国際研究論叢　23（2）：13～29，2010

Vocabulary Use in Children's Animated Films

Brad Visgatis *

Abstract

This study reports on the vocabulary level of eight popular children's animated movies from the Disney studios to determine the range of the vocabulary used in the films and the proportion of vocabulary from each frequency band. Comparisons are also made to the vocabulary distribution of the movie *Shrek*. Results indicate that these children's movies require a fairly high level of vocabulary, with up to 8000 words required for understanding of the four most lexically complex movies. Implications for use of children's movies for second language learning are discussed.

Keywords

vocabulary, lexis, animated films, BNC

Recent research has identified the range of vocabulary found in various types of text, including novels, newspapers, textbooks, and movies (Nation, 2006). Although a number of factors impact on native speaker comprehension of these texts (background knowledge, topic familiarity, genre familiarity, grammatical complexity, etc.), one major factor is knowledge of vocabulary.

Research into the lexical development of native speakers has shown that vocabulary size grows rapidly from childhood, especially after the onset of formal education, and has been estimated to range from between 1,000 and 3,000 words per year (Nagy, Anderson, & Herman, 1987; Nagy, Herman, & Anderson, 1985), but more realistically averaging out to approximately 1,000 words per year until adulthood. By extrapolation, this means that the representative native speaking adult should know upwards of 20,000 words (Nagy et al., 1985; Nation, March 4, 2007, personal communication; January 21, 2007; Nagy & Herman, 1984), although accurate testing of this is fraught with conceptual and technical difficulties (Nation, 1985, 1993).

Given that some estimates for the amount of time necessary for explicitly learning even one new vocabulary item can reach 15 minutes (under massed learning conditions) (Baddeley, 1990; Nation, 2001), only a small proportion of words are learned in this way.

＊ヴィスゲイテス　ブラッド：大阪国際大学人間科学部教授〈2009.10.1受理〉

Rather, the lion's share of vocabulary learning occurs incidentally through repeated exposure of the items within easily comprehensible contexts. The number of times an item must be seen during reading in order to establish recognition has been estimated at approximately seven (Pimsleur, 1967; Tinkham, 1993), and the easily comprehensible context is one where the large majority of the running words (e.g. 95–98%) is already known to the learner (Hseuh-chao & Nation, 2000, Nation, March 4, 2007, personal communication), which is the level of vocabulary knowledge thought to be necessary for students to be able to read unassisted. Of course, native speakers have access to more than just written textual input and benefit from language exposure geared to their L1 developmental pace.

Second language learners, however, do not have these same benefits, and so their progress is neither as rapid nor as extensive, and their normal developmental pathway is more problematic. Research has shown that explicit and deliberate learning of the most frequent 2,000 words from the *General Service List* (West, 1953) in English is a practical goal (Nation, 2006) and can provide the learner with familiarity of approximately 85% of non-simplified running text. Nevertheless, even this level of lexical recognition is only sufficient to insure content comprehension in the most limited of circumstances. Moreover, a vocabulary size of 2,000 word families is insufficient to yield adequate comprehension of the exemplars cited above, much less enable the learner to pick up new vocabulary implicitly.

To increase coverage up to a level where vocabulary can be implicitly learned through multiple contextualized exposures is more problematic. Beyond the most common 2,000 word families, vocabulary items become much less frequent, and the amount of input that must be accessed in order to meet the conditions for implicit learning grows voluminous. Some other well-defined word lists that focus on newspaper (Chung, 2007), academic (Coxhead, 1998, 2000), or technical vocabularies (Chung & Nation, 2003), for example, are also efficient for explicit study and can raise the learner's vocabulary coverage to almost 95% of the running words for non-simplified text. In numerical terms, addition of the items from the Academic Word List and Newspaper Word List would increase a person's vocabulary size to approximately 3,000 word families (Chung, 2007; Coxhead, 1998, 2000). Technical vocabulary appears to vary widely by field (Chung & Nation, 2003), but their addition may not confer much advantage when dealing with non-specialized texts where the percentage of low-frequency vocabulary included in the technical vocabulary lists only reaches approximately 20% (Nation, April 1, 2007, personal communication).

This leaves the second language learner with a challenge to bridge from a word base of approximately 3,000 word families (e.g. the most frequent 2,000 word families plus all the items from the academic and newspaper word lists) to the 8,000 word family

level necessary for 100% coverage of a novel (or 7,840 word families necessary for 98% coverage of that same novel) (Nation, 2006, Nation, January 21 and April 1, 2007, personal communication).

Extensive reading of simplified texts in the form of graded readers has also shown positive results in expanding learner's vocabulary size, provided a sufficient amount is actually covered (Hunt & Beglar, 2005; Waring & Takaki, 2003). Unfortunately, the vast majority of graded readers only reach up to the 3,000-word level (Nation & Wang, 1999), leaving learners without effective means for bridging the gap up to the 7,000 plus word level. Moreover, although books for native speaking children can serve as a source of easily comprehensible input, in order to stay within the appropriate lexical band, the books would have to be those directed at very young children, meaning that their content would also be directed at that level of maturity. Finally, movies for children may also be a possible source for input, although *Shrek* (Nation, 2006), for example, has been shown to contain over 1,000 word families, almost one third of which lie above the 2,000 word-level of the British National Corpus.

One question, however, is to what degrees are the exemplars cited above representative of those text types? In other words, how confident are we that the average book contains 8,000 word families? That the average newspaper contains 4,000 to 6,000? Or that the average children's film encompasses 1,000? It is specifically this last point that is the focus of this paper. In order to gain a better idea as to the range of vocabulary, this paper will analyze the vocabulary found in eight animated films for children from the Disney studios. The research questions are:

RQ1: What is the range of vocabulary found in these films?

RQ2: What proportion of vocabulary in these films are from each of the different frequency bands?

RQ3: To what degree is the vocabulary distribution of *Shrek* representative of animated films in general?

RQ4: What implications do these findings have for EFL teaching?

Method

Medium

Although movies provide visual and aural input to learners, this study uses written transcripts as data[1]. This is a problem as oral language contains many linguistic attributes that cannot be fully transcribed, such as partial utterances, overlapping dialogue, puns, interjections, ejaculations, and accented or affected pronunciation. The transcripts were downloaded from the Internet (http://animationarchive.net/Script/). In

all, eight Disney film scripts were selected for analysis: *Bambi* (Hand, 1942), *Cinderella* (Geronimi & Jackson, 1950), *Hercules* (Clements & Musker, 1997), *Lady and the Tramp* (Geronimi & Jackson, 1955), *The Lion King* (Allers & Minkoff, 1994), *The Little Mermaid* (Clements & Musker, 1989), *101 Dalmatians* (Geronimi & Luske, 1961) and *Snow White and the Seven Dwarfs* (Hand, 1937).

No specific selection criteria were applied, so these are not necessarily representative of Disney animated films in general, however, films from different time periods and lengths were deliberately chosen.

Data Preparation

Each script was downloaded in *html* format, converted into a text file, and then edited to either remove unneeded editorial comments left by the original transcribers or marked with triangular brackets so as to exclude them from analysis. Non-dialogic character names (i.e. the name of the character speaking each line) and stage directions were also marked for exclusion, but names used within the dialogue were left largely unmarked. However, in scripts where certain character names were derived from common nouns (e.g. *Lucky, Darling, Snow White*), those names were excluded from analysis so as to eliminate confounding of the counts. Interjections and ejaculations were either edited to standardize the spelling within each script, or marked for exclusion. Though the decision whether to standardize or exclude was not consistent, these elements are not germane to the ensuing analyses.

Next, each script was edited to restore contracted items to their long forms, to eliminate the stuttered or incomplete false starts to items that ultimately appeared in complete form, and to put words spelled orally rather than spoken into their correct written form (see Appendix for examples). These steps were taken to eliminate apostrophes and hyphens that had not been set off by spaces. Finally, proper nouns not excluded from analysis were compiled into a supplementary name file for later use as a secondary noun baseword file (see below).

Logs were kept that detailed the changes made during the text preparation stage (see the appendix for an example). Unfortunately, due to the evolving nature of the process and differences in the original scripts, logs were not updated consistently.

Software preparation and analysis

Two software programs were to be used in the analyses, Frequency version 3.4 (n.d.) and Range, version 3.2 (n.d.). The Frequency program can be used to generate word frequency lists for texts. The Range program compares vocabulary found in a text file with an array of baseword lists and returns data as to the number of items on each list in word tokens, types, and families. While these programs can use especially generated

word lists for reference, for the analyses in this paper only minimal modifications were made to the baseword lists that accompanied the program. These baseword lists included the most frequent 14,000 words adopted from the British National Corpus and supplemented by names and interjections.

For the analyses, in order to avoid counting non-excluded proper nouns derived from common nouns counted as common nouns, the supplemental noun list was inserted after the 2,000-word level baseword file, and the other baseword files shifted down (see Table 1 for baseword description). The files were then analyzed using the Frequency and Range programs.

Each of the files was processed first using the Frequency program (Nation & Heatley, 2002), and the vocabulary lists generated by the program checked for problematic entries. Once these were identified, the data texts were re-edited and then rerun through the program in an iterative process.

Results and Discussion

Scripts can be analyzed using various metrics, including tokens, word types, and word families. Each of these provides insights into the vocabulary range of text. Research questions 1 and 2 focus on the range and frequency level of vocabulary found in these films. Comparative figures for these metrics are presented in Table 2. Results indicate rather wide differences between the eight movies, with word families ranging from 479 to 1,351, word types from 611 to 1,792, and tokens from 2,708 to 10,147.

Table 3 provides the data about the distribution of word tokens, types, and families by level for each of the eight films. Though informative, a clearer representation of the data and the distribution of words by token, types, and families can be seen in the figures. Figure 1 gives the distribution of tokens by level. As can be seen, *Bambi* has far fewer tokens than *The Lion King*. However, the distribution in percent (see Figure 2) shows the films have similar distributions of word tokens from each of the frequency bands of the British National Corpus (BNC). In addition, the most frequent 2,000 words do not make up more than 90% of the tokens in any of the films.

Figures 3 and 4 show the same metrics for word types. Here, too, the films show a wide range from a minimum of 611 for *Bambi* and to a maximum of 1,792 for *The Lion King*. Distribution of these words by BNC level shows a distribution similar to that of the work tokens. However, the most frequent 2,000 word types make up less than 85% of the word types included, with *The Lion King* being the most difficult with only 66% of the word types within the first 2,000 of the BNC.

Finally, the same patterns hold true for word families, with *Bambi* showing again

the fewest number of total word families (479) and the highest percentage of those staying within the 2,000 frequency band (77%), while *The Lion King* includes 1,351 word families, with only 59% falling into the 2,000 and more frequent bands of the BNC (see Figures 5 & 6). However, four of the eight films were clustered near the 8000 word family level. Even if we suppose that 95% familiarity of the vocabulary families in the scripts is sufficient for enabling comprehension, then viewers would need a vocabulary reaching into the upper bands of the BNC.

Finally, if 98% coverage were assumed to be the level needed to understand new vocabulary from the context of the movies, learners would need to know 450 word families to understand *Bambi*, nearly 650 for *Snow White*, 900 for *Little Mermaid*, and more than 1,150 for either *Hercules* or *Lion King* (see Figure 7).

One other interesting facet of these eight films is that they seem to fall into three broad bands, with the older Disney films *Bambi*, *Cinderella*, and *Snow White and the Seven Dwarfs* at the easiest level, *The Little Mermaid*, *101 Dalmatians*, and *The Lady and the Tramp* at a middle level, and *The Lion King* and *Hercules*, the most recently released film, at the upper level.

While each of these films may contains some of the same items from each of the frequency bands in the BNC, it is difficult to estimate the number of items that are shared between one or more of the films. A rough estimate, though, can be found examining the number of films that contain each item. Figures 8 and 9 show the dispersions and indicate that only about a quarter of all families and types appear in four films or more. Unfortunately for learners, more than 50% of the items appear in only one of the films. Due to space limitations, it is not possible to include that list in this paper.

Research question 3 focuses on the degree to which *Shrek* is representative of animated films in general. Table 3 shows a comparison of *Shrek* with the eight films in this study. In general, *Shrek* seems to match up with the upper level of the Disney films. In this sense, we can say that it is indeed representative of the upper level of animated films to many L2 learners.

Finally, research question 4 focuses on the implications for EFL teaching. Judging solely by the vocabulary level of these films, students would need to have a rather large vocabulary to understand the contents, and only the easiest films would be accessible.

This is however, complicated by other factors. There are important differences between watching a movie and reading the script. First, the visuals in a movie usually convey a great deal of information that serves to support the script. Moreover, in some cases the unknown vocabulary will be repeated several times during the film. The combination of repetition with visual support may make the context such that the new words could be learned implicitly.

Disney is a major entertainment corporation that actively tries to penetrate markets with products (books, comics, talking toys, educational products) that are spun off from the animated films the studio releases. These products are widely available and help to create a situation where the viewer of the film may already be familiar with the characters. This may serve to make comprehension easier.

Unlike reading a book, watching a movie is often a social event enjoyed in the company of others. For children, the main target audience for Disney films, viewing is often done in groups together with peers, siblings, or parents. The presence of others during viewing offers opportunity for mediation (Bransford, Brown, & Cocking, 2000; Cole, John-Steiner, Scribner, & Souberman, 1978; Williams & Burden, 1997), and a certain amount of content may be understood after assistance from others. Finally, the medium of video (tape or disc) is such that multiple viewings are permitted, thereby giving the viewer multiple opportunities to engage with the text and provide a sort of fluency practice.

If these conditions are met in the EFL learning environment, it may be that watching animated films can help second language learners bridge the gap between where graded readers and vocabulary lists leave off and where authentic text begins.

Conclusion

Previous research has shown that animated films may contain several thousand word tokens, types, and families. This study also supports that finding. In general, there is a wide variation in the range and level of vocabulary found in animated films and there may be several films at difficulty levels appropriate to a range of second language learners. Moreover, these films may provide an alternative or supplement to graded readers and word lists. Unfortunately, many of the words appear in only one of the films, and so even someone familiar with all of the vocabulary in one of the films would not necessarily have an easy time understanding the vocabulary in another, even one at approximately the same difficulty level.

Two specific areas where more research is needed include examining other films to see if the patterns found here hold true and exploring how vocabulary learning can be accomplished by watching movies.

Footnotes

[1] This change in mode almost certainly impacts upon the results and a number of caveats are presented in the discussion section below.

References

Allers, R., & Minkoff, R. (Dir.). (1994). *The lion king*. U.S.A.: Disney Studios.

Baddeley, A. (1990). Massed and distributed practice. In *Human memory* (pp. 152–158). London: Lawrence Erlbaum Associates.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn*. Washington, D.C.: National Academy Press.

Chung, T. M. (2007). A specialised word list for reading newspapers. Paper presented at the Korea Association of Teachers of English, Annual Conference, July 7, Anyang, Korea, Retreived September 30, 2009 from http://kagee.withch.net/public_html/hwp/2007_16.hwp.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116.

Clements, R., & Musker, J. (Dir.). (1989). *The little mermaid*. U.S.A.: Disney Studios.

Clements, R., & Musker, J. (Dir.). (1997). *Hercules*. U.S.A.: Disney Studios.

Cole, M., John-Steiner, V., Scribner, S., & Souberman, E. (Eds.). (1978). *Mind in society: The development of higher psychological processes*. Cambridge, Massachusetts: Harvard University Press.

Coxhead, A. (1998). *An academic word list*. Wellington: Victoria University of Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.

Geronimi, C., & Jackson, W. (Dirs.). (1950). *Cinderella*. U.S.A.: Disney Studios.

Geronimi, C., & Jackson, W. (Dirs.). (1955). *Lady and the tramp*. U.S.A.: Disney Studios.

Geronimi, C., & Luske, H. (Dirs.). (1961). *101 dalmatians*. U.S.A.: Disney Studios.

Hand, D. (Dir.). (1942). *Bambi*. U.S.A.: Disney Studios.

Hand, D. (Dir.). (1942). *Snow White and the seven dwarfs*. U.S.A.: Disney Studios.

Hseuh-chao, M. H., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17*(1), 23–59.

Nagy, W. E., & Herman, P. A. (1984). *Limitations of vocabulary instruction*. Center for the Study of Reading: Bolt Beranek and Newman Inc.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24*(2), 237–270.

Nagy, W. E., Herman, P., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233–253.

Nation, I. S. P. (1985). Testing vocabulary size. *Proceedings of ATESOL 4th Summer School, Vol 3*, 1–10.

Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size: Essential, but rarely followed, procedures. *Language Testing, 10*(1), 27–40.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82.

Nation, I. S. P., & Heatley, A. (2002). Range and Frequency: programs for processing text. *LALS, Victoria University of Wellington, New Zealand*.

Nation, I. S. P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12*(2), 355–380.

Pimsleur, P. (1967). A memory schedule. *Modern Language Journal, 51*(2), 73–75.

Tinkham, T. (1993). The effects of semantic clustering on the learning of second language vocabulary. *System, 21*(3), 371–380.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 131–163.

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.

Williams, M., & Burden, R. L. (1997). *Psychology for language teachers*. Cambridge: Cambridge University Press.

Table 1. Description of baseword vocabulary lists

Wordlist File	Contents	Note
Basewrd1	Level 1, BNC 1,000	
Basewrd2	Level 2, BNC 2,000	
Basewrd3	Level 3, Names	Ths contained names from the scripts that had both proper and common noun functions.
Basewrd4	Level 4, BNC 3,000	
Basewrd5	Level 5, BNC 4,000	
Basewrd6	Level 6, BNC 5,000	
Basewrd7	Level 7, BNC 6,000	
Basewrd8	Level 8, BNC 7,000	
Basewrd9	Level 9, BNC 8,000	
Basewrd10	Level 10, BNC 9,000	
Basewrd11	Level 11, BNC 10,000	
Basewrd12	Level 12, BNC 11,000	
Basewrd13	Level 13, BNC 12,000	
Basewrd14	Level 14, BNC 13,000	
Basewrd15	Level 15, BNC 14,000	
Basewrd16	Level 16, Names	
Basewrd17	Level 17, Interjections	This contained interjections and ejaculations

Note: BNC = *British National Corpus*

Table 2. Word tokens, types, and families by level

	101 Dalmatians		Bambi		Cinderella		Hercules		Lady and the Tramp		The Lion King		Little Mermaid		Snow White	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Tokens																
Level 1, BNC 1,000	5,568	81	2,217	82	4,688	82	7,814	83	4,909	84	8,196	81	5,335	84	3,448	82
Level 2, BNC 2,000	297	4	202	7	265	5	442	5	283	5	537	5	308	5	318	8
Level 3, Names	280	4	79	3	180	3	280	3	99	2	323	3	149	2	0	0
Level 4, BNC 3,000	164	2	63	2	122	2	257	3	122	2	234	2	144	2	126	3
Level 5, BNC 4,000	60	1	22	1	52	1	132	1	82	1	198	2	102	2	48	1
Level 6, BNC 5,000	110	2	50	2	85	1	56	1	37	1	111	1	30	0	48	1
Level 7, BNC 6,000	12	0	9	0	18	0	46	0	31	1	66	1	25	0	20	0
Level 8, BNC 7,000	23	0	1	0	32	1	30	0	16	0	35	0	27	0	12	0
Level 9, BNC 8,000	10	0	3	0	13	0	16	0	21	0	40	0	12	0	2	0
Level 10, BNC 9,000	13	0	6	0	14	0	11	0	14	0	24	0	14	0	6	0
Level 11, BNC 10,000	8	0	2	0	13	0	12	0	14	0	39	0	10	0	8	0
Level 12, BNC 11,000	9	0	5	0	11	0	8	0	1	0	28	0	9	0	10	0
Level 13, BNC 12,000	5	0	3	0	6	0	10	0	4	0	9	0	2	0	6	0
Level 14, BNC 13,000	5	0	1	0	24	0	23	0	13	0	20	0	8	0	3	0
Level 15, BNC 14,000	2	0	1	0	0	0	5	0	2	0	5	0	6	0	0	0
Level 16, Names	9	0	0	0	1	0	10	0	19	0	17	0	5	0	1	0
Level 17, Interjections	189	3	17	1	132	2	124	1	117	2	143	1	94	1	113	3
Other words	80	1	27	1	44	1	86	1	66	1	122	1	60	1	56	1
Total	6,844		2,708		5,700		9,362		5,850		10,147		6,340		4,225	
Types																
Level 1, BNC 1,000	659	58	408	67	601	60	809	48	674	57	888	50	649	55	466	56
Level 2, BNC 2,000	165	15	84	14	133	13	251	15	178	15	286	16	167	14	132	16
Level 3, Names	35	3	3	0	22	2	47	3	25	2	9	1	26	2	0	0
Level 4, BNC 3,000	88	8	34	6	81	8	145	9	86	7	171	10	105	9	78	9
Level 5, BNC 4,000	35	3	17	3	34	3	81	5	48	4	90	5	55	5	36	4
Level 6, BNC 5,000	35	3	17	3	29	3	39	2	26	2	68	4	23	2	23	3
Level 7, BNC 6,000	11	1	7	1	10	1	37	2	19	2	43	2	21	2	13	2
Level 8, BNC 7,000	13	1	1	0	12	1	28	2	13	1	30	2	21	2	9	1
Level 9, BNC 8,000	5	0	3	0	7	1	16	1	15	1	23	1	12	1	2	0
Level 10, BNC 9,000	9	1	3	0	6	1	10	1	10	1	18	1	14	1	5	1
Level 11, BNC 10,000	7	1	1	0	10	1	10	1	10	1	19	1	9	1	7	1
Level 12, BNC 11,000	9	1	4	1	8	1	8	0	1	0	17	1	9	1	5	1
Level 13, BNC 12,000	3	0	3	0	3	0	7	0	4	0	6	0	2	0	3	0
Level 14, BNC 13,000	5	0	1	0	4	0	16	1	8	1	9	1	8	1	1	0
Level 15, BNC 14,000	2	0	1	0	0	0	3	0	2	0	5	0	3	0	1	0
Level 16, Names	3	0	0	0	1	0	8	0	11	1	17	1	2	0	9	1
Level 17, Interjections	5	0	6	1	7	1	12	1	10	1	0	0	9	1	0	0
Other words	43	4	18	3	27	3	142	9	51	4	93	5	51	4	37	4
Total	1,132		611		995		1,669		1,191		1,792		1,186		827	

	101 Dalmatians		Bambi		Cinderella		Hercules		Lady and the Tramp		The Lion King		Little Mermaid		Snow White	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Families																
Level 1, BNC 1,000	476	53	293	61	441	55	534	43	480	51	553	41	456	48	341	51
Level 2, BNC 2,000	140	15	76	16	119	15	214	17	150	16	242	18	150	16	112	17
Level 3: Names	34	4	3	1	20	3	42	3	24	3	9	1	24	3	0	0
Level 4, BNC 3,000	77	9	33	7	71	9	128	10	82	9	141	10	98	10	74	11
Level 5, BNC 4,000	34	4	14	3	31	4	73	6	42	4	81	6	51	5	36	5
Level 6, BNC 5,000	33	4	17	4	26	3	38	3	24	3	65	5	22	2	23	3
Level 7, BNC 6,000	11	1	6	1	10	1	37	3	19	2	40	3	20	2	13	2
Level 8, BNC 7,000	12	1	1	0	12	2	27	2	12	1	27	2	21	2	9	1
Level 9, BNC 8,000	4	0	2	0	7	1	16	1	12	1	21	2	12	1	2	0
Level 10, BNC 9,000	8	1	3	1	6	1	10	1	9	1	18	1	14	1	5	1
Level 11, BNC 10,000	7	1	1	0	9	1	9	1	10	1	16	1	8	1	7	1
Level 12, BNC 11,000	8	1	4	1	8	1	8	1	1	0	17	1	9	1	5	1
Level 13, BNC 12,000	3	0	3	1	3	0	5	0	4	0	6	0	2	0	3	0
Level 14, BNC 13,000	5	1	1	0	4	1	15	1	8	1	9	1	8	1	1	0
Level 15: BNC 14,000	2	0	1	0	0	0	3	0	2	0	5	0	3	0	0	0
Level 16: Names	3	0	0	0	1	0	8	1	11	1	4	0	2	0	1	0
Level 17: Interjections	4	0	3	1	4	1	4	0	4	0	4	0	4	0	3	0
Other words	43	5	18	4	27	3	71	6	51	5	93	7	51	5	37	6
Total	904		479		799		1242		945		1351		955		672	

Table 3. A comparison of word tokens, word types and word families

Metric	Shrek	101 Dalmatians	Bambi	Cinderella	Hercules	Lady and the Tramp	The Lion King	The Little Mermaid	Snow White
Tokens	9,984	6,844	2,708	5,700	9,362	5,850	10,147	6,340	4,225
Types	1,426	1,132	611	995	1,669	1,191	1,792	1,186	827
Families	1,097	904	479	799	1,242	945	1,351	955	672

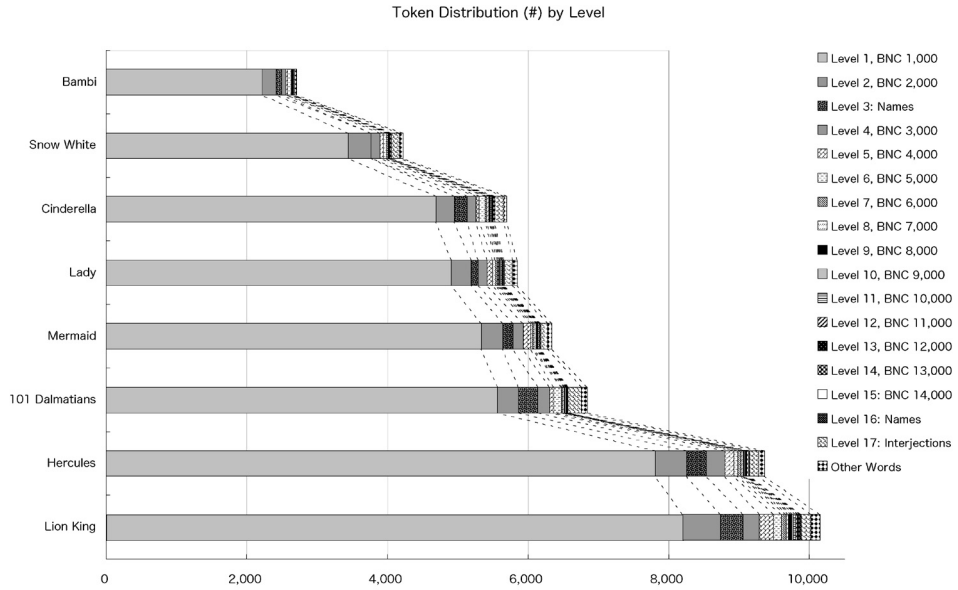

Figure 1. Token distribution by levels in number.

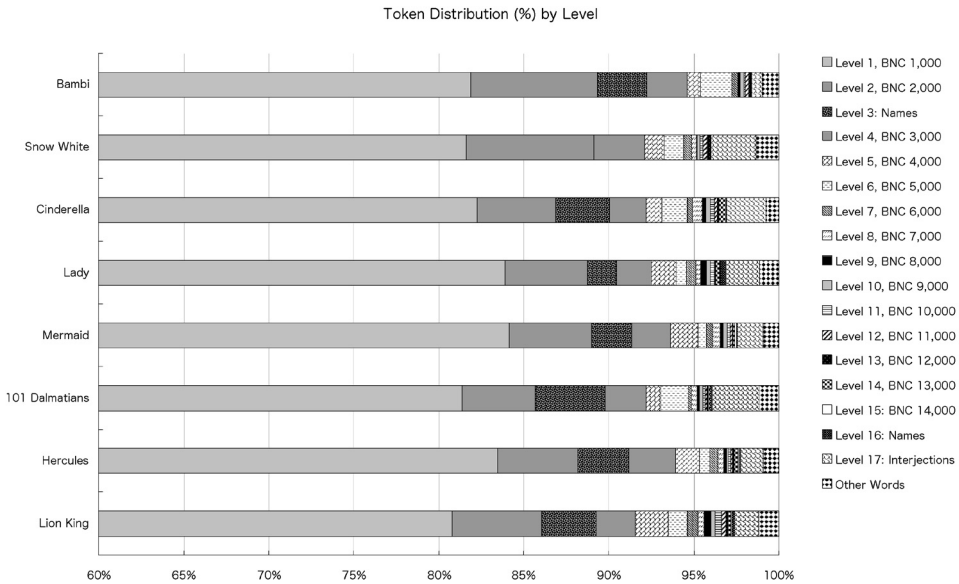

Figure 2. Token distribution by levels in percent

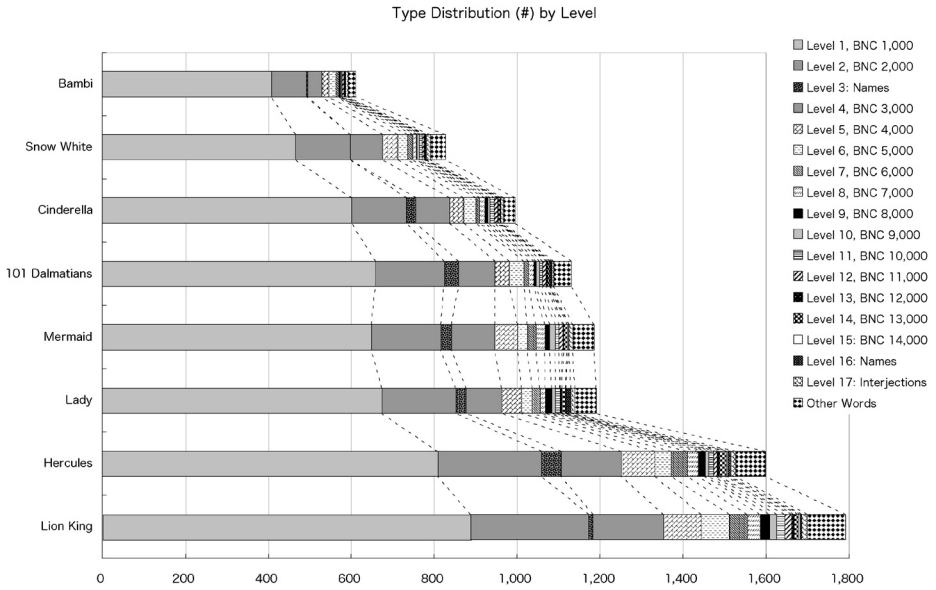

Figure 3. Word type distributions by levels in number.

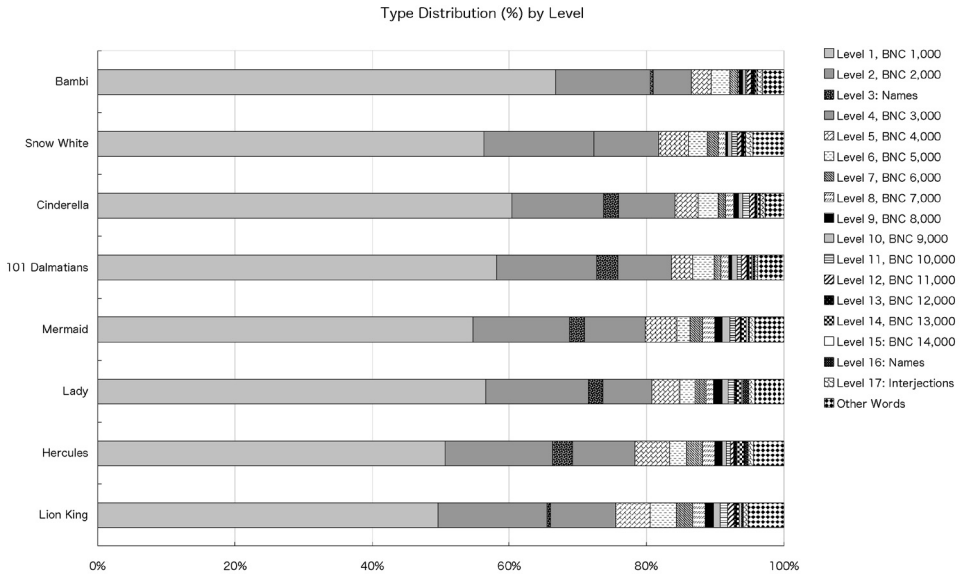

Figure 4. Word type distribution by levels in percent.

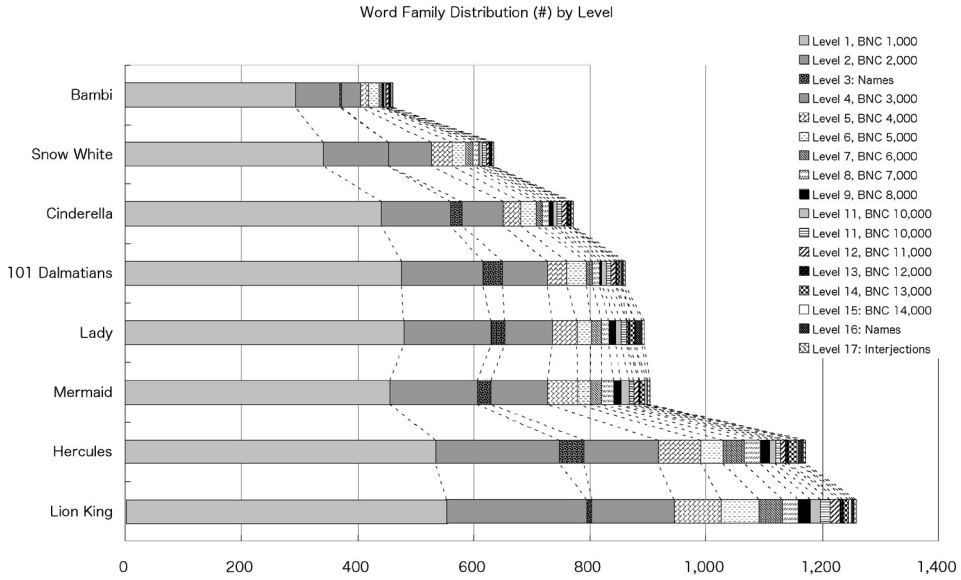

Figure 5. Word family distribution by level in number.

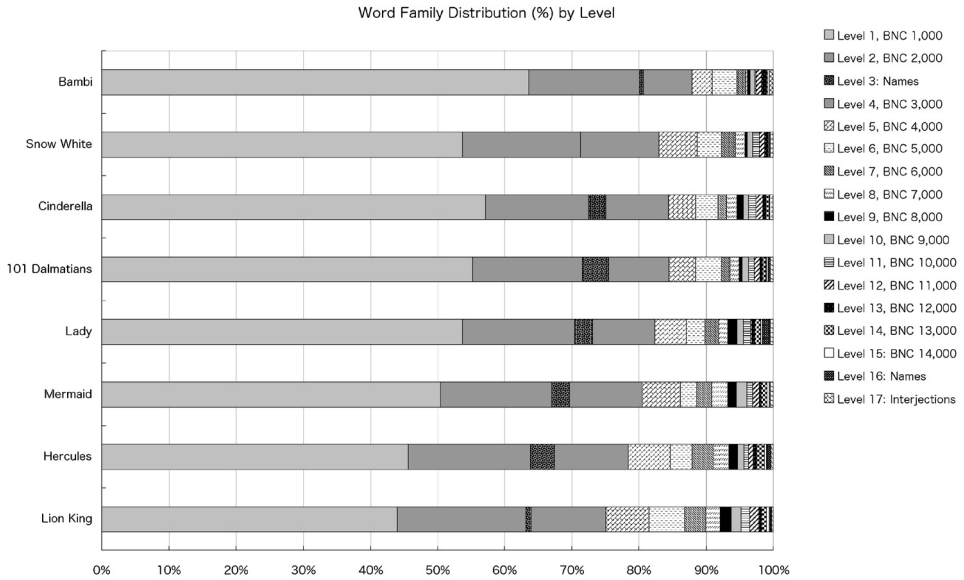

Figure 6. Word family distribution by level in percent.

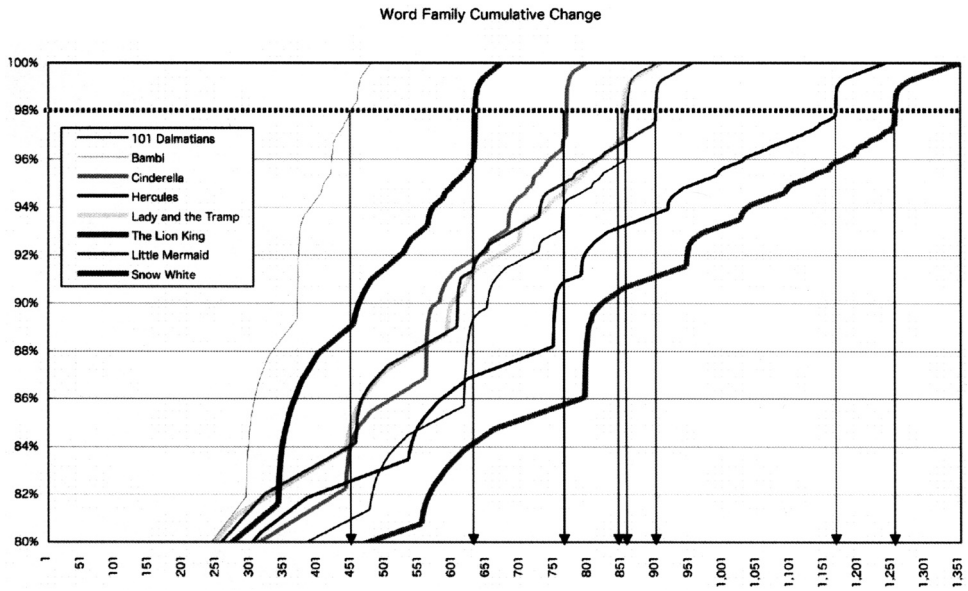

Figure 7. Change in word family number in cumulative percent of text with the 98% percent threshold for unassisted reading marked.

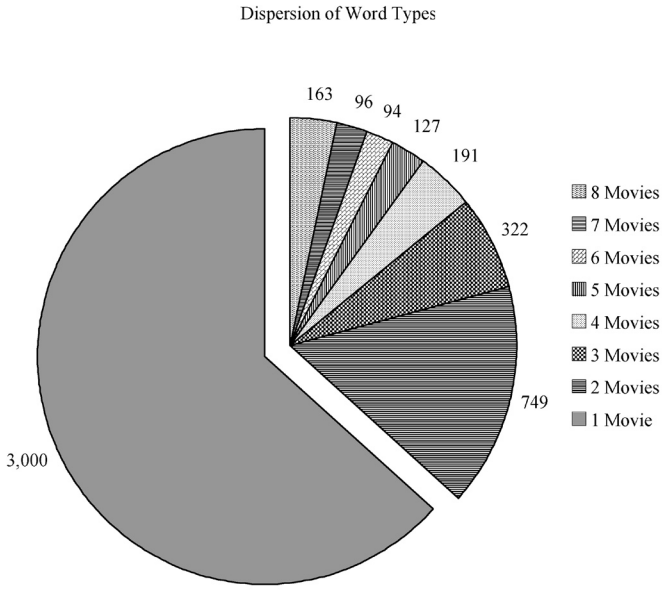

Figure 8. Dispersion of Word Types among Films

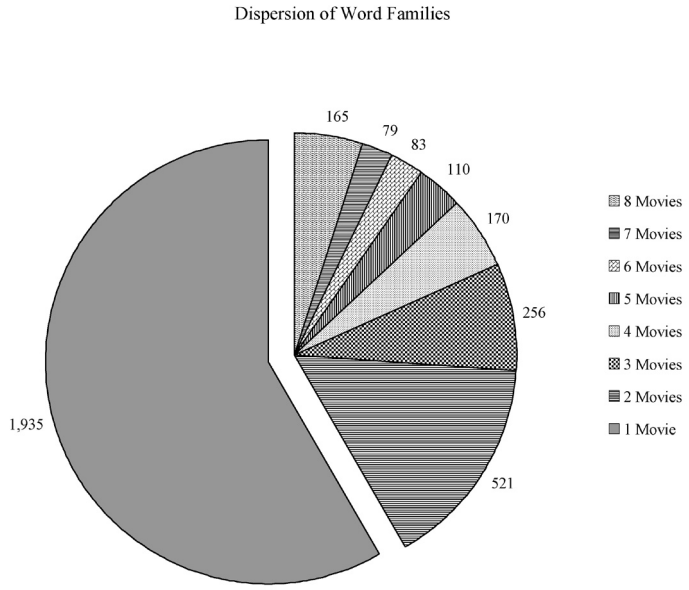

Figure 9. Dispersion of Word Families among Films

Appendix: Editing Log for 101 Dalmatians

Proper nouns added to the supplementary baseword list

- Anita
- Beethoven
- Birdwell
- Blimey
- Caruden
- Coco
- Cruella
- Cruella de Vil
- Dawson
- Dinsford
- Ducky
- Fauncewater
- George
- Hampstead
- Horace
- Jasper
- Jove
- Kanine
- Krunchies
- London
- Lucy
- Nellie
- Percival
- Perdy
- Pongo, Pongos
- Queenie
- Regents Park
- Roger Radcliff
- Rolly
- Suffolk
- Tibs
- Towser
- Vil
- Withermarsh
- de
- Christmas

DUAL USE
Names, marked for exclusion because they also have non-Proper noun functions:

- Bob
- Duchess
- Freckles
- Great Dane
- Hell Hall
- Lucky
- Meathead
- Nanny
- Pepper
- Primrose Hill
- Princess
- Scotland Yard
- Tartar
- Thunder
- Thunderbolt

Other changes

(~ in'	→ ~ ing)
Aaah	→ Ah
Ahem	left as is
Aw	excluded
Baduns	left as is
Cheerio	left as is
C'mon	→ Come on
D-do	→ Do
Dognapping	excluded
Eye-ther	→ either
Fiddle faddle	left as is
H-H-How	→ How
Huh	left as is
I-I-I-I	→ I
I's	→ I'd
N-n-not	→ Not
Oooh	→ Oh
Psst	left as is
Roof	left as is
Shhh	excluded
Ta-ta	left as is
Ta-tum-ti-ta-tum	tat um tit a tum
Taint	→ It ain't → it is not
That's witch	→ That witch
Ugh	left as is
Uh	excluded
Uh-oh	→ Uh oh
Woo-woo	left as is
Woof	left as is
a la mode!	left as is
ah	excluded
ain't	is not
all-dog	all dog
bloomin'	→ blooming
collywobbles	left as is
determinated	→ determined
eh	left as is
elen	→ eleven
git	→ got
gonna	→ going to
gotta	→ got to
ha-ha	→ ha ha
hmm	excluded
ho ho	left as is
hoodlums	left as is
mangry	→ mangy
ma'am	→ madam
missus	left as is
n-n-not	→ not
ol'	→ old
oo-oo-oo	excluded
righto	→ right
roo-roo-roog	→ roof roof roof
so's	→ so as
uh-oh	excluded
w-a-l-k	→ walk
wanna	→ want to
yip	left as is
'bout	→ about
'eh	→ her
'em	→ them
'til	→ until
"ee-ther"	→ either

eliminated all apostrophizes ['] except for possessives