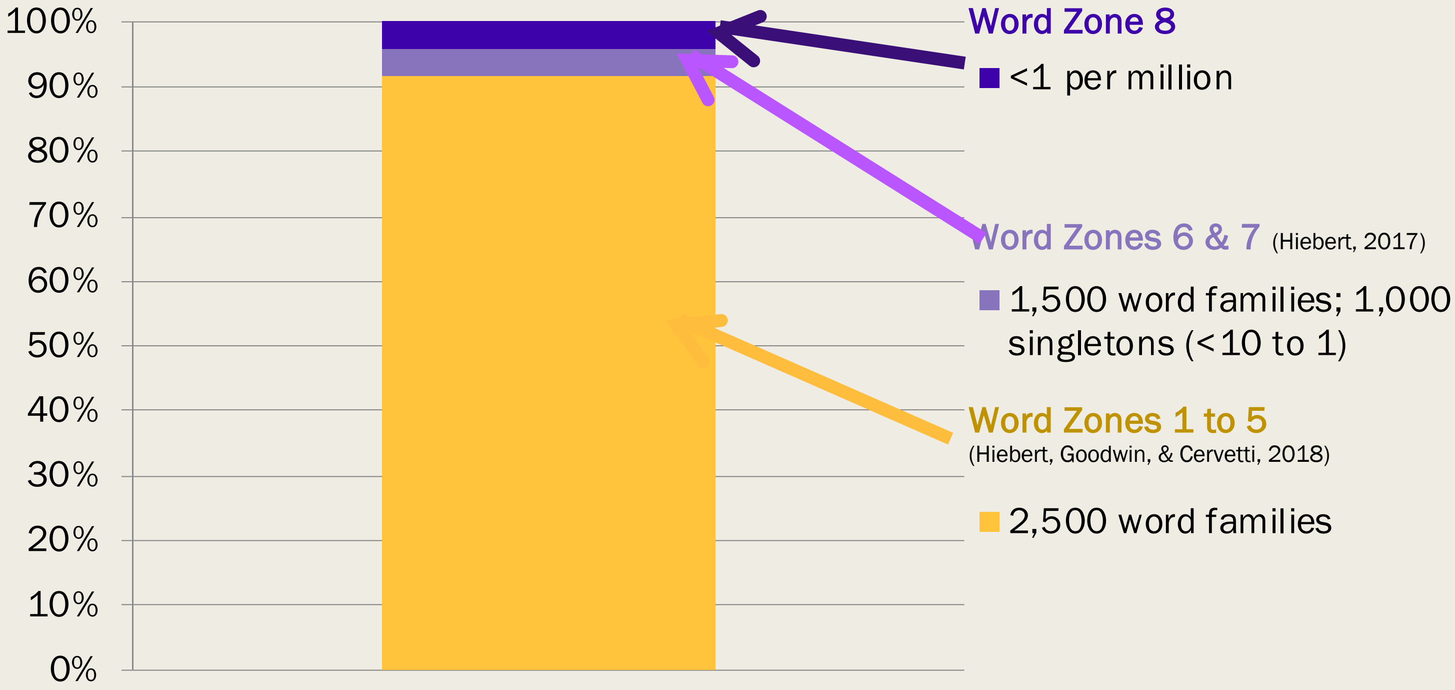
WHAT IS THE TASK REPRESENTED BY RARE VOCABULARY IN TEXTS?

Alia Pugh & Elfrieda H. Hiebert
TextProject

Framework for Study: What words should be taught?

- English has a substantial number of words (Nagy & Anderson, 1984)
- Estimate is that readers need to know approximately 95-98% of words for comprehension—at least when readers are young adults learning English as a second or third language ((Hsueh-Chao & Nation, 2000; Laufer, 1989; Schmitt et al., 2011).

Distribution of Vocabulary in English Lexicon



Word Zone 8

■ <1 per million

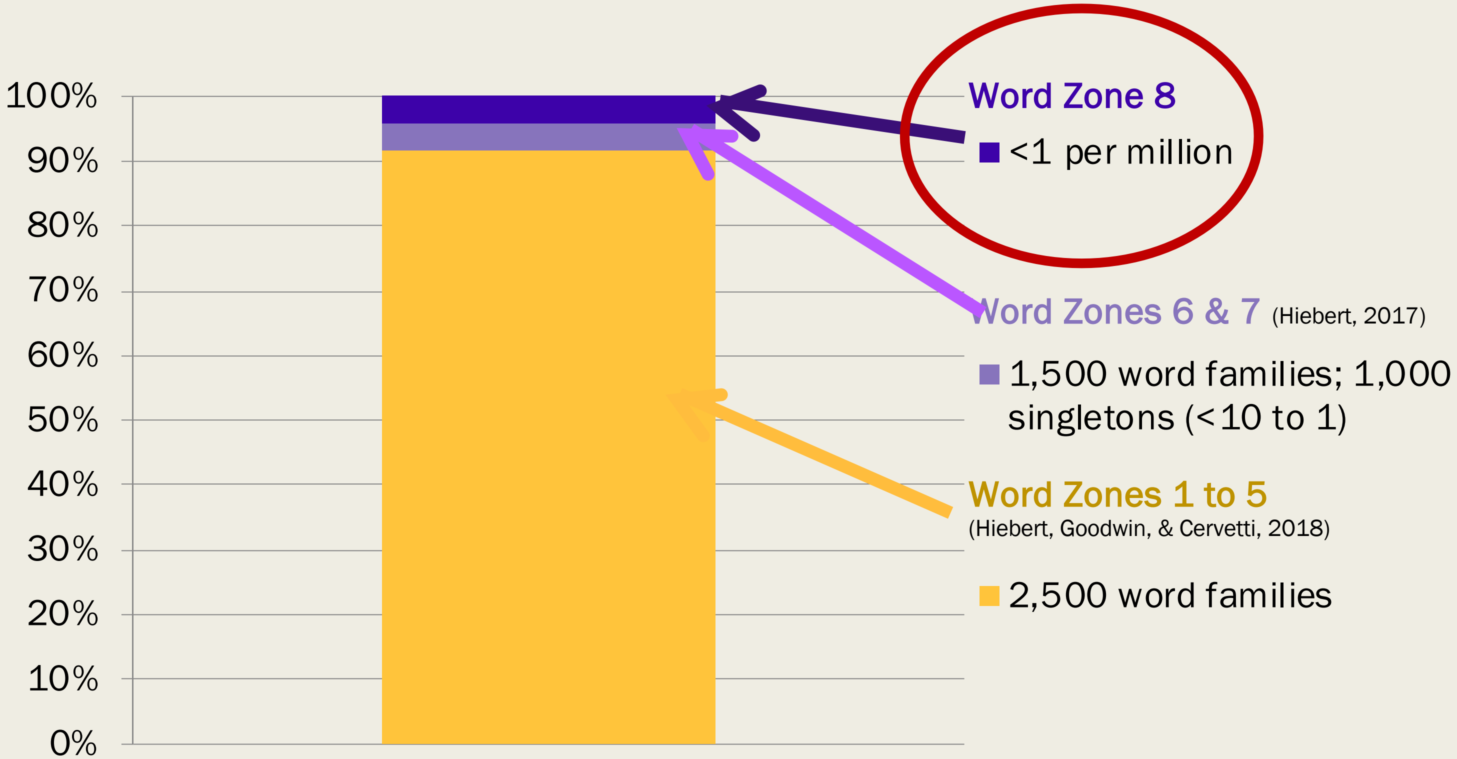
Word Zones 6 & 7 (Hiebert, 2017)

■ 1,500 word families; 1,000 singletons (<10 to 1)

Word Zones 1 to 5 (Hiebert, Goodwin, & Cervetti, 2018)

■ 2,500 word families

Distribution of Vocabulary in English Lexicon



Word Zone 8

■ <1 per million

Word Zones 6 & 7 (Hiebert, 2017)

■ 1,500 word families; 1,000 singletons (<10 to 1)

Word Zones 1 to 5
(Hiebert, Goodwin, & Cervetti, 2018)

■ 2,500 word families

Questions of Current Study

- 1. What is the nature of types and tokens of rare words across grade bands? What is the type to token relationship?

- 2. What are the main and variant categories of rare words in texts? Do these categories vary across grade bands and text types?

- 3. What proportion of rare words represent new morphological families? Do these proportions vary across grade bands and text types?

METHOD

Data Source

- TextBase database: Over 10,000 texts with narrative and informational school texts from K through Grade 11/College-and-Career Ready (CCR). The 10,000 texts consist of approximately 6.7 million words.
  - *Within the TextBase: texts are tagged as narrative or informational*
  - *Lexile levels & component information (i.e., Mean Sentence Length & Mean Log Word Frequency) are provided for every text.*
- The Common Core's staircase of text complexity does not have steps that are similar in size (e.g., Grades 2-3 texts cover a 400-Lexile span; Grades 4-5, 270; Grades 6-8, 260; Grades 9-10, 285; and Grades 11-CCR, 200. Further, steps overlap substantially.
  - *For this analysis, steps of equal size were established with the exception of the first two bands and the final one.*
    - *Grades K – 1 (<0L – 300L)*
    - *Grades 2 – 3 (310L – 600L)*
    - *Grades 4 – 5 (610L – 800L)*
    - *Grades 6 – 8 (810L – 1000L)*
    - *Grades 9 – 10 (1010L – 1200L)*
    - *Grades 11 – CCR (1210+)*

Text Sample

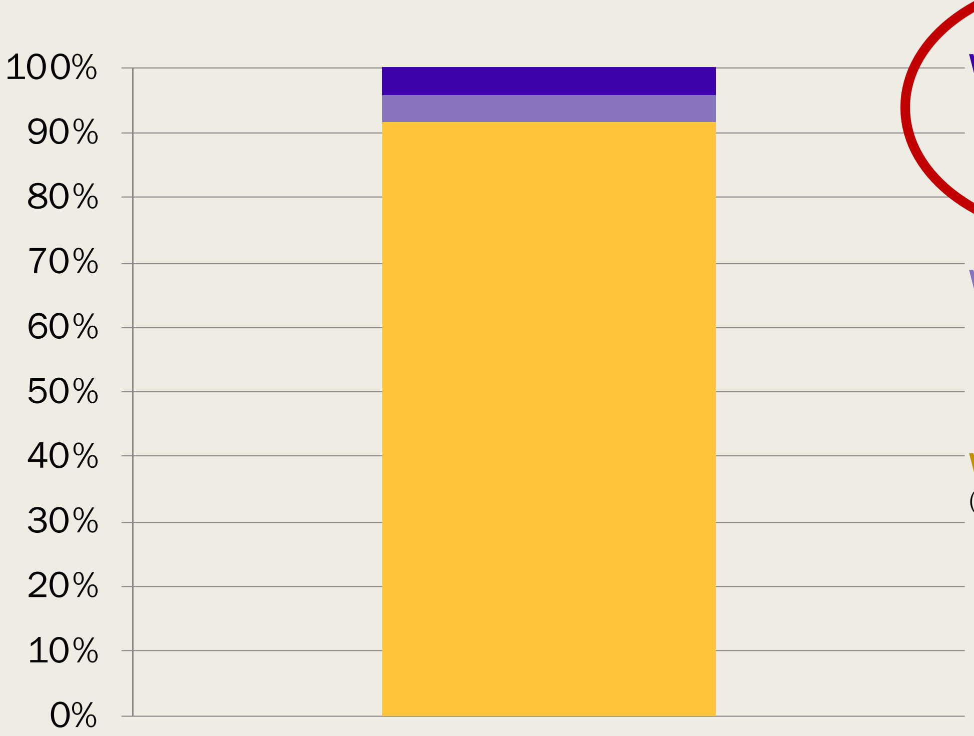
- Random sample: 20% of texts from each grade band and each genre

Band	Words in Sample
K-1 Info	9,819
K-1 Narr	50,110
2-3 Info	78,987
2-3 Narr	182,801
4-5 Info	97,869
4-5 Narr	180,935
6-8 Info	177,387
6-8 Narr	159,637
9-10 Info	109,783
9-10 Narr	115,147
11-CCR Info	83,830
11-CCR Narr	147,319
Combined Totals	1,393,624

Establishing Rare Words

- All texts were run through the Word Zone Profiler, which establishes U-function for each word type (Zeno, Ivens, Millard, and Duvuuri, 1995)
- Words grouped by *U*-function into Word Zones 1 – 8 (Hiebert, 2005)
  - *Types in Word Zone 8 are considered rare words.*

Distribution of Vocabulary in English Lexicon



Word Zone 8

■ <1 per million

Word Zones 6 & 7 (Hiebert, 2017)

■ 1,500 word families; 1,000 singletons (<10 to 1)

Word Zones 1 to 5
(Hiebert, Goodwin, & Cervetti, 2018)

■ 2,500 word families

Rare Word Sample Sizes

Grand Band	Genre	Sample Size
K-1	Info	35
	Narr	153
2-3	Info	536
	Narr	1330
4-5	Info	880
	Narr	1670
6-8	Info	1683
	Narr	1983
9-10	Info	2239
	Narr	1677
11-CCR	Info	2241
	Narr	3159

Band	Sample Size
Total Words Informational	7615
Total Words Narrative	9973
Grand Total	17588

Vocabulary Levels

- Based on Becker, Dixon, & Anderson-Inman’s (1980) classification, each rare word type was classified into one of four categories:
  - *Level 1: lead word in Word Zones 1 – 5*
  - *Level 2: lead word in Word Zones 6 – 7*
  - *Level 3: lead word in Word Zone 8 (i.e. a rare morphological family)*
  - *Not-Leveled words did not fit the following criteria:*
    - a recognizable form of a head word in an English dictionary;
    - recognizable in oral English—common exclamations/onomatopoeia; or
    - a full word—no word parts or separated affixes

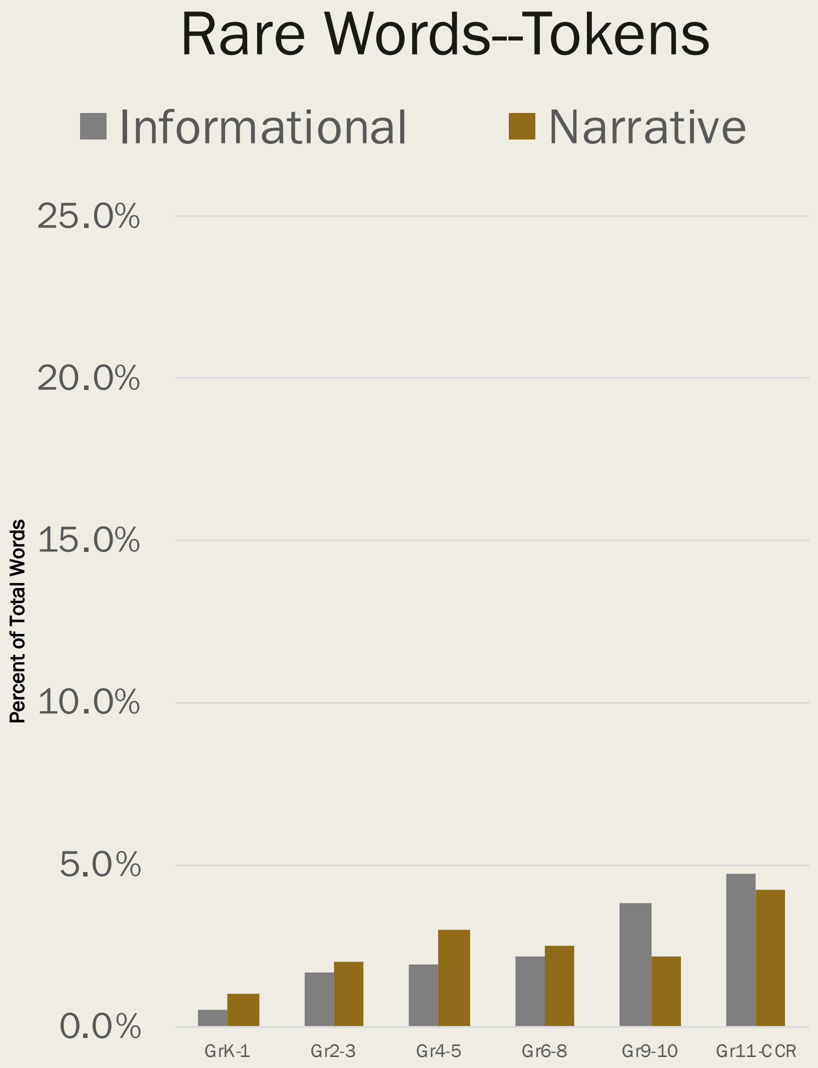
Word Categories

Primary Vocabulary	Examples
Roots	icon, periphery, sushi
Inflected Forms	tampered, orbs, plainest, pepper’s
Derived Forms	dishonor, hatless, unfeeling
Compounds and Contractions	seaside, it’s
Proper Names	Ahmad, Jacksons, Berlin’s, Latino

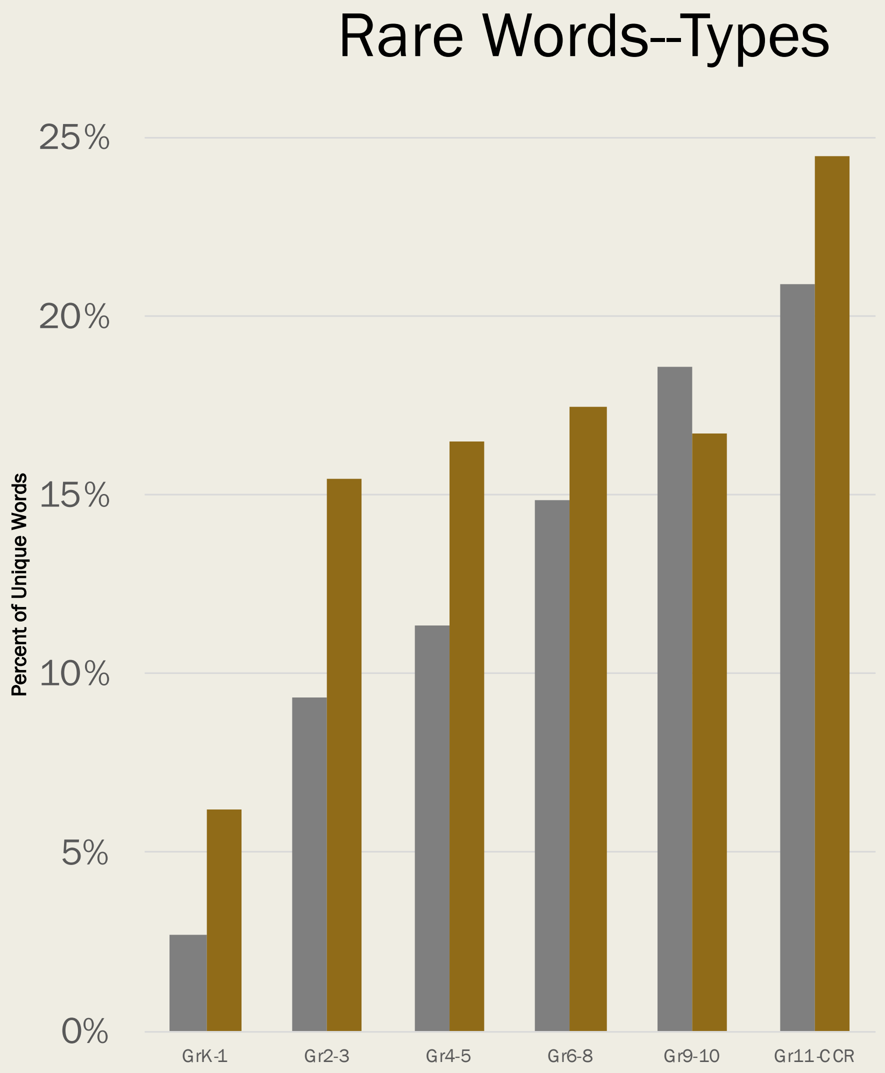
Word Variants	Examples
Abbreviations	USS, ATMs, ABA’s
Archaic, Dialect, Other Spellings	screeeches, didst, learnin’, worser
Exclamations, Onomatopoeia, Invented Words	ohhh, plink, sunflakes
Other Languages	vous, nalukataq
Affixes/Word Parts	mini-, co-, -nd

RESULTS

1. What is the relationship of types and tokens of rare words across grade bands and genres? What is the nature of word features of rare words across grade bands and genres?

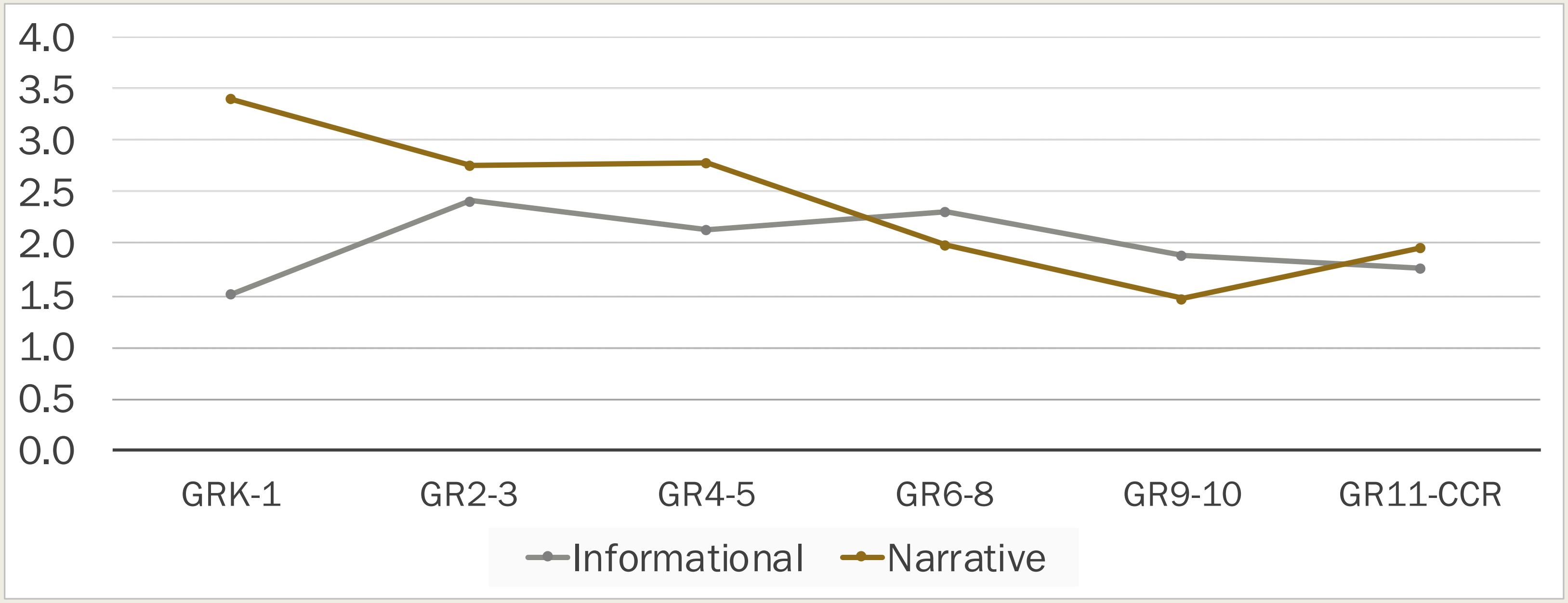


Rare Words--Tokens



Rare Words--Types

Repetitions of Rare Words



Features of Words: Grade Bands & Genres (Means)

Grade Band	# Appearances	Word Length	Age of Acquisition (Kuperman, Stadhagen-Gonzalez, & Brysbaert, 2012)	Concreteness (Brysbaert, Warrinner, & Kuperman, 2014)
2-3	2.6	6.8	9	3.8
6-8	2.2	7.7	10.5	3.4
11-CCR	1.9	8.2	11.8	2.8

Response to Q1:

- The percentage of rare words in texts is low but number of rare word types is high.

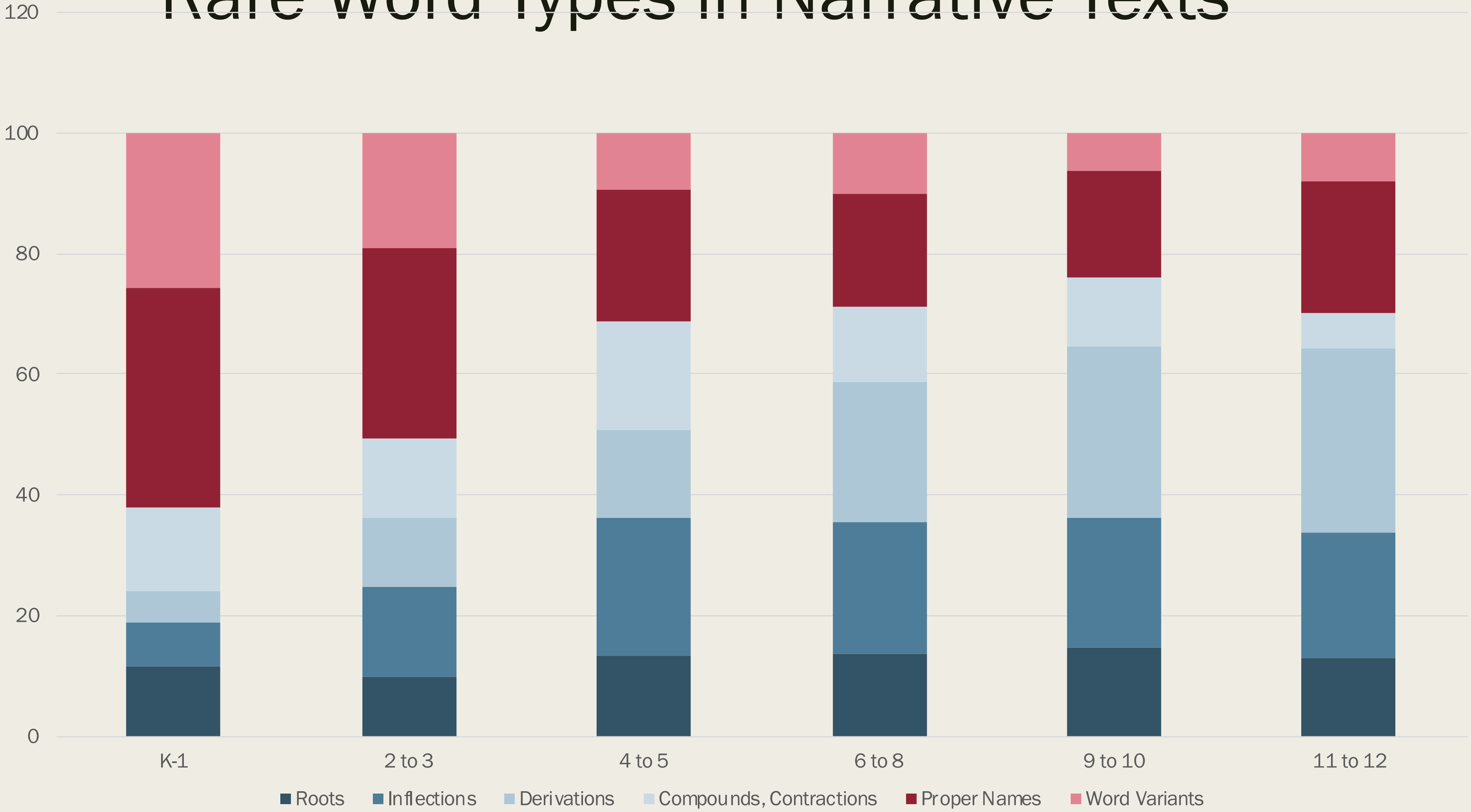
- There is a large increase in the diversity of words in narrative texts from Gr. K-1 to Gr. 2-3. The load of new words is not quite as high in informational text as it in narrative text until high school bands.

- A rare word is repeated an average of 2.3 times across texts of grade bands & genres.

- For both genres, average age of acquisition and word length increase steadily across grade bands, while average concreteness drops over time.

- 2. What are the categories of rare words in texts? Do these categories vary across grade bands and text types?

Rare Word Types in Narrative Texts

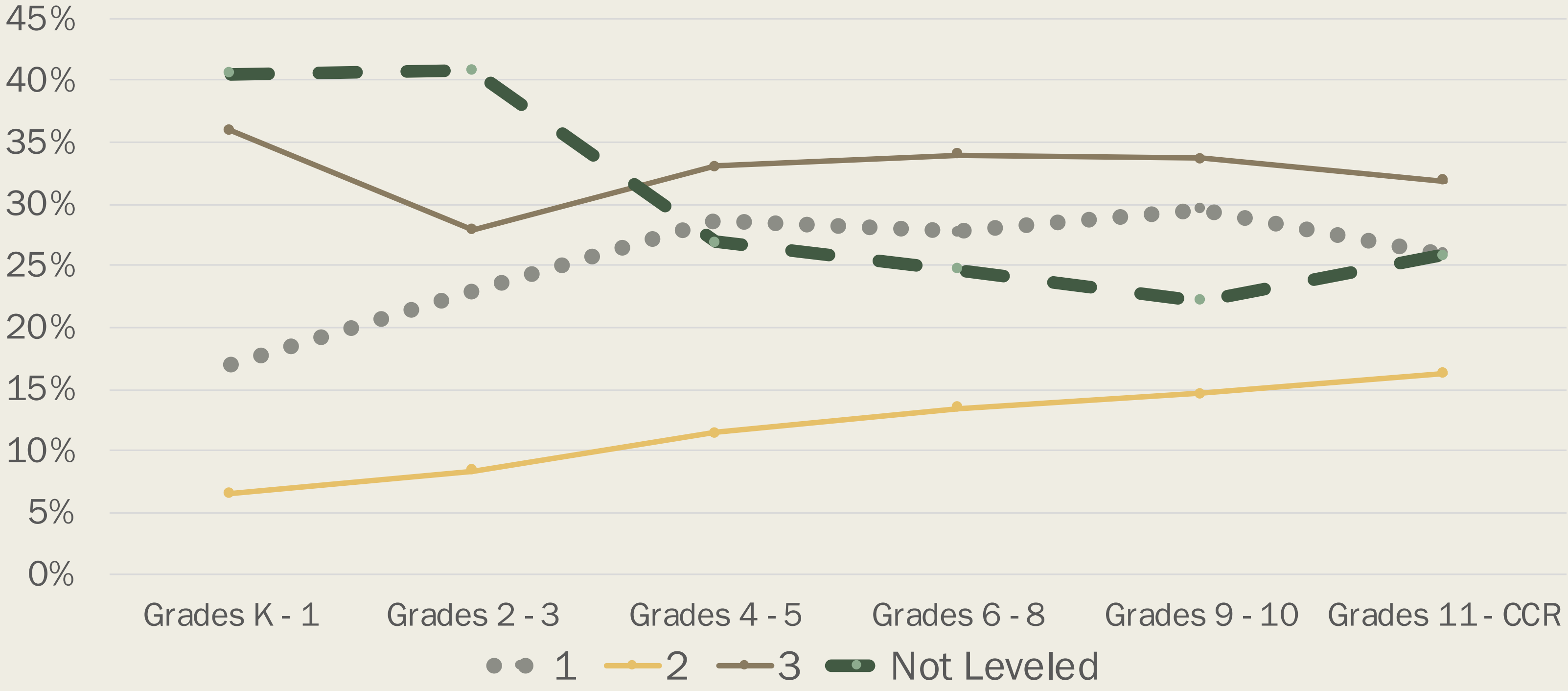


Response to Q2

- From Gr. 4 – 5 and up, inflections and derivations become the most common rare word categories in narrative texts.
- Proper names are the most common category of rare words in informational texts at all grade bands.
- Other primary vocabulary categories (Roots, Compounds & Contractions) maintain levels at roughly 10 – 20% for both genres in most grade bands.
- Word variants maintain a low proportion under 5% for both genres and all grade bands, with the exception of Exclamations, Onomatopoeia, and Invented words in Grs. K – 1 and 2 – 3 (both genres).

- 3. What proportion of rare words represent word families from Levels 1 and 2? New morphological families (i.e., Level 3 families)? Do these proportions vary across grade bands and text types?

Rare Types by Vocabulary Level in Narrative Texts



New Morphological Families: Level 3

Age of Acquisition Band	Rare Words (%)	Repetitions	Word Length	Mean Concreteness Rating	Examples
<7.1 -9	11.6	2.4	6.4	3.97	beeped, squirt
9.1-11	17	2.1	6.7	3.48	shrivel, scuttle
11.1-13	26	1.6	7.4	3.02	seismic, plummet
13.1-15	20	1.5	8.1	2.76	ecstatic, purveyor
15.1+	6	1.4	8.2	2.75	mitigate, salient
N.A.	19.5	1.2	7.6	3.40	omnivore, smolder

Response to Q3

- For both genres, rare words in Levels 1 & 2 comprise about 20 – 24% of rare word types in the early grades and increase steadily to about 40% in the highest grades.
- Nearly 50% of the words in narrative texts are from rare morphological families in Gr. K – 1, but this proportion is halved in Gr. 2 – 3 and remains consistent through the upper grades.
- In informational texts, about one third of the rare word types are from rare morphological families.
- Members of rare morphological families increase in word length and decrease in concreteness as age of acquisition rises.
- Mean family size in rare families is low—about 1.2 to 2.4 members (at least in this sample).

DISCUSSION

Discussion: What does this study say about which words should be taught?

- There is even more justification for instruction of the morphological families that account for 91.5% (Hiebert et al., 2018) of the total words in the CCSS sample.
- Of rare root words (i.e., Level 3), there are approximately 2,500 unique meanings; an estimated 20% form morphological families of notable size. These are families with at least two+ family members with a combined U function of at least 1 or more predicted appearances per million
  - Examples:
    - *redeem with 11 family members, U = 3*
    - *repent with 6 family members, U = 1.7*
    - *lament with 7 family members, U = 2.4*
  - Such word families range in age of acquisition, meaning that exposure/instruction could be spread across Grades 5-12.

Discussion (continued)

- But attention needs to go beyond morphological families:
  - *Many of the types represent concrete entities that are rare but often cluster into groups and can be represented with pictures/illustrations*
    - EXAMPLE: tapir, okapi, gnu, impala, wombat, vole, aviary, pachyderm, marsupial

      sarcophagus, morgue, inter, relics
  - *Attention also needs to be paid to the presence of onomatopoeia in the primary grades and abbreviations in the middle through high school grades.*

Discussion (continued)

- Proper names are highly prominent in school texts.
- Proper names differ substantially in their group memberships (Hiebert & Nagy, 2019).

Types of Proper names in EWFG
(1st ≈20,000 Words)

Type	#	Length in Letters
Character	216	6.9
First	653	5.6
Geography/Group	594	7.1
Surname	228	6.4
Misc. (Time, Object, Address)	98	6.7

hiebert@textproject.org