Rapport sur corpus de référence manuelle

Livrable D8

Projet PASSAGE

Référence: ANR-06-MDCA-013-02

Échéance : T36 - 31 décembre 2009

Anne Vilnat[1], Olivier Hamon[2], Patrick Paroubek[1]
ELDA
55-57, rue Brillat-Savarin
75013 Paris
hamon@elda.org[2]
LIMSI-CNRS
Bât. 508 Université Paris XI
91403 Orsay Cedex
tél. 01 69 85 80 04
email [anne,pap]@limsi.fr[1]

Table des matières

Chapitre 1

Corpus de référence manuelle

1.1 Introduction

L'ensemble des données annotées manuellement [1] [2] dans le projet PASSAGE est constitué des corpus de référence de la première campagne (livrable D5) et de la seconde campagne, objet de ce livrable.

1.2 Corpus de référence de la première campagne

Nous rappelons ici que le corpus de la première campagne contenait un ensemble de genres variés incluant : des textes de journaux, des débats parlementaires, des romans, des courriers électronique, un sous-corpus de textes médicaux, des transcription d'oral issues de la campagne ESTER, et des questions des campagnes Amaryllis et TREC [1]. À l'exception des romans, la plupart de ces corpus ne sont pas distribuables librement, mais peuvent être obtenus auprès de la société ELDA. La table 1.3

Genre	Énoncés	Mots	Groupes	Relations
Le Monde	432	11577	5386	5816
parlementaire	346	9678	4601	5010
litteraire	1054	28140	12350	14704
courriel	765	8250	3757	4202
medical	591	12738	6303	6218
oral	518	8192	3939	4624
questions	231	4644	2255	2421
WEB	91	2466	1173	1215
total	4028	85685	39764	21232

Tab. 1.1 – Caractéristiques du corpus de référence de la première campagne

1.3 Corpus de référence de la seconde campagne

Le second corpus de référence est essentiellement constitué de matériaux en distribution libre, issus d'un corpus de 100 millions de mots de différents genres :

- **Wikipedia Fr**, un corpus disponible librement couvrant différents domaines de connaissances, œuvre collective contenant essentiellement des descriptions de style variés.
- **Wikisources**, une collection de plusieurs milliers de textes en français librement disponible abordant différents thèmes.
- **Wikilivres**, une collection de textes éducatifs en français.
- **Monde Diplomatique**, un corpus journalistique à faible coût déjà utilisé par beaucoup de chercheurs qui aborde des thématiques très diverses.

[1] Les questions de la campagne TREC ont été traduites en français

– **Est-Répulicain**, petit échantillon d'un journal différent du Monde pour tester la variabilité en genre des analyseurs.
– **Europarl**, un corpus de textes multilingue parallèle de textes extraits des actes du parlement européén.
– **JRC-Acquis Multilingual Parallel Corpus** les textes de loi de l'Union Européenne.
– **Ester**, des transcriptions d'oral de radio et télévision francophone issues des campagnes d'évaluation nationales.

À l'exception des sous-corpus du Monde Diplomatique et Ester qui peuvent être obtenus par l'intermédiaire de la société ELDA, les autres corpus sont librement disponibles.

La table suivante donne le détail des annotations manuelles qui ont été réalisées, d'abord par ELDA, puis pour une part résiduelle par le LIMSI.

ELDA								
corpus	EP	Ester	FrWiki	JRC	LeMonde	EstRépu	WkS	total
énoncés	2122	2597	3714	2557	2664	0	4093	17747
mots	57934	59796	59606	61151	60464	0	59810	358761
groupes	27039	31648	28032	28403	32242	0	26834	174198
relations	30963	45000	29190	37403	44886	0	26714	214156
LIMSI								
corpus	EP	Ester	FrWiki	JRC	LeMonde	EstRépu	WkS	total
énoncés	0	2269	0	657	7754	169	694	11543
mots	0	56928	0	29553	107344	3062	13297	210184
groupes	0	24818	0	12698	33693	1317	5860	78386
relations	0	29153	0	9333	37079	1406	6622	83593
total								
corpus	EP	Ester	FrWiki	JRC	LeMonde	EstRépu	WkS	total
énoncés	2122	4866	3714	3214	10418	169	4787	29290
mots	57934	116724	59606	90704	167808	3062	73107	568945
groupes	27039	56466	28032	41101	65935	1317	32694	252584
relations	30963	74153	29190	46736	81965	1406	33336	297749

Tab. 1.2 – Caractéristiques du corpus de référence de la seconde campagne

Les annotateurs ayant travaillé sur une version différente du corpus de celle qui a été soumise aux participants pour les tests, à cause d'une désynchronisation partielle entre la tâche de constitution et normalisation du corpus et la tâche d'annotation manuelle, seulement une partie du corpus annoté manuellement a pu être utilisée pour faire les mesures de performance (voir les détails dans le livrable D16). Cette partie représente à ce jour : 5,842 énoncés, 130,163 mots, 60,562 groupes et 67,462 relations.

1.4 Rapport des annotations réalisées à Elda dans le cadre de la seconde campagne PASSAGE

1.5 Déroulement

Les annotations se déroulent en trois parties bien distinctes.

Tout d'abord, des corpus comparatifs sont annotés afin d'observer les différences entre annotateurs, mais aussi pour constater l'évolution des annotations au cours du temps (*tâche comparative*). Ces cinq corpus sont de différentes natures et issus des données EuroParl (EP), Est-Républicain (EstRep), Wikipedia (FrWiki), JRC-acquis (JRC) et Wikisource (Wks). Chacun des corpus ayant leurs particularités propres, il est intéressant d'en étudier les variations de comportement selon les annotateurs.

Ensuite, deux nouveaux corpus sont annotés sans aucune manipulation automatique (*tâche des corpus non préannotés*) . Il s'agit de corpus extrait du journal Le Monde et de transcriptions du projet Ester.

Finalement, six corpus préannotés sont validés et annotés à partir des préannotations du ROVER (*tâche des corpus préannotés*). Les six corpus retenus sont les corpus Europarl, Ester, Wikipedia, Le Monde, JRC-Acquis et Wikisource.

Les annotateurs sont libres d'annoter les constituants et les relations selon leur bon vouloir. Autrement dit, il est aussi bien possible d'annoter les constituants avant les relations ou inversement, ou constituants et relations en même temps. D'une manière générale, la première possibilité a été retenue par tous les annotateurs, même si certains énoncés, plus longs, ont plutôt été abordés avec la seconde méthode.

1.6 Annotateurs

Trois annotateurs ont été sélectionnés pour la tâche d'annotation, notés *A1*, *A2* et *A3* par la suite. Leurs cursus sont relativement hétérogènes, puisque l'un était doctorant en linguistique computationnelle (A1), l'autre linguiste-terminologue (A2) et le dernier détenteur d'un master en ingénierie linguistique et traitement de la communication (A3). Les trois annotateurs ont participé à la tâche comparative, mais seuls les annotateurs A1 et A3 ont participé aux tâches des corpus non préannotés et des corpus préannotés.

Dans un premier temps, les annotateurs travaillent ensemble et en se concertant, dans une même pièce. N'étant pas forcément, dès au début de la tâche d'annotation, très expérimentés pour réaliser cette tâche, le but est d'homogénéiser leur compréhension de la tâche et d'améliorer leurs performances par l'interactivité qu'il découle de leurs communications. De par une certaine cohésion, il s'avère que le gain de temps n'est pas négligeable. Toujours est-il qu'au bout d'un certain laps de temps, nous avons expérimenté le télétravail, puisque les concertations n'avaient plus lieu d'être, du moins elles s'en sont trouvés fortement réduites du fait de l'expérience acquise par les annotateurs au fur et à mesure des annotations.

De plus, un expert du guide d'annotation était à même de répondre à leurs questions, généralement par mail ou au téléphone.

1.7 Corpus comparatifs

Les corpus comparatifs ont pour objectif de comparer les annotateurs entre eux, d'une part afin de mesurer le potentiel risque d'erreur, mais surtout afin de comparer leur vitesse d'annotation et observer, entres autres choses, le gain sur les corpus préannotés. Pour ce faire, un échantillon de taille conséquente est sélectionné et annoté par chacun des trois annotateurs. Ainsi, les données reccueillies au cours de l'annotation du corpus serviront à comparer l'activité des annotateurs à celle des corpus préannotés.

Comme il a été dit précédemment, les annotateurs travaillent ensemble lors de l'annotation de ce corpus. Le but de ce travail n'est pas de comparer leurs connaissances respectives (les corpus annotés devant être homogènes et de bonne qualité), mais d'observer le taux d'erreurs inhérentes aux annotations, pouvant être dues tant aux erreurs d'innattention qu'à la fatigue, la mauvaise compréhension d'un type d'annotation étant limitée par l'homogénéisation de leurs méthodes de travail.

Toutefois, et ce afin d'éviter un biais trop important des résultats, les annotateurs n'ont pas travaillé en même temps sur un même corpus. Le tableau 1.3 présente pour chaque annotateur l'ordre dans lequel les corpus sont annotés.

Ordre	A1	A2	A3
1	EP	Wks	JRC
2	EstRep	EP	Wks
3	FrWiki	EstRep	EP
4	JRC	FrWiki	EstRep
5	Wks	JRC	FrWiki

Tab. 1.3 – Ordre d'annotation des corpus pour chaque annotateur.

Les aléas du projet d'annotation (bugs et mises à jour de l'interface d'annotation, réception des données, etc.) font que l'ordre n'a pas forcément été respecté à la lettre.

1.7.1 Données

Pour comparer les annotateurs entre eux, cinq corpus ont été sélectionnés parmi les données du CPCV. Chaque corpus contient environ 4 000 tokens, le nombre d'énoncés variant selon le corpus. Au total, 20 000 tokens ont donc

été annotés trois fois par trois différents annotateurs. Les statistiques sur les corpus annotés sont contenues dans la table 1.4.

Corpus	Enoncés	Tokens	T/E
EP	86	3 966	46,12
EstRep	141	4 147	29,41
FrWiki	13	4 232	325,54
JRC	102	3 744	36,71
Wks	358	4 817	13,46
Total	700	20 906	29,87

TAB. 1.4 – Statistiques des corpus annotés.

Parmi les cinq corpus, le corpus FrWiki se distingue particulièrement des autres, de par le nombre de tokens par énoncés. En effet, la segmentation ne s'étant pas correctement déroulée, la taille des énoncés s'en est retrouvée fortement augmentée. Les annotations s'en retrouvent forcément influencées, notamment concernant la vitesse d'annotation sur les relations (celle sur les constituants n'est pas réellement liée à la taille des énoncés). A l'autre extrême, le corpus Wks contient peu de tokens par énoncé.

1.7.2 Vitesse moyenne d'annotation

L'annotation des 700 énoncés a pris au total entre 42 et 44 heures de travail pour les constituants et entre 80 et 90 heures de travail pour les relations. Les tableaux 1.5 et 1.6 présentent le nombre d'énoncés annotés par heure, respectivement pour les constituants et les relations. Il faut tenir compte de la période d'adaptation des annotateurs, qui plus est différente selon les corpus puisque l'ordre d'annotation n'est pas le même selon les annotateurs.

Corpus	A1	A2	A3	Moyenne
EP	5,52	5,24	17,73	9,50
EstRep	16,89	17,63	11,86	15,46
FrWiki	16,49	12	11,75	13,41
JRC	16,66	12,82	12,25	13,91
Wks	27,23	33,94	22,63	27,93
Moyenne	16,56	16,33	15,24	16,04

TAB. 1.5 – Vitesse moyenne d'annotation en constituants sur les corpus comparatifs (E/h).

Pour l'annotation sur constituants, les annotateurs ont été constants sur les corpus EstRep, FrWiki et JRC, tandis que le corpus EP a donné plus de difficultés (pour les annotateurs A1 et A2) et le corpus Wks plus de facilités. Pour le corpus EP, l'explication est double : d'une part les énoncés du corpus sont particulièrement bien construits, d'autre part les deux annotateurs A1 et A2 débutaient par ce corpus, ceci complexifiant la tâche. Cela montre, entre autres choses, la marge de progression entre un annotateur débutant et un plus expérimenté. Le corpus Wks est quant à lui constitué de phrases courtes, d'où une vitesse d'annotation plus importante que celle sur les autres corpus. Il faut également noter que l'annotateur A3 a débuté par ce corpus, expliquant que sa vitesse d'annotation se retrouve en-deçà de celle des deux autres annotateurs. Ici l'importance de l'expérience des annotateurs nous paraît fondamentale.

L'annotation sur les relations montre des différences avec celle sur les constituants, notamment si l'on tient compte de la taille des énoncés. En effet, le corpus FrWiki contenant des énoncés disproportionnés, les conséquences sont immédiates, et se traduisent par des taux d'annotation très faibles. D'autre part, les performances sur le corpus EP sont amoindries, du fait de la présence de nombreux enchassés dans les énoncés. Le corpus Wks est plus rapide à annoter car il contient des énoncés de plus courte taille que les autres corpus.

Enfin, les remarques concernant l'évolution des vitesses d'annotation sur les constituants en fonction de l'ordre d'annotation sont similaires pour els relations. Par exemple, l'annotateur A3 a terminé par le corpus JRC et sa vitesse d'annotation sur ce corpus est plus importante que celle des autres annotateurs.

Corpus	A1	A2	A3	Moyenne
EP	2,84	5,06	6,32	4,74
EstRep	7,97	9,42	8,34	8,58
FrWiki	1,47	1,26	1,26	1,33
JRC	5,89	5,52	10,77	7,39
Wks	25,86	18,63	9,50	18,00
Moyenne	8,81	7,98	7,24	8,01

TAB. 1.6 – Vitesse moyenne d'annotation en relations sur les corpus comparatifs (E/h).

Corpus	A1 vs A2	A1 vs A3	A2 vs A3
EP	94,66	96,06	96,80
EstRep	91,10	91,81	96,16
FrWiki	91,13	92,78	95,18
JRC	94,78	95,91	96,92
Wks	95,62	95,60	97,08
Moyenne	93,46	94,43	96,43

TAB. 1.7 – Accords inter-annotateurs en constituants.

1.7.3 Correlations inter-annotateurs

Dans un premier temps, les corpus annotés ont été évalués deux-à-deux en prenant successivement les corpus d'un premier annotateur comme référence et ceux d'un second annotateur comme hypothèse. C'est-à-dire que le corpus EP de l'annotateur A1 a été pris comme référence lors de l'évaluation du corpus EP de l'annotateur A2, et ainsi de suite pour toutes les combinaisons. Les résultats sont présentés dans le tableau 1.7 pour les constituants et dans le tableau 1.8 pour les relations.

En premier lieu, les annotateurs se retrouvent souvent en désaccord et sont parfois loin d'un accord à 100% comme ce peut être le cas pour les annotations sur relations. En effet, environ 6% des annotations sur constituants divergent, en moyenne, selon les annotateurs. Pour les relations, cette moyenne monte à 25%, ce qui signifie que le quart des annotations posent problème et nécessite une révision.

Ensuite, il apparaît à la vue des résultats que les annotateurs A2 et A3 sont plus en accord que l'un et l'autre avec l'annotateur A1. C'est légèrement visible pour les annotations sur constituants, et c'est clairement le cas pour les annotations sur relations. Il est possible que cela soit dû à la méthode de travail (les annotateurs A2 et A3 ayant peut-être plus eu tendance à travailler en collaboration) ou tout simplement au cursus des annotateurs. Toutefois, une conclusion ne peut être effective qu'en observant les points de divergence et les erreurs ayant pu être commises.

Enfin, les résultats sont assez disparates selon les corpus, avec des différences pouvant varier jusqu'à 4% pour les constituants, mais surtout proche de 20% pour les relations sur certains corpus. Cela montre en particulier la difficulté d'annoter certains types de corpus. C'est en particulier le cas pour le corpus FrWiki qui, comme il en a déjà été fait mention, pose des problèmes au niveau du format, entraînant un sucroît de travail (et probablement de fatigue). Toutefois, l'accord inter-annotateurs ne semble ni lié à la vitesse d'annotation, ni à la proportion de tokens par énoncé, mis à part peut-être dans ce cas extrême. Il faut également noter que les accords inter-annotateurs selon

Corpus	A1 vs A2	A1 vs A3	A2 vs A3
EP	79,66	82,24	84,71
EstRep	66,61	66,06	78,59
FrWiki	60,29	62,52	78,96
JRC	77,83	77,46	84,82
Wks	76,83	77,90	82,03
Moyenne	72,44	72,24	81,82

TAB. 1.8 – Accords inter-annotateurs en relations.

	EP	EstR.	FrW.	JRC	Wks	$\bar{x}$
GN	95,5	90,7	93,1	94,8	96,3	93,1
	400±9	611±9	894±5	379±12	703±2	597±
NV	98,3	97,9	96,2	98,2	98,2	98,0
	446±5	251±3	132±5	392±4	624±5	369±
GA	95,0	90,8	88,6	94,7	87,5	92,2
	221±11	109±4	164±15	223±14	96±4	163±
GR	93,1	95,2	91,7	87,8	92,5	92,5
	129±8	62±3	8±1	49±2	208±9	91±
CP	95,1	93,3	93,7	96,0	95,9	94,8
	584±7	500±10	464±5	613±3	293±2	491
PV	96,2	100	93,9	97,9	98,5	97,6
	79±1	25±0	6±1	31±0	67±0	42±
$\sum$	95,8	93,0	93,0	95,9	96,1	94,9

TAB. 1.9 – Accords inter-annotateurs et détaillé par type de constituant. En haut : moyenne des accords ; En bas : moyenne des constituants.

les constituants et les relations semblent répondre de la même manière.

Les tableaux 1.9 et 1.10 présentent les résultats détaillés par type de constituants et de relations, en moyenne et indifféremment des comparaison entre annotateurs.

Les résultats détaillés montrent à la fois des inconsistances selon les types de constituants/relations et selon les corpus.

Pour les constituants, les accords sont relativement homogènes, mêmes si les GA, GR ou GN sont un peu en-deçà des performances. Les GA sont particulièrement bas avec les corpus FrWiki et Wks, les GR avec le corpus JRC. De même, les accords varient selon les corpus, en étant plus bas avec les corpus EstRep ou FrWiki.

Pour les relations, les variations sont plus marquées. Les relations Mod_N, MOD_A, MOD_R, MOD_P, Coord, Appos et Juxt obtiennent, sans surprise, des accords de 100%. Autrement, seul la relation Aux_V ressort avec plus de 94% d'accord en moyenne. Le reste des relations pose plus de problème, notamment les ATB_SO et MOD_V, atteignant même 43% d'accord pour le corpus FrWiki, tout comme sur les COMP (49%). Les relations SUJ_V, COD_V, voire CPL_V obtiennent légèrement un meilleur accord, même s'il reste loin des espérances. Concernant les corpus, le corpus FrWiki est le plus pénalisé, ce qui est justifié par son formatage. Les corpus EstRep n'est pas en reste, les accords baissant entre autre par les résultats faible en MOD_V, COMP et surtout ATB_SO. Compte-tenu de la moyenne, les autres corpus ont des accords corrects, mais insuffisants.

1.8 Corpus non préannotés

Les corpus non préannotés ont pour but de servir lors de l'évaluation des systèmes automatique de la seconde campagne PASSAGE. Ainsi, les données ne sont pas connues des systèmes avant l'évaluation elle-même (au contraire de ceux de la tâche sur corpus préannotés que les systèmes annotent au préalable, avant d'être vérifiés et corrigés par les annotateurs).

Au préalable, une estimation des délais d'annotation a été réalisée à partir du corpus comparatif, à savoir une moyenne de 16 énoncés par heure pour les annotations sur constituants et 8 énoncés par heure pour les annotations sur relations.

1.8.1 Données

Les corpus non préannotés sont au nombre de deux et constitués d'articles du journal Le Monde 2007 (la journée du 25 janvier 2007) et de transcriptions audio de la campagne d'évaluation ESTER (France Inter entre 8h et 9h le 7 octobre 2004 et France Info entre 18h et 18h30 le 8 octobre 2004). Ils contiennent respectivement 751 énoncés pour 28 139 tokens (environ 37,47 tokens par énoncé) et 854 énoncés pour 22 567 tokens (environ 26,43 tokens par énoncé). Chaque annotateur a annoté environ un tiers de chaque corpus.

	EP	EstR.	FrW.	JRC	Wks	$\bar{x}$
Suj_V	89,4 ±	93,2 ±	80,4 ±	85,4 ±	92,3 ±	89,6
Aux_V	96,2 ±	93,7 ±	96,5 ±	96,2 ±	89,9 ±	94,3
Cod_V	89,1 ±	85,5 ±	91,0 ±	93,2 ±	86,9 ±	88,7
Cpl_V	79,9 ±	80,1 ±	78,9 ±	79,7 ±	79,4 ±	79,7
Mod_V	78,1 ±	70,0 ±	43,3 ±	63,8 ±	78,1 ±	71,4
Comp	78,5 ±	73,9 ±	49,2 ±	69,9 ±	83,9 ±	75,2
Atb_So	65,0 ±	57,9 ±	73,8 ±	70,2 ±	62,5 ±	65,1
Mod_N	100 ±	100 ±	100 ±	100 ±	100 ±	100
Mod_A	100 ±	100 ±	100 ±	100 ±	100 ±	100
Mod_R	100 ±	100 ±	100 ±	100 ±	100 ±	100
Mod_P	100 ±	100 ±	100 ±	100 ±	100 ±	100
Coord	100 ±	100 ±	100 ±	100 ±	100 ±	100
Appos	100 ±	100 ±	100 ±	100 ±	100 ±	100
Juxt	100 ±	100 ±	100 ±	100 ±	100 ±	100
$\sum$	82,2	70,4	67,3	80,0	78,9	76,0

TAB. 1.10 – Accords inter-annotateurs détaillé par type de relation. En haut : moyenne des accords ; En bas : moyenne des constituants.

1.8.2 Vitesse moyenne d'annotation

Corpus Le Monde

L'annotation des 751 énoncés du corpus Le Monde a pris environ 46 heures pour les constituants et environ 92 heures pour les relations. Les tableaux 1.11 et 1.12 présentent le détail des annotations.

Annotateur	Temps (h)	Enoncés	E/h
A1	13	207	15,92
A2	20	305	15
A3	13	239	18,38
Tous	46	751	16,43

Tab. 1.11 – Annotations Le Monde pour les constituants.

Annotateur	Temps (h)	Enoncés	E/h
A1	29,5	207	7,02
A2	35,5	305	8,59
A3	27	239	8,85
Tous	92	751	8,15

Tab. 1.12 – Annotations Le Monde pour les relations.

Pour les constituants, les taux moyen d'annotations par heure sont conformes aux estimations planifiées, soit environ 16 E/h pour les constituants et 8 E/h pour les relations. Il y a toutefois, ici encore, des variations relativement importantes en fonction des annotateurs.

Corpus ESTER

L'annotation des 854 énoncés du corpus ESTER a pris environ 53 heures pour les constituants et environ 89 heures pour les relations. Les tableaux 1.13 et 1.14 présentent le détail des annotations.

Annotateur	Temps (h)	Enoncés	E/h
A1	20,5	306	14,93
A2	15	273	18,2
A3	17,25	273	15,83
Tous	52,75	854	16,32

Tab. 1.13 – Annotations ESTER pour les constituants.

Les taux d'annotations sur le corpus Ester restent fidèles aux estimations, voire légèrement supérieurs, avec 16E/h pour les constituants et 10E/h pour les relations.

Comparaison

En comparant les vitesses d'annotation des deux corpus non préannotés, il apparaît, avec surprise, que la vitesse d'annotation est plus importante sur le corpus Ester pour les relations, tandis qu'elle reste la même pour les constituants. Par contre, il faut signaler que les vitesses d'annotation sont plus homogènes sur le corpus Le Monde (écart type de 1,71 sur les constituants et 0,99 sur les relations) que sur le corpus Ester (écart type de 1,69 sur les constituants et 2,39 sur les relations), tant pour les constituants que pour les relations. Cela est sans doute dû à la plus grande variabilité de la forme des énoncés. Enfin, il est amusant de noter que cela ne corrobore pas avec le sentiment des annotateurs, trouvant l'annotation sur le corpus Ester beaucoup plus fastidieuse que celle sur le corpus Le Monde.

Annotateur	Temps (h)	Enoncés	E/h
A1	38	306	8,05
A2	21,5	273	12,7
A3	29	273	9,41
Tous	88,5	854	10,05

Tab. 1.14 – Annotations ESTER pour les relations.

1.9 Corpus préannoté

L'utilisation du corpus préannoté a pour objectif d'améliorer la vitesse d'annotation sur corpus en validant manuellement des données issues de ROVER, c'est-à-dire provenant de combinaison de systèmes. Cela permet notamment d'annoter une plus grande quantité de données, et d'en réutiliser les informations linguistiques dans les analyseurs.

Dans notre cas, l'annotation sur corpus préannoté nous permet aussi de déterminer l'efficacité d'une telle méthode en la comparant à une annotation directe (*from scratch*). Les annotations réalisées sur les corpus comparatifs décrites plus haut nous permettent de faire une telle comparaison. Le taux d'erreurs produites par le ROVER est également à prendre en compte, ce que nous ne manquerons pas de décrire par la validation de corpus pris à différentes étapes de développement du ROVER, avec une qualité *théorique* croissante.

1.9.1 Données

Six corpus ont été utilisés pour cette tâche, sélectionnés parmi les corpus de la *track Passage* de la première campagne d'évaluation, décris dans le tableau 1.15. Chacun des corpus contient environ 60 000 tokens.

Corpus	Enoncés	Tokens	T/E
EP	2 260	58 499	25,88
Ester	2 560	59 796	23,36
FrWiki	3 980	59 868	15,04
LeMonde	2 630	60 464	22,99
JRC	2 579	61 151	23,71
Wks	5 320	59 998	11,28
Total	19 329	359 776	18,61

Tab. 1.15 – Statistiques des corpus préannotés.

Les corpus sont quelques peu différents des corpus comparatifs. Mis à part leur taille beaucoup plus importante, le nombre de tokens par énoncés est dans tous les cas plus faible, de manière plus ou moins importante selon les corpus. Si la proportion diminue faiblement pour le corpus Wks (-17%), elle est conséquente pour les corpus EP (-44%) et JRC (-36%). Pour le corpus FrWiki, la perte est très importantes (-96%), ceci étant dû aux problèmes rencontrés pour la segmentation de ce corpus et déjà explicités précédemment. En théorie, cette baisse d'ensemble doit augmenter la vitesse d'annotation, sans même tenir compte du gain apporté par le ROVER.

L'ensemble des corpus représente une taille totale de 360 000 tokens et près de 20 000 énoncés, dont les constituants et les relations ont été préannotées par le ROVER[2].

1.9.2 Procédure et déroulement

Pour chacun des six corpus, les données sont annotées de manière séquentielle en constituants, puis en relations. Techniquement parlant, les corpus sont scindés en sous-corpus contenant moins de 100 énoncés, afin de ne pas surcharger l'utilisation d'EasyRef.

Concernant la validation des préannotations en relations, deux méthodes différentes ont été abordées, à savoir en cherchant à valider les relations pour chaque énoncé, ou bien de supprimer les relations lorsque les modifications

[2]Le détail des fichiers utilisés est donné en Annexe A

sur les constituants ont été trop importantes puis d'annoter directement. Il apparaît que la seconde méthode est plus adaptée pour le moment compte-tenu de la qualité des préannotations en relation.

+ Distribution des corpus avec quelle version du ROVER (avec distinction constituants / relations le cas échéant)

1.9.3 Vitesse moyenne d'annotation

L'objectif principal des annotations sur corpus préannoté est d'observer les différentes vitesses d'annotation d'après des paramètres divers. Une première étape consiste à le faire d'une manière générale sur l'ensemble des corpus. Dans un second temps, nous observons les différences d'annotation selon le lieu de travail, à domicile ou sur place.

Résultats généraux

Les tableaux 1.16 et 1.17 détaillent les vitesses d'annotation sur les corpus préannotés, respectivement en constituants et en relations.

maj 29/12/08

Corpus	Enoncés	Temps	E/h
EP	2 260	67,25	33,61
Ester	N/A	N/A	N/A
FrWiki	2 319	28,25	82,09
LeMonde	N/A	N/A	N/A
JRC	2 534	79,55	31,85
Wks	N/A	N/A	N/A
Moyenne			

TAB. 1.16 – Vitesse d'annotation sur corpus préannoté en constituants.

Corpus	Enoncés	Temps	E/h
EP	2 260	222,75	10,15
Ester	N/A	N/A	N/A
FrWiki	1052	35	30,06
LeMonde	N/A	N/A	N/A
JRC	N/A	N/A	N/A
Wks	N/A	N/A	N/A
Moyenne			

TAB. 1.17 – Vitesse d'annotation sur corpus préannoté en relations.

En attente de plus de données

Différents lieux d'annotation

Au cours de notre étude, les annotateurs ont progressivement réalisé le travail d'annotation à domicile. Il y a de multiples vertues à procéder de cette manière. Il faut tout d'abord considérer le travail fastidieux qu'ont à faire les annotateurs, qui prend, la plupart du temps, environ huit heures par jour. La fatigue pouvant influer sur le résultats des annotations, il faut tenir des pauses indispensables à leur bonne réalisation. Le simple fait d'être en télétravail permet aussi aux annotateurs d'organiser leur journée comme ils le souhaitent, en optimisant leurs activités, et par la même occasion les vitesses d'annotation. Cela comprend également le temps relatif au transport, pouvant parfois être assez long. Par ailleurs, cette méthode de travail ne perturbe en rien le déroulement des annotations. Le passage au télétravail s'est fait de façon progressive : si les premiers jours les questions et les débats entre les annotateurs étaient nombreux, cela n'était plus que mineur au bout de quelques semaines ; le recours aux e-mails étaient plus que suffisant. Toutefois, nous avons souhaité garder quelques jours de travail sur place, afin d'entretenir ces débats et d'assurer une bonne cohésion entre les annotateurs.

Le tableau 1.18 fait état des différentes vitesses d'annotation relatives au télétravail et au travail sur place. *maj 29/12/08*

Corpus	Consituants		Relations	
	Elda	Télé	Elda	Télé
EP	27,74	36,44	9,82	10,79
Ester	N/A	N/A	N/A	N/A
FrWiki	N/A	N/A	N/A	N/A
LeMonde	N/A	N/A	N/A	N/A
JRC	31,82	31,69	N/A	N/A
Wks	N/A	N/A	N/A	N/A
Moyenne				

TAB. 1.18 – Comparaison des vitesses d'annotation entre le télétravail et le travail sur place.

L'annotation en constituant du corpus JRC s'est partagée en deux, pour le corpus EP (constituants et relations) 2 jours sur 5 étaient réservé au travail sur place, descendu à 1 jour sur 4 pour les corpus FrWiki et Wks.

Même si les moyenne en télétravail sur le corpus EP sont légèrement plus faibles qu'en travail sur place, rien ne nous permet d'affirmer qu'il y a un réel gain de temps. La tendance est la même quel que soit l'annotateur.

Bibliographie

[1] Paroubek, P., de la Clergerie, E., Loiseau, S., Vilnat, A., Francopoulo, G. January 2009. The passage syntactic representation. *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)* Groningen, The Netherlands. 91–102. http ://lotos.library.uu.nl/publish/issues/12/index.html.

[2] Vilnat, A., Francopoulo, G., Hamon, O., Loiseau, S., Paroubek, P., de la Clergerie, E. V. August 2008. Large scale production of syntactic annotation to move forward. *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation (PE 2008) in conjuction with COLING* Manchester. pp 36–43.