LPC INTEGRATION TEST RESULTS

This document presents detailed test results of integration tests of Language Processing Chains for Bulgarian, German, Greek, English, Polish and Romanian implemented within ATLAS project.

For further explanations see *D4.1 – Language Processing Chains* ATLAS deliverable.

Definition of classes

This table has been repeated from section 3.6 of the D4.1 document and is presented here only to maintain internal coherence of this document.

Class	Number of tokens
c0	0 – 1000
c1	1001 – 2000
c2	2001 – 4000
c3	4001 – 8000
c4	8001 – 16000
c5	16001 – 32000
c6	32001 – 64000
c7	64001 – 128000
c8	128001 and more

Document count

The following table presents the number of documents used for performance testing, for all document classes and all project languages:

Class	Language						
	BG	**DE**	**EL**	**EN**	**PL**	**RO**	
c0	18462	16557	13743	59504	18375	16488	143129
c1	6302	7231	5158	21742	6023	6584	53040
c2	4087	4355	3318	12766	4231	4220	32977
c3	3829	4052	2980	10044	3866	3930	28701

Class	Language						
	BG	DE	EL	EN	PL	RO	
c4	1978	2271	1815	6038	2204	2768	17074
c5	883	1004	782	2959	967	1230	7825
c6	310	379	301	1074	352	512	2928
c7	100	132	86	362	123	155	958
c8	56	61	45	100	60	80	402
	36007	36042	28228	114589	36201	35967	287034

Average document size

The following table presents the average number of tokens in the documents used for performance testing, for all document classes and all project languages:

Class	Language						
	BG	DE	EL	EN	PL	RO	
c0	566	597	600	577	563	606	585
c1	1386	1361	1373	1374	1387	1387	1378
c2	2869	2868	2843	2857	2859	2867	2861
c3	5657	5737	5748	5734	5711	5751	5723
c4	11087	10985	10990	11176	10941	10955	11022
c5	21999	21916	22138	21956	22008	21944	21994
c6	42662	42220	43025	43416	42407	42946	42779
c7	84756	86542	87824	86229	85574	85725	86108
c8	241182	243250	233587	197995	241662	255502	235530
	412164	415476	408128	371314	413112	427683	407980

Performance results: average processing time

The following table shows the average processing time in seconds for all document classes and all integrated tools:

Lan-guage	Action	Class								
		c0	**c1**	**c2**	**c3**	**c4**	**c5**	**c6**	**c7**	**c8**
BG	Sentence splitter	0,07	0,14	0,28	0,52	0,91	1,72	3,15	6,30	12,44
	Tokenizer	0,04	0,09	0,23	0,29	0,62	0,83	5,84	137,72	37,89
	Lemmatizer	0,02	0,03	0,05	0,08	0,14	0,25	0,46	0,82	2,04
	POS tagger	0,86	1,05	1,57	2,23	3,96	6,59	12,59	23,50	51,42
	NP extractor	0,08	0,09	0,11	0,16	0,30	0,66	1,73	4,57	33,58
	NE recognizer	0,04	0,07	0,11	0,19	0,37	0,69	1,50	2,76	9,43
	NE recognizer – JRCNames	0,07	0,11	0,20	0,27	0,64	1,50	3,94	5,13	30,80
	LPC (total)[1]	**1,18**	**1,59**	**2,54**	**3,72**	**6,94**	**12,24**	**29,21**	**180,80**	**177,59**

[1] The total value includes execution times for all integrated tools together with pre- and post-processing tools.

DE	Paragraph splitter	0,02	0,01	0,01	0,02	0,02	0,01	0,01	0,02	0,03
	Tokenizer	2,98	2,20	3,62	6,54	11,92	25,47	98,66	715,20	4488,10
	ParZu	6,49	35,72	18,24	20,53	31,08	49,32	81,97	150,68	325,10
	NP fixer	0,01	0,03	0,06	0,26	0,93	3,18	13,46	71,21	1033,14
	NE recognizer – JRCNames	0,33	0,72	1,44	2,90	5,37	9,04	18,25	36,01	96,81
	NE fixer	0,01	0,01	0,02	0,06	0,14	0,34	1,26	7,57	108,60
	LPC (total)	**9,82**	**38,69**	**23,40**	**30,31**	**49,45**	**87,26**	**213,61**	**980,69**	**6051,78**
EL	Sentence splitter	4,81	11,67	25,14	53,94	102,55	196,71	401,36	747,54	2071,77
	Lemmatizer	0,04	0,07	0,13	0,24	0,46	0,89	1,74	3,46	10,89
	NP extractor	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,04	0,22
	NE recognizer	4,82	9,15	17,23	29,24	52,11	108,49	223,85	549,41	1410,93
	NE recognizer – JRCNames	0,05	0,07	0,15	0,20	0,36	1,13	2,61	7,85	25,57
	LPC (total)	**9,71**	**20,96**	**42,64**	**83,62**	**155,49**	**307,23**	**629,57**	**1308,30**	**3519,38**

EN	Paragraph splitter	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,02	0,03
	Sentence splitter (Open NLP)	0,00	0,00	0,00	0,00	0,01	0,02	0,05	0,13	1,46
	Tokenizer (Open NLP)	0,02	0,04	0,07	0,13	0,26	0,52	1,12	2,17	5,22
	Lemmatizer RASP	0,03	0,06	0,13	0,25	0,48	0,91	1,82	3,48	6,70
	NP extractor	0,13	0,14	0,14	0,15	0,24	0,53	1,69	6,38	40,20
	POS tagger (OpenNLP)	0,03	0,07	0,14	0,25	0,47	0,91	1,81	3,76	9,04
	NE recognizer (OpenNLP)	0,16	0,33	0,58	1,00	1,78	3,32	6,45	12,72	27,88
	NE recognizer – patterns	0,00	0,00	0,01	0,01	0,03	0,06	0,11	0,22	0,54
	NE recognizer – JRCNames	0,15	0,31	0,59	0,94	1,70	3,24	6,35	13,09	33,01
	LPC (total)	**0,54**	**0,96**	**1,67**	**2,75**	**4,98**	**9,52**	**19,41**	**41,97**	**124,07**
PL	Sentence splitter, tokenizer lemmatizer, POS tagger	4,79	5,02	5,11	5,65	6,56	8,61	12,20	22,30	58,90
	NP extractor	0,21	0,38	0,75	1,58	2,99	5,89	11,75	23,03	60,15
	NE recognizer	0,17	0,39	0,74	1,44	2,76	5,71	13,38	46,95	327,59
	NE recognizer – JRCNames	0,11	0,21	0,38	0,60	1,10	2,39	4,62	11,03	32,81

	LPC (total)	**5,29**	**5,99**	**6,98**	**9,28**	**13,41**	**22,59**	**41,96**	**103,30**	**479,44**
RO	Tokenizer	0,32	0,64	1,26	3,00	4,68	9,54	19,35	39,76	125,28
	NP extractor	0,36	0,82	1,70	5,03	6,49	12,86	25,70	48,42	144,20
	NE recognizer	1,58	1,04	0,79	1,74	5,97	4,09	15,01	11,46	49,07
	NE recognizer – patterns	0,00	0,01	0,01	0,02	0,04	0,08	0,14	0,26	0,71
	NE recognizer – JRCNames	0,13	0,24	0,42	0,65	1,12	2,25	4,71	10,87	31,22
	LPC (total)	**2,39**	**2,75**	**4,17**	**10,45**	**18,30**	**28,82**	**64,91**	**110,76**	**348,69**