

*Rapport sur corpus ROVER*  
Livrable D14  
Projet PASSAGE  
Référence: ANR-06-MDCA-013-02  
Échéance : T36 - 31 décembre 2009

Patrick Paroubek  
LIMSI-CNRS  
Bât. 508 Université Paris XI  
91403 Orsay Cedex  
tél. 01 69 85 80 04  
email pap@limsi.fr

# Contents

<b>1</b>	<b>Le corpus ROVER</b>	<b>2</b>
1.1	Les données des participants . . . . .	2
1.2	Les performances du ROVER . . . . .	4
1.2.1	Performance du ROVER sur le corpus de référence manuelle . . . . .	6

# Chapter 1

## Le corpus ROVER

### 1.1 Les données des participants

La table suivante décrit les données qui ont été annotées par les participants lors de la seconde campagne PASSAGE, chacun utilisant sa propre segmentation en mots et en énoncés.

Participant	Sous-corpus	fichiers annotés	énoncés	mots	groupes	relations
P1	Parlement Européen	73	1042918	27969087	13120081	0
P1	Ester	1	30857	233004	112650	26
P1	EstRépublicain	5	102009	1378377	636770	0
P1	FrWiki	67	2464543	23609501	11106351	96
P1	JRC	101	2022233	36656516	17150958	3636
P1	LeMonde	1	6810	150674	69095	0
P1	Wiki Source	70	1809651	28319391	12845338	69
P1	Total	318	7479021	118316550	55041243	3827
P2	Parlement Européen	73	1019106	28763977	13782876	14539360
P2	Ester	1	13909	236381	112166	117090
P2	EstRépublicain	5	87162	1440769	664003	644140
P2	FrWiki	67	1713115	24472996	11635311	10675647
P2	JRC	101	1461088	38929339	18597376	16703413
P2	LeMonde	1	5977	154151	71669	74391
P2	Wiki Source	70	1444718	28454317	13013514	13688984
P2	Total	318	5745075	122451930	57876915	56443025
P3	Parlement Européen	73	1091979	27411658	13582663	14039852
P3	Ester	1	14392	237338	109264	110601
P3	EstRépublicain	5	87571	1437964	646319	612441
P3	FrWiki	67	1661231	24087010	11276780	10488411
P3	JRC	101	1803447	38148975	18511317	17047604
P3	LeMonde	1	6137	150785	69364	70651
P3	Wiki Source	70	1524064	27647365	12897271	12889334
P3	Total	318	6188821	119121095	57092978	55258894

... données des participants cont.

Participant	Sous-corpus	fichiers annotés	énoncés	mots	groupes	relation
P4	Parlement Européen	73	983838	28784618	13873446	15304333
P4	Ester	1	24192	228025	115770	109234
P4	EstRépublicain	5	82291	1414895	678949	662210
P4	FrWiki	44	1210313	15712760	7498010	6877431
P4	JRC	5	101021	1877091	920752	894815
P4	LeMonde	1	6228	153019	73534	76785
P4	Wiki Source	0	0	0	0	0
P4	Total	129	2407883	48170408	23160461	23924808
P5	Parlement Européen	73	1013485	28087183	13707098	0
P5	Ester	1	29218	235347	115547	0
P5	EstRépublicain	5	93094	1419279	648925	0
P5	FrWiki	67	1923225	24069423	10782222	0
P5	JRC	101	2199205	38970092	18352466	0
P5	LeMonde	1	6502	153228	70728	0
P5	Wiki Source	70	1745543	27649729	12783082	0
P5	Total	318	7010272	120584281	56460068	0
P6	Parlement Européen	19	280772	7691358	3677439	3267387
P6	Ester	1	14976	228848	109294	88408
P6	EstRépublicain	1	21354	292259	137122	104080
P6	FrWiki	2	68358	779306	374771	276873
P6	JRC	5	82927	1854908	919287	685467
P6	LeMonde	1	6557	148124	69652	58631
P6	Wiki Source	0	0	0	0	0
P6	Total	29	474944	10994803	5287565	4480846
P7	Parlement Européen	19	265902	7636565	3473617	2792244
P7	Ester	1	12430	236892	103877	73971
P7	EstRépublicain	1	15834	307431	129157	96271
P7	FrWiki	2	58132	802264	348675	234391
P7	JRC	5	62118	1884462	794725	494089
P7	LeMonde	1	5772	153951	66557	52889
P7	Wiki Source	0	0	0	0	0
P7	Total	29	420188	11021565	4916608	3743855
P8	Parlement Européen	73	1009124	28248440	0	16181242
P8	Ester	1	9285	228783	0	133337
P8	EstRépublicain	5	90082	1416218	0	746139
P8	FrWiki	67	2021917	23837354	0	12039110
P8	JRC	98	2033240	36635186	0	19039085
P8	LeMonde	1	6548	152099	0	83040
P8	Wiki Source	70	1547955	28254616	0	15627799
P8	Total	315	6718151	118772696	0	63849752

... données des participants cont.

P9	Parlement Européen	73	1758654	25051799	14214079	13368957
P9	Ester	1	48900	201320	124118	92032
P9	EstRépublicain	5	160033	1152032	682761	531824
P9	FrWiki	67	3266169	19208808	11661773	8919749
P9	JRC	101	4816436	32191990	19051264	15015306
P9	LeMonde	1	11640	127920	73878	63927
P9	Wiki Source	0	0	0	0	0
P9	Total	248	10061832	77933869	45807873	37991795
P10	Parlement Européen	19	42758	393870	210463	209382
P10	Ester	1	8889	24540	15992	8320
P10	EstRépublicain	1	3960	17939	10556	7390
P10	FrWiki	1	0	0	0	0
P10	JRC	5	21025	56400	38449	19556
P10	LeMonde	1	911	7216	3919	3471
P10	Wiki Source	0	0	0	0	0
P10	Total	28	77543	499965	279379	248119
P11	Parlement Européen	73	1046389	29030724	0	13391152
P11	Ester	1	20124	409356	0	179976
P11	EstRépublicain	5	99050	1466570	0	528238
P11	FrWiki	67	2352997	23568494	0	8682271
P11	JRC	101	2133106	38443435	0	15178617
P11	LeMonde	1	6421	154473	0	63700
P11	Wiki Source	70	1768365	28616704	0	12339677
P11	Total	318	7426452	121689756	0	50363631

## 1.2 Les performances du ROVER

Le principe du ROVER consiste à combiner les données annotées par les participants pour améliorer automatiquement la qualité des annotations en croisant les informations et en sélectionnant les annotations en fonction des mesures de performance effectuées lors de la campagne d'évaluation. Cette opération se fait en deux étapes : d'abord on procède à l'alignement des données sur une segmentation en mots et en énoncés commune pour tous les analyseurs (une segmentation par défaut de granularité minimale, ou bien une segmentation obtenue par vote majoritaire ou bien la segmentation en mots du système ayant la meilleure performance pour les constituants), puis à la projection et fusion/sélection des annotations en groupes et en relations sur cette segmentation commune.

Les expériences pour produire un ROVER n'ont pas donné pour le moment les résultats escomptés initialement. Ils sont bien sûr meilleurs sur les groupes syntaxiques que sur les relations (cf les expériences initiales rapportées dans [1]). Les expériences qui ont pu être menées concernaient des stratégies grossières utilisant des algorithmes de vote majoritaire ou de sélection de maxima en considérant comme paramètre la valeur de performance d'un système pour une annotation donnée sur un sous-corpus donné. Les résultats de performances détaillées du ROVER par rapport au corpus de référence manuelle actuel sont donnés dans la sous-section 1.2.1. Pour ces mesures c'est la segmentation en mots du corpus de référence qui a été utilisée. Il s'agit ici d'une combinaison par sélection du maximum de performance en f-mesure pour une annotation donnée sur une sous-corpus donné. D'autres stratégies de vote majoritaire simple ou pondéré avec la même granularité (annotation et sous-corpus) ont été essayées avec des résultats comparables. Avec ces expériences, il n'a pas été possible d'améliorer la précision par rapport au meilleur score sans perdre en rappel. La table 1.1 montre les performances du ROVER présenté ici par rapport aux meilleures performances des systèmes. Le ROVER présenté ici se classerait donc 3<sup>ème</sup> pour les groupes avec une différence de performance de -0.144, 4<sup>ème</sup> pour les relations avec une différence de performance de -0.052 et 3<sup>ème</sup> en performance globale sur les groupe et les relations avec une différence de performance de -0.029. Clairement, une étude plus

poussée est nécessaire pour profiter de la combinaison des informations afin d’atteindre ou de dépasser les meilleurs scores individuels. En particulier, nous souhaitons tester une approche utilisant de l’apprentissage automatique et un paramétrage plus fin que le simple triplet système/annotation/sous-corpus.

Participant	Corpus	Annotation	Précision	Rappel	F-mesure
P4	ALL	ALL_GRPS	52554/58562=0.897	52246/74907=0.697	0.785
ROVER	ALL	ALL_GRPS	51008/57066=0.894	51286/74907=0.685	0.775
P3	ALL	ALL_RELS	32184/45554=0.707	32197/67462=0.477	0.570
ROVER	ALL	ALL_RELS	30056/48672=0.618	30015/67462=0.445	0.517
P4	ALL	ALL(G+R)	0.785	0.593	0.675
ROVER	ALL	ALL(G+R)	0.756	0.565	0.646

Table 1.1: Comparaison des performances du ROVER

### 1.2.1 Performance du ROVER sur le corpus de référence manuelle

Sous-corpus	Annotation	Précision	Rappel	F-mesure
FrWiki	GN	114/305=0.374	107/4677=0.023	0.043
FrWiki	NV	26/108=0.241	27/2234=0.012	0.023
FrWiki	GA	4/42=0.095	4/1418=0.003	0.005
FrWiki	GR	1/16=0.062	1/805=0.001	0.002
FrWiki	GP	183/278=0.658	180/3745=0.048	0.090
FrWiki	PV	1/12=0.083	1/255=0.004	0.007
FrWiki	ANONYMOUS	110/229=0.480	114/3073=0.037	0.069
FrWiki	ALL_CONSTITUENTS	439/990=0.443	434/16207=0.027	0.051
FrWiki	SUJ-V	6/34=0.176	6/1390=0.004	0.008
FrWiki	AUX-V	1/10=0.100	1/345=0.003	0.006
FrWiki	COD-V	0/22=0.000	0/1071=0.000	0.000
FrWiki	CPL-V	4/75=0.053	5/1402=0.004	0.007
FrWiki	MOD-V	0/10=0.000	0/666=0.000	0.000
FrWiki	COMP	0/4=0.000	0/307=0.000	0.000
FrWiki	ATB-SO	0/7=0.000	0/286=0.000	0.000
FrWiki	MOD-N	34/133=0.256	35/5997=0.006	0.011
FrWiki	MOD-A	0/2=0.000	0/257=0.000	0.000
FrWiki	MOD-R	0/0=0.000	0/62=0.000	0.000
FrWiki	MOD-P	0/0=0.000	0/16=0.000	0.000
FrWiki	COORD	2/24=0.083	2/1103=0.002	0.004
FrWiki	APPOS	0/4=0.000	0/289=0.000	0.000
FrWiki	JUXT	0/30=0.000	0/138=0.000	0.000
FrWiki	ALL_RELATIONS	47/355=0.132	49/13329=0.004	0.007
JRC	GN	475/720=0.660	471/973=0.484	0.558
JRC	NV	348/487=0.715	351/809=0.434	0.540
JRC	GA	173/379=0.456	173/508=0.341	0.390
JRC	GR	46/95=0.484	46/128=0.359	0.413
JRC	GP	729/878=0.830	756/1376=0.549	0.661
JRC	PV	31/46=0.674	31/71=0.437	0.530
JRC	ANONYMOUS	365/676=0.540	365/767=0.476	0.506
JRC	ALL_CONSTITUENTS	2167/3281=0.660	2193/4632=0.473	0.552
JRC	SUJ-V	95/233=0.408	95/435=0.218	0.284
JRC	AUX-V	44/77=0.571	44/136=0.324	0.413
JRC	COD-V	72/156=0.462	72/339=0.212	0.291

... Performance du ROVER sur le corpus de référence manuel cont.

Sous-corpus	Annotation	Précision	Rappel	F-mesure
JRC	CPL-V	96/264=0.364	95/489=0.194	0.253
JRC	MOD-V	31/79=0.392	31/162=0.191	0.257
JRC	COMP	16/52=0.308	16/96=0.167	0.216
JRC	ATB-SO	13/22=0.591	13/74=0.176	0.271
JRC	MOD-N	474/937=0.506	471/1851=0.254	0.339
JRC	MOD-A	9/28=0.321	9/97=0.093	0.144
JRC	MOD-R	1/3=0.333	1/13=0.077	0.125
JRC	MOD-P	0/13=0.000	0/3=0.000	0.000
JRC	COORD	22/107=0.206	22/209=0.105	0.139
JRC	APPOS	0/126=0.000	0/13=0.000	0.000
JRC	JUXT	2/14=0.143	2/115=0.017	0.031
JRC	ALL_RELATIONS	875/2111=0.414	871/4032=0.216	0.284
EstRépublicain	GN	397/433=0.917	398/425=0.936	0.927
EstRépublicain	NV	257/271=0.948	257/271=0.948	0.948
EstRépublicain	GA	81/94=0.862	82/93=0.882	0.872
EstRépublicain	GR	51/58=0.879	51/57=0.895	0.887
EstRépublicain	GP	386/398=0.970	386/421=0.917	0.943
EstRépublicain	PV	47/47=1.000	47/50=0.940	0.969
EstRépublicain	ANONYMOUS	322/374=0.861	322/352=0.915	0.887
EstRépublicain	ALL_CONSTITUENTS	1541/1675=0.920	1543/1669=0.925	0.922
EstRépublicain	SUJ-V	115/177=0.650	115/174=0.661	0.655
EstRépublicain	AUX-V	47/63=0.746	47/63=0.746	0.746
EstRépublicain	COD-V	60/110=0.545	60/112=0.536	0.541
EstRépublicain	CPL-V	106/185=0.573	105/201=0.522	0.547
EstRépublicain	MOD-V	39/71=0.549	39/78=0.500	0.523
EstRépublicain	COMP	12/22=0.545	12/27=0.444	0.490
EstRépublicain	ATB-SO	15/24=0.625	15/29=0.517	0.566
EstRépublicain	MOD-N	281/464=0.606	281/563=0.499	0.547
EstRépublicain	MOD-A	7/13=0.538	7/17=0.412	0.467
EstRépublicain	MOD-R	2/2=1.000	2/3=0.667	0.800
EstRépublicain	MOD-P	0/12=0.000	0/0=0.000	0.000
EstRépublicain	COORD	43/80=0.537	43/87=0.494	0.515
EstRépublicain	APPOS	4/12=0.333	4/20=0.200	0.250
EstRépublicain	JUXT	4/26=0.154	4/32=0.125	0.138

... Performance du ROVER sur le corpus de référence manuel cont.

Sous-corpus	Annotation	Précision	Rappel	F-mesure
EstRépublicain	ALL_RELATIONS	735/1261=0.583	734/1406=0.522	0.551
LeMonde	GN	3617/3816=0.948	3651/4070=0.897	0.922
LeMonde	NV	3173/3238=0.980	3185/3473=0.917	0.947
LeMonde	GA	1124/1297=0.867	1124/1342=0.838	0.852
LeMonde	GR	714/803=0.889	713/885=0.806	0.845
LeMonde	GP	4189/4284=0.978	4206/4595=0.915	0.946
LeMonde	PV	349/353=0.989	349/382=0.914	0.950
LeMonde	ANONYMOUS	3239/3899=0.831	3245/3610=0.899	0.863
LeMonde	ALL_CONSTITUENTS	16405/17690=0.927	16473/18357=0.897	0.912
LeMonde	SUJ-V	1472/2305=0.639	1474/2379=0.620	0.629
LeMonde	AUX-V	530/738=0.718	530/772=0.687	0.702
LeMonde	COD-V	836/1397=0.598	836/1591=0.525	0.560
LeMonde	CPL-V	1029/2047=0.503	1028/2043=0.503	0.503
LeMonde	MOD-V	415/765=0.542	415/957=0.434	0.482
LeMonde	COMP	177/389=0.455	177/380=0.466	0.460
LeMonde	ATB-SO	161/337=0.478	161/378=0.426	0.450
LeMonde	MOD-N	3378/5237=0.645	3377/6466=0.522	0.577
LeMonde	MOD-A	101/176=0.574	101/334=0.302	0.396
LeMonde	MOD-R	32/77=0.416	32/88=0.364	0.388
LeMonde	MOD-P	3/8=0.375	3/9=0.333	0.353
LeMonde	COORD	289/594=0.487	289/846=0.342	0.401
LeMonde	APPOS	39/131=0.298	39/219=0.178	0.223
LeMonde	JUXT	13/28=0.464	13/164=0.079	0.135
LeMonde	ALL_RELATIONS	8475/14229=0.596	8475/16626=0.510	0.549
Ester	GN	1215/1520=0.799	1237/1728=0.716	0.755
Ester	NV	1277/1401=0.911	1288/1412=0.912	0.912
Ester	GA	262/332=0.789	262/382=0.686	0.734
Ester	GR	561/644=0.871	561/736=0.762	0.813
Ester	GP	929/1067=0.871	960/1153=0.833	0.851
Ester	PV	122/136=0.897	125/143=0.874	0.885
Ester	ANONYMOUS	1013/1484=0.683	1017/1366=0.745	0.712
Ester	ALL_CONSTITUENTS	5379/6584=0.817	5450/6920=0.788	0.802
Ester	SUJ-V	738/1079=0.684	743/1263=0.588	0.633
Ester	AUX-V	162/229=0.707	162/233=0.695	0.701

... Performance du ROVER sur le corpus de référence manuel cont.

Sous-corpus	Annotation	Précision	Rappel	F-mesure
Ester	COD-V	348/596=0.584	349/868=0.402	0.476
Ester	CPL-V	274/652=0.420	276/711=0.388	0.404
Ester	MOD-V	289/706=0.409	289/663=0.436	0.422
Ester	COMP	111/233=0.476	112/276=0.406	0.438
Ester	ATB-SO	136/315=0.432	134/270=0.496	0.462
Ester	MOD-N	868/1518=0.572	868/1787=0.486	0.525
Ester	MOD-A	30/53=0.566	30/127=0.236	0.333
Ester	MOD-R	29/67=0.433	29/84=0.345	0.384
Ester	MOD-P	4/41=0.098	4/29=0.138	0.114
Ester	COORD	96/235=0.409	97/341=0.284	0.335
Ester	APPOS	3/21=0.143	3/60=0.050	0.074
Ester	JUXT	24/163=0.147	24/247=0.097	0.117
Ester	ALL_RELATIONS	3112/5908=0.527	3120/6959=0.448	0.484
Parl. Euro.	GN	4598/4984=0.923	4737/5076=0.933	0.928
Parl. Euro.	NV	5118/5173=0.989	5135/5382=0.954	0.971
Parl. Euro.	GA	2248/2388=0.941	2248/2460=0.914	0.927
Parl. Euro.	GR	1703/1834=0.929	1698/2071=0.820	0.871
Parl. Euro.	GP	5874/6217=0.945	5838/6159=0.948	0.946
Parl. Euro.	PV	727/727=1.000	727/797=0.912	0.954
Parl. Euro.	ANONYMOUS	4809/5523=0.871	4810/5177=0.929	0.899
Parl. Euro.	ALL_CONSTITUENTS	25077/26846=0.934	25193/27122=0.929	0.931
Parl. Euro.	SUJ-V	3096/4060=0.763	3097/3796=0.816	0.788
Parl. Euro.	AUX-V	662/766=0.864	662/784=0.844	0.854
Parl. Euro.	COD-V	2159/2947=0.733	2159/3326=0.649	0.688
Parl. Euro.	CPL-V	1725/3042=0.567	1727/2978=0.580	0.573
Parl. Euro.	MOD-V	1190/2138=0.557	1189/1836=0.648	0.599
Parl. Euro.	COMP	716/1185=0.604	716/1064=0.673	0.637
Parl. Euro.	ATB-SO	271/639=0.424	270/461=0.586	0.492
Parl. Euro.	MOD-N	5804/8113=0.715	5757/8178=0.704	0.710
Parl. Euro.	MOD-A	262/353=0.742	262/534=0.491	0.591
Parl. Euro.	MOD-R	99/177=0.559	99/234=0.423	0.482
Parl. Euro.	MOD-P	4/30=0.133	4/53=0.075	0.096
Parl. Euro.	COORD	797/1242=0.642	797/1570=0.508	0.567
Parl. Euro.	APPOS	8/66=0.121	8/113=0.071	0.089

... Performance du ROVER sur le corpus de référence manuel cont.

Sous-corpus	Annotation	Précision	Rappel	F-mesure
Parl. Euro.	JUXT	19/50=0.380	19/183=0.104	0.163
Parl. Euro.	ALL_RELATIONS	16812/24808=0.678	16766/25110=0.668	0.673
ALL	ALL_GRPS	51008/57066=0.894	51286/74907=0.685	0.775
ALL	ALL_RELS	30056/48672=0.618	30015/67462=0.445	0.517
ALL	ALL(G+R)	0.756	0.565	0.646

# Bibliography

- [1] de la Clergerie, E., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P., Vilnat, A. May 2008. Passage: from French parser evaluation to large sized treebank. ELRA, , (ed), *In proceedings of the sixth international conference on Language Resources and Evaluation (LREC)* Marrakech, Morocco.