

Training Data		Language	Genre	Unit	Minimum Targeted Volume	Volume Released to Date	R1 2.15.2007	R2 4.2.2007	R3 5.1.2007	Notes	
Collection	Arabic	NW	Words	23,000,000	See Notes					All collected NW to be released in Gigaword 3	
		BN	Hours	1,000	567	LDC2007E03	311	LDC2007E43		Includes both LDC- and web-harvested audio	
		BC	Hours	1,000	586	LDC2007E03	294	LDC2007E43		Includes both LDC- and web-harvested audio	
		WL	Words	5,000,000	5,571,378	LDC2007E04	2,480,147	LDC2007E44			
		NG	Words	5,000,000	35,715,439	LDC2007E04	13,824,636	LDC2007E44			
	Chinese	NW	Chars	89,000,000	See Notes						All collected NW to be released in Gigaword 3
		BN	Hours	1,000	487	LDC2007E03	365	LDC2007E43			Includes both LDC- and web-harvested audio
		BC	Hours	1,000	409	LDC2007E03	285	LDC2007E43			Includes both LDC- and web-harvested audio
		WL	Chars	5,000,000	8,183,366	LDC2007E04	5,345,417	LDC2007E44			Includes both LDC- and web-harvested audio
		NG	Chars	5,000,000	314,102,450	LDC2007E04	154,695,747	LDC2007E44			
	English	NW	Words	207,000,000	See Notes						All collected NW to be released in Gigaword 3
		BN	Hours	250	91	LDC2007E03	61	LDC2007E43			Includes both LDC- and web-harvested audio
		BC	Hours	250	100	LDC2007E03	70	LDC2007E43			Includes both LDC- and web-harvested audio
		WL	Words	5,000,000	9,206,512	LDC2007E04	5,996,083	LDC2007E44			Includes both LDC- and web-harvested audio
		NG	Words	5,000,000	83,400,142	LDC2007E04	4,401,928	LDC2007E44			
Transcription	Arabic	BN	Hours	200*	518	LDC2007E05	225	LDC2007E45	155	LDC2007E86	*Plus additional to make up for Phase 1 shortfall
		BC	Hours	300*	545	LDC2007E05	285	LDC2007E45	12	LDC2007E86	*Plus additional to make up for Phase 1 shortfall
	Chinese	BN	Hours	200	351	LDC2007E05	175	LDC2007E45	53	LDC2007E86	
		BC	Hours	300	317	LDC2007E05	206	LDC2007E45	14	LDC2007E86	
	English	BN	Hours	50	54	LDC2007E05	24	LDC2007E45	-	LDC2007E86	English transcripts are CCAP
		BC	Hours	50	58	LDC2007E05	31	LDC2007E45	-	LDC2007E86	English transcripts are CCAP
Translation	Arabic	NW	Words	150,000	180,900	LDC2007E06	-	LDC2007E46	105,300	LDC2007E87	
		BN	Hours	10	5	LDC2007E06	1	LDC2007E46	4	LDC2007E87	
		BC	Hours	23	53	LDC2007E06	50	LDC2007E46	3	LDC2007E87	
		WL	Words	150,000	700	LDC2007E06	-	LDC2007E46	700	LDC2007E87	
		NG	Words	150,000	43,200	LDC2007E06	-	LDC2007E46	8,300	LDC2007E87	
	Chinese	NW	Chars	225,000	298,800	LDC2007E06	-	LDC2007E46	113,900	LDC2007E87	
		BN	Hours	6	33	LDC2007E06	23	LDC2007E46	10	LDC2007E87	
		BC	Hours	14	23	LDC2007E06	11	LDC2007E46	13	LDC2007E87	
		WL	Chars	225,000	4,900	LDC2007E06	-	LDC2007E46	2,700	LDC2007E87	
		NG	Chars	225,000	109,500	LDC2007E06	-	LDC2007E46	73,900	LDC2007E87	