Word Frequency in Captioned Television

Carl Jensema, Ph.D.
Institute for Disabilities Research and Training, Inc.
2424 University Boulevard West
Silver Spring, MD 20902
(301) 942-4326 V/TTY
Email: IDRT@aol.com

Michele R. Rovins, M.A.
Institute for Disabilities Research and Training, Inc.
2424 University Boulevard West
Silver Spring, MD 20902
(301) 942-4326 V/TTY
Email: MROVINS@aol.com

Introduction

Reading is often one of the main ways deaf people gain information and develop independence in learning. In recent years, television captioning has become a prime source of reading material. As Koskinen, Wilson, and Jensema (1985) noted, "Captions are reading material...They can turn television into a moving story book, a steady stream of written language presented with both video and audio reinforcement. Viewers can see words on the screen, hear them spoken, and see them put into a visual context. One of the most exciting potential applications of closed captioning is its use as an educational tool."

The use of captioned television as reading material is difficult if there are no reading skills to begin with. People need some starting point, the ability to read at least some words. In this study, a relatively short list of frequently used words is presented. The authors believe that mastery of these words can greatly assist in expanding reading skills.

The report presented here is based on research by Jensema, McCann and Ramsey (1996). They obtained and analyzed caption data from 183 television programs and 22 music videos. The programs varied from thirty minutes to four hours, and the music videos were between two and five minutes in length. The research examined speed, word length, and similar characteristics of the captions. It was noted that relatively few distinct words accounted for a large proportion of the total words used in the captions.

In the present article, the observation that few words account for a large part of the total words used in captioning is carried further and the data is analyzed in more detail. The result is a caption word frequency list, the mastery of which is likely to provide an important assist to the reading skills of caption viewers.

Method

The caption scripts from all the programs in the study by Jensema, et al (1996) were combined into one large computer file. The file was edited to remove punctuation and anything else, which was not a word. The resulting file had 834.726 words.

It was decided that many words were merely variations of another word. Word endings of "s," "es," "ed," "ing," and "d" were deleted. On the other hand, certain endings created a new word which had a different meaning. It was decided to keep word endings of "ly," "t," "ive", "ion," "er," and "ie." Certain nonstandard "words", such as "uh,"

"mmmm," and "ahhh" were kept, since they are commonly used in captioning to indicate certain sounds in the audio.

The resulting edited 834.726 word list was sorted alphabetically, duplicate words were counted and then deleted, and the remaining list was sorted by frequency of occurrence. The final frequency list had 16,102 unique words, most of which were used only a few times.

Results

Table 1 presents a frequency count of the 250 words used most often in the television captions in this study. Out of 834,726 captioned words, 30,142 were the word "the," 22,600 were the word "you," and so on.

In Table 1, the word "the" accounted for 3.61% of the 834,726 captioned words. The words "the" and "you" together accounted for 6.32% of the 834,726 captioned words. Continuing in this manner, Table 1 shows that 250 unique words account for over 68% of all the words used in captioned television.

Discussion

The implications of Table 1 are striking. there are more than 500,000 words in the English language, but a person who masters the use of the 250 words in Table 1 will recognize more than two-thirds of all words shown in television captions. This is a tremendous advantage for any person with limited reading skills who attempts to read captioned television.

A beginning reader could be taught just 10 words (the, you, to, a, I, and, of, in, it, that) and would then recognize more than one out of every five words which appeared on a captioned television program. Being able to read 79 words means being able to read half of all words captioned. By using Table 1 as a guideline in teaching reading, a teacher can maximize the captioned words a student will recognize while watching television. It is suggested that teachers of deaf and hard of hearing students consider Table 1 carefully in planning their strategy for teaching reading.

The majority of the words on the list are everyday linking words, including many prepositions and pronouns. Prepositions, in particular, are traditionally problem areas for many deaf students because American Sign Language does not have prepositions.

Research shows that students learn vocabulary both definitionally and contextually (Stahl & Fairbanks, 1986). The words in Table 1 can be taught definitionally in context. Those students who develop a working knowledge of the 250 words will be able to apply them in a variety of situations and will be able to focus on other captioned television words that they many not understand.

References

Jensema, C., McCann, R. & Ramsey, S. (1996) Closed-Captioned Television Presentation Speed and Vocabulary. *American Annals of the Deaf.* 141(4), 284-292

Koskinen, P.S., Wilson, R.M. & Jensema, C.J. (1985). Closed-Captioned Television: A New Tool for Reading Instruction. *Reading World.* 24, 1-7.

Stahl, S.A. & Fairbanks, M.M. (1986) The Effects of Vocabulary Instruction. A model-based meta-analysis. *Review of Educational Research*, 56, 72-110.

Word	Freq	Cum%	Word	Freq	Cum%	Word	Freq	Cum%	Word	Freq	Cum%	Word	Freq	Cum%
the	30,142	4	if	2,751	42	us	1,381	54	let	863	61	am	598	65
you	22,600	6	want	2,730	43	I'll	1,369	54	life	859	61	long	593	65
to	22,161	9	as	2,714	43	yes	1,364	55	other	852	61	ask	587	65
a	20,023	11	now	2,696	43	he's	1,359	55	night	831	61	today	587	65
I	19,991	14	she	2,686	44	thank	1,352	55	they're	829	61	name	583	65
and	16,130	16	think	2,606	44	little	1,351	55	help	805	61	run	583	65
Of	13,914	17	her	2,591	44	love	1,340	55	happen	802	61	place	581	65
in	10,941	19	go	2,584	45	why	1,278	55	what's	800	62	stop	580	66
it	10,496	20	will	2,522	45	really	1,263	56	those	784	62	which	570	66
that	10,395	21	well	2,442	45	tell	1,256	56	than	782	62	sorry	566	66
is	8,764	22	going	2,428	45	over	1,249	56	find	776	62	friend	564	66
this	7,116	23	his	2,409	46	call	1,241	56	last	760	62	better	563	66
for	6,679	24	got	2,375	46	can't	1,192	56	world	760	62	through	562	66
on	6,411	25	from	2,373	46	where	1,179	56	after	756	62	house	559	66
was	5,945	25	that's	2,343	47	said	1,169	56	she's	743	62	does	558	66
have	5,804	26	look	2,324	47	day	1,163	57	Mr.	741	62	family	555	66
me	5,740	27	him	2,316	47	never	1,158	57	even	740	62	kind	554	66
we	5,521	27	you're	2,285	47	something	1,158	57	home	735	62	may	551	66
what	5,464	28	time	2,243	48	we're	1,155	57	again	727	62	most	548	66
be	5,449	29	when	2,231	48	then	1,140	57	made	719	63	god	530	66
he	5,218	29	see	2,230	48	two	1,133	57	big	718	63	woman	524	66
with	4,895	30	how	2,214	48	because	1,115	57	doing	718	63	many	512	66
my	4,834	31	say	2,200	49	their	1,089	58	please	712	63	hi	510	67
your	4,385	31	good	2,155	49	hey	1,087	58	put	711	63	nothing	509	67
do	4,375	32	by	2,115	49	first	1,065	58	lot	709	63	next	508	67
I'm	4,258	32	had	2,041	49	need	1,049	58	should	700	63	move	503	67
are	4,224	33	yeah	1,971	50	too	1,048	58	before	694	63	another	499	67
all	4,129	33	an	1,968	50	didn't	1,040	58	around	688	63	came	498	67
not	4,117	34	would	1,899	50	he	1,034	58	wait	688	63	tonight	495	67
it's	4,111	34	did	1,804	50	new	1,023	58	still	687	63	left	493	67
know	3,962	35	take	1,794	51	talk	1,020	59	start	684	64	turn	484	67
no	3,890	35	we're	1,765	51	into	1,012	59	live	680	64	doesn't	483	67

but	3,885	35	make	1,757	51	work	1,007	59	use	675	64	I'd	482	67
don't	3,859	36	back	1,739	51	play	1,006	59	sure	674	64	neither	481	67
get	3,739	36	who	1,719	51	try	998	59	keep	671	64	must	476	67
they	3,612	37	been	1,707	52	much	998	59	sir	670	64	kill	472	67
like	3,436	37	has	1,697	52	guy	987	59	old	667	64	hand	470	67
so	3,425	38	them	1,599	52	I've	980	59	maybe	657	64	stay	468	67
just	3,300	38	or	1,553	52	uh	976	60	we'll	653	64	watch	467	67
at	3,295	38	some	1,547	52	mean	954	60	thought	652	64	you've	467	68
here	3,197	39	man	1,529	53	there's	954	60	believe	650	64	children	465	68
out	3,117	39	very	1,510	53	only	938	60	boy	646	64	hear	463	68
up	3,074	40	our	1,475	53	give	924	60	three	644	64	hope	462	68
about	3,031	40	down	1,474	53	off	920	60	every	641	65	mother	455	68
one	2,998	40	thing	1,456	53	any	917	60	caption	639	65	nice	455	68
right	2,906	41	way	1,431	53	feel	907	60	ever	639	65	remember	454	68
come	2,904	41	year	1,420	54	these	905	60	show	636	65	own	453	68
there	2,886	41	people	1,409	54	great	884	60	away	635	65	won't	451	68
oh	2,781	42	could	1,408	54	let's	884	61	always	626	65	morning	449	68
can	2,772	42	more	1,383	54	prepare	871	61	anything	607	65	everything	446	68