In Druck *Language Variation and Change*

The Sociolinguistics of a short-lived Innovation: Tracing the development of quotative *all* across spoken and internet newsgroup data [i]

Isabelle Buchstaller, Newcastle University
John R. Rickford, Stanford University
Elizabeth Closs Traugott, Stanford University
Thomas Wasow, Stanford University and
Arnold Zwicky, Stanford University

ABSTRACT
This paper examines a short-lived innovation, quotative all, in real and apparent time. We used a two-pronged method to trace the trajectory of all over the past two decades: (i) Quantitative analyses of the quotative system of young Californians from different decades; this reveals a startling cross-over pattern: in 1990/4 all predominates, but by 2005 it has given way to like; (ii) Searches of internet newsgroups[ii]; these confirm that after rising briskly in the 1990s, all is declining. Tracing the changing usage of quotative options provides year-to-year evidence that all has recently given way to like. Our paper has two aims: We provide insights from ongoing language change regarding short-term innovations in the history of English. We also discuss our collaboration with Google Inc. and argue for the value of newsgroups to research projects investigating linguistic variation and change in real time, especially where recorded conversational tokens are relatively sparse.

1. INTRODUCTION

Until very recently, sociolinguistic research on the North American quotative system tended to focus on a few, by now well-researched, variants, such as like and go (Bakht-Rofheart, 2002; Buchstaller, 2004; Buchstaller and D'Arcy, 2009; Barbieri, 2005, 2007; Cukor-Avila, 2002; Romaine and Lange, 1991; Tagliamonte and D'Arcy, 2007).

(1) I'm **like** *"oh my uncle's calling me it must be important"*
(2) I **go** *"I seen you following me for a couple of miles now."*

Only in the last few years has the literature started to pick up on another, apparently new variant, quotative all, as in (3-5), (see Bayley and Santa Ana, 2004; Rickford, 2000; Singler, 2001)

(3) He's **all** *"well let me check em alright oh I'm sorry bout that"*
(4) I'm **all**, *"Dude, you're not helping your cause!"*
(5) She's **all** *"Ooh- he's so wonderful - I'm all in love with him - he's all in love with me."*

All's extension to quotative function is new. Quotative all is not in the *OED*, nor in any of the modern dictionaries except the 4th edition of the *American Heritage Dictionary*. The Switchboard Corpus I (collected in 1988-1992) and the Santa Barbara Corpus of spoken English (collected in 1988) each contain only one token of quotative all. The earliest report of quotative all we have found is in the fall 1982 issue of the newsletter "Not Just Words" edited by Danny Alford at the University of California at Berkeley. In terms of its regional pattern, quotative all has previously been attested primarily in California (Alford, 1982; Fought, 2003; Rickford, 2000; Rickford, Buchstaller, Wasow, & Zwicky, 2007; Waksler, 1991; Wimmer, 1990) but also in Arizona (Barbieri, 2005), Texas (Bailey and Santa Ana, 2004), New York (Singler, 2001) and Ontario, Canada (Tagliamonte and D'Arcy, 2004, 2007) and even in England (Buchstaller, 2004).

In earlier work, we discussed the relationship between all in intensifier and quotative function (Buchstaller and Traugott, 2006) as well as its social and linguistic constraints (Rickford et al., 2007). We have shown that the frequency of all in the quotative system decreases considerably in recent years and that the overall decline goes hand in hand with a shift in its constraint hierarchy. In this paper, we zoom in on the change of this relatively new variant. Using a combination of quantitative variationist and computational methodology, we focus on the recent history of the quotative variant in apparent and real time. As a first step, we trace the relative frequency of all in the set of quotative introducers used in recordings from California youth from 1990/4 until 2004/5. Moving beyond the Californian context, we then discuss the results of a collaborative research project with Google Inc., which allowed us to track the diachronic development of all versus other quotatives options in greater detail. Focusing on the distribution of quotative variants with different types of interpretations (speech, thought, or stereotypes) across time, we show that all has indeed taken on a quite specialized function within the quotative pool.

The investigation of both real and apparent time data leads us to conclude that quotative all is a rather short-lived innovation. It exhibits a steep drop-off, both in the comparison between the interviews conducted in the 1990s and those conducted in 2004/5, and in the Google corpus spanning the years 1982-2006, being replaced by like, which has been attested since the 1980s, in both instances. The extent of the shift from all to like also shows up in the proliferation of the intermediate form all like, as in (6) and (7):

(6) He's **all like** *"You know little punk. Say another word just keep on .."*
(7) She was **all like** um *"Yeah at my school knitting was banned"*

Looking specifically at the interaction between all and all like across real time we will detail the extent to which all has given way to like in the first few years of this century. The rise and fall of quotative all provides insight from language change in progress for similar short-term innovations and their actualization in earlier English (cf. *stinten* 'to stop V-ing' in Middle English). But before we get into our analysis we will first discuss the data-sets on which this study is based.

2. DATA

For this paper, we will report on three principal sources:

A) The 1990 -1994 Wimmer/Fought tape recorded corpus (WFTRC) collected in California from 12 high school and undergraduate students and young adults, who were all born in California and have never left the state for any protracted amount of time. The corpus consists of two sets of conversational recordings: one set was collected by Ann Wimmer for her Stanford senior honors thesis in 1990. It includes 6 middle class white speakers (ages 14-23), all from the San Francisco Bay area in Northern California. The second set, which includes 6 Chicano (Mexican American) speakers (ages 17-20) from the Los Angeles area (Southern California), was collected by Carmen Fought in 1994. These recordings, which yielded a total of 473 quotations, including 134 tokens of all (including all here and all like) and 97 tokens of like, served as a comparative base for our later corpus, recorded in Stanford in 2004/5.

B) The 2004-2005 Stanford tape recorded corpus (STRC) consists of sociolinguistic interviews with 17 Stanford University undergraduates (ages 17-22) and one graduate student (22 years old), 11 students from Gunn High School in Palo Alto, California (ages 14-18), and 3 young adults from San Francisco and Southern California (ages 24-27). The speakers were of various ethnicities but most of them could be counted as middle class (being children of highly educated parents, living in relatively affluent areas and attending a highly esteemed school / university). All speakers are native Californians and / or have spent most of their lives in California. By comparing this corpus with the earlier 1990/4 corpus, we were able to pinpoint how all has changed quantitatively, in terms of its relative frequency, and its internal constraints. This tape-recorded corpus yielded 1134 quotatives, including 26 tokens of all or all like and 820 tokens of like.

C) The Google Newsgroups corpus. In order to get a more fine-grained sense of the relative frequency of quotative all over the past two decades, we searched a massive archive of internet newsgroup postings hosted at Google. According to their webpage, when Google acquired the database from Deja.com in 2001 it contained about 500 million individual messages (http://www.google.com/press/pressrel/pressrelease48.html). Google Groups now exceeds one billion postings – hence many billion words – and it is steadily growing.[iii]

We now move on to the discussion of our findings. We first discuss the patterning of quotative all across time in the California data, and then the internet searches.

3. FINDINGS

3.1 VARIATIONIST ANALYSIS OF SPOKEN CALIFORNIA CORPORA

The overall distribution of the most frequent variants in the California corpora has shifted extensively within the last decade. For the California adolescents recorded in 1990/4, all is the most frequent single variant in the quotative system, being used by three quarters of the speakers in our corpus (9 out of 12) and making up the majority variant amongst these speakers. By 2004/5, however, the picture has changed

dramatically: Only about a third of the 32 adolescents and young adults we interviewed used the form at all and even amongst these speakers all was clearly a minority variant.

Given the inverse numerical relationship between all-users and non-users across the two corpora, we decided to represent our data split up by whether or not speakers used the quotative variant all. Table 1 includes the speakers in the 1990/4 data whose system contains all. For the California adolescents recorded in 1990/4, all is the most frequent single variant in the quotative system. While there is considerable variation across these speakers, all and all like make up about 37 % overall, with quotative like amounting to 20% and say and other (including unframed) quotes making up another 16% to 19% each. Table 2 shows the three speakers in the 1990/4 data who did not use all. What distinguishes the two groups, adopter vs. non-adopters,[iv] from one another is their age. Indeed, Ann Wimmer reported that age is the most important constraint in the 1990 corpus. "All of the high school students interviewed used it [all], but none of the college age speakers did. . . . No one in the study over the age of 19 was heard to use this variable at any time." (Wimmer, 1990: 10).

TABLE 1: Quotative variants of speakers in the Wimmer/Fought 1990/94 corpus who used all or all like

Speaker	Ethnicity, gender, age	Where from?	ALL (here)	ALL LIKE	SAY	GO	LIKE	Ø/Other	TOTAL
Mindy (MI)	WF 14	*Los Gatos*	*6 (.33)*	*0*	*5 (.28)*	*3 (.17)*	*2 (.11)*	*2 (.11)*	18
Robert (RO)	WF 14	*Los Gatos*	*15 (.48)*	*0*	*2 (.06)*	*4 (.13)*	*2 (.06)*	*8 (.26)*	31
Brandon (BG)	WM 15	*Los Gatos*	*69 (.57)*	*0*	*6 (.05)*	*5 (.04)*	*19 (.16)*	*23(.19)*	122
Carl (CW)	WM 14	*Los Gatos*	*1 (.02)*	*0*	*26 (.58)*	*7 (.16)*	*5 (11)*	*6 (.13)*	45
Damon (DH)	MAM 17	*Los Angeles*	*17 (.24)*	*3(.04)*	*8 (.11)*	*6 (.09)*	*15 (.21)*	*21 (.30)*	70
Erica	MAF17	*Los Angeles*	*13 (.45)*	*0*	*3 (.10)*	*2(.07)*	*9 (.31)*	*2 (.07)*	29
Veronica	MAF17	*Los Angeles*	*3 (.13)*	*0*	*3 (.13)*	*3 (.13)*	*13 (.54)*	*2 (.08)*	24
Christian	MAM18	*Los Angeles*	*2 (.25)*	*0*	*2 (.25)*	*0 (0)*	*2(.25)*	*2 (.25)*	8
Chuck	MAM17	*Los Angeles*	*5 (.28)*	*0*	*4 (.22)*	*0 (0)*	*6 (.33)*	*3 (.17)*	18
TOTAL			**131 (.36)**	**3 (.01)**	**59 (.16)**	**30 (.08)**	**73 (.20)**	**69 (.19)**	**365**

Notes: W=White, MA=Mexican American; M=Male, F=Female. Los Gatos (Wimmer's 1990 research site) is in the San Francisco Bay Area, Northern California; Los Angeles (Fought's 1994 research site) is in Southern California. "ALL" includes 52 tokens of "all here," used by Brandon.

TABLE 2: Quotative variants of speakers in the Wimmer/Fought 1990/94 corpus who did NOT use all or all like

Speaker	Ethnicity, Gender, age	Where from?	Tape	SAY	GO	LIKE	Ø/ OTHER	Total
Mia	WF 21	*Burlingame*	2A	36(.49)	22(.30)	16(.22)	0	74
Isadora	MAF 20	*Los Angeles*	2B	1(.04)	0	4(.17)	18(.78)	23
Kendall	WF 23	*Los Gatos*	2B	4(.36)	0	4(.36)	3(.09)	11
Total:				**41(.38)**	**22(.20)**	**24(.22)**	**21(.19)**	**108**

Notes: W=White, MA=Mexican American, M=Male, F=Female. Burlingame and Los Gatos are in the San Francisco Bay Area, N. California; Los Angeles is in Southern California.

TABLE 3: Quotative Variants of Speakers in Stanford Tape Recorded Corpus (STRC2004/5) who used all or all like

Speaker	Eth/Gen/ age/Cohort	Where from?	Tape	ALL	ALL LIKE	SAY	GO	LIKE	Ø/ OTHER	Totals
Kirsten	WF20C	S. Calif	A3	0(0)	4(.06)	0(0)	2(.01)	58(.87)	3(.04)	67
Sean	MAM19C	N. Calif	A8	6(.14)	0(0)	2(.05)	0(0)	27(.63)	8(.19)	43
Zinnia	WF20C	N. Calif	A8, A27	0(0)	1(.03)	0(0)	0(0)	30(.91)	2(.06)	33
Addison	WF16H	Calif	A14	0(0)	6(.15)	2(.05)	2(.05)	25(.63)	5(.13)	40
Eric	WM15H	Calif	A19	0(0)	1(.06)	2(.12)	0(0)	10(.59)	4(.24)	17
Isaiah	WM15H	Calif	A19	0(0)	1(.08)	0(0)	0(0)	8(.67)	3(.25)	12
Nadine	WF14H	N. Calif	A22	0(0)	2(.05)	6(.14)	1(.02)	33(.79)	0(0)	42
Fiona	WF20C	S. Calif	A22, A34	1(.05)	0(0)	0(0)	0(0)	15(.79)	3(.16)	19
Luis	MAM20C	S. Calif	A27	0(0)	2(.05)	2(.05)	0(0)	32(.78)	5(.12)	41
Jeremy	WM22JC	S. Calif	A34	2(.03)	0(0)	9(.13)	1(.01)	40(.57)	18(.26)	70
Total:				**9(.02)**	**17(.04)**	**23(.06)**	**6(.02)**	**278(.72)**	**51(.13)**	**384**

In our corpus collected from California adolescents and young adults a decade later, quotative all has decreased markedly in overall frequency as well as in the proportion of speakers who use it. Tables 3 and 4 show that in our 2004/5 corpus, all-users are clearly in the minority (10 out of 22 speakers). Note that even among those speakers

whose system contains all, it is like that has clearly established itself as the default form among the quotative introducers (72%) while all and *all like* amount to only 6%. Amongst the non-all users, like retains the same share in the system, 72%, with slightly higher frequencies of go and say.

TABLE 4: Quotative Variants of Speakers in Stanford Tape Recorded Corpus (STRC2004/5) who did NOT use all or all like

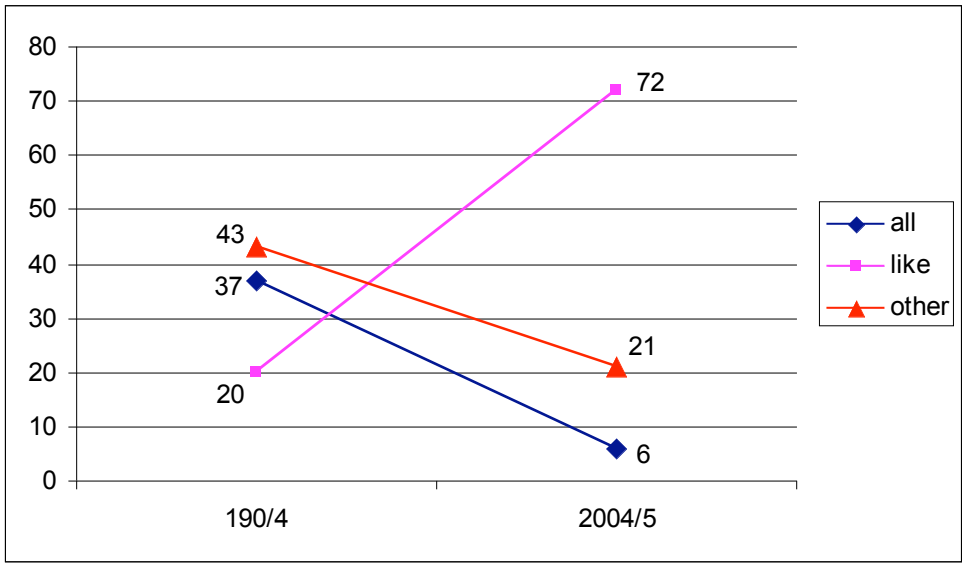
Name	Ethnicity, gender, age, cohort	Hometown	Tape	SAY	GO	LIKE	Ø/ OTHER	TOTAL
Leslie	WF17H	Palo Alto,CA	A2	0	0(0)	33(1)	0	33
Stacy (SS)	LF18C	SFO Bay Area,CA	A4	2(.15)	0(0)	11(.85)	0	13
Anna (AW)	WF18C	SFO Bay Area/LA,CA	A5	8(.18)	5(.11)	29(.66)	2(.05)	44
Jeffrey (JA)	WM21C	El Cerrito/La Jolla,CA	A6	0	0(0)	7(.88)	1(.13)	8
Kitty (KK)	WF17C	SFO Bay Area,CA	A7	0	0(0)	18(1)	0	18
Loraine (LG)	BF20C	North Ridge,CA	A9	0	1(.03)	30(.88)	3(.09)	34
Eve (EE)	WF18H	Palo Alto,CA	A10	0	0(0)	37(1)	0	37
Sergio (SE)	CRM17H	DC/Atl/LA/PaloAlto, CA	A11	0	0(0)	9(1)	0	9
Joseph (JW)	JM17C	Japan (3-6)/CA	A13	0	1(.09)	10(.91)	0	11
Mandy (MB)	WF15H	PaloAlto,CA (11+?)	A18	0	0(0)	8(1)	0	8
Jessica (JJ)	WF15H	Palo Alto,CA	A20	1(.01)	0(0)	82(.99)	0	83
Annette (AK)	WF15H	SFO Bay Area,CA	A21	0	1(.05)	20(.95)	0	21
Ellie (EE)	WF15H	NC(0-10)/PaloAlto,CA	A23	1(.02)	16(.25)	46(.72)	1(.02)	64
Sam (SB)	BM21C	NJ/MD/NC(13)/SFO BA,CA	A24	7(.33)	0(0)	11(.52)	3(.14)	21
Sandra (SE)	LF21C	Torrance,CA	A25	2(.07)	0(0)	25(.93)	0	27
Dale (DA)	WM21C	South Florida	A29	0(.11)	0(.25)	18(.56)	1(.08)	19
Kelly (KL)	PIM18C	Milpitas,CA	A29	11(.41)	9(.33)	5(.19)	2(.07)	27
Jeanine (JC)	CHF19C	San Jose,CA	A30	47(.8)	0(0)	3(.05)	9(.15)	59
Stephen (SS)	PM22G	San Diego,CA	A31	3(.1)	2(.07)	23(.79)	1(.03)	29
Guy (GG)	LM27N	LA(24)/SFO Bay Area,CA	A33	16(.18)	12(.13)	37(.42)	24(.27)	89
Rod (RP)	AM26N	Hawaii(19)/OR/SFO BA,CA	A35	7(.18)	0(0)	29(.74)	3(.08)	39
Cole (CJ)	WM24N	LA/SFO Bay Area,CA	A36	4(.12)	2(.06)	27(.82)	0(0)	33
Total::				**109 (.15)**	**49 (.07)**	**542 (.72)**	**50 (.07)**	**750**

A-Asian, B-Black, CH-Chinese, CR-Creole, J-Japanese, P-Punjabi, PI-Pacific Islander, W=White; F=Female, M=Male, C-College Student, H-High School Student, JC-Junior College Student, G-Graduate Student, N – Non-student

We decided to zoom in on the competition between all and like across time, concentrating on the speakers whose system contains quotative all. Figure 1 comparatively depicts the composition of the quotative system of the all-users in our 1990/4 and 2004/5 corpora. The cross-over pattern is evident: all, which in 1990/4

amounted to almost as large a share as all other quotatives together (mainly say, go and unframed) has been relegated to only 6% in 2004/5 when like clearly dominates the system. Indeed, all and like switch places as the primary quotative, with the overall frequency of the other variants (say, think, go, zero etc.) changing far less in overall proportion. This highly significant change ($\chi^2(2)= 217.851$, $p<.001$) is largely driven by the replacement of all with like as the preferred quotative.[v]

FIGURE 1: Relative frequency of all, like and other quotatives amongst the speakers who use all or all like in the 1990/4 and 2004/5 data sets (in %).



The overall trend across real time is also sustained when we look at individual speakers within these two data sets: Whereas at least four speakers in Wimmer's (1990) and Fought's (1994) recordings used 10 or more tokens of quotative all, the highest number used by any one speaker in our 2004/5 tape-recorded corpus was only 6. The movement away from all and towards like between the 1990s and the 2000s becomes even more dramatic if we re-consider the fact that in the 1990s, all was categorically constrained by age: in Wimmer's 1990 data, only the high school students used the new incoming form all (42% among the quotative options); none of the college-age speakers did.

Importantly, the extent of the shift from all to like also shows up in the development of a combined form: all like, as exemplified in (8) and (9).

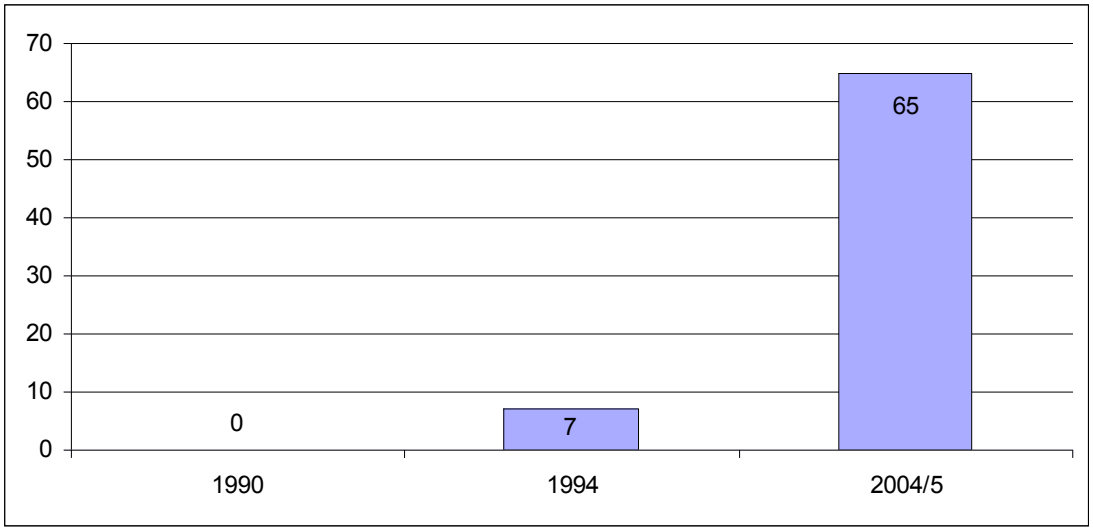
(8) I'd be **all like** *"You know I'm thirteen, right?"*
(9) He's **all like** *"You got any weapons in the car"*

There are no all like tokens whatsoever in Wimmer's (1990) corpus. By the mid-1990s, in Fought's corpus, three tokens of all like are found. In our 2004/5 corpus, all like is the primary sequence in which quotative all is used (17 out of 26 all tokens). As is evident in Figure 2, the increase in the proportional amount of all like in the three data sets 1990, 1994 and 2004/5 is quite dramatic.

It is furthermore remarkable that the only nine tokens of quotative all by itself in our 2004/5 tape recordings come from college students. All of our high school

students used all like instead.[vi] By our 2004/5 corpus, all like has become the primary sequence in which all is used as a quotative, and the only one used by the younger speakers.[vii]

FIGURE 2: All like ratio as calculated by the proportion of all like out of all quotatives introduced by all like and all (in %)



The demise of all within the set of quotative introducers used by the California youth represented in our corpora is also confirmed by the input probabilities of two separate VARBRUL runs on the two data-sets.[viii] Table 5 shows that all is much more likely to occur overall in the 1990/4 corpus (input probability.34) than in the 2004/5 corpus (input probability .04). But it is not only the input probabilities that have decreased sharply, showing that overall frequency of all has diminished; the constraints that govern quotative all have also changed across the two corpora. We discussed the constraint hierarchy of all across time in detail in Rickford et al. (2007).[ix] Here, we only briefly point to the major changes in the constraints that govern quotative all.

For both VARBRUL runs, we included a total of seven factor groups into the analysis: Tense and Modality (present non-modal, past non-modal or modal / quasi-auxiliaries), Subject Type (full singular or plural NPs, 1st, 3rd person pronouns including it and unframed quotes), Birds of a feather (priming effects with respect to the quotative choice in the preceding 5 turns, operationalised here as the occurrence of a different quotative (alternation), of the same quotative (perseverance) or of no quotative), Speech or Thought encoding, Drama/Animation (the (non)-occurrence of voice or sound effects), Gender and Ethnicity.

Table 5 shows that while ethnicity showed a significant effect in the 1990/4 data with white speakers slightly favoring quotative all over the Chicanos, none of the social factors tested for came out significant, in the 2004/5 data.[x] In 1990/4, the occurrence of all is conditioned by the tense / modality in the quotative frame, with present non modal contexts strongly favoring its occurrence (with a factor weight of .75). In the later corpus, however, tense / modality does not have a significant effect: The few tokens of all in the 2004/5 corpus occur with a broad range of tense and aspect markers (see examples 10 and 11).

TABLE 5: Varbrul analysis of the factor groups conditioning quotative <u>all</u> amongst the speakers who use the form in the 1990/4 corpus and the 2004/5 corpus (see Rickford et al., 2007)[xi]

	1990/4			2004/5		
Input probability or corrected mean:	.34			**.**04		
Overall %	42%			7%		
Total N	320			384		
	FW	**%**	**N**	**FW**	**%**	**N**
TENSE / MODALITY IN QUOTATIVE						
Present non modal	.75	63%	182	[ ]	7%	177
Past non modal	.41	21%	75	[ ]	7%	126
Other (*modal, conditional, no tense)*	.06	3%	63	[ ]	6%	81
Range	*.69*			*[ ]*		
BIRDS OF A FEATHER						
Alternation (Diff. quotative in 5 preceding lines)	.50	41%	127	.71	13%	47
No Quotative (in 5 preceding lines)	.39	31%	111	.61	8%	229
Perseverance (Quot. <u>all</u> in 5 preceding lines)	.65	57%	82	.21	2%	108
Range	*.26*			*.50*		
QUOTING SPEECH/THOUGHT						
Speech (external)	[…]	44%	269	.61	9%	241
Ambiguous or indeterminate	[…]	27%	15	.48	5%	79
Thought (internal)	[…]	28%	36	.17	2%	64
Range				*.44*		

SUBJECT TYPE								
3rd person pronoun		[…]	55%	149	.71	12%	139	
non-3rd person pronoun		[…]	41%	96	.57	5%	133	[xii]
full NP		[…]	16%	75	.20	2%	112	
	Range				*.51*			
ETHNICITY								
White		.58	53%	171	[…]	6%	300	
Chicano/Mexican/American		.41	29%	149	[…]	10%	84	
	Range	*17*						

Not significant: gender, sexual orientation, drama / animation.

In 1994/5 all is mainly used with past time reference

(10) I was **all** *"No I'm not giving you the keys to my car"*

In 2004/5 all is also used with future time reference and habitual *would*.

(11)a. He'd be **all** "*It's a it's a black guy."*

(11)b. I'll be **all like** *"Stop it. Don't text me."*

As regards the role of tense/modality, the numerical loss seems to go hand in hand with a loss of constraints, from a very high range of .69 in the 1994/5 dataset to a non significant outcome in 2004/5.

However, one other factor group continues to have a bearing on the occurrence of quotative all, albeit with varying strengths and directions. In the earlier corpus (see example 12a), all tended to cluster, since perseverance, which we define as the utterance of another token of all within the 5 preceding lines, favored its occurrence with a factor weight of .65 in 1990/4. Importantly, there are also several clustered examples in the corpus collected by Rachelle Waksler in Spring 1997 until Fall 2000 in San Francisco and which formed the basis for her (2001) article (e.g. 12b, below). This is worth noting because it extends the time period in which such sequences could be documented by another six years or so, which is potentially significant for a short lived trend (the rise and fall of all) that essentially lasted just 20 years.

(12) In 1990/4 all is mainly used in clusters

a. He's **all** *"What are you doing here?"*
I'm **all** *"You called me in."*
He's **all** *"For what? For what?"*

(12) Examples from Waksler (2001) collected 1997-2000

b. And so he's **all** *"NO, I'm not getting out of the car." ...*
And then I was **all** *"Well could you please give him a message for me, please?"*
He's **all** *"What?"*
I'm **all** *"Tell him to leave Mary alone."*
And he's **all** *"OK."*
And he's **all** *"Well I'm supposed to give YOU a message."*
And I was **all** *"Whatever!"*

By 2004/5, however, all mainly occurs in sequences where it is preceded by other quotative options (a context which we termed alternation, factor weight .71, see example 13) or where it is not preceded by reported activity at all (factor weight .61). In our 2004/5 corpus, all is very strongly disfavored in clustered contexts.[xiii]

(13) I asked some guys in Portuguese where the academy is
And they're **all** *"It's right here"*
And I went there and asked the lady when they trained
And she's **like** *"come back at eight"*

Finally, by 2004/5, all has acquired two constraints: the type of quote reported and type of subject. We will discuss both in turn. In the 1990/4 data set, all was used indiscriminately with speech and thought. However, by 2004/5, it has narrowed its

uses, being now mainly used for the introduction of reported speech rather than thought (consider examples 14a and b).

(14) In 2004/5 all is mainly used for the introduction of speech rather than thought
a. SPEECH: He's **all** *"Stay right there"*
b. THOUGHT: it was **all like** *"Oh my God I'm gonna fail"*

The second constraint that was significant only in the 2004/5 corpus is the subject type with which the quotative occurs. Importantly, this factor group harbors two intersecting constraints: full NP vs. pronoun and 1st vs. 3rd person. In 2004/5, full NPs strongly disfavor the occurrence of all (factor weight .20) whereas subject pronouns either favor it or have no effect. Amongst the pronouns we also notice a person-hierarchy: Whereas all is strongly favored by 3rd person pronouns (factor weight .71), which includes singular as well as plural forms (see example 15), 1st person pronouns *I* or *we* have a neutral effect on the occurrence of the form.[xiv] Interestingly, while the literature on quotation discusses the role of 3rd person it in the development of quotative like (see Tagliamonte and Hudson, 2001; Buchstaller. 2004), only one quotation in our corpus was framed by a form of it + all (see example 14b).

(15) In 2004/5 all is mainly used with 3rd person pronouns:
a) They're **all**, "*gotta get to the arcade*!"
b) So he's **all**, "*yeah, come over 'n' use it*."

The difference in constraint hierarchy and direction from 1990/4 to 2004/5 means that change has indeed taken place, both in relative frequency and in constraint patterning. As all decreases in frequency, it loses one constraint, namely tense and modality and gains two more, subject type and speech/thought representation. The birds of a feather effect continues to exert an influence on the occurrence of the form, albeit with a much larger range than in the earlier corpus. Overall, the development of this form seems to provide supporting evidence that all is a rather short-lived innovation that has ceded its territory to like and all like over the past years. After a high in the late 1990s, the overall use of quotative all is clearly in decline.

However, thus far, we have based our claims solely on California data. We are not in a position to state how robust and generalizable these findings are across geographical space. We also do not have any information about the more fine-grained temporal detail of what happened before and between the collection of the two data sets, a problem endemic to real-time analysis in sociolinguistics. As a second step, therefore, we set out to test the hypothesis that the frequency of all has dwindled in recent years in a larger, more finely time-differentiated corpus. We also wanted to give a wider geographical angle to our investigation.

Lacking large-scale corpora collected with the sociolinguistic research paradigm that span the full period since the first attestations of quotative *all*, while also exhibiting wide geographical coverage, we decided to work with data from the world wide web. More specifically, we drew on corpora culled from Google, the web-based search engine. It is worth noting here that most of the material in the Google corpus (as far as we can determine its provenance) is from the US. Hence, while the scope of the internet searches is indeed broader than California and does include a multitude of sources, it is still mainly based on US data. To what extent this is the case is

notoriously difficult to assess and cannot be determined here. The following sections detail our analysis of the Google corpus

3.2 AN ANALYSIS OF THE NEWSGROUPS DATA

The extent to which the language of internet newsgroups is comparable to spoken language is a point of contention (Androutsopoulos and Ziegler 2004, Crystal 2001, Tagliamonte and Denis 2005). Here, we do not intend to argue that newsgroups contain the same frequency and general distribution of quotatives as spoken interaction, although Jones and Schieffelin (2007) have shown that another type of new media, instant messaging, is very rich in quotations, which seem to be used for similar functions as reported speech in spoken interaction. The aim of this second section is rather to describe in some detail the methods and outcome of our collaborative project with Google which aimed at investigating the use of quotative all in internet newsgroups. We believe that the methods we employed for our work on quotative all can be applied to other kinds of linguistic research projects and therefore have the potential to substantially enrich the kinds of corpus-based analysis used in variation studies, sociolinguistics, and other linguistics subfields.

Our analysis of the Google corpus proceeded in two steps. The first step was a pilot study, which we reported in Rickford et al. (2007), so we provide only a brief summary of it. Here, we describe in some detail the second step, which builds on the findings of the preliminary analysis.

The pilot study, carried out in 2005, used Google's interface to the newsgroups corpus to search for examples of quotative all (<http://groups.google.com/advanced_search?q=&>). Google's search tool only allows simple string searches and ignores punctuation, so finding quotatives among the millions of occurrences of all in the newsgroups corpus was not straightforward.[xv] We thus constructed a number of strings containing all that we thought would have a good chance of matching quotative uses of all. In a nutshell, these consisted of a singular subject pronoun with a contracted present tense form of *be*, followed by all and a word that seemed likely to be the start of a quote, e.g. a *wh*-word, *yeah*, *no*, *shit, it,* or the like. For example, "*I'm all yeah"* or *"I'm all shit".*[xvi]

The resulting hits were examined and the quotatives culled, producing a total of 354 examples over the period 1982 - 2004. These were then grouped according to the year of posting. In order to determine whether the rate of quotative uses of all was changing during the period covered by the newsgroup archive, it was necessary to have some measure of the size of each year's archive. A crude metric of the rate of quotative all would be the number of instances we found in a given year, divided by the total size of that year's archive. Unfortunately, Google does not make publicly available the size of the newsgroup archives for each year. In our pilot study, we attempted to get a measure of the relative sizes of the archives on a year-by-year basis by searching for some very common words (such as *word, other, make, see, way, people, first, the*) and comparing the number of hits across years. The tentative conclusion of the first stage of our project, on the basis of this method, was that quotative all first appeared in the newsgroups in the mid-1990s, becoming rapidly more common until about 1999, and then declining precipitously in frequency (see Rickford et al. 2007:20). We could not be confident about this conclusion, however,

because of several methodological limitations, which we address in more detail below.

To advance our understanding of the development of all and to test our hypothesis that all has indeed dwindled in recent years, we collaborated with Thorsten Brants, a researcher at Google Inc., and David Hall, a Stanford undergraduate who was employed by Google for two months over the summer 2006. This collaboration allowed our searches of the Google Groups archive to improve on the standardly available tools we had employed previously in a variety of ways. The most serious limitation we had run into during the pilot study was that we needed a more reliable measure of the sizes of the newsgroup archives for each year. In order to test whether frequency of usage of any form is changing, the raw frequency of occurrences in the archive is useful information only if it is accompanied by information about how the size of the archive changed over time[xvii]. While Google remained reluctant to disclose absolute size of their newsgroup corpus, during our summer project, they provided us with numbers indicating the relative size of each year's archive[xviii], which allowed us to normalize our raw year-by-year counts of different quotatives. This was necessary to yield comparable data across time and thus to make the newsgroup searches a reliable source of data on the changing rates of quotative usage.

A second methodological problem that we had run into during our pilot study was that our pilot search tool was restricted to the search bar that Google makes available on its web site. Hence, the search mechanism was essentially just keyword search, with a few minor enhancements. However, because all is an extremely common word[xix], and only a tiny fraction of its uses are quotative, it was impossible to try to find all and only the quotative uses in the output yielded by keyword searches. As we mentioned above, in the pilot study we attempted to circumvent this problem by constructing linguistic environments that we hoped would yield a relatively high rate of quotative hits, and went through them by hand. But even with this method, the signal-to-noise ratio on our searches was relatively low. The 354 instances of quotative all that we found by this method had to be culled from thousands of hits by our search pattern. In our collaboration with Google, we were able to search in a way that was sensitive to punctuation and therefore reduce the amount of noise substantially.

Our Google partners developed a search tool allowing regular expressions[xx] in search patterns, which made the searches far more efficient. Preliminary attempts to find regular expressions that would yield all, or nearly all quotatives resulted in far too many hits to be analyzed individually. Moreover, an examination of random samples of those hits revealed a rather poor signal-to-noise ratio – that is, the vast majority of the hits were not quotative uses.[xxi] We therefore modified our strategy. We used our existing compilation of quotative examples including the 1990/4 and the 2004/5 California data, as well as other examples, such as those in Waksler's article, to look for words that were common as the first word in a quotative. Selecting the most common lexical items, their most common spelling variants, and a few closely related words, we constructed a regular expression that included a left context of a singular pronoun and contracted copula, followed by all, followed by optional comma and quotation marks, and finally one of our likely quotation-initial words. The regular expression can be summarized as in Figure 3.

The procedure for the regular expression search was as follows: First, we searched the newsgroup corpus using the regular expression in Figure 3. By including only these lexemes in the template (and thereby limiting hits to strings that contained

these exact sequences), we obviously missed many other quotes that did not start with these exact words.

FIGURE 3: Regular expression for the Google newsgroup search[xxii]

{I'm / he's / she's / it's} all (,) {" / '} W

W stands for one of our likely quotation-initial words, which are: *are*(*n't*), *blah, can*(*'t*)*, could, do, dude, fuck, gee, get, give, hey, hi, how, if, is*(*n't*)*, lets, look, no, oh, ok, OK, okay, ooh, shit, shut up, thank, uh, um, well, what, when, where, who, whoa, why, will, wow, yeah, yes.*

However, narrowing down our search to these typical quote beginnings also dramatically increased the ratio of quotatives in the output. Of the total 914 hits for all, only 162 (18%) were noise, and for the other quotative introducers, the noise rate was even less (see below).

A final methodological problem of the pilot study was that in 2004, we looked at only one quotative, all. But without checking the rates of other quotatives, our study of quotative all lacked adequate controls. Even if we could be confident that the rate of all was dropping, that could be the result of changes in what newsgroups were used for. Perhaps changing technologies were leading discourses rich in quotatives to migrate to other venues, such as blogs or instant messaging. In order to provide accountability in terms of the behavior of the competitor variants, we thus searched not only for all but for the quotatives say, go and like as well.

Using essentially the same method but exchanging the quotatives in the parametric slot (cf. Figure 3), we then searched the corpus for all like, like, say/go. The overall output can be seen in Table 6:[xxiii]

The searches for like and say/go yielded too many hits to be practically examined individually (10,938 for like and 132,036 for say/go). We thus decided to work with randomly selected samples of 1000 hits of each of them.

TABLE 6: Raw output from regular expression search on the Google newsgroup corpus

Quotative	Hits total:
all	914
all like	203
like	10,939
say/says/go/goes	132,036

Finally, all 3,118 hits in the corpus (914 all, 203 all like, 1000 like, 1000 say/go) were hand-coded into one of four categories. We now define these categories, exemplifying them with output from the Google searches.

SPEECH: This category consists of quotes in the traditional sense, namely reports of words said or written, as in (16).

(16) She **said** *"so you're* [sic] *baby juts* [sic] *turned one, I think I met her"*
and he's **all** *"yeah, you babysat her once, you were great*, like *$10 an hour"*

THOUGHT: These quotations appeared to be reports of thoughts that were not actually uttered or committed to writing, as in (17) and (18).

(17) No matter how many times I see this subject line, my first thought is that it's a score, and I'm **all** *"Who the hell could beat somebody 420-1?"*

(18) So, I been reading these posts and I'm **all**: *"Who's this Arrow Guy?"*

STEREOTYPES: These quotes are characterizations of a person or of a situation through a quote that might characteristically be produced by that person or in that situation. This category is exemplified in (19) and (20) with <u>all</u> and in (21) with <u>say</u>.[xxiv]

(19) What a bunch of whiner troops we have! It's **all** *"could we please have some body armor so our limbs aren't blown off"* and *"some metal shielding on our humvees might help us to die less."*

(20) You seem to be under the impression that we think that once you've sinned then it's **all** *'oh dear, game's over, that's us condemned to the eternal fires'.*

(21) When they **say** "*You'd better stay overnight for observation."*
It **means** *"I want everyone to get a good laugh at this one."*

The category 'stereotype' is new to discussions of quotatives.[xxv] In fact, it constitutes a relatively small fraction of the examples of all quotatives variants except for quotative <u>all</u>, so it is perhaps not surprising that it had not been noted before. But as we began examining the <u>all</u> data from the Google groups search, it was evident that a great many of the examples served to characterize people or situations through quotes without actually attributing words or specific thoughts to them. So we added this category to our study.

NON-QUOTATIVES: This category consists of examples that should not be considered quotes, such as cases of quotation marks used for emphasis, quotes around proverbs or clichés, or discussions of the use of non-standard quotatives (of which there were several in our data), as exemplified in 22-25.

(22) It's all "what ifs" but like it or not, Oct 4 was a big deal.

(23) it all depends on whether you consider reporting things 'too good to be true' has no grey area at all, or if it's all 'yes' and 'no' with great lines between them.

(24) Here in So. Calif. the most recent incarnation of "go" in lieu of "say" is And'm all "No Waaaaaaay!!" And then she's all "Yeah, waaaaaay!" Well all right, so there's a verb in there, but …

(25) I recall this even from elementary school. Two other annoying slang substitutes for "say" are "like" and "**all**".

The categorization of all quotatives into these four categories was carried out by Nick Romero, an undergraduate student at Stanford, and questionable cases were reviewed (and occasionally changed) by at least one of the authors. Full contexts from the newsgroups were available to us – and were consulted in the majority of cases – so that informed decisions could be made about the classifications. The raw data from our four searches of the Google corpus (for <u>all</u>, <u>all like</u>, <u>like</u> and <u>say/go</u>) are summarized in Tables 7-10. [xxvi]

TABLE 7: All-quotations by quotative category and year (raw data).

Year	Category Speech	Thought	Stereotypes	Non-quotatives	total
1982			1		1
1992	1		1	1	3
1993	5	1	3	5	14
1994	1		4	3	8
1995	5		6	7	18
1996	12	2	13	13	40
1997	18	1	13	10	42
1998	27	8	24	11	70
1999	58	9	45	20	132
2000	47	15	66	13	141
2001	39	8	44	29	120
2002	37	11	39	18	105
2003	27	9	41	11	88
2004	31	9	34	11	85
2005	12	3	17	10	42
2006	3		1		4
TOTAL	323	76	352	162	914

TABLE 8: Like-quotations by quotative category and year (raw data).

Year	Category Speech	Thought	Stereotypes	Non-quotatives	Total
1983				1	1
1991				1	
1992		1			1
1993	1	2	5	1	9
1994	1	1	1		3
1995	2	6	4		12
1996	21	10	13	3	47
1997	14	14	13	2	43
1998	58	33	17	2	110
1999	62	43	36		141
2000	55	62	30	5	152
2001	62	40	30	2	134
2002	29	43	25	11	108
2003	48	38	22	2	110
2004	28	23	22	2	75
2005	11	15	15	2	43
2006	7	3			10
TOTAL	399	334	233	34	1000

TABLE 9: All like-quotations by quotative category and year (raw data).

Year	Category Speech	Thought	Stereotypes	Non-quotatives	Total
1991	1			1	2
1993	1				1
1995	1				1
1996	2	1	1		4
1997	5	4	1		10
1998	13	3	2		18
1999	23	4	2		29
2000	23	6	3	1	33
2001	16	7	3		26
2002	16	6			22
2003	17	3	2		22
2004	16	8	3		27
2005	4	2	2		8
TOTAL	138	44	19	2	203

TABLE 10: Say / go -quotations by quotative category and year (raw data).

Year	Category Speech	Thought	Stereotypes	Non-quotatives	total
1985	1				1
1989	3				3
1990	4		1		5
1991	2				2
1992	5		1		6
1993	16	1			17
1994	20	1	2		23
1995	21		2		23
1996	65	3	1		69
1997	64	1	2	1	68
1998	97	1	8	4	1
1999	132	2	7	4	145
2000	107	4	8	3	122
2001	101	9	4	2	116
2002	69	2	3	5	79
2003	80	2	3	1	86
2004	60	1	1	2	64
2005	47	1	4		52
2006	8		1		9
TOTAL	902	28	47	23	1000

Note first of all that noise – category 4, the non-quotatives – constituted under 5% of the data in the like *(*34/1000), all like (2/203), and say/go (23/1000) searches but 18% in the all data (162/914). Hence, while the bulk of the material consisted of usable data from categories 1-3, the output for quotatives all nevertheless contained a sizeable ratio of noise. More importantly, note that all leads the way in the 'stereotype' category: 38.5% of the all-quotes are from the stereotype category (compared with only 23.3% for like and only 4.8% for the combined say/go tokens).

Hence, as we pointed out above, all seems to be fundamentally doing something different from the older quotatives say/go and also probabilistically from like.

In order to trace the development of the quotative variants across real time, we needed to normalize the raw output of our searches. Since we were not given the absolute word frequencies for the archive but only the relative sizes of the newsgroups on a year-by-year basis, we computed normalized numbers that take account of the fluctuations in newsgroup size per year in the following way: We took the numbers from Tables 7-10, excluded the non-quotatives, and adjusted for the relative size of each year's newsgroup archive by dividing the number of actual examples of all for each year by the percentage of the total newsgroups corpus contained in that year's archive. In the case of like and say/go, we also projected the rates based on the fact that we had only examined random samples of 1000 examples (by multiplying the like-rates by 10.939 and the say/go-rates by 132.036).[xxvii] Finally, we plotted the normalized rates of each quotative over the years in Figures 4-7. Due to the fact that token numbers for all of the quotative variants were generally very low in the pre-1995 newsgroup postings, we collapsed these age bands into one composite figure. The reader is advised to refer to the (non-normalized) frequencies in Tables 7-10 for information about the patterning of the individual quotatives in these earlier age bands. We turn now to the results of these manipulations, one quotative at a time, starting with all.

The earliest occurrence of all in the newsgroup corpus is a category 3 quote, a stereotype, which occurred in 1982. It is given in example (26).

(26) Those mercenaries sure lead a life, don't they? **It's all** "*What Ho! Roger, we've been double-crossed! Let's take over the country!"* and "Aargh, I'm hit-kill me ...

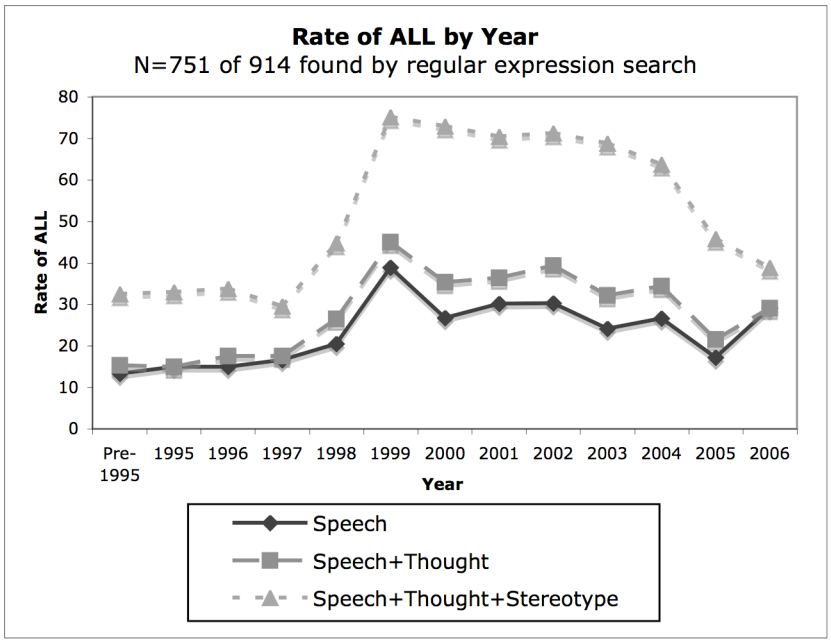
After this lone occurrence, we did not find another token of quotative all in the newsgroup corpus until 1993. Let us now consider the development of all across real time in the Google newsgroups from this point (represented by "pre-1995") on.

The lines in Figure 4 represent the year-by-year distribution of quotative all broken down into the categories speech, thought and stereotypes. The topmost line (dotted, with triangles) represents the total occurrences of quotative all in our newsgroup corpus. The two lines below indicate how the total is divided among speech (the area below the lowest line), thought (the area between the lowest line and the second line) and stereotypes (the area between the second line and the top line). The fact that the top line is relatively far above the other two shows that the category "stereotype" constitutes a substantial fraction of the occurrence of quotative all.

Overall, Figure 4 shows that all is used mainly for speech and stereotypes. The category "thought" does not contribute much to its overall frequency of occurrence. And, as we pointed out before, the main locus of occurrence of all is the introduction of stereotypes. This is especially the case in the period when it is the most frequent, between 1999 and 2005.

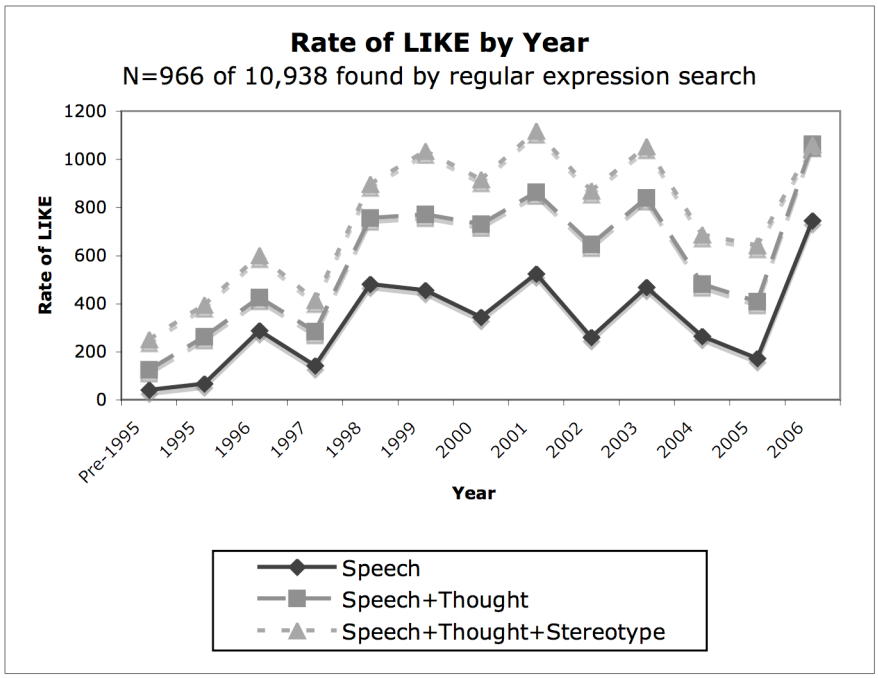
Returning to the question of whether quotative all is in decline, the data presented here supports the conclusions we drew on the basis of our pilot project: Quotative all usage increased during the 1990s, peaked in 1999 and has been declining rapidly in the past six or seven years.

FIGURE 4: Rate of all in the Google newsgroups, computed by taking the totals of quotative categories 1-3 and adjusting for the size of each year's newsgroup archive (frequency count)



And our larger, more recent study also allows us to see that it is especially in the stereotypes category that all first expands and then dwindles in frequency, whereas the speech and thought categories, while declining slightly since 1999, stay relatively stable. Importantly, this rate of decline is not matched by other quotatives. Let us first discuss quotative like, which is depicted in Figure 5.

FIGURE 5: Rate of like in the Google newsgroups, computed by taking the totals of quotative categories 1-3 and projecting the rates based on the fact that we had only examined random samples of 1000 examples (frequency count).

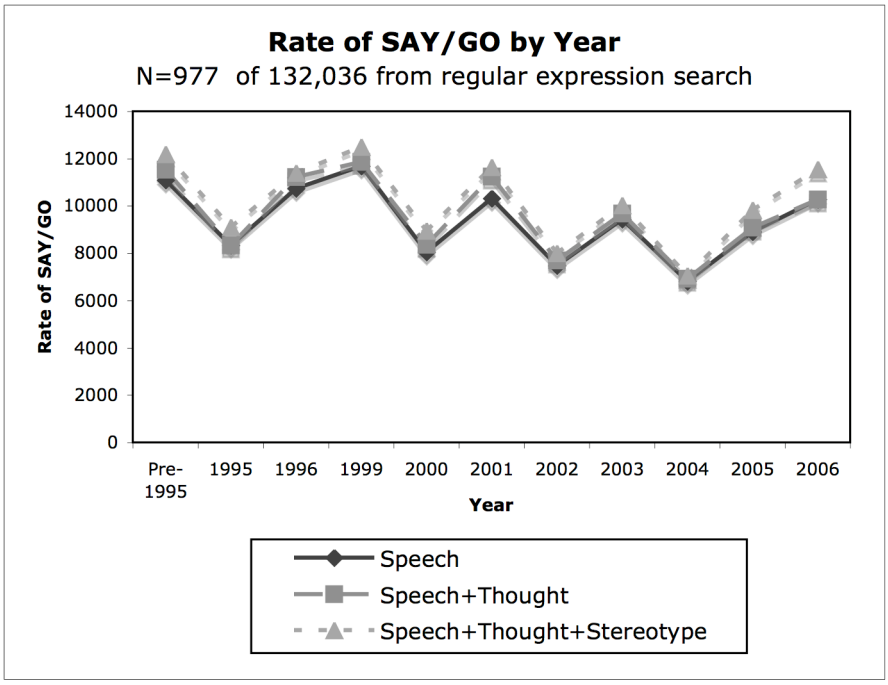


Rising sharply after 1995, like is fluctuating in frequency of occurrence across time but seems to have hit a high in 2006 after a steady rise (discounting an inexplicable trough in 2004 and 2005). Importantly, like seems to be used in almost equal proportions for the introduction of speech, thoughts and stereotypes.
Furthermore, the overall proportion of these categories seems to stay relatively stable across time, except for 2006. However, since our database for 2006 was relatively small (including newsgroup postings from only the first six weeks of the year), we treat the 2006 figure with caution.

Hence, like and all seem to be fundamentally distinguished by their propensity to introduce reported thought: While like occurs with speech, thought and stereotypes in equal measure (see Buchstaller 2004, who also found that like is used in equal proportions with quotes of various epistemic stances), the fraction of reported thought framed by all is negligible. But it is important to note that both like and all introduce speech and stereotypes, which sets them apart from the traditional quotatives say and go. Consider now Figure 6, which plots say and go across time.

Clearly, say/go are used virtually exclusively for true quotes. The categories "thought" and "stereotype" do not add much to their overall frequency of occurrence. This fact is also reflected in the low numbers in the columns for categories 2 and 3 in Table 10. In terms of the development of say/ go, we note that the curve exhibits considerable year-by-year variation and very slow long-term decline, but nothing like the rapid drop-off of all.

FIGURE 6: Rate of say/go in the Google newsgroups, computed by taking the totals of quotative categories 1-3 and projecting the rates based on the fact that we had only examined random samples of 1000 examples (frequency count).

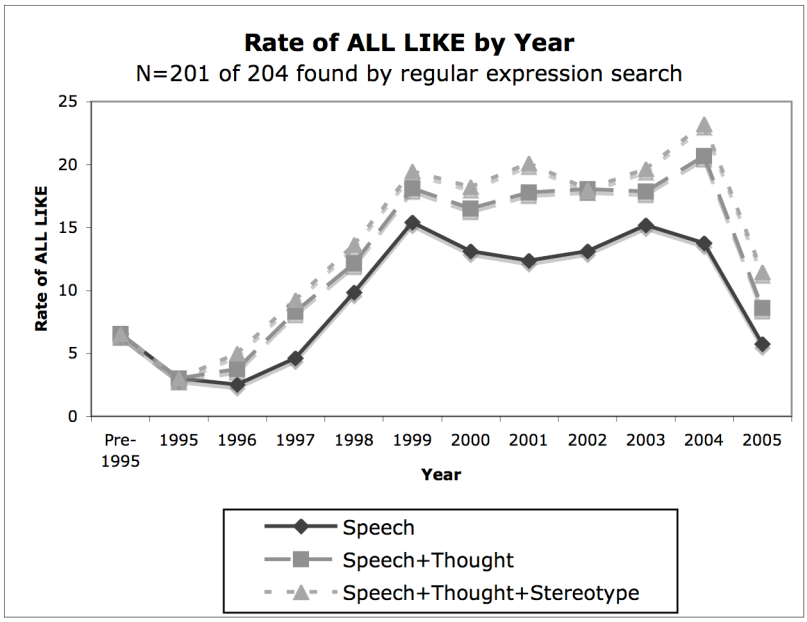


These findings thus lend support to our earlier claim that all has declined in the last few years. Its lower frequencies of occurrence in the years since 1999 cannot be attributed entirely to the fact that populations that use a lot of quotation have left the

newsgroups and migrated to a newer, possibly hipper medium such as blogs. If that were the case, we would see a similar trend for like and say/go. But this is clearly not the case. The curve for all looks so different from all the others that it seems safe to rule out any attempt to explain its shape as a function of more general changes in what people use newsgroups for.

Finally, we discuss the figure for all like. The number of examples of all like is so much smaller than the other quotatives that we are hesitant to draw any conclusions from the recent dearth of examples (consider Table 9). However, we need to address one point in particular: Above, we noted that the move from all to like is accompanied by the development of the form all like. If this is the case, we would expect all like to rise in frequency at the point in time when the transition actually happens, namely around 1999. Figure 7 shows that this is indeed the case. All like starts at low frequencies (under 8), picks up until 1999, plateaus while steadily increasing between 1999 and 2004 until the last two years, when examples are almost non-existent.

FIGURE 7: Rate of all like in the Google newsgroups, computed by taking the totals of quotative categories 1-3 and adjusting for the size of each year's newsgroup archive (frequency count).



Hence, on the basis of these findings – which are admittedly based on relatively low token numbers – we conjecture that all like developed in tandem with all and continued to rise during the demise of all. Finally, in the last two years, when all is clearly ousted by like, all like also almost disappeared.

Figure 7 fits well with the pattern found in the California data (see Figure 2): In 1990, the California adolescents did not produce any tokens of all like. By 1994, some tokens of all like had developed in California. The 2004/5 data collected in California, seems to have caught it at its high point, just before it dropped dramatically in frequency. Obviously, it would be interesting to follow up the California study and add another time slice to see whether the drop in all like frequency in the Google data is also replicated in California.

4. CONCLUSION

In this paper we have investigated the change of quotative all, using two different data sources: traditional sociolinguistic interviews and a web-based newsgroups corpus. In both our California data and in the Google newsgroup data all has dramatically declined in real time. Importantly, its numerical decline has also significantly impacted the constraints it is governed by, both in terms of the direction of constraints as well the types of factor groups.

The trajectory of quotative all discussed here is interesting from the perspective of language change in progress, as it provides a direct window on what has often been observed in historical texts: the short-term flourishing of a linguistic form or usage. In the case of quotative all, we clearly have an instance not simply of innovation in the individual, but of change in the sense of spread to many users (Milroy, 1992, 2003; Weinreich, Labov, & Herzog, 1968). Earlier examples of such changes that were relatively short-lived in the textual evidence for Standard English include the use of auxiliary do in affirmative clauses such as [*T*]*here I did see the whole Consent of the Realm against it* (1554, Throckmorton; Nevalainen 2004: 202), and of several aspectualizers such as stinten and finen, both meaning 'finish', and both short-lived in Middle English (Brinton, 1988: 151).[xxviii] In some cases, like all and do, the form becomes realigned with other uses, in others, like finen, the form ceases to be used. Emergent structures are unstable in nature (Bybee and Hopper, 2001), so it is no surprise that this kind of phenomenon of development and dissolution occurs, despite a tendency for analysts to expect a new phenomenon, especially one of a grammatical nature, to persist. (Contrast this with the loss of the verbal coda in topic restricting as far as constructions, a change that has been in process since the 19th century, and appears to be moving forward in terms of frequency and linguistic environments affected (Rickford, Wasow, Mendoza-Denton and Espinoza 1995)).

Our newsgroups study has added an interesting angle to our earlier findings. Perhaps the most remarkable thing to emerge is that there are some important differences among quotatives in the distribution of the three subtypes we identified (speech, thought, and stereotypes). Clearly, say/go are used virtually exclusively for true quotes. Like, on the other hand, is used as much to introduce thoughts or stereotypes as to introduce speech. All is unique in its frequent use to introduce stereotypes, particularly during its peak period of use, from 2000 to 2004. This indicates that all is functionally somewhat different from the other quotatives examined here.

Also, we hope to have shown that Google newsgroups (and similar data to the extent that they exist and are made available at other sites) is a valuable source for studying recent trends in language variation and change (see also Hundt et al 2007, Hoffmann 2007). The collaboration with Google has given us the opportunity to search a huge amount of chronologically organized data using punctuation-sensitive regular expressions, a more powerful tool than the search methods Google makes available to everyone. In principle, we could have done our searches using the standard Google search tools, but it would have been vastly more time-consuming and error-prone. But the one thing we got from the collaboration that we absolutely couldn't have had without it is accurate data on the relative sizes of the archives year-by-year. The web provides linguists with a corpus so large that it would have been unimaginable just a few years ago. Unfortunately, its very size and the variety of its contents make it unwieldy as a source of linguistic data. The newsgroups provide a much smaller, but still immense, corpus, with a modicum of useful organization built

in. Two particularly attractive features of the newsgroups archives are that they can be searched by language and that they are organized chronologically. The latter property allowed us to study change in language usage over a time span far shorter than those usually considered in diachronic linguistics. We recommend this tool to others interested in studying ongoing changes that are detectable in the written form of language.[xxix]

References:

Alford, Daniel Moonhawk Alford. (1982-83). A new English language quotative. *Not Just Words. The Newsletter of Transpersonal Linguistics* 2: 6.

Androutsopoulous, Jannis, & Ziegler, Evelyn. (2004). Exploring language variation om the Internet: Regional speech in a chat community. In B.-L. Gunnarsson et al. (eds.), *Papers from ICLaVE 2 Uppsala Papers from the Second International Conference on Language Variation in Europe, ICLaVE 2*. Department of Scandinavian Languages, Uppsala University, Uppsala, Sweden

Bakht-Rofheart, Maryam. (2002). Avoidance of a new standard: Quotative use among Long Island Teenagers. Paper presented at NWAVE 31, Stanford University.

Barbieri, Frederica. (2005). Quotative use in American English. *Journal of English Linguistics* 33: 222-256.

Barbieri, Federica. (2007). Older men and younger women: A corpus-based study of quotative use in American English. *English World-Wide* 28: 23-45.

Bayley, Robert, & Santa Ana, Otto. (2004). Chicano English: Morphology and syntax. In B. Kortmann et al. (eds.), *A handbook of varieties of English, vol. 2*. Berlin/New York: Mouton de Gruyter. 374-390.

Buchstaller, Isabelle. (2001). *He goes* and *I'm like*: The new quotatives re-visited. Paper presented at New Ways in Analyzing Variation 30, University of North Carolina, Raleigh.

Buchstaller, Isabelle. (2004). *The sociolinguistic constraints on the quotative system - British English and US English compared*. Doctoral dissertation, University of Edinburgh.

Buchstaller, Isabelle, & D'Arcy, Alexandra. (2009). Localized globalization: A multi-local, multivariate investigation of *be like*. *Journal of Sociolinguistics* 13: 291-331.

Buchstaller, Isabelle, & Traugott, Elizabeth Closs. (2006). *The Lady was al demonyak:* Historical aspects of adverbial ALL. *English Language and Linguistics* 10(2): 345-370.

Bucholtz, Mary. (2004). From Stance to Style: Innovative Quotative Markers and Youth Identities in Discourse. Paper presented at the Sociolinguistic Perspectives on Age, NYU.

Brinton, Laurel J (1988). *The development of English aspectual systems*: (Cambridge Studies in Linguistics 49). Cambridge: Cambridge University Press.

Bybee, Joan, & Hopper, Paul J. eds. (2001). *Frequency and the emergence of linguistic structure.* Amsterdam/Philadelphia: Benjamins.

Chen, Wenghong, Boase, Jeffrey, & Wellman, Barry. (2002). The global villagers: Comparing internet users and uses around the world. In B. Wellman and C. Haythornthwaite (eds.), *The internet of everyday life*. Malden, MA: Blackwell.

Crystal, David. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Cukor-Avila, Patricia. (2002). *She say, she go, she be like*: Verbs of quotation over time in African American Vernacular English. *American Speech* 77: 3-31.

Dailey-O'Cain, Jennifer. (2000). The sociolinguistic distribution and attitudes towards focuser <u>like</u> and quotative <u>like</u>. *Journal of Sociolinguistics* 41: 60-80.

Ferrara, Kathleen, & Bell, Barbara. (1995). Sociolinguistic variation and discourse function of constructed dialogue introducers: The case of be + like. *American Speech* 70 (3): 265-290.

Fought, Carmen. (2003). *Chicano English in context*. New York: Palgrave/MacMillan Publishers.
Guy, Gregory. (1988). Advanced VARBRUL Analysis. In K. Ferrara et al. (eds.), Language change and contact: Proceedings of the sixteenth annual conference on New Ways in Analysing Variation. Austin: Department of Linguistics, University of Texas at Austin: 124-136
Hoffmann, Sebastian. (2007). Processing Internet-Derived Text – Creating a Corpus of Usenet Messages. *Literary and Linguistic Processing* 22(2): 151-65
Hopcroft, John E., Motwani, Rajeev, & Ullman, Jeffrey D (2001). *Introduction to automata theory, language, and computation*. Boston: Addison-Wesley.
Hundt, Marianne, Nesselhauf, Nadja, & Biewer, Carolin (eds.) (2007). *Corpus Linguistics and the web.* Amsterdam, Rodopi.
Jones, Graham, & Schieffelin, Bambi. (2007). Enquoting voices, accomplishing talk: Uses of be + like in instant messaging. *Language and Communication* 29(1): 77-113
Katz, James, Rice, Ronald, & Aspden, Philip. (2001). The Internet, 1995-2000: Access, civic involvement, and social interaction. *American Behavioral Scientist* 45(3): 404-418.
Labov, William. (1972). *Language in the Inner City*. Philadelphia, PA: University of Pennsylvania Press
Milroy, James. (1992). *Linguistic variation and change: On the historical sociolinguistics of English*. Oxford: Blackwell.
Milroy, James. (2003). On the role of the speaker in language change. In R. Hickey (ed.), *Motives for language change*. Cambridge: Cambridge University Press. 143-157.
Nevalainen, Terttu. (2004). Mapping change in Tudor English. In L. Mugglestone (ed.), *The Oxford history of English.* Oxford/New York: Oxford University Press. 198-211.
Rickford, John R (2000). Variation and change in our living language. *The American heritage dictionary of the English language*, 4th edition. Boston/New York: Houghton Mifflin Company. xxii-xxv.
Rickford, John R., Buchstaller, Isabelle, Wasow, Thomas & Zwicky, Arnold. (2007). Intensive and quotative *all*: Something old, something new. *American Speech* 82(1): 3-31.
Rogers, Everett M (1983) [1962]. *Diffusion of innovations*, 3rd edition. New York: The Free Press.
Romaine, Suzanne & Lange, Deborah. (1991). The use of <u>like</u> as a marker of reported speech and thought: A case of grammaticalization in progress. *American Speech* 66: 227-279.
Singler, John. (2001). Why you can't do a VARBRUL study of quotatives and what such a study can show us. *University of Pennsylvania Working Papers in Linguistics* 7: 257-278.
Tagliamonte, Sali, & D'Arcy, Alexandra. (2004). *He's like, she's like*: The quotative system in Canadian youth. *Journal of Sociolinguistics* 8(4): 493-514.
Tagliamonte, Sali, & D'Arcy, Alexandra. (2007). Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change* 19(2): 199-217.
Tagliamonte, Sali, & Derek, Denis. (2005). OMG, its so PC! Instant Messaging and Teen Language. Paper presented at NWAVE 34. New York City, New York, 20 October – 23 October, 2005.

Vincent, Diane, & Dubois, Sylvie. (1996). A study of the use of reported speech in spoken language. In J. Arnold et al. (eds.), *Sociolinguistic variation: Data, theory, and analysis*. Stanford, CA: CSLI. 361-374.
Waksler, Rachelle. (2001). A new ALL in conversation. *American Speech* 76: 128-138.
Weinreich, Uriel, Labov, William, & Herzog, Marvin I (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (eds.), *Directions for historical linguistics.* Austin: University of Texas Press. 95-189.
Wimmer, Ann. (1990). *Be* + *all* and other new quotative introducers in California English. Senior honors thesis. Stanford: Stanford Linguistics Department.

http://www.google.com/press/pressrel/pressrelease48.html
http://en.wikipedia.org/wiki/Regular_expression.)

Notes:

---

[i] Acknowledgements: We are grateful to John Singler and other reviewers of this paper for their helpful feedback on an earlier draft. We are also thankful to Google Inc. for the opportunity to collaborate on this exciting project, drawing both on their personnel and facilities. Many thanks go to Thorsten Brants for his enthusiasm for and support of the project as well as for his enormous input in terms of computational methods. We are also indebted to David Hall, for developing and implementing the tools needed to do the searches we requested, and for responding swiftly and extensively to all our queries and suggestions. Thanks are due to Carmen Fought, Rachelle Waksler, and Ann Wimmer for allowing us to use their data on quotative all and other forms from the 1980s and 1990s as well as to Bob Bayley and Mackenzie Price for guidance with statistical analysis. Finally we are grateful to Stanford faculty colleagues for their input, and to several Stanford students who provided substantial assistance with data collection and analysis between 2004 and 2010, especially Zoe Bogart, Crissy Brown, Kayla Carpenter, Tracy Conner, Kristle McCracken, Rowyn McDonald, Cybelle Smith, Francesca Smith, Laura Whitton, Kayla Carpenter and Cybelle Smith.

[ii] "A usenet **newsgroup** is a repository usually within the Usenet system, for messages posted from many users in different locations. … Newsgroups are technically distinct from, but functionally similar to, discussion forums on the World Wide Web." *Wikipedia* March 2008.

[iii] The Groups archive is cumulative, so it is always growing, even though, as far as we can make out, its rate of annual growth has been slowing recently.

[iv] Rogers (1983:246 ff) differentiates adopters into several categories depending on when they adopt an innovation: *Innovators*, among the first 2.5% to adopt an innovation, *Early Adopters*, among the next 13.5%, *Early Majority*, in the next 34% of adopters, *Late Majority*, among the next 34%, and *Laggards*, in the last 16%. From the evidence that they were among the very earliest users of quotative *all*, the adopters in our Table 1 must be considered either Innovators or Early Adopters. Rogers has also written revealingly (1983:20 ff) about the innovation-decision process, which involves five steps to adoption—knowledge, persuasion, decision, implementation and confirmation. However, since we did not have the opportunity to interview the quotative *all* innovators and early adopters about this issue directly, we cannot tell whether they went through a relatively conscious innovation-decision process like this, or each of its component steps. This is something that all of us interested in the study of linguistic innovations might include in future research designs.

[v] "The Chi square test allows us to determine whether there is a statistically significant association between two variables" (Acton and Miller 2009:144), in this case, speakers' quotative choice and time of data collection. The chi square reveals that quotative choice has significantly changed between the two time slices sampled.

[vi] For the calculations on which Figure 1 is based, we decided to count the all like cases as tokens of quotative all rather than like. This is due to two facts: (i) VARBRUL runs that collapsed all like and all achieved a better log likelihood (as a measure of the fit of the model to the data) (ii) the percentage used for the speakers' thoughts in the Google data (to be discussed below) is very similar for all and all like, and much lower in both cases than is the case for like. But were we to count all like tokens as instances of like or as a totally separate form, the decrease in the relative frequency of all would be even more dramatic.

[vii] For simplicity, we will refer to the variant as all in the rest of the discussion of our California data, bearing in mind that in 2004/5 the variant contains a considerable amount of the combined form, all like.

[viii] Singler (2001) has argued that multivariate analysis programmes like VARBRUL that rely on the concept of the sociolinguistic variable cannot be used for the analysis for quotatives since they do not satisfy the criterion of semantic equivalence. Bearing this shortcoming in mind, he nevertheless goes on to show that a variationist analysis of quotatives can offer important insights into the patterning of the system of reported speech and thought introducers. Like Singler, we feel that a multivariate analysis of the quotative system post *all* presents an exciting opportunity to investigate the constraints on a change in progress occurring in a complex variable. Unlike Singler, though, our analysis relies on a functional definition of the variable as "all strategies used to introduce reported speech, sounds, gesture and thought by self or other" (Buchstaller 2006: 5, see also the discussion there). However, one problem we need to acknowledge is the very low token number of quotative all in our 2004/5 data set. Yet again, in line with Singler, we have decided to present the analysis in the hope that it will shed comparative light on the constraint hierarchy of *all* vs. its competitor variants in the later as well as the earlier data set (see Guy 1988 who sets the threshold for analysis for 5%). Furthermore, by analysing the data produced by speakers whose quotative system contains the form, we have maximised the occurrence of all in our data.

[ix] Careful readers will note that the VARBRUL results reported for both data sets in this paper differ somewhat from the results reported in our 2007 paper. The most substantial change is in the number of tokens used for the VARBRUL run in this paper, which increased from 245 to 320 for the 1990/94 data set (as we excluded Carl from Wimmer's 1990 data set, and added five speakers from Fought's 1994 data set), and decreased from 544 to 384 for the 2004/05 data set (as we appropriately deleted speakers who used no tokens of all or all like). Interestingly enough, however, changes in the factor group weights were generally minimal, and the significance and relative ordering of the factors in the primary Tense/Modality and Birds of a Feather factor groups were unchanged. However, in the 1990/94 VARBRUL run published in our (2007) paper, Quoting Speech and Thought was marginally significant, with Speech favoring all at .56, while this factor group is non-significant in the revised run prepared for this paper and presented in Table 5. Moreover, with the addition of more Chicano/Mexican speakers from Fought's corpus, ethnicity becomes significant where it was not before. For the 2004/04 corpus, the only difference is that Subject Type becomes significant where it was not before.

[x] We are grateful to Mary Bucholtz for pointing out that all seems to continue to flourish among her middle school Latinas in LA (in November 2006). Further research is needed in order to investigate whether this ethnicity and gender effect holds outside of southern California. In our 2004/05 data, the Latino speakers do use more all (like) than the White speakers do, but in the multivariate VARBRUL analysis, the difference is not statistically significant.

[xi] For the 1990/4 data we have excluded Carl, a marginal all user who only produced a single token of all in his 45 quotatives. With him included, the results of our VARBRUL run would look slightly different with N= 365 and an Input Probability of .29. Only one factor comes out as significant, namely tense (Present= .77 Past = .33 Other (future, conditional, etc.) = .07), with all other factor groups chosen as significant.

[xii] The category non-3rd person also includes one token of the very rare 2nd person generic *you* in the string *you're just all, "I can't do this."*

[xiii] Obviously, the higher proportional frequency of all in 1990/4 (42% as compared to only 7% in the 2004/5 corpus) makes it more probable for all to cluster in the earlier corpus. Even so, we observe that rows of consecutive all-tokens, as exemplified in examples 12a and 12b, are notably absent in the 2004/5 corpus.

[xiv] We have conducted cross-correlation analyses in order to test for interaction effects between the reporting of speech versus thought and the person in the quotative frame. Generally, it seems that speech reproduction tends to occur in 3rd person contexts while thought representation is much less clearly distributed by person. While this interaction came out significant for the other quotative forms ($\chi 2(2)$ 17.439, $p < .001$), it was not chosen as significant for all ($\chi 2(2)$: 1.555, $p = .460$).

[xv] According to the OED, all can function as an adjective (*with all my heart*), a noun (*whatever it was it was their all*), and an adverb (*all at once*) and it also occurs in a number of special constructions.

[xvi] As we pointed out above, Google is not punctuation sensitive so we did not include quotation marks around the quoted passage. However, we needed to search for the exact string pronoun + *be* + word, which can be achieved in Google searches by putting it in quotes.

[xvii] Finding twice as many occurrences of quotative all in one year than in the preceding year would only indicate a doubling in the frequency of usage if the archives for the two years were the same size. If the archive from the later year were twice as big as the one from the earlier year, a doubling in the occurrence of quotative all would indicate no change in the usage frequency.

[xviii] More specifically, Google provided information on the relative sizes of each year's corpus in both words and postings, starting in May 1980 and going to February 2006. We used the word-based sizes. The relative sizes were given as numbers between 0 and 1, such that the total for all the years added up to 1. Thus, for example, the number associated with 1990 was 0.00204509 and the number associated with 2000 was 0.13503676. From this we could deduce that the archive from 2000 was slightly over 66 times as big as the archive from 1990.

[xix] According to http://www.edict.com.hk/lexiconindex/frequencylists/words2000.htm all is the 36th most frequent word in the Brown corpus (1,015,945 words), occurring with a frequency of .2954 %.

[xx] Regular expressions are patterns allowing optionality, wild cards, and arbitrary repetitions. See Hopcroft, Motwani, & Ullman (2001) or http://en.wikipedia.org/wiki/Regular_expression.)

[xxi] Our initial attempt involved searching for forms of be immediately followed by all, followed by (single or double) quotation marks (with an optional comma after all. Inspection of a small portion of the huge output file from this search revealed that only a tiny fraction of the hits involved quotative all. Minor variations on that pattern (such as stipulating a pronominal subject) did not noticeably improve the results.

[xxii] As in Figure 3, the curly brackets indicate paradigmatic alternatives. The parenthesis around the comma indicates optionality.

[xxiii] So while in our California corpus we subsumed the tokens of all and all like in one category (mainly due to low tokens numbers but also due to methodological decisions detailed earlier), we decided to treat all and all like separately here, hoping that such a separate treatment would give us some information about the diachronic patterning of all like vis a vis all.

[xxiv] Bucholtz (2004) has argued that all is used evaluatively and we agree (see also Labov 1972). Thus, when a speaker uses all evaluatively, they are adding an attitude and thereby assessing the person being quoted (usually negatively). Bucholtz' analysis is entirely comparable with our discussion of stereotypes.

[xxv] Buchstaller (2001, 2004) and Vincent and Dubois (1996) discuss habitual / iterative quotations, a category that is related in that it characterises people or situations via typically occurring quotation.

[xxvi] One problem of our method is that the nature of the usernet is likely to have changed quite dramatically since 1982, conditioned by a range of variables, such as age, education, media use, locality, ethnicity and gender (see Chen, Boase and Wellman 2002 and Katz, Rice and Aspden 2001). Initially, it was mainly restricted to computer wizards, followed by some academics and the army. Indeed, in Rickford et al. (2007:20), we pointed out that "in the early years newsgroups were primarily the province of expert computer users, and much of their content consisted of information exchanges about computers, which might not invite quotation. Later, newsgroups also became a forum for discussions of popular culture by a much wider group of users." As one anonymous reviewer has rightly pointed out, in more recent years, newsgroups have returned to being only the domain of computer aficionados while more casual users would use message boards, facebook groups, etc. We are grateful to David White for suggesting that it might be the case that specific quotatives rich literary genres such as certain genres of creative writing might have left the newsgroups completely. This is likely to have influenced the style of the postings and potentially also the choice and use of quotatives.

[xxvii] To illustrate the calculations used in producing our figures, consider the uses of like to introduce speech in the year 2000. Table 8 shows that there were 55 examples in our sample of 1000 sentences. In order to account for year-to-year variation in the archive sizes, we multiplied .13503676 (the fraction of the total corpus coming from the year 2000) by 13 (the number of years in our sample); the product, 1.7554779, tells us that the archive size for 2000 was about one and three-quarters times as big as the average year's archive, so we normalize by dividing 55 by 1.7554779, yielding 31.3305. We then multiplied this by 10.939 (because our 55 examples came from a sample of only 1000 out of 10939 total) yielding 342.72434. This is the number that appears in Figure 5.

[xxviii] They did, however, have somewhat longer histories than the all-quotative, sometimes as much as a hundred years. This may be a function of the textual record. We may simply not know that a certain form spread for a short while because we do not have the manuscripts to show that. Furthermore, in many older texts we have only one or two examples of any form, not the larger numbers that contemporary databases give access to.

[xxix] Thorsten Brants (brants@google.com), our principal collaborator at Google, has expressed interest in working with other linguists interested in using the newsgroups for research.