | 69 | 61 |
| ess | 106 | 293 | 140 | 3060 | 5828 | 2454 |
| est | 57 | 25 | 37 | 1634 | 488 | 649 |
| et | 11 | 37 | 39 | 313 | 736 | 685 |
| ette | 7 | 13 | 62 | 209 | 266 | 1080 |
| eur | 1 | 3 | 6 | 35 | 53 | 102 |
| ful | 1153 | 1246 | 984 | 33171 | 24772 | 17252 |
| hood | 37 | 50 | 84 | 1078 | 985 | 1467 |
| i | 1 | 4 | 22 | 35 | 86 | 388 |
| ia | 4 | 76 | 82 | 104 | 1504 | 1438 |
| ial | 0 | 1 | 6 | 0 | 30 | 111 |
| ian | 140 | 361 | 382 | 4033 | 7176 | 6690 |
| ic | 274 | 636 | 757 | 7893 | 12647 | 13279 |
| ice | 86 | 178 | 259 | 2469 | 3533 | 4540 |
| id | 73 | 101 | 59 | 2086 | 2003 | 1030 |
| ide | 0 | 3 | 10 | 0 | 58 | 170 |
| ie | 73 | 93 | 55 | 2086 | 1856 | 971 |
| ify | 147 | 177 | 246 | 4242 | 3522 | 4311 |
| ile | 0 | 2 | 7 | 0 | 48 | 118 |
| in | 0 | 4 | 9 | 0 | 89 | 154 |
| ine | 57 | 77 | 123 | 1634 | 1532 | 2161 |
| ion | 1399 | 3774 | 5733 | 40229 | 75001 | 100518 |
| ious | 575 | 917 | 988 | 16551 | 18230 | 17316 |
| ish | 189 | 375 | 526 | 5424 | 7451 | 9215 |
| ism | 4 | 69 | 68 | 104 | 1371 | 1188 |
| ison | 2 | 5 | 15 | 70 | 96 | 265 |
| ist | 50 | 297 | 392 | 1426 | 5899 | 6874 |
| ite | 167 | 207 | 239 | 4798 | 4114 | 4182 |
| itis | 0 | 2 | 18 | 0 | 48 | 320 |
| ity | 393 | 1046 | 1606 | 11300 | 20795 | 28163 |
| ium | 19 | 120 | 106 | 556 | 2377 | 1853 |
| ive | 191 | 482 | 657 | 5494 | 9581 | 11510 |
| ize | 37 | 130 | 171 | 1078 | 2578 | 2991 |
| le | 19 | 10 | 13 | 556 | 208 | 220 |
| less | 158 | 304 | 300 | 4555 | 6038 | 5266 |
| let | 2 | 50 | 68 | 70 | 986 | 1188 |
| ling | 29 | 76 | 10 | 834 | 1514 | 168 |
| ly | 8319 | 9780 | 9303 | 239256 | 194365 | 163109 |
| ment | 383 | 948 | 1276 | 11022 | 18850 | 22369 |
| most | 0 | 9 | 7 | 0 | 186 | 116 |
| n | 190 | 386 | 377 | 5459 | 7666 | 6618 |

(*Continued*)

**Table 4.** (Continued).

| Suffix | Suffix frequency per million tokens | | | Suffix frequency per million suffixed tokens | | |
|---|---|---|---|---|---|---|
| | KS1 | KS2 | KS3 | KS1 | KS2 | KS3 |
| ness | 324 | 767 | 966 | 9318 | 15234 | 16933 |
| o | 66 | 58 | 25 | 1912 | 1151 | 435 |
| oid | 0 | 11 | 14 | 0 | 221 | 249 |
| on | 12 | 79 | 68 | 348 | 1579 | 1200 |
| or | 579 | 784 | 801 | 16655 | 15590 | 14045 |
| ory | 387 | 814 | 941 | 11127 | 16176 | 16495 |
| ship | 22 | 93 | 166 | 626 | 1851 | 2908 |
| some | 109 | 51 | 47 | 3129 | 1006 | 819 |
| st | 31 | 55 | 50 | 904 | 1087 | 882 |
| ster | 52 | 12 | 8 | 1495 | 244 | 136 |
| t | 92 | 151 | 217 | 2643 | 3011 | 3806 |
| teen | 81 | 91 | 112 | 2330 | 1805 | 1957 |
| th | 73 | 164 | 278 | 2086 | 3263 | 4878 |
| tude | 2 | 11 | 19 | 70 | 226 | 340 |
| ure | 491 | 829 | 866 | 14117 | 16484 | 15184 |
| ward | 158 | 271 | 264 | 4555 | 5378 | 4629 |
| wise | 25 | 51 | 77 | 730 | 1018 | 1343 |
| y | 6296 | 4482 | 4715 | 181085 | 89070 | 82665 |

plots each morphological category as a proportion of total complex words by Key Stage (panel a) and genre (panel b). Standardized residuals are reported in Table 2, and show that compounds were over-represented in Key Stage 1 and 2 texts and underrepresented in Key Stage 3 texts, while derived and compound-derived words were associated with Key Stage 3 texts and underrepresented at Key Stages 1 and 2. The analysis by genre showed that compounds were strongly associated with fiction and under-represented in nonfiction, while derived and compound-derived words were associated with nonfiction.

### *Type frequencies*

The distribution of complex word type categories was associated with Key Stage ($\chi^2$ [4] = 21.24, $p$ < .001), but not genre ($\chi^2$ [2] = 1.00, $p$ = .61). Standardized residuals for both analyses are reported in Table 3, and proportions of complex word types are plotted in Figure 3. These show that the distribution of complex word type frequencies across morphological categories was much in line with expected values. In the analysis by Key Stage, the largest deviation was an underrepresentation of compound-derived words at Key Stage 1.

### *(iii) How do frequencies of individual suffixes vary with target age and genre?*

Suffix token and type frequencies were highly correlated (all Pearson's coefficients above 0.9, $p$ < .001). We first present token and type frequency analyses for Key Stage before turning to genre.

### *Key stage*

Table 4 shows suffix token frequencies per million words and per million suffixed words by Key Stage for the 80 target derivational suffixes. Table 5 shows raw suffix type frequencies (based on our random sampling method outlined above), and suffix type frequencies per thousand suffixed words.

We ran generalized linear mixed effects models testing the main effect of Key Stage on a) suffix frequency per million words, and b) raw type frequency based on randomly selected size-matched subcorpus samples. Models included by-suffix random intercepts and slopes for the effect of Key Stage. Key Stage was coded using successive differences contrasts from the MASS package (Venables & Ripley, 2002), which allowed us to compare overall suffix frequencies between adjacent Key Stages.

**Table 5.** Suffix type frequencies by key stage, normalized by random sampling and per thousand suffixed word types.

| | Suffix type frequency | | | Suffix type frequency per thousand unique suffixed words | | |
|---|---|---|---|---|---|---|
| Suffix | KS1 | KS2 | KS3 | KS1 | KS2 | KS3 |
| ment | 53 | 77 | 92 | 16.67 | 16.64 | 17.71 |
| ate | 52 | 104 | 129 | 16.35 | 22.44 | 24.89 |
| ion | 206 | 356 | 444 | 64.78 | 76.87 | 85.37 |
| ity | 60 | 111 | 146 | 18.87 | 24 | 28.03 |
| al | 73 | 184 | 235 | 22.96 | 39.69 | 45.17 |
| able | 59 | 100 | 126 | 18.55 | 21.59 | 24.16 |
| ly | 532 | 660 | 664 | 167.3 | 142.5 | 127.77 |
| ance | 58 | 81 | 99 | 18.24 | 17.41 | 18.96 |
| ant | 73 | 104 | 126 | 22.96 | 22.41 | 24.17 |
| ive | 42 | 71 | 92 | 13.21 | 15.25 | 17.63 |
| ia | 1 | 10 | 12 | 0.31 | 2.12 | 2.37 |
| ic | 43 | 80 | 107 | 13.52 | 17.31 | 20.54 |
| y | 359 | 400 | 428 | 112.89 | 86.48 | 82.36 |
| acy | 5 | 7 | 9 | 1.57 | 1.45 | 1.65 |
| an | 4 | 15 | 19 | 1.26 | 3.21 | 3.6 |
| or | 39 | 67 | 77 | 12.26 | 14.41 | 14.8 |
| ess | 4 | 8 | 9 | 1.26 | 1.78 | 1.71 |
| er | 240 | 332 | 353 | 75.47 | 71.6 | 67.99 |
| in | 0 | 1 | 2 | 0 | 0.19 | 0.3 |
| hood | 2 | 8 | 10 | 0.63 | 1.78 | 1.92 |
| ious | 66 | 98 | 112 | 20.75 | 21.09 | 21.62 |
| ure | 22 | 34 | 37 | 6.92 | 7.34 | 7.16 |
| ory | 47 | 72 | 87 | 14.78 | 15.49 | 16.74 |
| ize | 12 | 26 | 35 | 3.77 | 5.66 | 6.74 |
| n | 9 | 23 | 25 | 2.83 | 4.96 | 4.84 |
| ward | 16 | 18 | 18 | 5.03 | 3.93 | 3.48 |
| arian | 1 | 1 | 2 | 0.31 | 0.21 | 0.37 |
| less | 31 | 56 | 56 | 9.75 | 12.07 | 10.85 |
| i | 1 | 1 | 3 | 0.31 | 0.14 | 0.58 |
| ful | 62 | 79 | 82 | 19.5 | 17.07 | 15.79 |
| ine | 4 | 8 | 10 | 1.26 | 1.73 | 1.91 |
| ium | 4 | 14 | 16 | 1.26 | 3.03 | 3.15 |
| st | 3 | 3 | 3 | 0.94 | 0.63 | 0.58 |
| ian | 11 | 24 | 27 | 3.46 | 5.23 | 5.19 |
| ify | 13 | 25 | 30 | 4.09 | 5.46 | 5.68 |
| ist | 4 | 23 | 32 | 1.26 | 4.9 | 6.19 |
| age | 15 | 25 | 29 | 4.72 | 5.46 | 5.53 |
| ar | 19 | 25 | 27 | 5.97 | 5.39 | 5.18 |
| ie | 9 | 8 | 10 | 2.83 | 1.81 | 1.99 |
| wise | 4 | 3 | 3 | 1.26 | 0.71 | 0.53 |
| ism | 2 | 15 | 18 | 0.63 | 3.15 | 3.46 |
| ship | 9 | 16 | 21 | 2.83 | 3.47 | 4.05 |
| let | 1 | 6 | 7 | 0.31 | 1.28 | 1.27 |
| itis | 0 | 1 | 2 | 0 | 0.16 | 0.37 |
| en | 68 | 93 | 91 | 21.38 | 20.08 | 17.48 |
| oid | 0 | 2 | 3 | 0 | 0.4 | 0.54 |
| ee | 4 | 8 | 10 | 1.26 | 1.84 | 1.97 |
| eer | 2 | 5 | 5 | 0.63 | 1.07 | 0.99 |
| ness | 57 | 93 | 104 | 17.92 | 20.18 | 20.1 |
| some | 10 | 10 | 10 | 3.14 | 2.13 | 1.89 |
| ish | 23 | 28 | 29 | 7.23 | 6.03 | 5.56 |
| et | 2 | 6 | 6 | 0.63 | 1.21 | 1.09 |
| ite | 9 | 12 | 14 | 2.83 | 2.52 | 2.71 |
| est | 3 | 6 | 5 | 0.94 | 1.2 | 0.93 |
| aire | 1 | 1 | 2 | 0.31 | 0.26 | 0.33 |
| ade | 1 | 2 | 1 | 0.31 | 0.49 | 0.26 |
| dom | 3 | 5 | 6 | 0.94 | 0.97 | 1.11 |
| most | 0 | 2 | 2 | 0 | 0.34 | 0.3 |
| on | 2 | 6 | 6 | 0.63 | 1.3 | 1.13 |
| t | 6 | 8 | 9 | 1.89 | 1.77 | 1.77 |
| ling | 3 | 5 | 4 | 0.94 | 1.01 | 0.78 |
| ette | 2 | 2 | 3 | 0.63 | 0.4 | 0.61 |

(*Continued*)

**Table 5.** (Continued).

| Suffix | Suffix type frequency | | | Suffix type frequency per thousand unique suffixed words | | |
|---|---|---|---|---|---|---|
| | KS1 | KS2 | KS3 | KS1 | KS2 | KS3 |
| ison | 1 | 1 | 1 | 0.31 | 0.21 | 0.19 |
| ice | 4 | 6 | 6 | 1.26 | 1.22 | 1.15 |
| ial | 0 | 1 | 2 | 0 | 0.2 | 0.42 |
| ard | 1 | 2 | 2 | 0.31 | 0.39 | 0.36 |
| ern | 4 | 6 | 8 | 1.26 | 1.37 | 1.46 |
| th | 9 | 19 | 24 | 2.83 | 4.17 | 4.67 |
| teen | 6 | 6 | 6 | 1.89 | 1.29 | 1.15 |
| ster | 1 | 2 | 3 | 0.31 | 0.51 | 0.49 |
| eur | 1 | 1 | 1 | 0.31 | 0.16 | 0.19 |
| id | 3 | 3 | 3 | 0.94 | 0.66 | 0.62 |
| o | 4 | 3 | 3 | 1.26 | 0.68 | 0.53 |
| ile | 0 | 1 | 2 | 0 | 0.22 | 0.3 |
| le | 3 | 2 | 2 | 0.94 | 0.52 | 0.43 |
| ac | 1 | 1 | 1 | 0.31 | 0.19 | 0.15 |
| erie | 1 | 1 | 0 | 0.31 | 0.15 | 0.08 |
| tude | 1 | 2 | 2 | 0.31 | 0.41 | 0.37 |
| ide | 0 | 0 | 1 | 0 | 0.1 | 0.18 |
| esque | 1 | 1 | 1 | 0.31 | 0.2 | 0.18 |

### *Token Frequencies*

Estimated coefficients from the full model revealed that suffix frequencies per million words increased by approximately 2.41 tokens between Key Stages 1 and 2 ($\beta = 0.88$, SE = 0.12, z = 7.17, $p < .001$) and 1.18 tokens between Key Stages 2 and 3 ($\beta = 0.16$, SE = 0.05, z = 3.18, $p < .01$). To examine whether changes in suffix frequency across Key Stage varied by suffix, we compared the full model described above to a reduced model in which we removed the random slope. This analysis revealed that including by-suffix random slopes for the effect of Key Stage significantly improved model fit (LRT: $\chi^2$ [5] = 7964.8, $p < .001$). Figure 4 shows random intercepts plotted against random slopes for each suffix, split by Key Stage comparison. These plots indicate that the increase in suffix frequency from Key Stage 1 to 2 and from Key Stage 2 to 3 was greater for lower frequency suffixes.

### *Type frequencies*

Suffix type frequencies similarly showed an increase of 1.65 word types between Key Stages 1 and 2 ($\beta = 0.50$, $SE = 0.05$, $z = 10.73$, $p < .001$) and 1.15 word types between Key Stages 2 and 3 ($\beta = 0.14$, $SE = 0.03$, $z = 5.28$, $p < .001$). The inclusion of by-suffix random slopes for the effect of Key Stage significantly improved model fit (LRT: $\chi^2$ [2] = 68.73, $p < .001$). Figure 4 shows random intercepts plotted against random slopes for each suffix, split by Key Stage comparison. These plots show a similar pattern to the token frequency analysis: lower frequency suffixes showed the greatest growth in type frequency between Key Stages 1 and 2 and Key Stages 2 and 3.

### *Genre*

Suffix frequencies by genre were analyzed in the same way. Table 6 shows suffix token frequencies per million words and per million suffixed words, and Table 7 shows raw suffix type frequencies (based on random sampling from fiction documents, which comprised the larger subcorpus for genre) and suffix type frequencies per thousand suffixed word types. Generalized linear mixed effects models were run to test the effect of genre on a) suffix frequency per million words, and b) type frequency. For both analyses, the model included by-suffix random intercepts and slopes for the effect of genre.

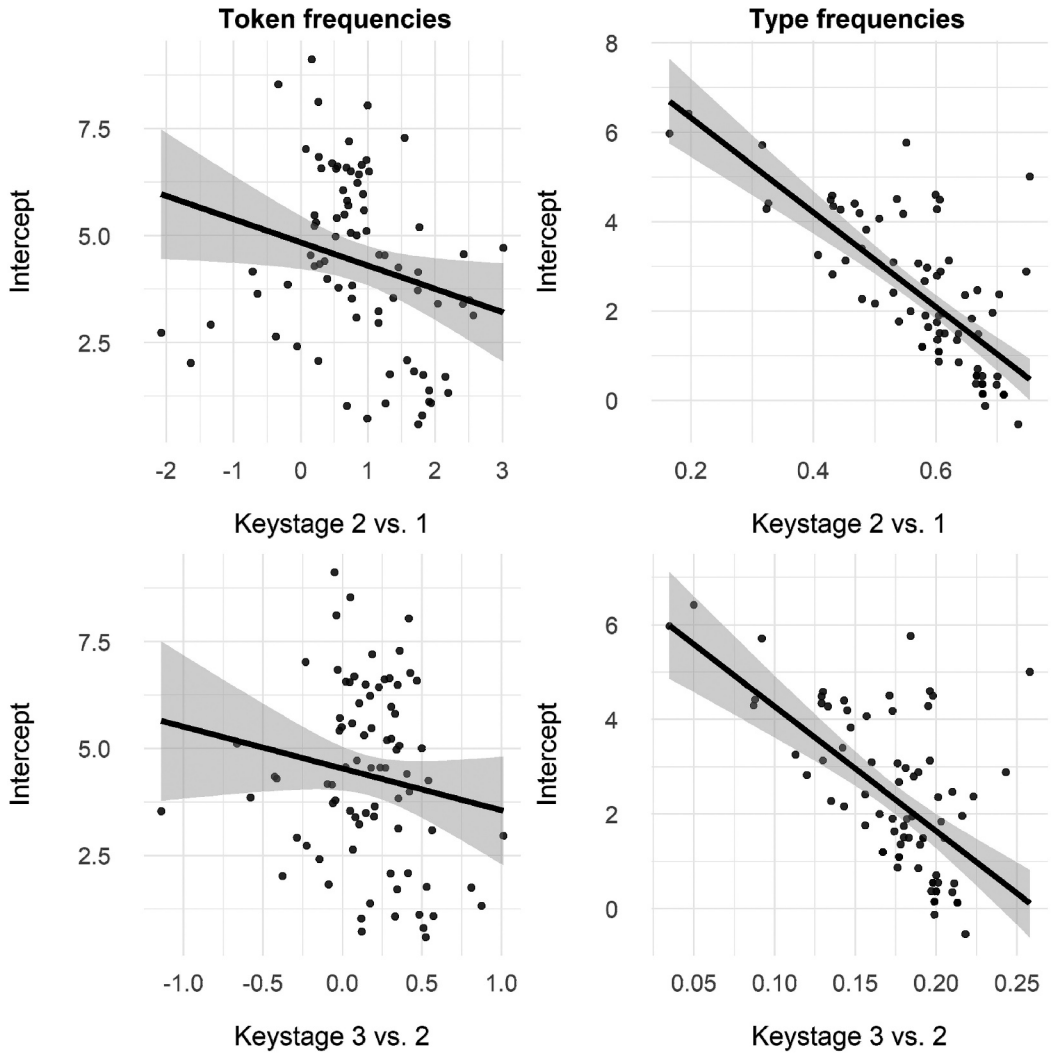


**Figure 4.** Relationship between by-suffix random intercepts and slopes for key stage 2 vs. 1 (top) and key stage 3 vs. 2 (bottom) for suffix token (left panel) and type (right panel) frequencies.

### *Token frequencies*

Estimated coefficients revealed that suffix frequencies per million were greater overall for nonfiction vs. fiction ($\beta = 0.77$, SE = 0.12, $z = 6.33$, $p < .001$), an effect size of 2.15 tokens. Including the by-suffix random slope for the effect of genre significantly improved model fit compared to a reduced model in which the random slope was removed (LRT: $\chi^2$ [2] = 10004, $p < .001$). Figure 5 shows the relationship between intercepts and slopes, indicating that slopes for genre did not vary as a function of overall frequency.

### *Type frequencies*

Suffix type frequencies did not differ significantly between nonfiction and fiction ($\beta = -0.01$, SE = 0.02, $z = -0.41$, $p = .680$). Additionally, the inclusion of by-suffix random slopes did not significantly improve model fit compared to a reduced model without the random slope (LRT: $\chi^2$ [1] = 1.16, $p = .281$).

**Table 6.** Suffix token frequencies by genre, normalized per million tokens and per million suffixed tokens.

| Suffix | Suffix frequency per million tokens | | Suffix frequency per million suffixed tokens | |
|---|---|---|---|---|
| | Fiction | Non-fiction | Fiction | Non-fiction |
| able | 896 | 937 | 20604 | 11752 |
| ac | 3 | 0 | 66 | 5 |
| acy | 24 | 45 | 546 | 570 |
| ade | 10 | 12 | 235 | 154 |
| age | 246 | 289 | 5656 | 3628 |
| aire | 8 | 9 | 176 | 116 |
| al | 1349 | 5942 | 31020 | 74531 |
| an | 50 | 1079 | 1138 | 13529 |
| ance | 864 | 1174 | 19859 | 14729 |
| ant | 1212 | 3243 | 27852 | 40676 |
| ar | 340 | 1101 | 7806 | 13813 |
| ard | 10 | 17 | 226 | 208 |
| arian | 4 | 22 | 81 | 276 |
| ate | 684 | 1731 | 15715 | 21711 |
| dom | 92 | 284 | 2119 | 3565 |
| ee | 35 | 137 | 800 | 1724 |
| eer | 41 | 95 | 944 | 1187 |
| en | 1055 | 887 | 24248 | 11124 |
| er | 2948 | 5813 | 67764 | 72906 |
| erie | 4 | 1 | 100 | 18 |
| ern | 58 | 687 | 1325 | 8613 |
| esque | 4 | 2 | 91 | 23 |
| ess | 201 | 315 | 4612 | 3957 |
| est | 34 | 19 | 787 | 235 |
| et | 37 | 37 | 843 | 463 |
| ette | 21 | 63 | 485 | 792 |
| eur | 4 | 2 | 103 | 20 |
| ful | 1187 | 1009 | 27289 | 12651 |
| hood | 59 | 71 | 1356 | 891 |
| i | 0 | 44 | 10 | 549 |
| ia | 14 | 268 | 323 | 3365 |
| ial | 2 | 6 | 53 | 78 |
| ian | 161 | 988 | 3701 | 12395 |
| ic | 308 | 1802 | 7082 | 22600 |
| ice | 155 | 360 | 3567 | 4522 |
| id | 53 | 180 | 1225 | 2263 |
| ide | 0 | 22 | 0 | 278 |
| ie | 65 | 119 | 1502 | 1494 |
| ify | 189 | 242 | 4349 | 3030 |
| ile | 3 | 7 | 68 | 89 |
| in | 2 | 17 | 57 | 210 |
| ine | 49 | 233 | 1137 | 2924 |
| ion | 3324 | 7856 | 76405 | 98531 |
| ious | 840 | 1215 | 19300 | 15243 |
| ish | 207 | 1110 | 4756 | 13919 |
| ism | 22 | 203 | 514 | 2552 |
| ison | 8 | 12 | 176 | 144 |
| ist | 92 | 1051 | 2118 | 13180 |
| ite | 171 | 364 | 3929 | 4559 |
| itis | 1 | 32 | 16 | 403 |
| ity | 921 | 2208 | 21163 | 27694 |
| ium | 40 | 332 | 921 | 4170 |
| ive | 348 | 1128 | 8006 | 14152 |
| ize | 90 | 303 | 2071 | 3805 |
| le | 13 | 7 | 297 | 94 |
| less | 342 | 154 | 7854 | 1937 |
| let | 44 | 89 | 1006 | 1111 |
| ling | 62 | 10 | 1416 | 132 |
| ly | 9943 | 8278 | 228545 | 103825 |
| ment | 807 | 1813 | 18545 | 22734 |
| most | 9 | 4 | 213 | 48 |
| n | 125 | 1165 | 2864 | 14613 |

(*Continued*)

**Table 6.** (Continued).

| Suffix | Suffix frequency per million tokens | | Suffix frequency per million suffixed tokens | |
|---|---|---|---|---|
| | Fiction | Non-fiction | Fiction | Non-fiction |
| ness | 928 | 492 | 21340 | 6175 |
| o | 17 | 137 | 392 | 1716 |
| oid | 4 | 36 | 95 | 453 |
| on | 52 | 139 | 1187 | 1742 |
| or | 598 | 1366 | 13738 | 17137 |
| ory | 577 | 1690 | 13256 | 21197 |
| ship | 85 | 221 | 1952 | 2767 |
| some | 51 | 52 | 1178 | 656 |
| st | 55 | 44 | 1254 | 557 |
| ster | 13 | 9 | 304 | 111 |
| t | 126 | 324 | 2901 | 4061 |
| teen | 111 | 57 | 2560 | 711 |
| th | 124 | 454 | 2846 | 5699 |
| tude | 18 | 2 | 410 | 25 |
| ure | 554 | 1701 | 12730 | 21332 |
| ward | 289 | 184 | 6638 | 2309 |
| wise | 65 | 43 | 1495 | 537 |
| y | 4199 | 6043 | 96513 | 75790 |

## Discussion

We documented morphological complexity in a large corpus of fiction and nonfiction reading materials targeted at children and young adolescents. Our analyses showed firstly that the proportion of morphologically complex words in texts increased in line with the target age of the text, and that this trend was driven in particular by the proportion of derived words. Patterns by genre were slightly

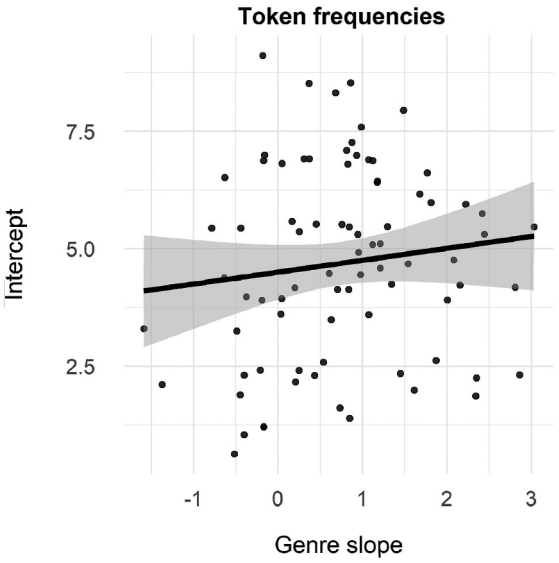


**Figure 5.** Relationship between by-suffix random intercepts and slopes for fiction vs. non-fiction for suffix token frequencies.

**Table 7.** Suffix type frequencies by genre, normalized by random sampling and per thousand suffixed word types.

| Suffix | Suffix type frequency | | Suffix type frequency per thousand unique suffixed words | |
|---|---|---|---|---|
| | Fiction | Non-fiction | Fiction | Non-fiction |
| able | 153 | 131 | 24.11 | 21.12 |
| ac | 1 | 1 | 0.16 | 0.16 |
| acy | 9 | 10 | 1.43 | 1.61 |
| ade | 3 | 3 | 0.46 | 0.48 |
| age | 38 | 36 | 5.93 | 5.8 |
| aire | 3 | 3 | 0.47 | 0.48 |
| al | 266 | 291 | 42.05 | 46.91 |
| an | 14 | 32 | 2.2 | 5.16 |
| ance | 113 | 104 | 17.82 | 16.77 |
| ant | 152 | 143 | 23.93 | 23.05 |
| ar | 28 | 30 | 4.38 | 4.84 |
| ard | 3 | 3 | 0.46 | 0.48 |
| arian | 2 | 2 | 0.32 | 0.32 |
| ate | 155 | 160 | 24.41 | 25.79 |
| dom | 7 | 6 | 1.03 | 0.97 |
| ee | 12 | 16 | 1.95 | 2.58 |
| eer | 8 | 8 | 1.21 | 1.29 |
| en | 115 | 101 | 18.16 | 16.28 |
| er | 488 | 494 | 77.11 | 79.64 |
| erie | 1 | 1 | 0.16 | 0.16 |
| ern | 7 | 8 | 1.16 | 1.29 |
| esque | 1 | 1 | 0.16 | 0.16 |
| ess | 12 | 12 | 1.95 | 1.93 |
| est | 8 | 5 | 1.18 | 0.81 |
| et | 10 | 8 | 1.55 | 1.29 |
| ette | 4 | 4 | 0.56 | 0.64 |
| eur | 1 | 1 | 0.16 | 0.16 |
| ful | 97 | 80 | 15.27 | 12.9 |
| hood | 13 | 11 | 2 | 1.77 |
| i | 1 | 4 | 0.09 | 0.64 |
| ia | 12 | 16 | 1.91 | 2.58 |
| ial | 3 | 3 | 0.44 | 0.48 |
| ian | 31 | 39 | 4.95 | 6.29 |
| ic | 115 | 136 | 18.09 | 21.92 |
| ice | 6 | 7 | 0.95 | 1.13 |
| id | 3 | 4 | 0.47 | 0.64 |
| ide | 0 | 1 | 0 | 0.16 |
| ie | 14 | 15 | 2.14 | 2.42 |
| ify | 35 | 33 | 5.48 | 5.32 |
| ile | 3 | 2 | 0.43 | 0.32 |
| in | 1 | 2 | 0.16 | 0.32 |
| ine | 11 | 12 | 1.78 | 1.93 |
| ion | 505 | 518 | 79.74 | 83.51 |
| ious | 134 | 114 | 21.19 | 18.38 |
| ish | 37 | 34 | 5.8 | 5.48 |
| ism | 22 | 35 | 3.41 | 5.64 |
| ison | 1 | 1 | 0.16 | 0.16 |
| ist | 35 | 50 | 5.52 | 8.06 |
| ite | 15 | 20 | 2.38 | 3.22 |
| itis | 2 | 3 | 0.33 | 0.48 |
| ity | 168 | 163 | 26.46 | 26.28 |
| ium | 19 | 21 | 3.04 | 3.39 |
| ive | 106 | 104 | 16.8 | 16.77 |
| ize | 45 | 52 | 7.06 | 8.38 |
| le | 3 | 3 | 0.47 | 0.48 |
| less | 72 | 58 | 11.32 | 9.35 |
| let | 8 | 7 | 1.19 | 1.13 |
| ling | 9 | 7 | 1.47 | 1.13 |
| ly | 817 | 701 | 128.92 | 113.01 |
| ment | 103 | 105 | 16.32 | 16.93 |
| most | 2 | 2 | 0.36 | 0.32 |
| n | 25 | 43 | 4 | 6.93 |

(*Continued*)

**Table 7.** (Continued).

| Suffix | Suffix type frequency | | Suffix type frequency per thousand unique suffixed words | |
|---|---|---|---|---|
| | Fiction | Non-fiction | Fiction | Non-fiction |
| ness | 125 | 108 | 19.8 | 17.41 |
| o | 4 | 4 | 0.62 | 0.64 |
| oid | 2 | 4 | 0.35 | 0.64 |
| on | 6 | 10 | 0.98 | 1.61 |
| or | 95 | 103 | 15.01 | 16.6 |
| ory | 102 | 108 | 16.13 | 17.41 |
| ship | 24 | 25 | 3.74 | 4.03 |
| some | 13 | 12 | 1.99 | 1.93 |
| st | 3 | 3 | 0.47 | 0.48 |
| ster | 4 | 3 | 0.62 | 0.48 |
| t | 11 | 11 | 1.7 | 1.77 |
| teen | 6 | 6 | 0.95 | 0.97 |
| th | 24 | 27 | 3.78 | 4.35 |
| tude | 2 | 2 | 0.32 | 0.32 |
| ure | 43 | 42 | 6.77 | 6.77 |
| ward | 21 | 21 | 3.29 | 3.39 |
| wise | 4 | 4 | 0.63 | 0.64 |
| y | 539 | 506 | 85.07 | 81.57 |

more complex: nonfiction contained a higher proportion of complex words overall relative to fiction, and this was again mainly attributable to the distributions of derived words. However, as a proportion of all unique word types, fiction contained a higher proportion of complex words compared to nonfiction, and genre was not associated with type frequency distributions across different complex word categories (compound, derived, compound-derived). Secondly, we examined the frequency distributions of 80 derivational suffixes, showing that the distribution of suffix categories was associated with both Key Stage and genre. Below we discuss our findings in relation to developmental trends and genre before turning to the broader implications of these statistics for the role of print exposure in establishing orthography-meaning links.

The increase in proportion of complex words (specifically derivations) with intended target age reflects the more advanced properties of these word types. The number of morphemes in a word is associated with other lexical statistics such as frequency, concreteness (how abstract a word is) and age of acquisition: words containing more morphemes occur less frequently and are rated as more abstract and later acquired (see Supplemental Materials for details of these analyses: https://osf.io/4xurw/?view_only=acd838535db344d3a78911f53b1b9e9c). Derived words in particular are more abstract and have a higher age of acquisition relative to other complex word types, such as compounds, even though compounds are less frequent. Derivational suffixes frequently result in changes to word class, which offers greater flexibility in how complex ideas and information are expressed, and often corresponds to denser information content in more advanced or formal texts. The increase in morphologically complex words across Key Stage is therefore unsurprising, given that texts targeted at older readers contain a higher proportion of longer and more sophisticated word types (Kearns et al., 2014; Nippold, 2018). These morphological characteristics of written language parallel developmental trends in children's morphological knowledge when measured in the context of vocabulary and morphological awareness, with understanding of derivational relationships emerging more slowly relative to knowledge of inflections and compounds (Anglin, 1993; Berko, 1958; Berninger et al., 2010; Carlisle, 1988; Nagy et al., 1993). Linguistic experience may be key to developing sensitivity to overlaps in form and meaning between morphologically related words, and particularly so for derived words, which vary in semantic, phonological, and orthographic transparency and incorporate a much larger set of affixes relative to inflection (Lieber, 2004; Reichle & Perfetti, 2003). However, the growth in derivation we report here is not universal across languages: in their examination of German children's texts, Segbers

and Schroeder (2017) found that compounds increased most as lexicon size grew and formed the most dominant morphological category for older age groups.

Aside from growth in children's understanding and use of word formation processes more generally, exposure to written morphology has the potential to highlight form-meaning links that are less salient in spoken language (Rastle, 2019a, 2019b). In particular, work by Berg and Aronoff (2017) and Ulicheva et al. (2020) shows that while pronunciations of derivational suffixes may overlap with non-morphological word endings (e.g., *bonus*, *nervous*), their spellings often diverge, such that morpho-orthographic information provides more reliable cues to meaning than morpho-phonemic information. Given these considerations, our second aim was to document derivational suffix distributions in children's texts as a function of developmental stage and genre. We found that suffix token frequencies varied by both target Key Stage and genre. Comparing suffix frequencies normalized by number of suffixed words and word types (right-hand columns of Tables 4–7) allows us to examine which suffix categories are disproportionately associated with a given Key Stage or genre, accounting for differences in overall complex word frequency. Suffixes most associated with texts for younger children were typically neutral (they attach to free-standing base words and do not alter the pronunciation of the base; Tyler & Nagy, 1989) and productive (they attach to a wide range of base words; Sánchez-Gutiérrez et al., 2018). Examples include *–ly*, which forms adverbs (*quickly*), and *–y*, which has several functions, including the formation of adjectives (*sleepy*) and diminutives (*doggy*; see also *–ie*). These suffixes are among those showing the most growth in children's spontaneous language production between 2 and 5 years, and some of the most common in caregiver speech (Laws, 2019).

Other high-frequency suffixes representative of Key Stage 1 texts were *–ful*, which typically forms adjectives (*playful*) but also nouns (*mouthful*), and *–er*, which forms nouns. While *–er* was among the suffixes showing significant expansion in children's production between 2 and 5 years in Laws' (Laws, 2019) dataset, *–ful* was not, and nor was it one of the suffixes that increased significantly in caregiver speech. While *–ful* is both neutral and productive, it is possible that it occurs more commonly in written compared to spoken language. Indeed, recent corpus comparisons of spoken and written language indicate that adjectives occur more frequently in writing, even in texts targeted at pre-school children (Dawson, Hsiao, et al., 2021). While noun-forming suffixes *–aire* and *–ster* were also common in Key Stage 1 texts, examination of their overall type frequencies indicated that these suffixes only appeared in a limited range of contexts (namely *millionaire*, *billionaire, questionnaire*; *gangster*, *spinster*, *trickster*, *youngster*).

By contrast, high-frequency suffixes occurring commonly in Key Stage 3 texts were *–ion* (often used in nominalization of verbs, e.g., *act-action*), *–ance*, *–ness*, and *–ity*, which also form nouns (e.g., *disturbance*, *happiness* and *security*, respectively) and adjective-forming *–able* (e.g., *reliable*). These suffixes are mostly typical of more advanced vocabulary. In particular, *–ion* and *–ity* are less transparent in that they often attach to bound morphemes, frequently result in a pronunciation shift in the stem, and tend to refer to abstract concepts (Biber et al., 1998; Laws, 2019; Tyler & Nagy, 1989). These examples are of Latinate origin, and typically produce nominalized forms of verbs and adjectives (e.g., *expression*, *sincerity*). Nominalizations are around four times more common in academic prose relative to fiction and speech (Biber et al., 1998) and reflect a more abstract and depersonalized style (e.g., *she expressed her frustration* vs. *an expression of frustration*). Meanwhile, *–ness* and *–able* are examples of neutral, productive suffixes that are nevertheless representative of more advanced texts. The addition of *–ness* forms abstract nouns, often relating to personal qualities or states of mind (e.g., *awareness*, *willingness*, *shyness*). The affix *–able* may be conceptually more advanced than the adjectival suffixes most strongly associated with texts targeted at younger children (e.g., *–ful* and *–y*), as it typically attaches to verbs and refers to the capacity of being subjected to the action denoted by the stem (*inhabitable*, *predictable*, *valuable*, *enjoyable*), rather than being characterized by the property of the stem (e.g., *playful*, *sleepy*). Note that both *–ness* and *–able* are underrepresented in children's speech relative to caregiver and adult production, and underrepresented in caregiver speech compared to the adult baseline (Laws, 2019).

Our finding that individual suffix frequencies vary in accordance with the target age range of the text has implications for studies of morphological processing in reading development. This is particularly so in the context of the unique relationships between suffix spellings and their meanings: if exposure to written morphology facilitates the development of morpho-orthographic representations, then how those suffixes are represented in texts at different stages of reading development should have a bearing on children's processing of words containing those suffixes. For example, a suffix such as *–ion* shows a stepwise increase in frequency across Key Stage, and it is also highly diagnostic, meaning that words ending in *–ion* will almost always be an noun (Ulicheva et al., 2020). Therefore, we may predict that morphological effects for words containing *–ion* will emerge at a later stage in reading development compared to morphological effects for words containing a suffix like *–aire*, which is also diagnostic of nouns, but is proportionately more common in texts aimed at younger children than older children. Based on input from spoken language, the function of some suffixes (e.g., *–ist; –ous*) may be comparatively opaque given that their phonological form is shared with other suffixes or word endings forming alternate parts of speech (e.g., *longest; bonus*). Seeing *–ist* or *–ous* in written form disambiguates similar-sounding word endings, but clearly this will depend on opportunities to encounter words containing these suffixes in print. Whether these predictions play out in children's reading behavior has yet to be directly tested, but evidence from the spelling literature indicates that younger children often spell complex words phonetically (e.g., *kist* for *kissed*), while consistent and appropriate use of morphological spellings develops over time (Nunes et al., 1997).

Our analysis of genre indicates that the types of texts children read may also have a bearing on their experiences of written morphology. Nonfiction texts contained a higher proportion of complex word tokens relative to fiction, and again this was attributable to the frequency of derived words. The suffixes most typical of fiction tended to overlap with those occurring commonly in texts targeted at younger children (for example, *–ly*, *–y*, and *–ful*), although this trend was stronger for token frequencies than for type frequencies. Two exceptions to the above pattern were *–ance* and *–ness*. While these suffixes were associated with fiction (*–ness* in particular), they were also underrepresented in texts targeted at younger children. These are used in the formation of abstract nouns, often referring to personal qualities or mental states (e.g., *dominance*, *annoyance*, *happiness*, *politeness*) which predominate in fiction. As described above, more formal registers tend to adopt alternate nominalized forms (Biber et al., 1998; Snow, 2010). Many of the suffixes associated with nonfiction are characteristic of academic language (e.g., *–al*, *–ate*, *–ic*, *–ist*, *–ism*, *–ide*, *–itis*), deriving from Latinate forms and often producing nominalized forms of verbs. In contrast to many suffixes seen in fiction texts, these often produce changes to the phonology and stress pattern of the base word (*atom – atomic; produce – production*; Tyler & Nagy, 1989). They also include many examples of bound morphemes as base words (*quant – quantify*), such that the meaning of the complex word is less transparently a sum of its parts. While we found that words containing these suffixes occurred more frequently in nonfiction overall, the number of unique word types did not differ significantly across genre, likely reflecting the limited productivity of these suffix categories.

The diverging properties of neutral and nonneutral suffixes are reflected in children's learning. Studies indicate that neutral suffixes undergo a period of overgeneralization, while the same developmental trend is not observed for nonneutral suffixes (Tyler & Nagy, 1989). For example, before distributional knowledge is fully acquired, children will accept nonwords in which a neutral suffix combines with an incorrect part of speech (e.g., *snapness*), but do not accept an equivalent incorrect combination featuring a nonnetural suffix (e.g., *wheelic*). Given these different developmental trajectories, it is interesting to note that our data indicated that differences by genre may depend on the target age of the texts. In Key Stage 1 texts, the percentage of complex word types (as a proportion of all unique words) was roughly equivalent across fiction and nonfiction, while frequency differences by genre emerged in Key Stages 2 and 3 texts. This suggests that for Key Stage 1 texts only, nonfiction looks quite similar to fiction in terms of how many unique complex words a child is likely to encounter. One important caveat to this finding is that the nonfiction subsample of the Key Stage 1 texts consisted of only 14 documents that were all part of the same book series, albeit on different

topics, so further work is needed to confirm that the patterns we observed are not down to idiosyncrasies of these particular texts.

### Limitations

The present study has several limitations that should be taken into consideration. Firstly, our analysis did not take into account words such as *corner* which contain the orthographic form of a suffix (*–er*). While such words are not true-derived forms, skilled readers are sensitive to their surface morphological composition (Rastle et al., 2004). Secondly, while our analysis of morphological complexity covers a substantial developmental period, Key Stage 3 is the equivalent of up to 14 years of age, which is far from the end point of written linguistic experience. This means that the patterns we report may not generalize to texts targeted at older students, and we will not have captured any additional features that emerge beyond this point. Although the original corpus included data from Key Stage 4 texts, this comprised only lengthy fiction texts, so comparisons by genre were not possible and comparisons by Key Stage were confounded with differences across genre. The Key Stage 1–3 corpus reported in this study did contain both fiction and nonfiction at each Key Stage, but these were not evenly balanced (for example, Key Stage 1 contained only 14 nonfiction documents). While we used normalized frequencies to account for variation in size of these subcorpora, the imbalance in genre representation across Key Stage limits our conclusions about the relative contribution of target age and genre to morphological complexity in children's texts.

These issues reflect a wider point on the utility and limitations of large-scale naturally occurring language datasets. There are many advantages to this approach, such as the size and representativeness of the sample, and the potential to interrogate language use to make predictions about language processing (Jackson et al., 2022). However, many of these advantages translate to disadvantages in relation to the amount of control a researcher has over the content and composition of the language they analyze, and this can mean a trade-off with fine-grained accuracy and interpretation. Additionally, while corpus-based approaches reveal general trends in language use across different sources, this is only an approximation of human experience and it cannot account for individual-level variation. By combining analysis of natural language data with experimental approaches, we can achieve a more nuanced understanding of how language input shapes language development.

## Conclusions

In summary, this large-scale study of morphological complexity in texts targeted at children and adolescents aged 5–14 years reveals an increase in the proportion of complex words in line with the target age of the text, and also highlights differences across genre. In both instances, the frequency of derived words was the main driving force behind this variation, which parallels the protracted development of derivational knowledge, evident in both morphological awareness (Anglin, 1993; Nippold & Sun, 2008) and processing (Dawson et al., 2018) tasks in English. Our analyses of complex word and individual suffix frequencies offer a window into children's potential experience of written morphology at different developmental stages and across different types of text. However, further work is needed to link these patterns to reading behavior and development. If access to written morphology does indeed play an important role in establishing direct links between spelling and meaning, then it should be possible to link suffix frequencies from different developmental stages of the corpus with children's word recognition performance at those ages. Furthermore, these developmental patterns differ across languages (Segbers & Schroeder, 2017), which is an important consideration in understanding cross-linguistic differences in the development of morphological processing (Beyersmann et al., 2012; Dawson, Rastle, et al., 2021; Quémart et al., 2011) and

making direct comparisons across those languages (Beyersmann et al., 2020; Casalis et al., 2015; Mousikou et al., 2020). Finally, while our corpus-based approach details how morphology is represented in children's texts, it cannot fully capture how children's reading is shaped by exposure to this input. In particular, while it is highly probable that children will regularly encounter high-frequency written suffixes, there is likely to be substantial variation in exposure to suffixes occurring more rarely in print, depending on a child's reading preferences and habits. Recent developments in corpus-based methods have addressed this intersection between text and individual characteristics by proposing an alternate measure of frequency that estimates prevalence, thus capturing the likelihood that an individual will have encountered a given word or structure (Johns et al., 2020). While this approach has yet to be applied to suffix frequencies, combined with experimental data, it could further our understanding of how variation in an individual's exposure to written morphology combines with properties of the texts themselves to support the emergence of skilled reading behavior.

## Acknowledgments

The work for this paper was supported by a grant from the Nuffield Foundation (EDO/43392) to Kate Nation, and resources made available to Nilanjana Banerji by the department of Children's Dictionaries and Children's Language Data at Oxford University Press. Yaling Hsiao is supported by a British Academy Post-Doctoral Fellowship (PF2/180013). We would like to thank William Thurwell for his contribution to coding the morphology database.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The work was supported by the British Academy [PF2/180013]; Nuffield Foundation [EDO/43392]

## References

Amenta, S., & Crepaldi, D. (2012). Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in Psychology*, *3*(JUL), 1–12. https://doi.org/10.3389/fpsyg.2012.00232

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58*(10), 58(10. https://doi.org/10.2307/1166112

Baayen, R. H. (2008). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 899–919). Walter de Gruyter.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Berg, K., & Aronoff, M. (2017). Self-organization in the spelling of english suffixes: The emergence of culture out of anarchy. *Language*, *93*(1), 37–64. https://doi.org/10.1353/lan.2017.0000

Berko, J. (1958). The child's learning of english morphology. *WORD*, *14*(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

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*(2), 141–163. https://doi.org/10.1007/s10936-009-9130-6

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*(3), 390–405. https://doi.org/10.1006/jmla.1999.2681

Beyersmann, E., Castles, A., & Coltheart, M. (2012). Morphological processing during visual word recognition in developing readers: Evidence from masked priming. *The Quarterly Journal of Experimental Psychology*, *65*(7), 1306–1326. https://doi.org/10.1080/17470218.2012.656661

Beyersmann, E., Mousikou, P., Javourey-Drevet, L., Schroeder, S., Ziegler, J. C., & Grainger, J. (2020). Morphological processing across modalities and languages. *Scientific Studies of Reading*, *24*(6), 500–519. https://doi.org/10.1080/10888438.2020.1730847

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*(JUL), 1–11. https://doi.org/10.3389/fpsyg.2016.01116

Burani, C., Marcolini, S., & Stella, G. (2002). How early does morpholexical reading develop in readers of a shallow orthography? *Brain and Language*, *81*(1–3), 568–586. https://doi.org/10.1006/brln.2001.2548

Burani, C., & Thornton, A. M. (2003). The interplay of root, suffix and whole-word frequency in processing derived words. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 157–208). Mouton de Gruyter. https://doi.org/10.1515/9783110910186.157

Carlisle, J. F. (1988). Knowledge of derivational morphology and spelling ability in fourth, sixth, and eighth graders. *Applied Psycholinguistics*, *9*(3), 247–266. https://doi.org/10.1017/S0142716400007839

Carstairs McCarthy, A. (2002). *An introduction to english morphology: Words and their structure*. Edinburgh University Press.

Casalis, S., Quémart, P., & Duncan, L. G. (2015). How language affects children's use of derivational morphology in visual word and pseudoword processing: Evidence from a cross-language study. *Frontiers in Psychology*, *6*(MAR), 1–10. https://doi.org/10.3389/fpsyg.2015.00452

Crepaldi, D., Rastle, K., & Davis, C. J. (2010). Morphemes in their place: Evidence for position-specific identification of suffixes. *Memory & Cognition*, *38*(3), 312–321. https://doi.org/10.3758/MC.38.3.312

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*. https://doi.org/10.34842/5we1-yk94

Dawson, N., Rastle, K., & Ricketts, J. (2018). Morphological effects in visual word recognition: Children, adolescents, and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 645–654. https://doi.org/10.1037/xlm0000485

Dawson, N., Rastle, K., & Ricketts, J. (2021). Finding the man amongst many: A developmental perspective on mechanisms of morphological decomposition. *Cognition*, *211*(January), 1–15. https://doi.org/10.1016/j.cognition.2021.104605

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*(4–5), 329–365. https://doi.org/10.1080/01690960050119625

Duke, N. K. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. *Reading Research Quarterly*, *35*(2), 202–224. https://doi.org/10.1598/RRQ.35.2.1

Ford, M. A., Davis, M. H., & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, *63*(1), 117–130. https://doi.org/10.1016/j.jml.2009.01.003

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Grainger, J., & Beyersmann, E. (2017). Edge-Aligned embedded word activation initiates morpho-orthographic segmentation. *Psychology of Learning and Motivation*, *67*. https://doi.org/10.1016/bs.plm.2017.03.009

Hasenäcker, J., Beyersmann, E., & Schroeder, S. (2016). Masked morphological priming in German-speaking adults and children: Evidence from response time distributions. *Frontiers in Psychology*, *7*(JUN), 1–11. https://doi.org/10.3389/fpsyg.2016.00929

Hasenäcker, J., Schröter, P., & Schroeder, S. (2017). Investigating developmental trajectories of morphemes as reading units in German. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1093–1108. https://doi.org/10.1037/xlm0000353

Jackson, J. C., Watts, J., List, J. -M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *The Quarterly Journal of Experimental Psychology*, *73*(6), 841–855. https://doi.org/10.1177/1747021819897560

Kearns, D. M., Steacy, L. M., Compton, D. L., Gilbert, J. K., Goodwin, A. P., Cho, E., Lindstrom, E. R., & Collins, A. A. (2014). Modeling polymorphemic word recognition: exploring differences among children with early-emerging and late-emerging word reading difficulty. *Journal of Learning Disabilities*, *49*(4), 368–394. https://doi.org/10.1177/0022219414554229

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading polymorphemic dutch compounds: Toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 876–895. https://doi.org/10.1037/a0013484

Laws, J. (2019). Profiling complex word usage in the speech of preschool children: Frequency patterns and transparency characteristics. *First Language*, *39*(6), 593–617. https://doi.org/10.1177/0142723719872669

Lázaro, M., Camacho, L., & Burani, C. (2013). Morphological processing in reading disabled and skilled spanish children. *Dyslexia*, *19*(3), 178–188. https://doi.org/10.1002/dys.1458

Lieber, R. (2004). *Morphology and lexical semantics*. Cambridge University Press.

Mousikou, P., Beyersmann, E., Ktori, M., Javourey-Drevet, L., Crepaldi, D., Ziegler, J. C., Grainger, J., & Schroeder, S. (2020). Orthographic consistency influences morphological processing in reading aloud: Evidence from a cross-linguistic study. *Developmental Science*, *23*(6), 1–19. https://doi.org/10.1111/desc.12952

Nagy, W. E., & Anderson, R. C. (1984). How Many Words Are There in Printed School English? *Reading Research Quarterly*, *19*(3), 304. doi:10.2307/747823.

Nagy, W. E., Diakidoy, I. -A., & Anderson, R. (1993). The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of Literacy Research*, *25*(2), 155–170. https://doi.org/10.1080/10862969309547808

Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill. *Npj Science of Learning*, *2*(1), 1–4. https://doi.org/10.1038/s41539-017-0004-7

Nation, K., Dawson, N., & Hsiao, Y. (2022). Book language and its implications for children's language, literacy, and development. *Current Directions in Psychological Science*, *31*(4), 375–380. https://doi.org/10.1177/09637214221103264

Nippold, M. A. (2016). Later language development: School-age children, adolescents, and young adults (4th ed.). Pro Ed.

Nippold, M. A. (2018). The literate lexicon in adolescents: Monitoring the use and understanding of morphologically complex words. *Perspectives of the ASHA Special Interest Groups*, *3*(1), 211–221. https://doi.org/10.1044/persp3.sig1.211

Nippold, M. A., & Sun, L. (2008). Knowledge of morphologically complex words: A developmental study of older children and young adolescents. *Language, Speech, and Hearing Services in Schools*, *39*(3), 365–373. https://doi.org/10.1044/0161-1461(2008/034)

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology*, *33*(4), 637–649. https://doi.org/10.1037/0012-1649.33.4.637

Plag, I., Dalton-Puffer, C., & Baayen, H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, *3*(2), 209–228. https://doi.org/10.1017/S1360674399000222

Quémart, P., Casalis, S., & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of Experimental Child Psychology*, *109*(4), 478–496. https://doi.org/10.1016/j.jecp.2011.02.008

Rastle, K. (2019a). EPS mid-career prize lecture 2017: Writing systems, reading, and language. *The Quarterly Journal of Experimental Psychology*, *72*(4), 677–692. https://doi.org/10.1177/1747021819829696

Rastle, K. (2019b). The place of morphology in learning to read in english. *Cortex*, *116*, 45–54. https://doi.org/10.1016/j.cortex.2018.02.008

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*(7/8), 942–971. https://doi.org/10.1080/01690960802069730

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090–1098. https://doi.org/10.3758/BF03196742

Reichle, E. D., & Perfetti, C. (2003). Morphology in word identification: A word- experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading*, *7*(3), 219–237. https://doi.org/10.1207/S1532799XSSR0703_2

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 english words. *Behavior Research Methods*, *50*(4), 1568–1580. https://doi.org/10.3758/s13428-017-0981-8

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 131–154). Lawrence Erlbaum.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*(1), 118–139. https://doi.org/10.1006/jmla.1997.2510

Segbers, J., & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing*, *34*(3), 297–320. https://doi.org/10.1177/0265532216641152

Snow, C. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*(5977), 450–452. https://doi.org/10.1126/science.1182597

Taft, M. (1979). Recognition of affixedwords and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272. https://doi.org/10.3758/BF03197599

Taft, M., & Ardasinski, S. (2006). Obligatory decomposition in reading prefixed words. *The Mental Lexicon*, *1*(2), 183–199. https://doi.org/10.1075/ml.1.2.02taf

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39. https://doi.org/10.1016/j.cogpsych.2015.03.003

Tyler, A., & Nagy, W. (1989). The acquisition of english derivational morphology. *Journal of Memory & Language*, *28*(6), 649–667. https://doi.org/10.1016/0749-596X(89)90002-8

Ulicheva, A., Harvey, H., Aronoff, M., & Rastle, K. (2020). Skilled readers' sensitivity to meaningful regularities in english writing. *Cognition*, *195*, 103810. https://doi.org/10.1016/j.cognition.2018.09.013

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (*Fourth edition* ed.). Springer.

Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The oxford children's corpus: Using a children's corpus in lexicography. *International Journal of Lexicography*, *26*(2), 190–218. https://doi.org/10.1093/ijl/ecs017

Xu, J., & Taft, M. (2015). The effects of semantic transparency and base frequency on the recognition of english complex words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 904–910. https://doi.org/10.1037/xlm0000052

Yopp, R. H., & Yopp, H. K. (2006). Informational texts as read-alouds at school and home. *Journal of Literacy Research*, *38*(1), 37–51. https://doi.org/10.1207/s15548430jlr3801_2

# Appendix

## Appendix A: Characteristics of subcorpora with and without Key Stage information

| Measure | Genre | Subcorpus with Key Stage data | Subcorpus without Key Stage data |
|---|---|---|---|
| Document count | Fiction | 895 | 1,313 |
| | Non-fiction | 8,704 | 10,169 |
| Word count | Fiction | 17,405,759 | 17,760,303 |
| | Non-fiction | 4,786,455 | 5,989,650 |
| % complex lemma tokens | Fiction | 7.61 | 8.92 |
| | Non-fiction | 11.91 | 12.82 |

## Appendix B: Target Suffixes with Example Words

| Suffix | Example word |
|---|---|
| able | reliable |
| ac | maniac |
| acy | accuracy |
| ade | blockade |
| age | drainage |
| aire | millionaire |
| al | regional |
| an | European |
| ance | disturbance |
| ant | vigilant |
| ar | similar |
| ard | Spaniard |
| arian | vegetarian |
| ate | originate |
| dom | freedom |
| ee | employee |
| eer | engineer |
| en | frighten |
| er | teacher |
| erie | menagerie |
| ern | northern |
| esque | grotesque |
| ess | lioness |
| est | darkest |
| et | packet |
| ette | statuette |
| eur | grandeur |
| ful | playful |
| hood | childhood |
| i | alkali |
| ia | academia |
| ial | tutorial |
| ian | civilian |
| ic | magnetic |
| ice | cowardice |
| id | vivid |
| ide | oxide |
| ie | auntie |
| ify | classify |
| ile | juvenile |
| in | insulin |
| ine | heroine |
| ion | action |
| ish | childish |

(*Continued*)

| Suffix | Example word |
|---|---|
| ism | symbolism |
| ison | comparison |
| ist | cyclist |
| ite | favourite |
| itis | bronchitis |
| ity | security |
| ium | aquarium |
| ive | active |
| ize/ise | organize |
| le | nozzle |
| less | endless |
| let | booklet |
| ling | duckling |
| ly | quickly |
| ment | agreement |
| most | foremost |
| n | African |
| ness | happiness |
| o | volcano |
| oid | asteroid |
| on | electron |
| or | actor |
| ory | directory |
| ous/ious | nervous |
| ship | partnership |
| some | tiresome |
| st | amongst |
| ster | youngster |
| t | joint |
| teen | sixteen |
| th | growth |
| tude | multitude |
| ure | moisture |
| ward | outward |
| wise | likewise |
| y | sleepy |