

Broadcast	Language	Genre	Unit	P3 Target	P3 To Date	% P3 To Date	P3R1 (LDC2007E99)					P3R2 (LDC2008E38)					Notes		
Broadcast	Arabic	BC	hours	1000	951	95%									435				516
	Arabic	BN	hours	1000	1104	110%									373				731
	Chinese	BC	hours	1000	698	70%									305				393
	Chinese	BN	hours	1000	621	62%									377				243
	English	BC	hours	250	121	48%									60				60
	English	BN	hours	250	121	48%									60				60
Newswire				P3 Target	P3 To Date	% P3 To Date	P3R1					P3R2					Notes		
	Arabic	NW	tokens (words)		See Notes		See Notes					See Notes							
	Chinese	NW	tokens (chars)		See Notes		See Notes					See Notes							
	English	NW	tokens (words)		See Notes		See Notes					See Notes							
Web Text				P3 Target	P3 To Date	% P3 To Date	P3R1 (LDC2007E102)					P3R2 (LDC2008E41)					Notes		
	Arabic	NG	tokens (words)	10,000,000	70,784,544	708%	27,838,327					42,946,217							
	Arabic	WL	tokens (words)	10,000,000	98,729,851	987%	26,296,828					72,433,023							
	Chinese	NG	tokens (chars)	10,000,000	271,207,196	2712%	164,658,186					106,549,010							
	Chinese	WL	tokens (chars)	10,000,000	206,361,199	2064%	185,550,342					20,810,857							
	English	NG	tokens (words)	10,000,000	827,115,809	8271%	400,201,443					426,914,366							
	English	WL	tokens (words)	10,000,000	4,515,219	45%	3,024,785					1,490,434							
Transcription				P3 Target	P3 To Date	% P3 To Date	P3R1 (LDC2007E100)					P3R2 (LDC2008E39)					Notes		
							QTR	QRTA	Web	CCAP	P3R1 Total	QTR	QRTA	Web	CCAP	P3R2 Total			
	Arabic	BC	hours	300	543	181%		241	140		381		123	40		163			
	Arabic	BN	hours	200	450	225%		306			306		144			144			
	Chinese	BC	hours	300	280	93%	72	149			220	38	12	10		60			
	Chinese	BN	hours	200	341	171%	89	124	80		293	12	8	27		48			
	English	BC	hours	0	87	n/a				38	38				49	49			
English	BN	hours	0	100	n/a				47	47				53	53				
Translation				P3 Target	P3 To Date	% P3 To Date	LDC2007E101	LDC2007E103	P3R1 Total		P3R2 (LDC2008E40)					Notes			
	Arabic	NW	tokens (words)	See notes	10,319,327	n/a	102,377	10,216,950	10,319,327										
	Arabic	WEB	tokens (words)	See notes	0	n/a			-										
	Arabic	BN	tokens (words)	See notes	220,797	n/a	159,480		159,480		61,317								
	Arabic	BC	tokens (words)	See notes	442,641	n/a	245,789		245,789		196,852								
	Chinese	NW	tokens (chars)	See notes	13,310,949	n/a	146,573	13,164,376	13,310,949										
	Chinese	WEB	tokens (chars)	See notes	0	n/a			-										
	Chinese	BN	tokens (chars)	See notes	501,852	n/a	307,516		307,516		194,336								
Chinese	BC	tokens (chars)	See notes	485,475	n/a	444,373		444,373		41,102									

Includes both LDC- and web-harvested audio

All collected NW to be released in Gigaword 4 (2009) and/or ad hoc as required for specific tasks

Includes manual and web-harvested transcripts

Targets unspecified; data committee instructions are to translate up to 10Kw/source for new or previously under-represented sources. NW Volumes includes both manually translated and found parallel text.