Procura-PALavras (P-PAL): A web application for a new European Portuguese lexical database

Ana Paula Soares¹, Montserrat Comesaña¹, Álvaro Iriarte², José João de Almeida³, Alberto Simões³, Ana Costa¹, Patrícia França¹ & João Machado¹ {¹Psycholinguistics Research Group, ²Centre for Humanistic Studies, ³Computer Science and Tecnology Center}, University of Minho

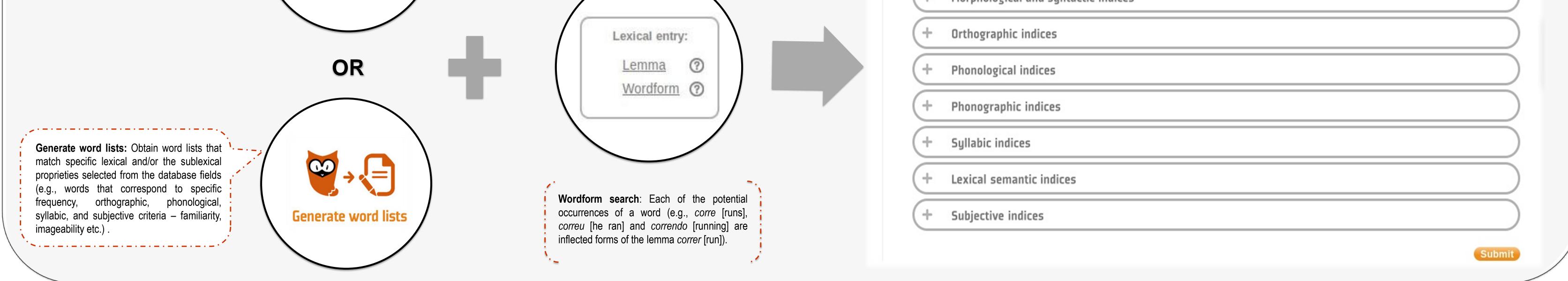
Procura-PALavras (P-PAL) is a web application for a new European Portuguese (EP) lexical database that provides a series of objective (lexical and sublexical) and subjective indices for ≈160.000 nonlemmatized and ≈42.000 lemmatized EP words. Based on a *corpus* of over 225 million EP words, the P-PAL enables users to obtain a broad range of statistics concerning the properties of word stimuli, including several measures of word frequency, syllable frequency, bigram and biphone frequency, orthographic and phonological structure, morphological and syntactic structure, orthographic and phonological similarity, lexical semantic indexes, concreteness, familiarity, imageability, valence, arousal, and dominance measures. In order to obtain these statistics the user should decide between a lemma or wordform search in the application and between two word-based queries: (i) analyze word lists in specific characteristics (objective and/or subjective); or (ii) generate lists of words with specific characteristics (objective and/or subjective). In this work we present the lemma and wordform frequency indices available, as well as some structure and similarity orthographic measures such as word length, neighborhood density and frequency. Letters and bigram type and token frequencies are also presented.

Analyze word lists: select specific database		Lemma search: The canonical form chosen as representative of all inflected forms of a	🗑 P-PAL	Universidade do Minho Universidade do Minho	idade do Minho
fields to analyze your own vocabulary lists according to the lexical and/or sublexical proprieties of words (e.g., uniqueness point, phonologic features, number of syllables,	$ \rightarrow \odot $	word (e.g., the lemma <i>ir</i> [go] represents the inflected forms <i>vou</i> [I go], <i>vai</i> [goes], <i>indo</i> [going], <i>fui</i> [I went] and <i>ido</i> [gone]).	Back to menu		e Letras e Ciéncias Humanas
parts of speech, homographs, neighborhood statistics, Levenshtein distance).	Analyze word lists		+ Word frequency indices + Morphological and syntactic indices		\leq

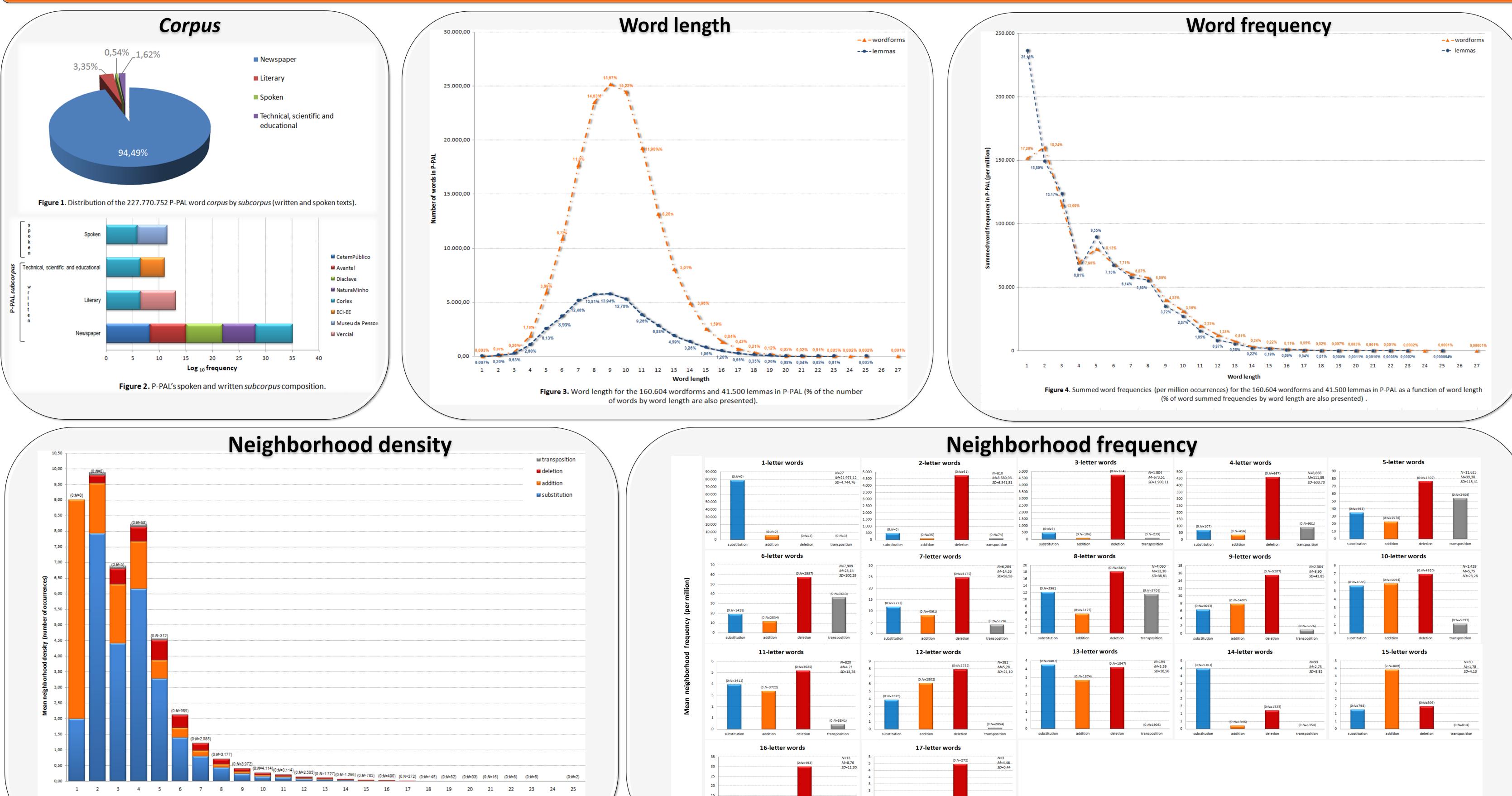


University of Minho, School of Psychology Research Group in Psycholinguistics





Procura-PALavras: http://p-pal.di.uminho.pt/



Word length

Figure 5. Mean neighborhood density (number of occurrences) for the 41.500 lemmas in P-PAL as a function of word length (*N* for words with no substitution, addition, deletion and transposition neighbors by word length presented in parentheses above each column).

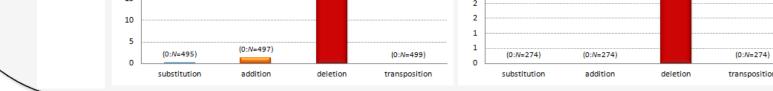


Figure 6. Mean neighborhood frequency (per million occurrences) for 1-17 word length lemmas in P-PAL as function of Neighborhood type (*N* for words with no neighbors by type of neighborhood presented in parentheses above each column).

Letters and Bigrams

le	tters	Туре	Token	P	redecess	aor A	В	C D	Е	FG	Н	I J	K L	М	N	0 P	Q	R	S	T	U \	/ W	ХҮ	Z	Predec	essor	A	В		6	D	Β	F		G	H]	1	K	L	М	N	0	P	Q	R	S	Т	U	٧	W	X	γ	Z
			frequency	M	A	4	1.039 3.5	530 5.370) 73 F	42 1.075	10	479 128	3 7 4.53	5 2.972	4.656 3.	016 97	165	9.027 1	1.495 2	415 5	21 1.2	95 1	86 14	282		Α	0,8	7.005	5,6 23.9	52,3 21	.222,4	389,8	2.195	i,4 5.	821,8 2	29,5 10.	06,2 3	370,4	10,9	32.401,2	2.844,0 3	1.912,5	36.726,2	7.536.853,0	1.870,2	89.489,2	2 15.448,	1 14.288,2	3.954,3	6.764	4,4 0,8	415,9	25,4 4.	4.260,0
	A	32.804		E E	В	1.2	06 3 3	0 28	682	5 3	0 1	1.064 39	0 141	19	9 7	71 6	1	859	160	28 4	04 1	2 1	0 14	1		В	6.499,6	6 16,0	2	,8	33,0	6.257,8	0,5	i	1,0	0,0 3.4	21,4 4	467,3	0,0	2.674,4	68,0	2,9	3.715,8	2,0	0,0	6.815,7	893,9	265,1	1.369,5	92,9	9 2,4	0,0	23,6	0,1
	в	5.353	32.574,83	as i	C	6.3	74 0 14	48 2	1.645	0 0	845 2	2.752 0	38 389	1	45 5.	168 0	0	842	1 6	656 9	50 1	0	0 0	5		С	46.804,	,8 0,0	1.4	30,3	27,9	14.830,2	0,0)	0,0 3.	009,7 23.	150,1	0,0	12,5	3.044,2	0,5	472,3	54.473,5	0,0	0,0	5.277,1	1,9	6.583,6	7.557,9	0,4	, 0,0	0,0	0,0	5,1
	C	17.112	147.254,22	E	D	2.8	41 1 1	1 1	4.657	0 3	0 2	2.071 30	1 3	34	1 4.	318 0	5	384	6	0 4	37 2	3 2	06	0		D	21.013,	4 0,6	0	0	0,0	122.632,0	0,0	r.	1,5	0,0 18.	463,3	86,5	2,2	0,8	597,8	0,2	26.759,5	0,0	95,8	1.020,2	2,8	0,0	3.576.575,0	0 246,	,3 3,4	0,0	4,0	0,0
	D	12.587	183.401,81	e e	E	88	37 331 1.8	318 956	164 4	420 832	2 1	1.961 251	1 2.59	8 1.758	6.783 3	53 556	218	4.116 4	4.610 1	.465 2	57 47	75 3	683 7	232		E	3.690,8	5 1.703	8,4 18.3	06,5 6	.267,6	552,9	3.766	5,1 9.0	645,5	4,2 12.	95,8 6	582,7	0,0	15.029,6	3.101,4 4	7.800,6	1.241,9	4.310,4	2.144,6	80.796,3	49.722,	6.925,5	5.401,6	5.728	3,9 1,5	5.686,0 2	2.788,0 2	2.970,5
	E.	25.869		1.50	F	87	6 0 /	0 0	816	19 0	0 1	1.272 0	1 250	0	0 7	53 <mark>0</mark>	0	497	1	18 3	56 0) ()	0 1	0		F	7.122,7	7 0,0	0	0	0,0	6.409,2	38,	7	0,0	0,0 10.)84,5	0,0	0,9	804,2	0,0	0,0	4.744,9	0,0	0,0	2.295,9	0,0	29,1	2.515,4	0,0	0,0	0,0	1,6	0,0
	F	4.805	00.004,00	е 4	G	1.5	95 2 /	0 2	945	0 5	6	918 <mark>0</mark>	0 140	53	172 7	00 0	0	875	2	1 7	30 C	0 (0 0	0		G	8.820,2	2 0,1	0	0	0,6	5.766,3	0,0		8,1 6	62,7 4.8	32,0	0,0	0,0	352,1	91,2	748,8	7.258,6	0,0	0,0	6.515,5	3.205,0	0,0	9.292,8	0,0	0,0	0,0	0,0	0,0
F	G	6.039	43.528,67	는 다	н	1.5	/10 1 /	0 0	549	0 0	0	371 0	0 0	8	2 7	63 0	0	1	0	9 1	45 0	0 (0 0	0		H	10.098,	,2 1,2	0	0	0,0	5.173,0	0,0)	0,0	0,0 8	5,7	0,0	0,0	0,0	3,3	2,5 6	.331.118,0	0,0	0,0	3,3	0,0	60,8	998.364,0	0,0	0,0	0,0	0,0	0,0
4-	н.	3.330	23.606,45	¥. ₽	1	3.4	04 455 4.2	205 2.26/	4 394 F	35 726	2	6 55	9 1.71	2 1.751	4.359 2.	429 45	109	2.416 3	3.118 1	.882 5	56 1.4	75 1	173 0	1.295		1	24.033,	6 2.193	8,8 25.0	18,3 15	5.344,2	2.878,2	3.188	,9 <u>5.</u>	661,4	1,2	,1	70,5	10,6	10.145,1	4.004,3 2	8.659,8	16.180,0	3.522.943,0	250,3	26.435,3	31.087,	3 19.021,7	155.468,0	9.410	0,3 0,8	1.582,1	0,0 5	j.958,5
as in		24.389	197.173,62	, indi	J	40	05 0 0	0 0	165	0 0	0	15 <mark>0</mark>	0 0	0	0 2	13 0	0	0	0	0 2	66 0	0 (0 0	0		J	3.445,8	8 0,0	0	0	0,0	2.207,7	0,0		0,0	0,0 1	5,7	0,0	0,0	0,0	0,0	0,0	2.219,5	0,0	0,0	0,0	0,0	0,0	2.606,7	0,0	0,0	0,0	0,0	0,0
Ĩ.		1.060	10.492,29 274,52	edin	С к	< 1	3 2 /	0 0	27	0 1	2	25 0	0 1	1	1	8 3	0	1	3	1	2 0	2	0 2	0	ζ.	К	19,0	0,5	0	0	0,0	38,5	0,0	1	0,6	0,6 4	9,9	0,0	0,0	0,1	0,5	2,9	2.283,0	2,6	0,0	9,2	4,8	5,0	2,4	0,0	483,0	0,0	5,1	0,0
0 le	.	130 14.277	96.368,29	E L	Ξ L	2.9	11 39 19	90 146	1.589 1	100 151	1.019 2	3.218 5	5 28	439	9 1.	610 76	17	20	102 8	514 5	90 18	31 0	0 3	3	TEN	L	13.980,	6 122,	6 50	4,2 60	6. <mark>114</mark> ,0	12.724,0	114,	5 2.	235,0 7.	525,6 17.	97,9	0,4	6,8	27,7	2.203,3	26,0	8.323,3	193,5	650,5	4,2	342,6	5.320.052,	3.885,6	1.403	3,5 0,0	0,0	0,8	0,1
1.50		14.277	168.904,10	B	S M	2.5	34 605 0	0 0	4.059	0 0	0 1	1.600 0	0 2	2	23 2.	210 1.02	9 0	0	0	0 4	59 C	0 (0 0	0	ğ	М	33.620,	0 4.854	i,1 0	0	0,0	28.650,9	0,0		0,0	0,0 13.	80,3	0,0	0,0	9,6	1.181,0	74,6	14.315,7	9.589,8	0,0	0,0	0,0	0,0	6.539,0	0,0	0,0	0,0	0,0	0,0
he 4	N	12.034	153.943,86	q L	N N	2.6	50 1 2.0	JO2 1.38 ⁴	1 1.266 4	491 710	761 1	1.846 131	10 30	0	3 1.	534 2	131	72 1	1.319 6	.622 2	96 27	77 0	32 2	68	Ë	N	23.146,	,5 0,7	17.7	80,6 16	6.761,0	6.209,7	2.626.5	21,0 2.	925,8 5.	926,7 10.	58,0 498	8.348,0	51,8	29,0	0,0	30,9	16.414,9	1,1	1.383,0	129,3	9.573,0	50.541,3	3.159,1	2.593	3,5 0,0	18,3	1,3 7	291,6
ort	0	26.748	350.639,13	toke	80	36	38 574 97	78 458	192 7	923	4	311 71	12 2.05	1 1.360	3.309	75 600	96	4.730 1	1.7 5 0 8	871 3	71 39	90 17	92 6	66	EN	0	1.365,4	4 5.798	8,9 8.0	16,9 7	.983,0	717,5	1.701	,4 3.	705,2 2	21,2 9.7	57,5 1.	516,5	7,5	10.709,6	3.524,1 2	9.638,7	309,4	4.753,4	216,8	42.834,0) 10.161,	2 3.736,2	4.796,8	6.120),4 29,5	925,9	8,4 5	576,4
ies	P	8.637	108.547,98	and	≩ P	1.9	03 0 4	4 0	1.835	0 1	0	929 0	0 479	0	20 1.	385 9	0	1.700	85 2	208 4	06 0	0 (0 1	0	ŏ	P	26.005,	,3 0,0	41	6,7	0,0	15.364,5	0,0)	0,8	0,0 2.9	70,6	0,0	0,0	3.457,2	0,0	25,3	35.032,0	17,0	0,0	24.121,3	3 137.331	0 692,9	3.187,2	0,0	0,0	0,0	49,5	0,0
Juan	0	1.309	41.112,76	e d	Q	2 0	/ 0 /	0 0	0	0 0	0	1 0	0 0	0	0	0 0	0	0	0	0 1.3	308 0	0 (0 0	0	100	Q	0,0	0,0	0	0	0,0	0,0	0,0	1	0,0	0,0	,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	41.112,7	0,0	0,0	0,0	0,0	0,0
requ	R	25.800	288.525,39	5	R	6.3	10 220 5	51 579	4.794 1	19 360	0 4	4.260 25	8 63	618	442 3.	575 17	106	1.345	257 9	965 6	15 23	37 1	5 5	13		R	49.826,	,0 580,	6 5.2	72,6 3	.670,9	45.872,6	296,	6 2.	674,1	0,0 26.	44,3	13,1	24,1	342,2	5.813,6	.368,8	27.283,0	652,2	1.675,5	5.603,7	2.212,8	10.956,4	3.231,0	1.932	2,5 0,0	10,7	9,9	21,9
ter	s	14.678	181.643,79	Sa	S	2.0	164 103 1.3	313 24	1.744 2	/10 108	32 1	1.774 5	13 84	1.025	68 1.	776 854	139	25 1	1.229 3	.346 9	65 7	63	0 1	0		S	19.201,	,5 110,	5 6.0	27,8	664,5	51.821,4	802,	4 1	74,8 8	54,9 14.)51,9	0,6 19	.063,0	470,5	2.856,5	77,3	16.099,1	6.883,0	500.227,0	103,2	15.854,	31.386,6	10.318,9	147,4	,4 3,6	0,0	1,9	0,0
let	т	18.698	173.419,57	-suc	т	4.9	92 3 1	,3 1	5.571	2 0	9 3	3.753 <mark>0</mark>	0 19	20	27 <mark>3</mark> .	667 1	0	2.500	16	21 1.0	063 1	2	0 1	9		T	45.781,	,5 1,0	1	,5	1,4	49.374,8	1,2	19	0,0	12,6 23.	404,6	0,0	0,0	08.596,0	123,7	55,3	33.832,6	0,2	0,0	25.283,7	8,7	10,7	10.858,6	33,3	3 20,0	0,0	0,7	3,8
le ke	U	9.949	144.371,92	de la	U	75	6 476 3	9 <mark>4 4</mark> 13	834 1	/01 1 97	1 /	923 37	0 1.37	4 540	717 1	09 349	22	1.499	680 8	804	0 9	91	50 0	98		U	16.620,	9 3.262	2,2 3.1	13,7 3	.555,6	34.166,3	246,	6 2.	213,1	0,0 10.	66,5 4	129,3	0,0	7.796,6	5.342,6	.466,0	404,5	2.201,4	17,2	9.116,6	4.694,9	8.064,7	0,0	701,2	,2 2,2	137,8	0,0 8	873,8
ndt	v	5.473	46.730,38	esso	V	1.2	39 0 /	0 C	2.085	0 0	0 1	1.240 0	0 0	0	1 9	87 0	0	48	0	0 6	50 C) ()	0 1	0		V	7.534,2	2 0,0	0	0	0,0	20.409,3	0,0	0	0,0	0,0 10.	52,1	0,0	0,0	0,0	0,0	1,6	8.664,8	0,0	0,0	659,8	0,0	0,0	300,4	0,0	/ 0,0	0,0	0,4	0,0
ype a	w	54	89,37	Sec	W	/ 1	j 2 '	1 0	5	0 0	1	7 0	0 1	0	3	6 0	0	1	2	1	0 0	0 (0 0	0		W	31.422,	,0 3,3	0	1	0,0	8,9	0,0		0,0	4,1	,6	0,0	0,0	0,4	0,0	1,2	6,3	0,0	0,0	0,1	2.283,0	0,1	0,0	0,0	/ 0,0	0,0	0,0	0,0
1	x	1.164	8.806,88	ă	X	23	.30 6	<i>i</i> 6 0	109	2 0	0	234 0	0 0	0	0 1	36 152	0	0	0	159 4	48 0	0 (0 1	0		X	1.711,2	2 0,0	52	7,9	0,0	1.097,5	0,1		0,0	0,0 2.2	35,7	0,0	0,0	0,0	0,0	0,0	672,3	1.586,5	0,0	0,0	0,0	839,1	112,6	0,0	0,0	0,0	1,4	0,0
able 1	Y	71	140,35	ble 2	Y	4	3 /) O	7	0 0	0	2 0	0 2	0	1	3 1	0	3	3	3	2 0) 1	0 0	0		Y	2,8	1.642	2,0 0	0	0,0	2,1	0,0		0,0	0,0	,9	0,0	0,0	0,6	0,0	0,4	1,9	0,0	0,0	49,6	0,6	2,9	2,9	0,0	/ 0,5	0,0	0,0	0,0
Ta	Z	2.160	15.439,20		Z	1.4	27 0 /) O	199	0 0	0	186 0	0 1	11	0 1	09 0	0	0	0	1 3	38 (0 (0 0	7		Z	4.597,	1 0,0	0	0	0,0	5.600,2	0,0		0,0	0,0 9	6,7	0,0	0,0	3,5	49,1	0,0	790,3	0,0	0,0	0,0	0,0	0,5	85,4	0,0	0,0	0,0	0,0 <mark>5</mark> 5	5.029,0

For more information: <u>asoares@psi.uminho.pt</u>



UNIÃO EUROPEIA FEDER COMPETE



Research project [PTDC/PSI-PCO/104679/2008] funded by FCT (Fundação para a Ciência e Tecnologia), and FEDER (Fundo Europeu de Desenvolvimento Regional) through the European programs QREN (Quadro de Referência Estratégico Nacional), and COMPETE (Programa Operacional Factores de Competitividade).