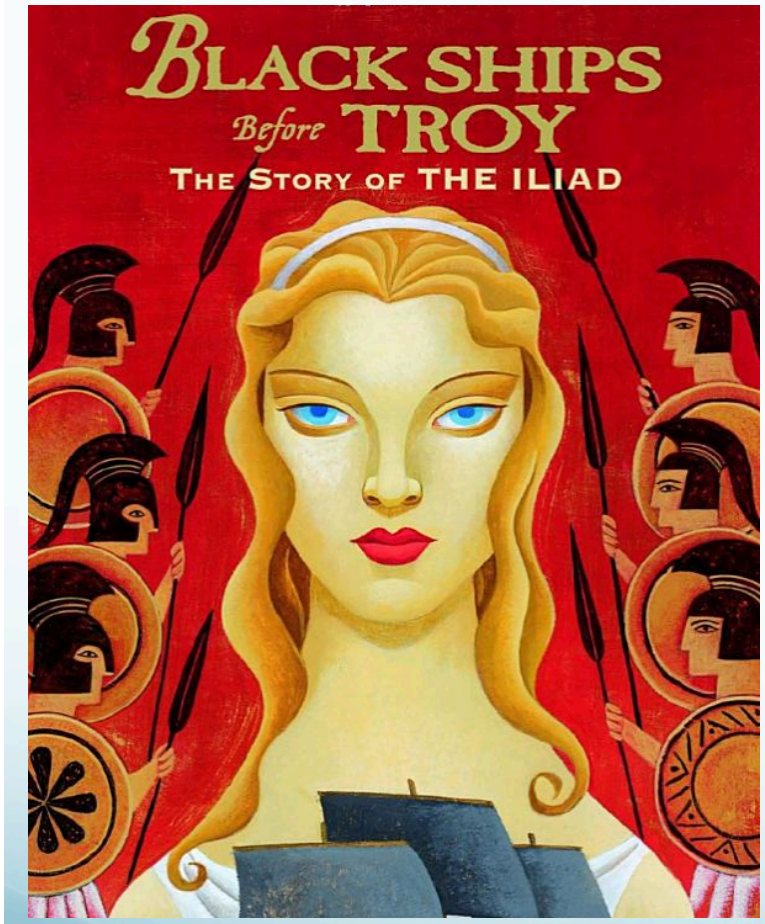
Development and Application of A Morphological Family Database In Analyzing Vocabulary Patterns In Text

Elfrieda H. Hiebert
TextProject &
University of California, Santa Cruz

Vocabulary within Current Text Analyzers



The Lexile Framework:
1300L
3.55 MLWF

Reading Maturity Metric:
RMM score: 8

Words:

keels	breach
striving	gateways
buckler	outstrip
stead	galleys
foremost	valiant

Chapter: Battle for the Ships

Aim of the Word Zone Profiler

What the Word Zone Profiler is intended to do:

- Establish vocabulary demands of texts, especially for beginning and struggling readers

What the Word Zone Profiler does not do:

- Provide a comprehensive measure of text complexity

In this study:

I. Establishing the database

   Including: Establishing the size of the morphological families of the "core vocabulary" (i.e., the 2,522 complex word families) across all word zones, not simply the word zones of moderate and high frequency

II. Describing features of the words within the database

III. Examining the validity of the database

IV. Illustrating potential extensions and applications

1. Establishing the Database

Decision 1: Choose a database to use as foundation

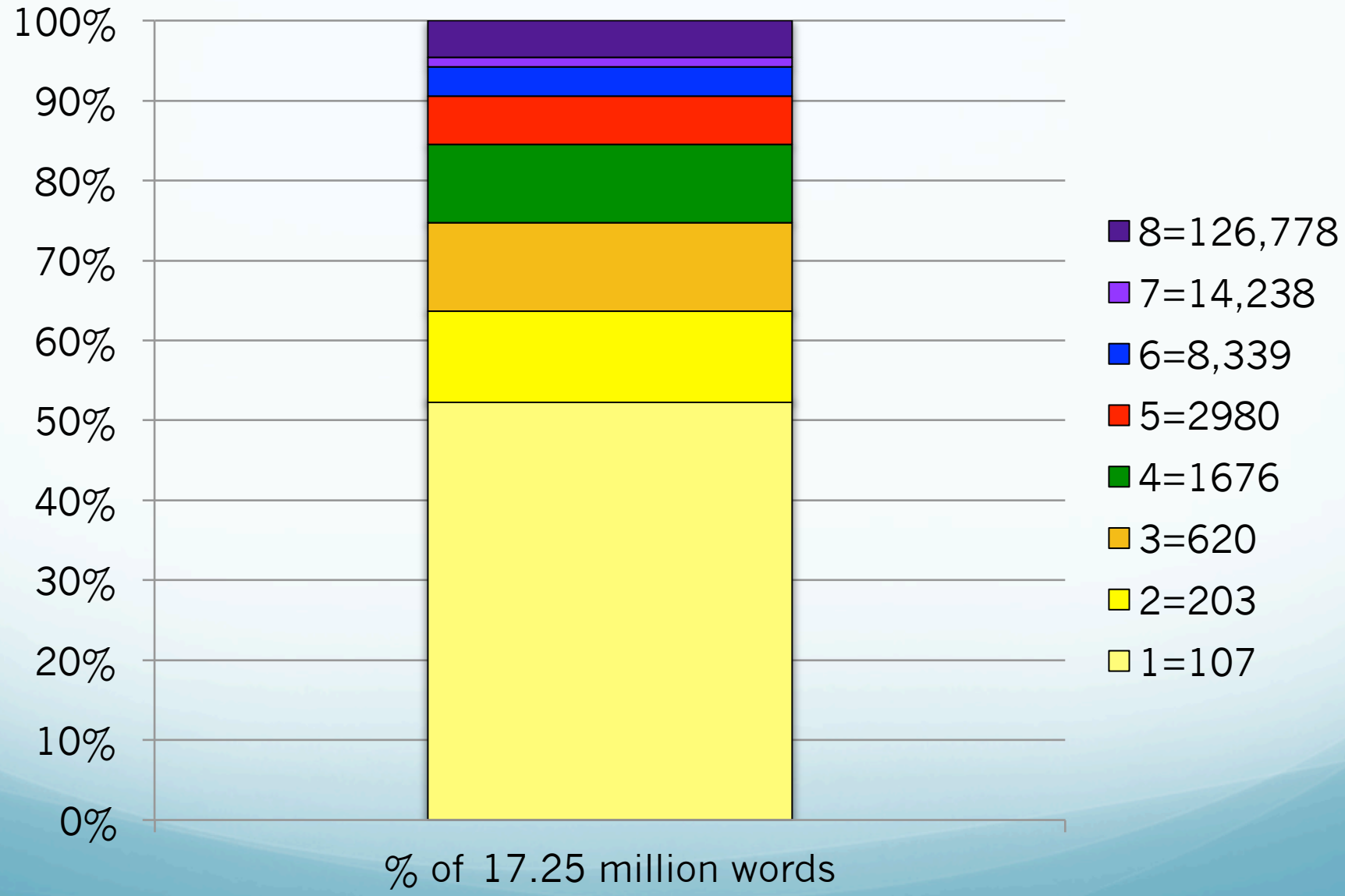
Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995)

- 154,941 types based on 17,272,580 tokens
- Approximately 1,000 words have been added (primarily derivatives of words already in the database) as a result of analyses of 3,500 digitized texts

Decision 2: Identify parameters for word zones of high, moderate, and rare frequencies

- Highly frequent words: 100+ occurrences per million
- Moderately frequent words: 10-99 occurrences per million
  - McKeown, M., I. Beck, R. Omanson, and M. Pople. 1985. Some Effects of the Nature and Frequency of Vocabulary Instruction on the Knowledge and Use of Words. *Reading Research Quarterly 20:* 222-35.
  - Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46-65.
- Rare words: 9 or fewer

The Word Zones that Form the Word Zone Profiler



Decision 3: Establish Morphological Families Within Word Zones 1-5

- Began with 5,586 words (Zones 1-5)
- Based word family membership on:
  - Becker, W. C., Dixon, R., & Anderson-Inman, L. (1980). *Morphographic and root word analysis of 26,000 high frequency words*. University of Oregon Follow Through Project, College of Education.
- Resulted in 2,522 complex word families; henceforth called the 2,500 complex word families

Decision 4: Define Critical Word Factor

Critical Word Factor (CWF):

- Number of rare words per 100
- A rare word does not imply that readers cannot decode it nor that readers do not know the meaning of the word; A rare word is simply a word that readers have likely not encountered frequently in text.

Decision 5: Identify Word Features In Addition to Frequency

- **Word Length**
- **Age of Acquisition** (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012)
- **Concreteness** (Brysbaert, Warriner, & Kuperman, 2013)
- **Academic Vocabulary List** (Gardner, D., & Davies, M.)
- **Dispersion** *(Zeno et al., 1995)*
- **Semantic Superclusters & Megaclusters** (Marzano & Marzano, 1988; & Hiebert, 2011)

Decision 5: Identify Word Features In Addition to Frequency

- **Word Length**
- **Age of Acquisition** (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012)
- **Concreteness** (Brysbaert, Warriner, & Kuperman, 2013)
- **Academic Vocabulary List** (Gardner, D., & Davies, M.)
- **Dispersion** *(Zeno et al., 1995)*
- ~~**Semantic Superclusters & Megaclusters** (Marzano & Marzano, 1988; & Hiebert, 2011)~~

Word Zone Profile Output: Words in a Text

Word	Frequency	Word Leng	UFunction	WordZone	Age of Acq	Concreten	Core Acad	Dispersion
hector	23	6	4	6	0	0	0	0.4657
trojans	18	7	2	6	0	0	0	0.4508
patroclus	16	9	0	8	0	0	0	0
achilles	14	8	2	6	0	0	0	0.5867
armor	13	5	7	6	7.17	4.76	0	0.6759
zeus	10	4	6	6	0	0	0	0.4255
ditch	8	5	7	6	6.22	4.5	0	0.6472
myrmidons	8	9	0	8	0	0	0	0
flung	7	5	9	6	0	3.29	0	0.3508
chariot	7	7	4	6	9.11	4.86	0	0.5348
chariots	6	8	4	6	9.11	4.86	0	0.4861
charioteer	6	10	0.0326	8	12.57	0	0	0
spears	5	6	8	6	7.22	5	0	0.6554
trojan	5	6	2	6	0	0	0	0.51
ajax	5	4	0.5951	8	0	0	0	0.4753
sarpedon	5	8	0	8	0	0	0	0
spear	4	5	8	6	7.22	5	0	0.6392
hurling	4	7	3	6	0	3.81	0	0.514
comrades	4	8	2	6	10.38	3.69	0	0.444
automedon	4	9	0	8	0	0	0	0
allies	3	6	9	6	9.61	3.48	1859	0.3685
plunged	3	7	7	6	8.61	3.04	0	0.5328
hector's	3	8	4	6	0	0	0	0.4099
foremost	3	8	3	6	11.78	2	0	0.8325
mortal	3	6	3	6	9.84	1.96	0	0.4544
poseidon	3	8	1	7	0	0	0	0.3726

Decision 6: Identify Features for Summarizing the Words in Entire Texts

- # of repetitions of words
- Distribution of words
  - Total types
  - Types as a function of tokens

Word Zone Profiler: Results for ("Battle for the Ships" in *Black Ships Before Troy)*

Total words: 3037
Total unique 829

WordZone	Unique Word	Ratio of Total	Percent	Total Words	Ratio of Total	Percent
1	105	105/829=	12.6658625	1559	1559/3037=	51.3335528
2	126	126/829=	15.199035	356	356/3037=	11.7220942
3	160	160/829=	19.3003619	335	335/3037=	11.0306223
			47.1652593			74.0862693
4	149	149/829=	17.972362	255	255/3037=	8.39644386
5	115	115/829=	13.8721351	165	165/3037=	5.43299309
			31.8444971			13.8294369
6	100	100/829=	12.0627262	243	243/3037=	8.00131709
7	22	22/829=	2.67792521	28	28/3037=	0.92196246
8	52	52/829=	6.32726176	96	96/3037=	3.16101416
	174		21.0679131			12.0842937

Repeated W	Unique Word	Ratio of Total	Percent	Total Words	Ratio of Total	Percent
Singletons	479	479/829=	57.7804584	479	479/3037=	15.7721436
2 to 4 times	240	240/829=	28.9505428	623	623/3037=	20.5136648
5 to 9 times	60	60/829=	7.23763571	379	379/3037=	12.4794205
10-plus time	50	50/829=	6.03136309	1556	1556/3037=	51.2347712

Decision 7: Establish Members of Morphological Families within the Entire Dataset

- Identifying the additional members of the 2,500 word families from the 149,355 words in Word Zones 6-8 + inflected endings (not on Zeno et al. list)
- Becker et al.'s (1980) classification of root words as the basis for the inclusion/ exclusion of words

II. Describing Features of the Words Within the Database

Features of the 2522 Base Words on Core Vocabulary List

	Mean (SD)	% Of Entire List
Word Length	5.75 (1.99)	
Age of Acquisition	6.87 (2.45)	
Concreteness	3.44 (.89)	
Core Academic List		18%
Dispersion		>.85: 46% <.65: 14%

Morphological Family Size

- The 2,500 word families have an average of 7.7 members: 19,419 words in all.
- How much does the addition of these morphological family members add to the frequency of families?
  - Mean: 31.97 (SD=22.60)

III. Examining the Validity of the Database

Exemplar Texts of the Common Core

- Appendix B of the Common Core has lists of texts for 5 grade bands; some titles are repeated in the Standards themselves where texts are also given for Grs. K-1
- Aims of Appendix B:
  - Exemplify the complexity and quality of texts required by Standards
  - Breadth of text types

Sample Size

1. Size of samples

   In Nelson et al. study: Average of 475.5 words

   - To Kill a Mockingbird: 220-word sample of 99,121 (.2%)
   - The Grapes of Wrath: 608-word sample of 169,481 (.3%)
   - The Gettysburg Address: 264-word sample of 264 (100%)
   - Tops and Bottoms: 189-word sample of 878 (22%)

2. Choices about how large a sample of a text to use:
   - When texts were 40 pages or less: Entire text
   - When texts were >40 pages: 10%
     - 10% was taken from the middle of the text

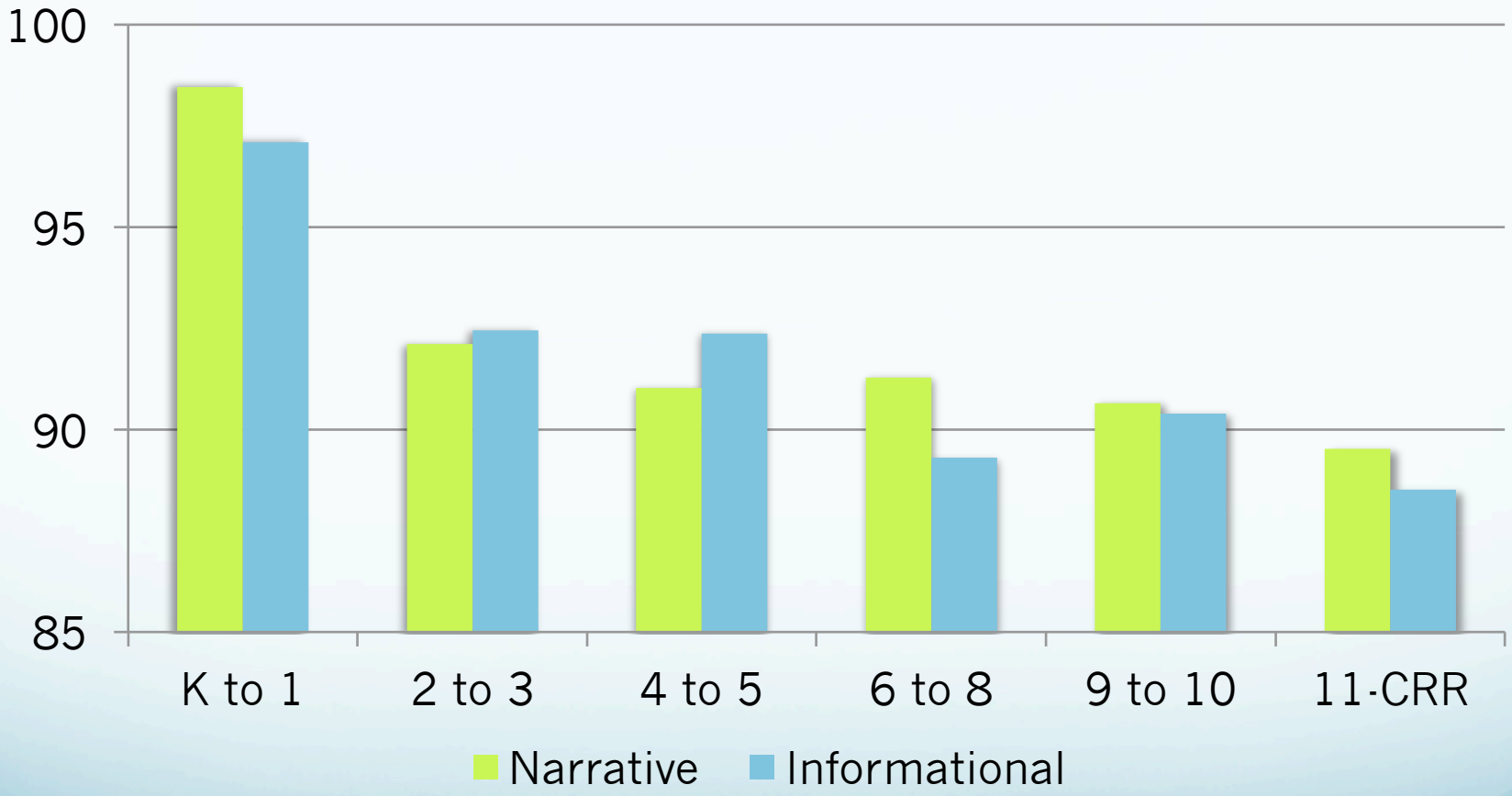
Grade Band	Text	CCSS sample (X = 475.5 words)		Current sample (X = 4,169.05 words)	
		CWF	Lexile	CWF	Lexile
2-3	Tops & Bottoms	6.95	640	8.27	700
4-5	Black Stallion	13.44	1250	7.72	680
6-8	Black Ships Before Troy	9.37	1050	12.18	1300
9-10	The Book Thief	10.10	560	9.25	750
11-CCR	Pride & Prejudice	5.90	770	8.26	1130

Grade Band	Text	CCSS sample (X = 475.5 words)		Current sample (X = 4,169.05 words)	
		CWF	Lexile	CWF	Lexile
2-3	Tops & Bottoms	6.95	640	8.27	700
4-5	Black Stallion	13.44	1250	7.72	680
6-8	Black Ships Before Troy	9.37	1050	12.18	1300
9-10	The Book Thief	10.10	560	9.25	750
11-CCR	Pride & Prejudice	5.90	770	8.26	1130

Sample: Texts & Words

Grade Band	# of Texts	Total Words
K-1	17	7,394
2-3	25	32,089
4-5	29	61,743
6-8	33	155,240
9-10	45	249,662
11-CCR	54	343,237
Total	203	849,365

Percentage of Total Words Within 2,500 Morphological Families



IV. Illustrating potential extensions and applications

1. Variable definitions of CWF that take into account what we know about students' vocabulary learning

Illustration with the 175 rare words from chapter of *Black Ships Before Troy*

Word	Frequency	Word Leng	UFunction	WordZone	Age of Acq	Concreten	Core Acad	Dispersion
[illegible]	23	6	4	6	0	0	0	0.4657
[illegible]	18	7	2	6	0	0	0	0.4508
[illegible]	16	9	0	8	0	0	0	0
[illegible]	14	8	2	6	0	0	0	0.6867
[illegible]	13	5	7	6	7.17	4.76	0	0.6759
[illegible]	10	4	6	6	0	0	0	0.4255
[illegible]	8	5	7	6	6.22	4.5	0	0.6472
[illegible]	8	9	0	8	0	0	0	0
flung	7	5	9	6	0	3.29	0	0.3508
chariot	7	7	4	6	9.81	4.86	0	0.5348
chariots	6	8	4	6	9.11	4.86	0	0.4861
charioteer	6	10	0.0326	8	12.57	0	0	0
spears	5	6	8	6	7.22	5	0	0.6554
trojan	5	6	2	6	0	0	0	0.51
ajax	5	4	0.5951	8	0	0	0	0.4753
sarpedon	5	8	0	8	0	0	0	0
spear	4	5	6	6	7.22	5	0	0.6392
[illegible]	4	7	0.3159	6	0	3.81	0	0.5144
[illegible]	4	8	0.1688	6	10.38	3.69	0	0.4440
[illegible]	4	9	0.0488	8	0	0	0	0
[illegible]	3	6	0.9162	6	9.61	3.48	0.1859	0.3685
[illegible]	3	7	7	6	8.61	3.04	0	0.5328
[illegible]	3	8	4	6	0	0	0	0.4099
foremost	3	8	3	6	11.78	2	0	0.8325
mortal	3	6	3	6	9.84	1.96	0	0.4544
poseidon	3	8	1	7	0	0	0	0.3726
arched	2	6	9	6	9.89	3.82	0	0.5299
grief	2	5	9	6	8.39	2.7	0	0.7069
ax	2	2	8	6	9.31763	4.9	0	0.702
helmet	2	6	8	6	5.71	4.92	0	0.5891
retreat	2	7	7	6	10.53	3.03	0	0.6273
stakes	2	6	7	6	0	0	0	0.6157
shattered	2	9	6	6	9.33	3.96	0	0
[illegible]	2	6	5	6	9.22	3.93	0	0.7071
[illegible]	2	4	5	6	0	0	0	0.4559
[illegible]	2	6	3	6	10.22	4.75	0	0.6585
[illegible]	2	5	3	6	10.67	1.92	0	0.7054
[illegible]	2	6	3	6	0	3.52	0	0.5752
[illegible]	2	6	3	6	9.56	1.67	0	0.7173
[illegible]	2	4	2	6	0	0	0	0.3459
clanging	2	8	2	6	9.53	3.96	0	0.587
[illegible]	1	5	4	7	10.47	2.48	0	0.5959
[illegible]	1	8	4	7	10.47	3.7	0	0.2852
[illegible]	1	5	4	7	8.96	4.19	0	0.4302

Parsing the Hard Words According to Knowledge about Word Recognition

	Black Ships Before Troy	
Beginning	175 words	
Morphological Families		11 words
Short words (5 letters or fewer)		81 words
Other Languages		0
Proper Names		24
Compound words		12
Highly Concrete		7
Known Before age 7		1
Ending	39	

RMM

- keels
- breach
- striving
- gateways
- buckler
- outstrip
- stead
- galleys
- foremost
- valiant

Word Zone Profile process

- mortal/immortal, amends, sacrifice
- parapet, ramparts, galleys
- frenzy, retreat, trampled

2. Comparing the vocabulary demands of complex texts at different grade bands

	Pride & Prejudice (11-CCR)		The Book Thief (9-10)		Black Ships Before Troy (6-8)		Black Stallion (4-5)		Tops & Bottoms (2-3)	
Start	434 words (CWF: 8.26)		584 words (CWF: 9.25)		175 words (CWF: 12.23)		185 words (CWF=7.72)		20 words (CWF=8.27)	
		proportion		proportion		proportion		proportion		proportion
Morphological Families		.18		.13		.06		.13		0
Short words (5 letters or fewer)		.09		.32		.46		.36		.55
Other Languages		.002		.02		0		0		0
Proper Names		.04		.03		.14		.02		0
Compound words		.02		.06		.07		.10		.05
Highly concrete		.03		.04		.04		.08		.40
Known before age 7		.02		.02		.01		.01		0
End	.61 (263; CWF=5)		.38 (219; CWF=3.52)		.22 (39; CWF=2.69)		.31 (58; CWF=2.40)		0	

2. Comparing the vocabulary demands of complex texts at different grade bands

	Pride & Prejudice (11-CCR)		The Book Thief (9-10)		Black Ships Before Troy (6-8)		Black Stallion (4-5)		Tops & Bottoms (2-3)	
Start	434 words (CWF: 8.26)		584 words (CWF: 9.25)		175 words (CWF: 12.23)		185 words (CWF=7.72)		20 words (CWF=8.27)	
		proportion		proportion		proportion		proportion		proportion
Morphological Families		.18		.13		.06		.13		0
Short words (5 letters or fewer)		.09		.32		.46		.36		.55
Other Languages		.002		.02		0		0		0
Proper Names		.04		.03		.14		.02		0
Compound words		.02		.06		.07		.10		.05
Highly concrete		.03		.04		.04		.08		.40
Known before age 7		.02		.02		.01		.01		0
End	.61 (263; CWF=5)		.38 (219; CWF=3.52)		.22 (39; CWF=2.69)		.31 (58; CWF=2.40)		0	

www.textproject.org

https://independent.academia.edu/ElfriedaHiebert