Les résultats de la campagne EASY d'évaluation des analyseurs syntaxiques du français

Patrick PAROUBEK[1], Anne VILNAT[1], Isabelle ROBBA[1], Christelle AYACHE[2]
[1] LIMSI-CNRS Bât. 508 Université Paris XI, BP 133 - 91403 ORSAY Cedex
[2] ELRA-ELDA 55-57, rue Brillat Savarin 75013 Paris
{pap,anne,isabelle}@limsi.frayache@elda.fr

Résumé. Dans cet article, nous présentons les résultats de la campagne d'évaluation EASY des analyseurs syntaxiques du français. EASY a été la toute première campagne d'évaluation comparative des analyseurs syntaxiques du français en mode boîte noire utilisant des mesures objectives quantitatives. EASY fait partie du programme TECHNOLANGUE du Ministère délégué à la Recherche et à l'Éducation, avec le soutien du ministère de délégué à l'industrie et du ministère de la culture et de la communication. Nous exposons tout d'abord la position de la campagne par rapport aux autres projets d'évaluation en analyse syntaxique, puis nous présentos son déroulement, et donnons les résultats des 15 analyseurs participants en fonction des différents types de corpus et des différentes annotations (constituants et relations). Nous proposons ensuite un ensemble de leçons à tirer de cette campagne, en particulier à propos du protocole d'évaluation, de la définition de la segmentation en unités linguistiques, du formalisme et des activités d'annotation, des critères de qualité des données, des annotations et des résultats, et finalement de la notion de référence en analyse syntaxique. Nous concluons en présentant comment les résultats d'EASY se prolongent dans le projet PASSAGE (ANR-06-MDCA-013) qui vient de débuter et dont l'objectif est d'étiqueter un grand corpus par plusieurs analyseurs en les combinant selon des paramètres issus de l'évaluation.

Abstract. In this paper, we present the results of the EASY evaluation campaign on parsers of French. EASY has been the very first black-box comparative evaluation campaign for parsers of French, with objective quantitative performance measures. EASY was part of the TECHNOLANGUE program of the Delegate Ministry of Research, jointly supported by the Delegate Ministry of Industry and the ministry of Culture and Communication. After setting EASY in the context of parsing evaluation and giving an account of the campaign, we present the results obtained by 15 parsers according to syntactic relation and subcorpus genre. Then we propose some lessons to draw from this campaign, in particular about the evaluation protocole, the segmenting into linguistic units, the formalism and the annotation activities, the quality criteria to apply for data, annotations and results and finally about the notion of reference for parsing. We conclude by showing how EASY results extend through the PASSAGE project (ANR-06-MDCA-013), which has just started and whose aim is the automatic annotation of a large corpus by several parsers, the combination of which being parametrized by results stemming from evaluation.

Mots-clés : analyseur syntaxique, évaluation, français.

Keywords: parser, evaluation, french.

1 L'évaluation des analyseurs syntaxiques

Les premières tentatives d'évaluation des analyseurs ont été le fait d'experts qui fondaient leur appréciation d'un analyseur sur les observations qu'ils avaient faites de ses sorties sur différentes phrases de test, parfois aidés d'une grille d'analyse (Blache & Morin, 2003). Pour le français, à notre connaissance la première tentative d'évaluation comparative a été faite par A. Abeillé (Abeillé, 1991). Dans le souci de réduire la part de subjectivité dans le processus d'évaluation et pour réutiliser les connaissances acquises lors d'une évaluation, les chercheurs se sont ensuite tournés vers des jeux de test prédéfinis, dont TSNLP (Oepen *et al.*, 1996), qui contient des exemples d'analyses correctes et erronées classés par type de constructions linguistiques, est un archétype. Cependant les jeux de test ne peuvent pas rendre compte de la distribution des phénomènes dans un corpus. De plus leur utilité à des fin d'évaluation dans des campagnes ouvertes est limitée dès lors qu'ils sont rendus publics. En effet, il sont de petite taille et paramétrer un analyseur en fonction d'un jeu de test donné devient alors une tâche aisée.
Avec le développement conjoint des standards pour les méta-données et des capacités des ordinateurs, nous avons vu appraître les corpus arborés (*treebanks*), dont le plus célèbre est certainement le Penn Treebank (Marcus *et al.*, 1993). Depuis sa création de nombreux développements pour différents formalismes et pour différentes langues ont vu le jour, dont certains pour le français (Brant *et al.*, 2002) (Abeillé *et al.*, 2000). Cependant, si les corpus arborés peuvent apporter un élément de réponse en ce qui concerne la représentativité des différents genres de texte et la distribution des phénomènes linguistiques, ils n'apportent pas de réponse au problème du formalisme pivot, pour lequel il n'existe à ce jour aucun standard [1].

Comparer des analyseurs implique donc de pouvoir projeter leurs annotations dans une représentation unique, ce qui en général ne peut se faire sans perte d'information. Pour résoudre ce problème, certains (Gaizauskas *et al.*, 1998) proposent de définir une fonction entre systèmes d'annotation, d'autres de tenir compte de la quantité d'information (Musillo & Sima'an, 2002) (méthode qui a le désavantage de nécessiter la construction d'un corpus parallèle par formalisme d'annotation), d'autres encore proposent d'utiliser des mécanismes d'apprentissage grammatical ou des mesures basées sur la distance d'édition (Roark, 2002). En remontant un peu plus dans le passé, (Black *et al.*, 1991) fut le premier à proposer une mesure d'évaluation fondée sur les limites des constituants pour comparer les analyseurs en mesurant le taux de croisement des frontières avec les annotations de référence (*crossing brackets*) et le rappel. En ajoutant la précision aux deux mesures précédentes, on obtient le protocole GEIG (Grammar Evaluation Interest Group) (Srinivas *et al.*, 1996), ou mesures PARSEVAL (Carroll *et al.*, 2002). Cependant ces mesures ont été appliquées uniquement sur des constituants non étiquetés, car il était impossible alors de définir un jeu d'étiquettes commun (Black *et al.*, 1991).

À part quelques tentatives ponctuelles, de comparaisons d'analyseurs syntaxiques, comme celle du projet SPARKLE qui a comparé des analyseurs syntaxiques pour déterminer le plus approprié pour une tâche d'extraction terminologique, ou encore les expériences développées récemment sur des transcriptions orales (Roark *et al.*, 2006), le paradigme d'évaluation n'a jusqu'à présent pas été appliqué à l'analyse syntaxique sur une grande échelle, à l'exception du projet EASY (Vilnat *et al.*, 2004) (Paroubek *et al.*, 2005) qui concerne les analyseurs du français.

[1] Une proposition est en cours d'élaboration à l'ISO.

2 La campagne EASY

La campagne EASY était une des 8 campagnes d'évaluation des technologies de la langue du projet EVALDA du programme TECHNOLANGUE (décembre 2002 - avril 2006). Dans cette campagne, 15 analyseurs provenant de 13 participants différents : ERSS, FT R&D, INRIA, LATL, LIC2M, LIRMM, LORIA, LPL, STIM, SYNAPSE, SYSTAL, TAGMATICA, VALORIA et XRCE ont été évalués sur les données fournies par les 5 fournisseurs de corpus que sont l'ATILF, le LLF, le DELIC, le STIM et ELDA. La tâche des fournisseurs de corpus a consisté en la collecte du corpus de différents genres de textes et en leur annotation. Le rapport entre la portion de texte annoté et la taille totale du corpus est choisie de manière à décourager une annotation manuelle de l'intégralité du corpus. Le corpus contient des articles de journaux (*Le Monde*), des textes littéraires (issus de la base *Frantext* de l'ATILF), des textes médicaux (pathologies et traitements), des questions (issues de la campagne EQUER de TECHNOLANGUE), des transcriptions de débats parlementaires (Sénat français et Parlement Européen), des pages WEB du site ELDA, des courriers électroniques et des transcriptions de parole[2]. On pourra trouver dans le tableau 4 plus loin dans l'article, les tailles respectives de ces différents corpus.

Le protocole d'évaluation EASY suppose que tous les participants adoptent la même segmentation en mots et en énoncés (voir (Roark, 2002) pour les problèmes que cela pose). Le formalisme inspiré de (Carroll *et al.*, 2002) et défini en collaboration avec les participants doit permettre d'exprimer l'essentiel d'une annotation syntaxique quelle que soit son type (de surface ou profonde, complète ou partielle), ceci sans privilégier une approche particulière. Le formalisme d'annotation EASY permet d'annoter des constituants continus et non-récursifs ainsi que des relations représentant les fonctions syntaxiques. Les relations (binaires pour la plupart ou ternaires) peuvent associer indifféremment des formes individuelles ou des constituants. Notons, qu'EASY ne connait pas la notion de *tête* lexicale (Gendner *et al.*, 2003) (Vilnat *et al.*, 2004).

Dans EASY, il y a 6 types de constituants : (1) nominal, (2) adjectival, (3) prépositionnel, (4) adverbial, (5) verbal et (6) prépositionnel-verbal, le dernier étant utilisé pour les verbes à l'infinitif introduits par une préposition, et 14 types de relations de dépendance : (1) sujet-verbe, (2) auxilliaire-verbe, (3) c-o-d, (4) complément-verbe, (5) modifieur de non, (6) modifieur de verbe, (7) modifieur d'adjectif, (8) modifieur d'adverbe, (9) modifieur de préposition, (10) complémenteur, (11) attribut du sujet/objet, (12) coordination, (13) apposition, (14) juxtaposition. Le choix de ces constituants et de ces relations a été fait à la suite de discussions avec l'ensemble des participants à la campagne. Il a ensuite fait l'objet d'une description plus d'étaillée à la fois pour les participants et pour les annotateurs dans un guide[3], Ils sont également décrits dans (Vilnat *et al.*, 2004). La figure 1 donne un exemple d'annotation d'une phrase issue du corpus littéraire.

Pour comparer les résultats des différents analyseurs, les mesures d'évaluation sont la précision et le rappel (ainsi que la f-mesure qui les combine) sur lesquelles nous avons expérimenté 15 relâchements de contrainte différents (Paroubek *et al.*, 2006), obtenus en combinant les 5 manières présentées dans la table 1 de comparer les empans de textes correspondant soit aux constituants soit aux cibles de relations, avec les 3 façons de considérer les définitions des constituants (ceux de l'hypothèse, ceux de la référence, ou ceux de l'hypothèse lorsqu'ils existent sinon ceux de la

[2]Les transcriptions d'émission radio-télévisées fournies par le projet ESTER de TECHNOLANGUE sur l'évaluation de la transcription de parole automatique n'ont finalement pas été prises en compte dans le calcul des performances en raison d'un problème dans la segementation des énoncés.

[3]Le guide d'annotation est disponible à l'URL www.limsi.fr/Recherche/CORVAL/easy

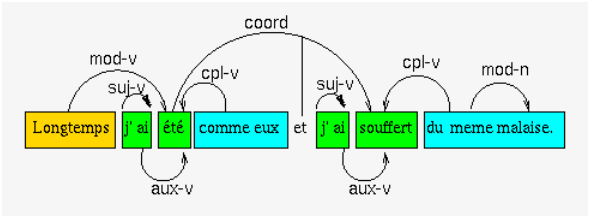

FIG. 1 – Exemple d'annotation d'un énoncé extrait du corpus littéraire (Coppé).

référence). L'évaluation a été menée indépendamment sur les constituants et les relations. Les résultats ont été calculés individuellement pour chaque constituant, chaque relation et chaque type de corpus ainsi que de manière globale.

Fonction	Formule
ÉGALITÉ	$H = R$
FLOU UNITAIRE	$\|H \backslash R\| \leq 1$
INCLUSION	$H \subset R$
INTERSECTION	$R \cap H \neq \emptyset$
BARYCENTRE	$\frac{2*\|R \cap H\|}{\|R\|+\|H\|} > 0.25$

avec :
H Empan de texte hypothèse
et
R Empan de texte référence,

TAB. 1 – Comparaison des empans correspondant aux constituants et aux cibles des relations.

3 Les résultats de la campagne EASY

Dans tout cette partie qui illustre les résultats des participants, nous ne donnerons pas directementleurs noms, nous y ferons référence par le biais de noms *anonymisés* Pi. Notre but n'est pas de donner un classement de ces participants mais d'indiquer les performances obtenues, ainsi que les écarts observés entre ces performances dans les différents domaines de l'évaluation.

3.1 Les mesures sur les constituants

Pour les constituants c'est le système P10 qui obtient les meilleurs résultats pour les 3 mesures (précision, rappel, f-mesure), tous constituants et tous genres de corpus confondus avec la comparaison barycentre pour les empans de texte des constituants (voir table 1 pour la définition de ces notions). La figure 2 illustre les résultats obtenus par ce participant, avec les différents corpus et les constituants annotés sur le plan horizontal (respectivement axe des x et des y) et la performance calculée en vertical (axe des z). Le graphe de gauche correspond à une vue avant, celui de droite à une vue arrière, comme l'illustrent les petits schémas au-dessus des graphes..

Nous avons utilisé la mesure barycentrique, car c'est celle qui, tout en permettant un certain relâchement des contraintes imposées sur les frontières de constituant (qui sont parfois le résultat d'un choix arbitraire), sans toutefois être aussi laxiste que l'intersection (où il suffit qu'un seul mot soit partagé).

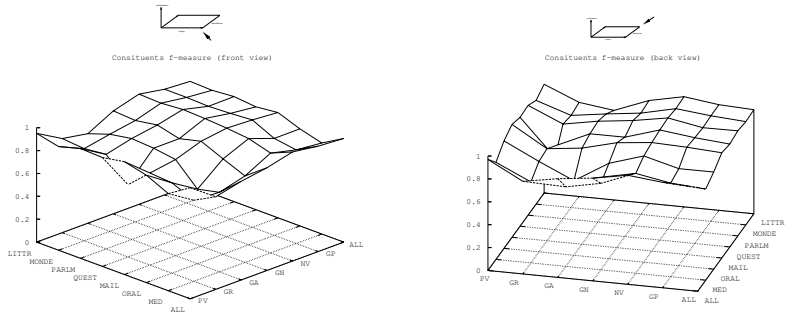

FIG. 2 – Vue avant et arrière sur les performances en f-mesure de P10 pour les constituants.

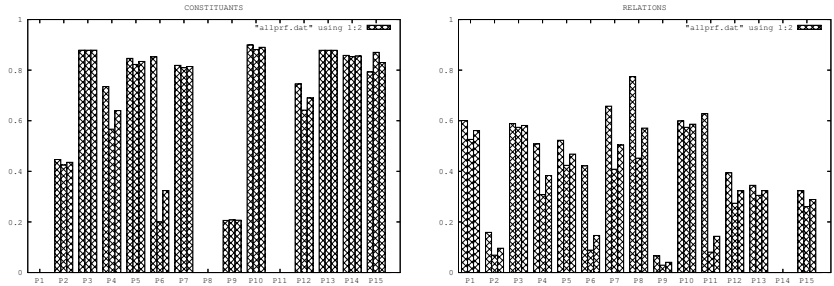

FIG. 3 – Résultats des 15 analyseurs pour les constituants (à gauche) et les relations (à droite) en précision/rappel/f-mesure, tous corpus et toutes annotations confondus

La table 2 donne les résultats de tous les analyseurs par type de corpus pour tous les constituants, en précision et f-mesure, pour distinguer les analyseurs visant à la correction de ceux visant à l'exhaustivité. Comme nous pouvions nous y attendre, les mesures de performance sur les constituants s'apparentent fortement aux types de résultat que l'on obtient avec un simple étiquetage morpho-syntaxique, les problémes étant assez similaires. Le profil des résultats est assez plat et dépend peu du type d'annotation ou du type de corpus traité, au contraire de ce qui se passe pour les relations comme nous le verrons plus loin.

La figure 3 illustre les résultats des différents analyseurs en combinant tous les corpus et toutes les annotations, à la fois en précision, rappel et f-mesure. Sur la figure de gauche, on peut observer 12 colonnes, car trois participants n'ont pas fourni de résultats pour les annotations en constituants mais uniquement l'annotation des relations de dépendance. De même sur la figure de droite, on voit que l'un des participants n'a pas fourni d'annotation en relations de dépendance.

	lemonde	littéraire	médical	oral_delic	parlement	questions	web
P1	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P2	p=0.717 f=0.690	p=0.329 f=0.320	p=0.332 f=0.312	p=0.612 f=0.591	p=0.702 f=0.644	p=0.395 f=0.373	p=0.719 f=0.679
P3	p=0.920 f=0.926	p=0.901 f=0.912	p=0.907 f=0.913	p=0.752 f=0.760	p=0.923 f=0.930	p=0.931 f=0.935	p=0 f=0
P4	p=0.813 f=0.660	p=0.802 f=0.770	p=0.459 f=0.436	p=0.787 f=0.717	p=0.808 f=0.653	p=0.877 f=0.856	p=0.841 f=0.696
P5	p=0.883 f=0.878	p=0.847 f=0.824	p=0.882 f=0.873	p=0.714 f=0.713	p=0.876 f=0.868	p=0.901 f=0.894	p=0.877 f=0.880
P6	p=0.837 f=0.782	p=0 f=0	p=0 f=0	p=0 f=0	p=0.849 f=0.803	p=0 f=0	p=0.903 f=0.893
P7	p=0.832 f=0.832	p=0.838 f=0.845	p=0.825 f=0.805	p=0.784 f=0.743	p=0.833 f=0.831	p=0.826 f=0.822	p=0.739 f=0.734
P8	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P9	p=0.141 f=0.137	p=0.145 f=0.152	p=0.191 f=0.183	p=0.336 f=0.334	p=0.175 f=0.159	p=0.305 f=0.301	p=0.856 f=0.866
P10	p=0.904 f=0.904	p=0.910 f=0.909	p=0.909 f=0.902	p=0.849 f=0.794	p=0.921 f=0.917	p=0.913 f=0.902	p=0.924 f=0.922
P11	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P12	p=0.737 f=0.685	p=0.714 f=0.681	p=0.806 f=0.733	p=0.605 f=0.562	p=0.712 f=0.649	p=0.832 f=0.767	p=0.801 f=0.749
P13	p=0.888 f=0.884	p=0.901 f=0.910	p=0.903 f=0.892	p=0.803 f=0.763	p=0.907 f=0.909	p=0.910 f=0.903	p=0.913 f=0.911
P14	p=0.855 f=0.855	p=0.887 f=0.895	p=0.879 f=0.869	p=0.775 f=0.731	p=0.867 f=0.867	p=0.873 f=0.866	p=0.879 f=0.875
P15	p=0.802 f=0.836	p=0.795 f=0.839	p=0.835 f=0.870	p=0.770 f=0.747	p=0.835 f=0.868	p=0.860 f=0.878	p=0.808 f=0.843

TAB. 2 – Mesures en précision (p) et f-mesure (f) par type de corpus pour tous les constituants

3.2 Les mesures sur les relations

Pour les relations, c'est le système P8 qui obtient la meilleure précision, le système P3 qui obtient le meilleur rappel et le système P10 qui obtient la meilleure f-mesure toutes relations et tous genres de corpus confondus en tenant compte des constituants de l'hypothèse lorsqu'ils existent sinon de ceux de la référence et avec la comparaison barycentre pour les empans de texte des constituants (voir table 1). On voit dans le figure 4 les graphes respectifs de ces trois participants, avec les mêmes conventions que dans la figure 2 .

Le tableau 3 présente les résultats de tous les analyseurs en précision et f-mesure, pour toutes les relations, par type de corpus.

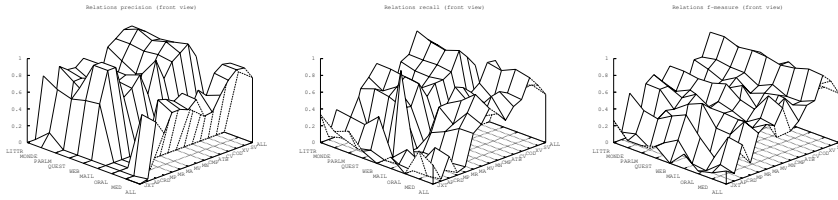

FIG. 4 – Vues avant sur les performances toutes relations et tous genres de corpus confondus pour les meilleures performances en précision (P8), rappel (P3) et f-mesure (P10)

	lemonde	littéraire	médical	oral_delic	parlement	questions	web
P1	p=0.571 f=0.543	p=0.611 f=0.576	p=0.599 f=0.561	p=0.608 f=0.544	p=0.579 f=0.546	p=0.683 f=0.648	p=0.594 f=0.549
P2	p=0.319 f=0.173	p=0.083 f=0.054	p=0.068 f=0.046	p=0.333 f=0.144	p=0.29 f=0.163	p=0.158 f=0.089	p=0.418 f=0.226
P3	p=0.628 f=0.616	p=0.577 f=0.596	p=0.641 f=0.634	p=0.555 f=0.513	p=0.593 f=0.590	p=0.662 f=0.635	p=0 f=0
P4	p=0.583 f=0.409	p=0.529 f=0.429	p=0.277 f=0.231	p=0.563 f=0.459	p=0.551 f=0.400	p=0.669 f=0.607	p=0.554 f=0.415
P5	p=0.562 f=0.508	p=0.507 f=0.456	p=0.564 f=0.524	p=0.514 f=0.425	p=0.529 f=0.472	p=0.447 f=0.412	p=0.553 f=0.489
P6	p=0.419 f=0.377	p=0 f=0	p=0 f=0	p=0 f=0	p=0.410 f=0.372	p=0 f=0	p=0.463 f=0.433
P7	p=0.663 f=0.521	p=0.681 f=0.524	p=0.652 f=0.527	p=0.633 f=0.434	p=0.644 f=0.498	p=0.665 f=0.521	p=0.608 f=0.472
P8	p=0.762 f=0.656	p=0.797 f=0.651	p=0.790 f=0.699	p=0 f=0	p=0.746 f=0.644	p=0.771 f=0.696	p=0.795 f=0.686
P9	p=0.004 f=0.003	p=0.023 f=0.015	p=0.042 f=0.026	p=0.257 f=0.128	p=0.003 f=0.002	p=0.110 f=0.065	p=0.688 f=0.416
P10	p=0.610 f=0.599	p=0.640 f=0.624	p=0.605 f=0.597	p=0.522 f=0.502	p=0.582 f=0.568	p=0.635 f=0.622	p=0.595 f=0.573
P11	p=0.604 f=0.131	p=0.640 f=0.160	p=0.622 f=0.169	p=0.646 f=0.175	p=0.597 f=0.137	p=0.605 f=0.161	p=0.670 f=0.111
P12	p=0.406 f=0.338	p=0.389 f=0.320	p=0.433 f=0.375	p=0.337 f=0.258	p=0.365 f=0.289	p=0.483 f=0.402	p=0.406 f=0.337
P13	p=0.355 f=0.338	p=0.429 f=0.404	p=0.359 f=0.343	p=0 f=0	p=0.337 f=0.321	p=0.354 f=0.330	p=0.268 f=0.255
P14	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0	p=0 f=0
P15	p=0.336 f=0.312	p=0.381 f=0.340	p=0.326 f=0.302	p=0 f=0	p=0.335 f=0.311	p=0.358 f=0.319	p=0.337 f=0.329

TAB. 3 – Mesures en précision (p) et f-mesure (f) par type de corpus pour toutes les relations

4 Les leçons à tirer

Tout d'abord, rappelons que ce n'est pas parce qu'un système a une valeur de performance 0 pour un sous-corpus ou une relation particulière qu'il a de mauvaises performances, il peut

genre	énoncés nb total	mots nb total	relations nb total	énoncés erronés/testés	relations erronées/testées
WEB	77	2104	113	3/7 = 43%	4/77= 03%
LE MONDE	380	10081	5072	12/39= 30%	22/519= 04%
PARLEMENT	276	7551	3884	14/28= 50%	57/366= 15%
LITTÉRATURE	892	24358	12725	36/93= 38%	92/1196= 07%
EMAILS	852	9243	3960	21/75= 28%	30/421= 07%
MÉDICAL	554	11799	5595	16/54= 29%	28/518= 05%
ORAL_DELIC	505	8117	4591	10/50= 20%	14/462= 03%
QUESTIONS	203	4116	2165	9/20= 45%	20/217= 09%

TAB. 4 – Nombres d'énoncés et de mots par genre de sous-corpus dans la référence.

s'agir d'un choix délibéré de son concepteur de ne pas traiter un phénomène particulier ou de ne retourner qu'une sorte d'annotation, par exemple seulement les relations. Ensuite, de mauvaises performances peuvent provenir de problèmes d'alignement entre les données du participant et celles de références et non d'un mauvais analyseur. Rappelons que dans EASY, contraitement à ce qui avait été fait dans GRACE (Adda *et al.*, 1999) ou dans (Roark *et al.*, 2006), il n'y a pas de procédure de réalignement automatique des données du participant, celui-ci doit respecter la segmentation en mots et en phrases des données qu'il traite.

Concernant les résultats, nous constatons, comme cela était à prévoir, une plus grande variabilité et de moins bonnes performances pour les relations que pour les constituants. Bien entendu, ces résultats ne sont qu'un point de vue ponctuel et sont à relativiser (comme dans toute évaluation quantitative) en fonction des facteurs décrits ci-après. Tout d'abord, la qualité des annotations de référence : nous avons réalisé une première estimation du taux d'erreur d'annotation sur les relations, par type de corpus en demandant à un expert d'examiner à la main un échantillon représentant environ un dixième de chaque corpus annoté. Les résultats sont donnés dans la table 4. Un énoncé est considéré comme erroné s'il contient au moins une erreur d'annotation en relation.

Pour les sous-corpus ayant un taux d'erreur en relation supérieur à 6%, nous avons effectué des corrections systématiques des erreurs les plus fréquentes avant de lancer les calculs de performance [4]. L'estimation du taux d'erreur d'annotation pour tous les sous-corpus permettra de déterminer des classes de performance parmi les différents systèmes sans prendre en compte des différences de performance inférieures aux taux d'erreur estimé.

Le second point dont il faut tenir compte concerne les erreurs de segmentation en mots/phrases encore présentes dans la référence et qui nous ont conduit en particulier à abandonner le traitement du corpus oral provenant de la campagne ESTER. Ces erreurs (auxquelles parfois s'ajoutent les erreurs de format des données des participants) sont à notre avis le résultat de divers facteurs : l'absence de tests *à blanc* du protocole (par manque de temps) et le fait d'avoir imposé une segmentation en mots et phrases de la référence, qui se heurte au problème de déterminer une définition acceptable par tous.

Dans le projet PASSAGE, qui regroupe certains des participants d'EASY, nous annoterons un

[4] Pour le moment nous n'avons pas estimé le taux d'erreur d'annotation pour les sous-corpus web et emails, ni effectué une nouvelle estimation pour les sous-corpus dont les erreurs les plus fréquentes ont été corrigées.

grand corpus en combinant automatiquement des analyseurs syntaxiques. Pour les deux campagnes d'évaluation prévues, nous envisageons de recourir à des procédure d'alignement automatique à partir du texte comme dans GRACE (Adda *et al.*, 1999) ou (Roark *et al.*, 2006). Les participants pourront ainsi conserver leurs propres algorithmes de segmentation en mots et phrases. La phrase dans les données de référence sera déterminée à partir des annotations elles-mêmes, une phrase étant constituée par l'empan de texte sur lequel un arbre syntaxique se projette, comme cela a déjà été fait dans EASY pour le sous-corpus ORAL-DELIC.

Bien entendu, le formalisme d'annotation EASY s'il semble suffisament abouti pour les relations les plus fréquentes comme la relation sujet-verbe, nécessite d'être approfondi pour les autres ; ce qui sera fait dans le cadre du projet PASSAGE, où cette fois nous considèrerons des constitutants admettant plusieurs niveaux de récursivité.

5 Conclusion

EASY a permis de poser les bases d'un protocole d'évaluation des analyseurs syntaxiques du français en mode boîte noire avec des mesures quantitatives objectives. Il a surtout été l'occasion de former un groupe autour du problème de l'évaluation comparative des technologies d'analyse syntaxique et d'acquérir une première expérience dans le cadre d'une campagne d'envergure qui déjà trouve des prolongements dans le projet PASSAGE. Concernant les mesures de performances proprement dites, l'image ponctuelle qu'elles donnent des performances des analyseurs syntaxiques à un instant particulier, nous montre qu'il reste encore un fort potentiel de développement dans la combinaison des approches pour l'annotation de relations syntaxiques, car ce sont 3 systèmes différents qui obtiennent chacun les meilleurs résultats pour la précision, le rappel et la f-mesure. Ce qui laisse à penser que ces systèmes ont des caractéristiques complémentaires, il reste encore à les identifier et à trouver le moyen de les combiner harmonieusement.

Références

ABEILLÉ A. (1991). Analyseurs syntaxiques du français. *Bulletin Semestriel de l'Association pour le Traitement Automatique des Langues*, **32**, 107–120.

ABEILLÉ A., CLÉMENT L. & KINYON A. (2000). Building a treebank for french. In *Proceedings of the 2nd International Conference on Language Ressources and Evaluation (LREC)*, p. 1251–1254, Athen, Greece.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M. & LECOMTE J. (1999). L'action grace d'évaluation de l'assignation des parties du discours pour le français. *Langues*, **2**(2), 119–129.

BLACHE P. & MORIN J. (2003). Une grille d'évaluation pour les analyseurs syntaxiques. In *Acte de l'atelier sur l'Evaluation des Analyseurs Syntaxiques dans les actes de la 10e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, Batz-sur-Mer.

BLACK E., ABNEY S., FLICKENGER D., GDANIEC C., GRISHMAN R., HARISON P., , HINDLE D., INGRIA R., JELINECK F., KLAVAN J., LIBERMAN M., MARCUS M., ROUKOS S., SANTORINI B. & STRZALKOZSKIJL (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, p. 306–311, Pacific Grove, California : Morgan Kaufman.

BRANT S., DIPPER S., HANSEN S., LEZIUS W. & SIMTH G. (2002). The tiger treebank. In *Proceedings of the 1st Workshop on Treebank and Linguistics Thories (TLT)*, Sozopol, Bulgaria.

CARROLL J., LIN D., PRESCHER D. & USZKOREIT H. (2002). Proceedings of the workshop beyond parseval - toward improved evaluation measures for parsing systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

GAIZAUSKAS R., HEPPLE M. & HUYCK C. (1998). Modifying existing annotated corpora for general comparative evaluation of parsing. In *Proceedings fo the Workshop on Evaluation of Parsing Systems in the Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

GENDNER V., ILLOUZ G., JARDINO M., MONCEAUX L., PAROUBEK P., ROBBA I. & VILNAT A. (2003). Peas the first instanciation of a comparative framework for evaluating parsers of french. In *Proceedings of the 10th Conference of the European Chapter fo the Association for C omputational Linguistics*, p. 95–98, Budapest, Hungarie. Companion Volume.

MARCUS M., SANTORINI B. & MARCINKIEWCIZ M. (1993). Building a large annotated corpus of english : The penn treebank. *Computational Linguistics*, **19**, 313–330.

MUSILLO G. & SIMA'AN K. (2002). Toward comparing parsers from different linguistic frameworks - an information theoretic approach. In *Proceedings of the Workshop Beyon Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

OEPEN S., NETTER K. & KLEIN J. (1996). Test suites for natural language processing. In *CSLI Lecture Notes*. Center for the Study of Language and Information.

PAROUBEK P., POUILLOT L.-G., ROBBA I. & VILNAT A. (2005). Easy : Campagne d'évaluation des analyseurs syntaxiques. In *Proceedings of the 12e conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 3–12, Dourdan, France.

PAROUBEK P., ROBBA I., VILNAT A. & AYACHE C. (2006). Data, annotations and measures in EASY - the evaluation campaign for parsers of French. In ELRA, Ed., *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, p. 315–320, Genoa, Italy : ELRA.

ROARK B. (2002). Evaluating parser accuracy using edit distance. In *Proceedings of the Workshop Beyon Parseval - Toward improved evaluation measures for parsing systems at the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

ROARK B., HARPER M., CHARNIAK E., DORR B., JOHNSON M., KAHN J., LIN Y., OSTENDORF M., HALE J., KRANYANSKAYA A., LEASE M., SHAFRAN I., SNOVER M., STEWARD R. & YUNG L. (2006). Sparseval : Evaluation metrics for parsing speech. In *Proceedings of the 5th International Conference on Language Ressources and Evaluation (LREC)*, Genoa, Italy.

SRINIVAS B., DORAN C., HOCKEY B. & JOSHI K. (1996). An approach to robust partial parsing and evaluation metrics. In *Proceedings of the Workshop on Robust Parsing*, Pragues : ESSLI.

VILNAT A., PAROUBEK P., MONCEAUX L., ROBBA I., GENDNER V., ILLOUZ G. & JARDINO M. (2004). The ongoing evaluation campaign of syntactic parsing of french : Easy. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, p. 2023–2026, Lisboa, Portugal.