Présentation de l'édition 2009 du DÉfi Fouille de Textes (DEFT'09)

Cyril Grouin[(1)], Béatrice Arnulphy[(1)], Jean-Baptiste Berthelin[(1)], Sarra El Ayari[(1)], Anne García-Fernandez[(1)], Arnaud Grappy[(1)], Martine Hurault-Plantet[(1)], Patrick Paroubek[(1)], Isabelle Robba[(1)] et Pierre Zweigenbaum[(1)]

[(1)]LIMSI–CNRS
BP 133 – F-91403 Orsay Cedex
prenom.nom@limsi.fr

Résumé – Abstract

La cinquième édition du défi fouille de textes (DEFT) porte sur la fouille d'opinion. Deux corpus multilingues (français, anglais et italien) ont été produits, le premier composé d'articles de journaux, le second de débats parlementaires européens. Trois tâches sont proposées : 1^o Identifier le caractère globalement objectif ou subjectif d'un article de journal, 2^o Identifier les passages subjectifs dans des articles de journaux et dans des interventions parlementaires, et 3^o Identifier le parti politique d'appartenance d'un parlementaire à partir d'interventions. Cet article présente le déroulement de la campagne, de la constitution des corpus aux mesures d'évaluation utilisées en passant par les évaluations humaines.
The fifth edition of the DEFT (DÉfi Fouille de Textes) Text Mining Challenge focuses on opinion mining. Two multilingual corpora (French, English and Italian) were produced, the first composed of newspaper articles, the second of parliament debates. Three tasks are proposed : 1st Identify the overall objectiveness or subjectiveness in a newspaper article, 2nd Identifying subjective passages in newspaper articles and speeches in parliament, and 3rd Identify the political party affiliation of a parliamentarian from interventions. This article describes the campaign, the creation of corpus, the evaluation measures used and the human scores.

Mots-clefs – Keywords

Corpus multilingues, fouille d'opinion, référence par votes majoritaires
Majority vote reference, multilingual corpora, opinion mining

1 Introduction

Pour cette cinquième édition du DÉfi Fouille de Textes, nous avons fait le choix de proposer une nouvelle tâche en fouille d'opinion. Il s'agit d'un thème intéressant à plus d'un titre : des entreprises en vivent, parfois même en complément de sondages d'opinion plus classiques, et le Web fournit des données en abondance, issues de blogs, de réseaux sociaux, de sites d'évaluation de produits, ou encore de journaux en ligne. Les applications concernent l'analyse et le suivi d'une « image » publique ou médiatique, avec des sphères d'application dans le commerce (image d'un produit, d'un service, d'une société), la vie publique (image d'une personnalité médiatique) ou politique (perception d'un projet politique).

Une analyse d'opinion commence par la détection du caractère plus ou moins subjectif d'un texte ou d'un passage, c'est-à-dire par déterminer s'il est porteur d'un « sentiment », d'un jugement, d'une opinion, ou au contraire de données essentiellement factuelles. Les parties de texte qui contiennent une opinion sont ensuite analysées pour donner une valeur à l'opinion exprimée, soit suivant une polarité positive/négative, soit suivant une échelle de valeurs[1]. Enfin, le jugement exprimé sur un sujet particulier peut être influencé par, ou laisser transparaître, des opinions d'un type plus général comme par exemple une opinion politique.

[1]C'était le thème retenu pour l'édition 2007 de DEFT : `http://deft07.limsi.fr`.

Pour cette campagne d'évaluation, nous avons proposé une approche multilingue de l'analyse d'opinion (français, anglais et italien) sur les trois tâches complémentaires suivantes :

– La détection du caractère objectif ou subjectif global d'un texte depuis un corpus d'articles de journaux ;
– La détection des passages subjectifs d'un texte, sur deux corpus : articles de journaux et débats parlementaires ;
– Enfin, la détermination du parti politique d'appartenance de chaque intervenant dans le corpus parlementaire.

Préalablement au lancement du défi, un groupe de juges humains a réalisé ces différentes tâches, sur un petit échantillon du corpus. L'objectif de ces évaluations humaines consistait, d'une part, à tester la faisabilité des tâches du défi, et d'autre part, à disposer d'un ordre de grandeur sur les résultats auxquels il était possible de prétendre pour chacune des tâches (Berthelin *et al.*, 2008). Précisons que le rôle de ces juges humains se limitait à une stricte participation aux tâches qui leur ont été soumises. En aucune manière ils n'ont eu pour charge de produire les données de référence des différents tâches.

Les données de référence utilisées pour chacune des tâches ont été constituées suivant deux méthodes différentes. Alors qu'il est possible de définir automatiquement les données de référence des tâches 1 et 3 (voir sections 4.1 et 6.1), nous ne disposons d'aucune référence de passages subjectifs annotés à l'intérieur d'un document. Pour la tâche 2, nous avons donc pris pour principe que la référence serait constituée par le croisement des résultats des participants à cette tâche. Nous reviendrons plus en détail sur ce choix dans la section 5.1.

Dans cet article, nous nous proposons de présenter dans un premier temps les principaux éléments du déroulement du défi, puis les corpus que nous avons rassemblés. Ensuite, pour chacune des trois tâches, nous montrerons d'abord comment nous avons constitué les données de référence, puis nous présenterons l'évaluation humaine de la tâche et enfin les résultats des participants.

2 Déroulement du Défi

2.1 Calendrier de la campagne

L'ouverture des déclarations d'intention de participation a été réalisée le 1er décembre 2008. Les corpus d'apprentissage ont été distribués à partir du 7 janvier aux équipes s'étant inscrites et ayant retourné signé le contrat d'utilisation des corpus. Comme lors des précédentes éditions, nous avons offert la possibilité aux participants de choisir leur période de test, soit trois jours complets à définir dans un intervalle d'un mois, du 18 mars au 17 avril.

Les résultats obtenus par les participants ont été diffusés, équipe par équipe, le 24 avril. Contrairement aux éditions antérieures, nous n'avons délivré aucun indice de comparaison entre participants (moyenne ou écart-type). Ce choix repose sur le modèle des campagnes TREC où les résultats globaux ne sont présentés que le jour de l'atelier de clôture.

2.2 Participations

Pour la première fois de l'existence du défi, nous avons inscrit cette campagne sous le signe du multilinguisme. À cet effet, nous avons donc proposé de travailler sur trois langues (français, anglais et italien). Cependant pour chaque tâche, nous n'avons formulé qu'une seule obligation aux participants, celle de travailler sur le français, les deux autres langues étant optionnelles. Par ailleurs, chaque équipe a pu librement choisir les tâches pour lesquelles elle souhaitait concourir sans obligation de nombre minimum de tâches.

Du fait du multilinguisme affiché pour cette campagne, plusieurs équipes internationales se sont inscrites au défi. Nous avons donc reçu les inscriptions de six équipes francophones (dont une de Belgique et une du Québec) et deux équipes non francophones (en provenance d'Allemagne et du Royaume-Uni).

Les équipes qui ont poursuivi leur travail jusqu'aux phases de tests et qui ont soumis des résultats sont les suivantes :

– CHART, *Cognition Humaine et ARTificielle* (Paris, France) : D. Legros, A. El Ghali, Y. V. Hoareau
– LINA, *Laboratoire d'Informatique Nantes Atlantique* (Nantes, France) : M. Vernier, B. Daille, N. Hernandez, L. Monceaux, S. Pena-Saldarriaga, F. Poulard
– LIPN, *Laboratoire d'Informatique de Paris Nord* (Villetaneuse, France) : M. Généreux, Th. Poibeau
– UCL, *Université Catholique de Louvain-la-Neuve* (Belgique) : G. Lories, Y. Bestgen
– UdeM, *Université de Montréal* (Canada) : D. Forest, M. Bélanger, D. Létourneau, A. van Hoeydonck
– UKP, *Ubiquitous Knowledge Processing* (Darmstadt, Allemagne) : C. Toprak, I. Gurevych

La tâche 1 (détection du caractère objectif/subjectif global d'un texte) a eu le plus de participants avec les laboratoires du CHART (sur le français, l'anglais et l'italien), du LINA (sur le français), de l'UCL (sur le français et l'anglais), de l'UdeM (sur le français) et de l'UKP (sur le français et l'anglais).

La tâche 2 (détection des passages subjectifs d'un texte) n'a eu que deux participants, le LINA et le LIPN, tous deux sur le français uniquement. Il faut souligner la particulière difficulté de cette tâche ainsi que des possibles controverses sur les principes de constitution des données de référence (voir section 5.1).

Finalement, un seul participant a poursuivi la tâche 3 (détermination du parti politique auquel appartient l'orateur) jusqu'au bout. Trois participants s'y étaient inscrits, mais deux ont renoncé à donner leurs résultats. Il est vrai que les résultats des logiciels sont faibles sur cette tâche, mais néanmoins tout à fait conformes à ce que laissait prévoir notre évaluation humaine (voir section 6.2).

2.3 Mesures d'évaluation des résultats

Les différentes tâches peuvent être considérées comme des tâches de classification, un élément à classer étant alors :

- Pour la tâche 1 : un document (article de journal) avec les classes OBJECTIF et SUBJECTIF ;
- Pour la tâche 2 : un passage de texte d'un document avec la classe SUBJECTIF, les parties de texte non étiquetées appartenant par défaut à la classe OBJECTIF ;
- Pour la tâche 3 : un document (intervention dans les débats parlementaires) avec les classes Verts-ALE, GUE-NGL, PSE, ELDR, PPE-DE.

Chaque fichier de résultat pour une tâche a été évalué en calculant la F-mesure sur toutes les classes de cette tâche avec $\beta = 1$, ce qui ne privilégie ni la précision ni le rappel, mais un équilibre entre les deux.

$$\mathrm{F}_{\text{mesure}}(\beta) = \frac{(\beta^2 + 1) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

La précision et le rappel sur les classes d'une tâche sont ici calculés suivant la macro-moyenne (Nakache & Métais, 2005) dans laquelle chaque classe compte à égalité avec les autres, qu'elle ait un fort ou un faible effectif.

F-mesure pondérée Dans la F-mesure classique, une seule classe peut être attribuée à chaque document. Cependant, un système de classification automatique peut attribuer à un document une distribution de probabilité sur les différentes classes au lieu de lui attribuer une seule classe. L'indice de confiance est cette probabilité pour un document d'appartenir à une catégorie donnée.

La F-mesure pondéré par l'indice de confiance sera utilisé à titre indicatif pour des comparaisons complémentaires entre les méthodes mises en place par les équipes.

Dans la F-mesure pondéré, la précision et le rappel pour chaque classe sont pondérés par l'indice de confiance. Ce qui donne :

$$\text{Précision}_i = \frac{\sum_{\text{attribué correct. } i=1}^{\text{Nombre attribué correct. } i} \text{indice de confiance}_{\text{attribué correct.}_i}}{\sum_{\text{attribué } i=1}^{\text{Nombre attribué } i} \text{indice de confiance}_{\text{attribué}_i}}$$

$$\text{Rappel}_i = \frac{\sum_{\text{attribué correct. } i=1}^{\text{Nombre attribué correct. } i} \text{indice de confiance}_{\text{attribué correct.}_i}}{\text{nombre de documents appartenant à la classe } i}$$

Avec :

- Nombre attribué correct.$_i$: nombre de documents attribué correct.$_i$ appartenant effectivement à la classe i et auxquels le système a attribué un indice de confiance non nul pour cette classe ;
- Nombre attribué$_i$: nombre de documents attribués$_i$ auxquels le système a attribué un indice de confiance non nul pour la classe i.

La F-mesure pondéré est ensuite calculé à l'aide des formules de la F-mesure classique.

Macro-moyenne

$$\text{Précision} = \frac{\sum_{i=1}^{n}\left(\frac{TPi}{(TPi+FPi)}\right)}{n} \quad \text{Rappel} = \frac{\sum_{i=1}^{n}\left(\frac{TPi}{(TPi+FNi)}\right)}{n}$$

Avec :
– TPi = nombre de documents correctement attribués à la classe i ;
– FPi = nombre de documents faussement attribués à la classe i ;
– FNi = nombre de documents appartenant à la classe i et non retrouvés par le système ;
– n = nombre de classes.

3 Présentation des corpus

Nous avons rassemblé deux types de corpus dans chacune des trois langues du défi. Le premier type de corpus concerne des articles de journaux tandis que le second se compose d'interventions parlementaires au Parlement européen.

Le corpus d'articles de journaux est destiné à servir aux tâches 1 (identification du caractère globalement objectif ou subjectif de l'article) et 2 (identification des passages subjectifs d'un article), tandis que le corpus des interventions parlementaires est utilisé dans le cadre des tâches 2 et 3 (détermination du parti politique auquel appartient l'orateur).

3.1 Articles de journaux

Le premier corpus intègre des articles de journaux provenant de trois titres européens :

– Le corpus en français est issu du quotidien *Le Monde*, 42 000 articles sur les années 2003 à 2006 ;
– Le corpus en anglais provient du quotidien économique *The Financial Times*, 13 000 articles de l'année 1993 ;
– Le corpus en italien comprend 2 500 articles du journal économique *Il Sole 24 Ore* sur la période 1992/1993.

Les corpus des journaux *The Financial Times* et *Il Sole 24 Ore* ont été rassemblés dans le cadre du projet MLCC (MultiLingual Corpora for Co-operation). Ce projet visait deux objectifs principaux. En premier lieu, permettre la réalisation de travaux sur des corpus comparables (à partir d'une collection d'articles de journaux en 6 langues[2] d'Europe de l'Ouest). En second lieu, fournir les bases pour des travaux de traduction (corpus parallèles multilingues dans 9 langues[3] européennes provenant de questions écrites et de débats parus au Journal Officiel de la Communauté Européenne).

Tous ces corpus d'articles de journaux sont disponibles auprès de l'agence ELDA qui les commercialise[4] sous les références ELRA-W0015 pour le corpus du *Monde* et ELRA-W0023 pour le corpus MLCC.

3.2 Débats parlementaires européens

Le corpus de débats parlementaires européens a été constitué en récupérant, depuis le site Internet du Parlement européen[5], les archives multilingues des 313 séances parlementaires qui se sont tenues entre 1999 et 2004. Dans ces archives, chacune des séances a été intégralement retranscrite et traduite dans les 11 langues officielles[6] de l'Union européenne. Chaque intervention est par ailleurs enrichie de plusieurs méta-données : le nom du parlementaire, la langue dans laquelle il s'exprime (qui n'est pas nécessairement celle de son pays d'origine) et le nom du groupe politique européen[7] duquel il relève.

[2]Corpus comparable MLCC d'articles de journaux : allemand (*Handelsblatt*), anglais (*The Financial Times*), espagnol (*Expansíon*), français (*Le Monde*), italien (*Il Sole 24 Ore*) et néerlandais (*Het Financieele Dagblad*).

[3]Corpus parallèle MLCC : allemand, anglais, danois, espagnol, français, grec, italien, néerlandais et portugais.

[4]Voir le site Internet `http://catalog.elra.info/` pour plus de précisions.

[5]Le site Internet du Parlement européen `http://www.europarl.europa.eu/` propose un libre accès à ses archives.

[6]Entre 1999 et 2004, l'UE comptait onze langues officielles : allemand, anglais, danois, espagnol, finnois, français, grec, italien, néerlandais, portugais et suédois. Depuis 2005, le nombre total de langues officielles a été porté à vingt-trois.

[7]Il existe neuf groupes politiques européens : EDD (Europe des Démocraties et des Différences), ELDR (parti Européen des Libéraux, Démocrates et Réformateurs), GUE/NGL (groupe confédéral de la Gauche Unitaire Européenne et Gauche Verte Nordique), NI (les non inscrits), PPE-DE (Parti Populaire Européen (démocrates chrétiens) et Démocrates Européens), PSE (Parti Socialiste Européen), TDI (groupe Technique des Députés Indépendants), UEN (Union pour l'Europe des Nations) et enfin, les Verts/ALE (Verts, Alliance Libre Européenne).

4 Tâche 1 : Détection du caractère objectif/subjectif global d'un texte

4.1 Constitution des données de référence

4.1.1 Préparation des données

Chaque corpus de journal enrichit ses articles de méta-données[8]. Afin de constituer automatiquement la référence de cette tâche, nous avons étudié les méta-données disponibles en nous focalisant sur deux aspects principaux : la disponibilité des méta-données de manière transversale au corpus et la possibilité d'effectuer une catégorisation sur la base des méta-données retenues.

Concernant le journal *Le Monde*, nous avons utilisé le secteur de rédaction[9] sous lequel a paru chacun des articles. Le secteur de rédaction est une subdivision du journal qui correspond à la maquette. Nous avons ainsi considéré comme étant objectifs les articles relevant des secteurs de rédaction « France » et « International » (autrement dit, des articles traitant de politique nationale et internationale) tandis que les articles des secteurs « Éditorial – Analyses » et « Débats – Décryptages » ont été qualifiés d'articles subjectifs.

Pour ce qui concerne le corpus d'articles du journal britannique *The Financial Times*, nous avons pris en compte les éléments d'indexation de chaque article en nous intéressant à deux descripteurs : les articles indexés « CMMT Comment & Analysis » ont été qualifiés d'articles subjectifs alors que ceux indexés « NEWS General News » ont été enregistrés parmi les articles objectifs. Les autres descripteurs étant plus spécifiques (*COMP Company News, MGMT Management & Marketing, RES Capital expenditures,* etc), il n'a guère été possible d'en faire usage.

Enfin, nous avons classé les articles du journal italien *Il Sole 24 Ore* en étudiant les descripteurs d'indexation : les articles ont été qualifiés d'article subjectifs s'ils étaient indexés par le descripteur « Opinioni e commenti » et objectifs dans les autres cas. Nous n'avons pas utilisé les autres descripteurs du fait de leur trop grande spécificité (*Attività immobiliari, Inchieste e notizie giudiziarie, Statistiche monetarie e finanziarie,* etc).

Journal	Objectifs	Subjectifs	Répartition
Le Monde	34 761	7 232	83%/17%
The Financial Times	5 708	7 403	44%/56%
Il Sole 24 Ore	1 559	936	62%/38%

TAB. 1 – Nombre d'articles objectifs et subjectifs pour chaque corpus de journal.

4.1.2 Exemples

Le Monde – objectif. SOUPÇONNÉ par la direction de la surveillance du territoire (DST) d'être l'un des informateurs anonymes des juges Renaud Van Ruymbeke et Dominique de Talancé, qui enquêtent sur l'affaire des frégates de Taïwan, Imad Lahoud, informaticien chez EADS, rompt le silence par la voix de son avocat. Dans un communiqué adressé au Monde, Me Olivier Pardo assure qu'« en dépit de la multiplicité des assertions aucune preuve d'une quelconque participation de -son- client à cette affaire n'existe ». « Dans une ronde sans fin, la calomnie s'installe et risque de devenir extrêmement préjudiciable », ajoute Me Pardo, qui précise que son client « n'a jamais rencontré le député Alain Marsaud », contrairement à ce qu'a laissé entendre la DST dans un rapport (Le Monde du 20 juillet). « M. Lahoud demande que cessent ces assertions et l'instrumentalisation de sa personne, conclut l'avocat. Il ne réclame qu'une chose : qu'il puisse continuer à animer son équipe de recherche dans la grande entreprise européenne à laquelle il a l'honneur d'appartenir, dans la sérénité et la quiétude. »

[8] Un document du *Financial Times* comprend les éléments suivants : titre, date de publication, signature, article, indexation par un thésaurus, lieu d'édition et numéro de page.
Un document du *Monde* comprend les éléments suivants : date de publication, secteur de rédaction, titraille (têtière, sur-titre, titre, sous-titre), chapô, nom et localisation géographique du journaliste, article, éléments d'indexation (titre complémentaire, catégories) et mots-clés (de type France, Étranger et Personne) issus d'un thésaurus interne.
Un document du journal *Il Sole 24 Ore* comprend les éléments suivants : date de publication, rubrique, chapô, signature, article, et éléments d'indexation (de type descripteurs, aire géographique et didascalies).

[9] Une étude de la répartition des articles dans les différents secteurs de rédaction du journal entre 1987 et 2006 est disponible sur le document `http://perso.limsi.fr/grouin/rubriques_lemonde_1987-2006.html` d'après un travail de S. Loiseau et C. Grouin.

Le Monde – subjectif. Avec la campagne présidentielle qui se profile, voici revenu le temps des effets chocs et des idées chics. Il en est ainsi du « dialogue social », serpent de mer du débat public que les politiques défendent surtout quand ils n'ont pas à le pratiquer. Jacques Chirac en a parlé le 14 juillet. Dominique de Villepin redécouvre son existence après l'avoir superbement ignoré. L'UMP va achever, en septembre, des rencontres avec tous les partenaires sociaux, CGT comprise. Le PS recevra, à son Université d'été de La Rochelle, du 25 au 27 août, tous les syndicats et le numéro deux du Medef, Denis Gautier-Sauvagnac.

The Financial Times – objectif. THE Health and Safety Executive yesterday accused the European Community of imposing a mass of health and safety legislation on member states without adequate thought or consultation, Diane Summers writes. The executive is concerned that it is becoming a focus of complaints that UK businesses are increasingly subject to excessively bureaucratic regulation. It is anxious to remind the government that much of the regulation and perceived red tape originates from Brussels and not from it. Mr John Rimington, the executive's director-general, said the directives had been constructed in 'smoke-filled rooms in Brussels'. He blamed, in particular, the French for trying to get the directives to reflect their own domestic laws and said some of the provisions were 'incomprehensible'.

The Financial Times – subjectif. Denmark has a new government. The foreign minister has pledged himself to secure a Yes to Maastricht in the second Danish referendum on the subject, to be held probably 'before June'. All's right with the world. Some moaning Euro-minnies are still muttering about the dangerous precedent set by Denmark's special 'opt-outs', negotiated at last month's Edinburgh summit. Won't Conservative backbenchers try to obtain the same deal for Britain, as the price of ratification ? Won't candidates for EC membership, with three of whom formal negotiations are to start on Monday, demand that the same exemptions apply to them ? Isn't this the beginning of the a la carte union so dreaded by Mr Jacques Delors, president of the European Commission ?

Il Sole 24 Ore – objectif. Sarà la presidenza del Consiglio a dire l'ultima parola sulla riforma dei fondi pensione. Il ministro del Lavoro Nino Cristofori ha già annunciato per il mese di dicembre la presentazione del suo progetto di regolamentazione delle casse integrative aziendali, in attuazione della recente legge delega sul riordino della previdenza. Quello di Cristofori, però, sarà soltanto il materiale preparatorio per la stesura definitiva del decreto legislativo che darà il via libera alla riforma. A Palazzo Chigi si è costituito in questi giorni un gruppo di esperti con il compito di analizzare le proposte del Lavoro confrontandole con quelle provenienti da altri ministeri (Tesoro, Industria e Finanze), parti sociali e gruppi economici.

Il Sole 24 Ore – subjectif. L'economia va male e quindi i tassi di interesse si riducono. Questa apparentemente banale relazione è tornata a essere vera in questi giorni un pò in tutta Europa e quindi anche in Italia. Dopo i fuochi d'artificio di un mese fa, quando sembrava che le banche centrali di tanti Paesi europei fossero solo impegnate nella nobile gara a chi alzava di più i tassi di interesse, il buonsenso economico è tornato a prevalere.

4.2 Évaluation humaine de la tâche

Le test humain de la tâche 1 a été réalisé sur un ensemble de sept articles du *Monde*. Les résultats obtenus (voir tableau 2) se révélèrent assez élevés mais doivent sans doute être relativisés en tenant compte du nombre réduit de documents constituant notre corpus d'évaluation humaine.

Il faut cependant noter que le caractère objectif ou subjectif de chaque article a toujours été bien reconnu par la majorité des juges humains. Sur un vote majoritaire, le groupe des six juges humains auraient donc obtenu 100% de réussite.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Rappel	0,71	0,83	0,67	0,88	1,00	0,88
Précision	0,71	0,90	0,83	0,88	1,00	0,88

TAB. 2 – Rappel et précision obtenus par les testeurs humains sur la tâche de qualification globale des articles.

4.3 Résultats

La première tâche proposée concernait donc la caractérisation globale d'articles de journaux parmi deux classes possibles : objectif ou subjectif. Cinq équipes se sont essayées à cette tâche, chacune sur le français, trois d'entre elles sur l'anglais, une seule pour l'italien.

Nous présentons dans le tableau 3 la F-mesure obtenue par chaque équipe, pour chacune des soumissions effectuées dans chacune des langues de cette tâche. Une F-mesure suivie d'une étoile renvoie à une soumission utilisant des indices de confiance (pondération des valeurs de résultat).

Langue	Équipe	F-mesures par soumission
Anglais	CHART	0,676 – 0,652*
Anglais	UCL	0,851
Anglais	UKP	0,822 – 0,769 – 0,814
Français	CHART	0,771 – 0,715*
Français	LINA	0,850
Français	UCL	0,925
Français	UdeM	0,757 – 0,778 – 0,781
Français	UKP	0,662 – 0,769
Italien	CHART	0,716 – 0,691*

TAB. 3 – F-mesures obtenues pour chaque soumission de chaque équipe dans chacune des langues sur la tâche 1. L'étoile indique qu'il s'agit d'une F-mesure pondérée.

Afin de faciliter la comparaison des résultats obtenus entre équipes, nous ne représentons dans le graphique suivant que la meilleure soumission obtenue par chaque équipe.

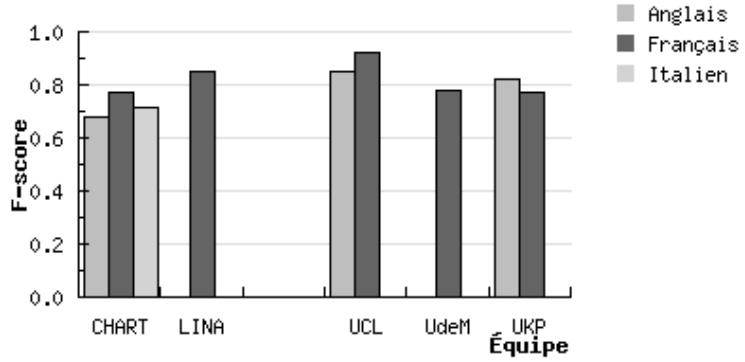

FIG. 1 – F-scores obtenus sur la meilleure soumission de chaque équipe sur la tâche 1.

Il apparaît que les meilleurs résultats ont été obtenus, pour les équipes francophones, sur le corpus en français tandis que l'équipe allemande (UKP), non francophone, a obtenu de meilleurs résultats sur l'anglais.

5 Tâche 2 : Détection des passages subjectifs d'un texte

5.1 Constitution des données de référence

5.1.1 Principe

Dans la mesure où nous ne disposions pas de données annotées en passages objectifs ou subjectifs, nous avons retenu le principe de constitution des données de référence par vote majoritaire : les données de référence d'un corpus donné sont constituées a posteriori, à partir d'un vote majoritaire entre les résultats des participants. Ce principe a déjà été expérimenté dans des campagnes d'évaluation d'analyseurs syntaxiques (Paroubek *et al.*, 2008).

Dans une tâche de classification, les données de référence sont donc constituées par les catégorisations sur lesquelles la majorité des participants à cette tâche sont tombés d'accord. Par exemple si la phrase « *L'affaire ne devrait pas améliorer les relations entre Séoul et Pyongyang.* » a été considérée comme un passage subjectif par la majorité des participants, alors cette phrase sera annotée comme subjective dans les données de référence. Et tous les mots de cette phrase, considérés seulement dans cette phrase évidemment, seront comptés comme mots subjectifs. Dans le cas contraire, cette phrase sera annotée comme objective.

Les données de référence sont donc, suivant ce principe, les données qui ont été classées de la même manière par les logiciels en compétition. Elles sont intéressantes car elles donnent un état des capacités des logiciels, mais nous sommes conscients qu'elles peuvent être contestées en tant que références. L'évaluation des logiciels par rapport à cette référence donne une bonne idée des écarts entre les logiciels, sans pour autant donner un classement fiable entre participants. Par ailleurs, la faible participation à cette tâche affaiblit également la valeur des données de référence.

5.1.2 Exemples

Nous avons expérimenté ce principe de constitution des données de référence lors de l'évaluation humaine de la tâche (voir la section suivante 5.2). L'exemple de texte qui suit (tiré d'un article du *Monde*) est extrait des données de référence constituées à partir des résultats de cette évaluation humaine sur le petit corpus de la tâche 1. Il est annoté en passages subjectifs et objectifs.

SUBJECTIF : Pourquoi ne pas fonder sans complexes une critique des feux d'artifice ? Trop vulgaires ? Puérils ? Naïfs ? Manque de dignité esthétique, en somme ? Allons, allons.
OBJECTIF : De Chantilly, en juin, à Saint-Sébastien (jusqu'au 17 août), en passant par Biarritz le 15 août, et toutes les nuits olympiques en Chine,
SUBJECTIF : la matière ne manque pas. Les trois vertus théologales de la critique moderne (affluence, âge, berlue) s'y hisseraient à leur zénith.
OBJECTIF : Exemple : plus personne ne se risque, en festival pyrotechnique, au bleu de Chine, aperçu pour la dernière fois en 1957.
SUBJECTIF : Chimiquement trop complexe et bien aléatoire. A Pékin, peut-être ?
OBJECTIF : Pendant les
SUBJECTIF : fameuses
OBJECTIF : fêtes de Pampelune (7-14 juillet), Mikel Pagola Erviti exerce la fonction de critique de feux d'artifice au Diario de Navarra. Quand une prestation déçoit Mikel Pagola Erviti - la Pirotecnia Vicente Caballer de Valence, cette année -, il égrène quatre motifs :
SUBJECTIF : composition peu discernable, manque de rythme, faiblesses des effets, banalité des couleurs.
OBJECTIF : Sur place, un jury
SUBJECTIF : pointilleux
OBJECTIF : s'aligne sur ces critères. Le lendemain à 9 heures, la chaîne Navarra 6 retransmet en boucle le festival pyrotechnique de la veille.
SUBJECTIF : Son infernal bruit de guerre s'y noie évidemment. Or, la vérité d'un feu, c'est son bruit.
OBJECTIF : A la fin des années 1950, la chaîne locale de la radio d'Etat diffusait en direct le feu d'artifice de Biarritz.
SUBJECTIF : Un feu d'artifice à la radio, formidable. Presque mieux qu'un concours de mime. A Pampelune, Mikel Pagola Erviti misait sur la Pirotecnia Zaragozana. Laquelle ressent la gloire qu'on lui ait confié l'ouverture et la clôture de l'Expo universelle.
OBJECTIF : Le gérant de la Zaragozana occupe la chaire de chimie à l'université de la ville.
SUBJECTIF : Ses cours sont très marrants. Mais le favori de Mikel Pagola Erviti, c'est " Gori ",
OBJECTIF : le fondateur de la Pirotecnia Gori à Valence. Gregorio Juan Moreno, dit " El Gori ", a effectué sa première et dernière présentation à Pampelune le 11 juillet.
SUBJECTIF : Le Gori ne fait rien comme les autres. Il se signale par un inimitable parfum des enchaînements, des durées et des surprises du temps. Le Gori est une légende. Il attendait de Pampelune la consécration pour se retirer dans le bouquet, tel un Cincinnatus de la belle bleue. Gare au Gori.

5.2 Évaluation humaine de la tâche

Pour la tâche 2, chaque testeur humain a encadré les passages objectifs et subjectifs de balises `<obj>...</obj>` et `<sub>...</sub>`, dans chaque article du corpus d'évaluation humaine de la tâche 1. Un passage est un

extrait de texte dont nous avons volontairement laissé indéfinies les limites. Celui-ci pouvait donc aller d'un mot à plusieurs phrases suivant l'estimation personnelle du juge humain. Parfois, c'est seulement un modifieur qui a été qualifié de subjectif, mais très souvent, c'est une proposition complète qui constitue un passage subjectif. Les annotations ont ensuite été alignées au niveau du mot. Pour chacun des mots du corpus, la valeur majoritairement attribuée par les testeurs à ce mot a constitué la référence. Les annotations des testeurs ont ensuite été évaluées sur la base de cette référence. Les résultats obtenus par les testeurs humains sur cette tâche se sont révélés très bons et encourageants pour l'organisation du défi.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Rappel	0,81	0,92	0,82	0,90	0,89	0,78
Précision	0,77	0,81	0,73	0,79	0,78	0,70

TAB. 4 – Rappel et précision obtenus par les testeurs humains sur la tâche de détection des passages subsjectifs.

L'un des enseignements que nous pouvons tirer de cette expérience concerne le caractère personnel de ce qui constitue un élément subjectif par opposition à un élément objectif. Ainsi, le corpus que les six testeurs humains ont eu pour charge d'annoter comprenait 5036 mots. Sur ces 5036 mots, on observe une concordance stricte entre les six testeurs sur 2053 mots seulement, soit 41% du corpus. Parmi ces concordances, il existe 1447 concordances sur des mots objectifs et 606 concordances sur des mots subjectifs. Malgré ce désaccord apparent, les bons résultats des testeurs humains (voir tableau 4) montrent que chacun d'entre eux a finalement peu d'écart avec l'accord majoritaire.

Nous donnons ci-après le nombre de mots catégorisés « objectif » et « subjectif » par chacun des testeurs.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Mots objectifs	3635	3050	3263	2936	2455	2186
Mots subjectifs	1401	1986	1773	2100	2581	2850

TAB. 5 – Nombre de mots catégorisés « objectif » et « subjectif » par chacun des testeurs humains.

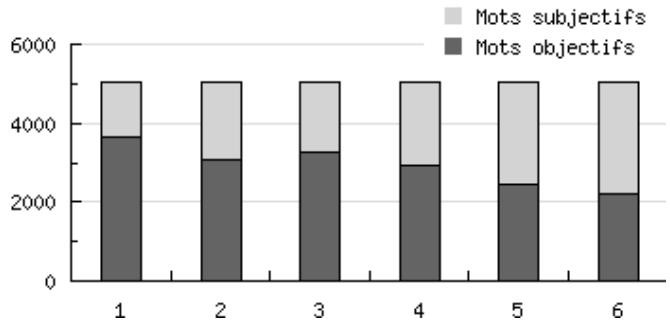

FIG. 2 – Variations personnelles des mots catégorisés « objectif » (blocs foncés) et « subjectif » (blocs clairs) pour chacun des testeurs humains (barres numérotées de 1 à 6), sur les 5036 mots du corpus testé.

Il apparaît ainsi que le testeur humain n° 1 considère le corpus comme étant majoritairement objectif (72% des mots du corpus sont catégorisés « objectifs ») alors qu'à l'opposé, le testeur humain n° 6 envisage le corpus sous un aspect nettement plus subjectif (57% des mots du corpus relèvent de passages catégorisés « subjectifs »).

Lien entre les tâches 1 et 2 Nous avons voulu évaluer également en quoi la qualification globale de l'article en subjectif/objectif d'une part, et la détection des passages subjectifs d'autre part, pouvaient se rejoindre. Nous avons là en effet deux points de vue sur la subjectivité d'un texte, l'un global et l'autre local. Pour chaque article, nous avons donc comparé la proportion de mots qualifiés de subjectifs en moyenne par l'ensemble des juges humains, à la référence de qualification globale de cet article en objectif ou subjectif. Le tableau 6 qui rassemble ces comparaisons montre qu'à l'exception de l'article 1656, et donc pour 6 articles sur les 7 du corpus donné aux juges humains, un article considéré comme objectif comporte une majorité de mots classés objectifs par les juges humains, et un article considéré comme subjectif comporte une majorité de mots classés subjectifs par les

juges humains. Cela montre une certaine cohérence, concernant la différenciation entre objectif et subjectif, entre l'impression globale sur un texte et la somme des impressions locales.

L'article 1656 en revanche constitue un contre-exemple. Cet article, qui est un éditorial à propos d'une biographie, a été jugé globalement de caractère subjectif par la majorité des juges humains (quatre juges sur les six). Malgré cela, ces mêmes juges ont trouvé localement peu de passages subjectifs dans cet article. Les liens entre le caractère subjectif global d'un texte et les passages subjectifs qu'il contient sont parfois complexes et mériteraient une analyse approfondie.

Article (id)	**4415**	**2628**	**1662**	**1935**	**1656**	**3998**	**4123**
Caractère de l'article	objectif	subjectif	objectif	objectif	subjectif	subjectif	objectif
Mots subjectifs	12%	67%	22%	33%	28%	71%	39%

TAB. 6 – Proportion de mots subjectifs dans chaque article, suivant les juges humains, comparé à sa qualification globale de référence en objectif/subjectif.

5.3 Résultats

Dans la présentation proposée sur le site Internet du défi, nous avions définie cette tâche de la manière suivante : « *Un texte peut être segmenté en passages objectifs, qui donnent des faits, ou le thème du texte, et en passages subjectifs qui délivrent une opinion, un sentiment, sur ces faits, concernant ce thème. Cette deuxième tâche consiste donc à repérer les passages subjectifs d'un texte, que ce texte soit globalement subjectif ou objectif. Un passage peut aller d'un mot (par exemple un modifieur) à plusieurs phrases* ». Un passage est donc un extrait de texte pouvant aller d'un mot à plusieurs phrases. Dans la section 5.1.2, nous avons présenté un exemple de la variabilité de la taille d'un passage, extrait du test humain de la tâche.

Les deux participants à cette tâche ont pris, concernant la taille d'un passage, des options fixes et radicalement opposées. L'un a systématiquement pris une phrase comme passage, et l'autre a pris comme passage ce que nous appellerions un *déclencheur* de subjectivité, c'est-à-dire en général un mot, modifieur ou pronom personnel par exemple.

Pour pouvoir extraire des données de référence de ces résultats, nous avons dû harmoniser leurs notions différentes d'un passage. Pour cela, nous avons systématiquement étendu les passages du deuxième participant à une portion de texte comprise entre deux ponctuations.

Cette expérience nous a montré que cette tâche était sans doute trop ambitieuse par rapport aux moyens que nous pouvions mettre en œuvre. Car si les juges humains ont intuitivement annoté des passages de tailles bien différentes, ils l'ont fait suivant des critères difficiles à expliciter. Et il est clair que les logiciels demandent, en revanche, des critères bien définis.

Résultats des participants. Deux équipes ont participé à cette tâche et ont soumis chacune trois fichiers de résultats. Chaque soumission comporte à la fois le corpus des débats parlementaires et le corpus des articles de journaux. Le tableau 7 rassemble les résultats de ces soumissions. Les données de référence pour chaque corpus résultant de l'accord majoritaire entre les six soumissions, elles constituent un sous-ensemble de toutes les soumissions. Plus une soumission est proche de ce sous-ensemble, plus sa F-mesure sera élevée. Les résultats pour chaque soumission marquent donc avant tout son écart par rapport à cet accord.

Corpus	**Équipe**	**F-mesures par soumission**	**Précisions**	**Rappels**
Journaux	LINA	0,670 – 0,623 – 0,863	0.928 – 0.623 – 0.808	0.524 – 0.623 – 0.926
Journaux	LIPN	0,777 – 0,714 – 0,775	0.701 – 0.929 – 0.699	0.871 – 0.579 – 0.869
Parlement	LINA	0,648 – 0,648 – 0,909	0.805 – 0.804 – 0.903	0.543 – 0.543 – 0.916
Parlement	LIPN	0,799 – 0,678 – 0,797	0.806 – 0.816 – 0.805	0.791 – 0.580 – 0.789

TAB. 7 – F-mesures obtenues pour chaque soumission de chaque équipe dans chacune des langues sur la tâche 2.

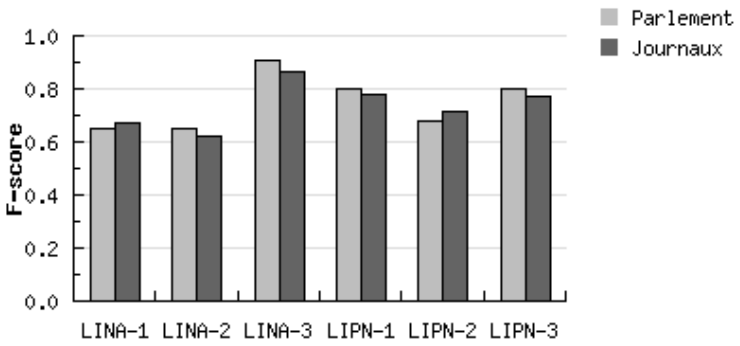

FIG. 3 – F-mesures obtenues sur les deux corpus de la tâche 2 pour chaque soumission de chaque équipe.

Accords et désaccords. La représentation graphique des résultats obtenus par les deux équipes permet d'apprécier les différences entre les différentes soumissions. Ces différences semblent légèrement plus accentuées sur le corpus du Parlement que sur le corpus des journaux. Quoi qu'il en soit, les données de référence constituées à partir des résultats marquent un état des lieux intéressant à analyser. En effet, une analyse de ces données permet de voir ce qui est actuellement repéré (avec d'ailleurs d'éventuelles erreurs), et ce qui a été omis par les logiciels.

Par exemple la phrase « *Madame la Présidente, il est indiqué à l'ordre du jour : vote de 12 heures à 13 heures, suite à 18 heures 30.* » a bien été considérée comme objective dans la majorité des soumissions, et la phrase « *Je pense que l'on devrait s'en tenir à l'ordre du jour.* » a bien été annotée comme subjective. En revanche la phrase « *Monsieur le Président, il est capital, selon moi, de disposer dorénavant d'une législation commune en matière de responsabilité environnementale.* » a été considérée comme objective dans la majorité des soumissions.

6 Tâche 3 : Détermination du parti politique auquel appartient l'orateur

6.1 Constitution des données de référence

6.1.1 Traitements des données

Ce corpus de débats a été constitué en récupérant, pour chaque séance parlementaire de la période de 1999 à 2004, l'ordre du jour, à partir duquel nous avons récupéré la retranscription des débats, dans les trois langues prévues pour le défi (pour rappel, français, anglais et italien).

Nous avons ensuite extrait de ces retranscriptions le nom de l'intervenant, son parti politique d'appartenance, la langue dans laquelle il s'est exprimé ainsi que son intervention. Le résultat de ces extractions a ensuite été aligné dans les trois langues de manière à disposer d'un corpus parallèle, chaque intervention dans une langue ayant sa traduction dans les deux autres langues, dans le même ordre des interventions pour chacune des trois langues.

Nous avons procédé à un nettoyage minimal de ce corpus aligné (suppression des fonctions de personnes, des codes de langues, et des caractères étranges) ainsi qu'à une anonymisation de base (par le remplacement des noms de groupes politiques[10] dans les textes par une balise <anonyme />).

Nous avons alors segmenté le corpus en deux parties, l'une réservée à la constitution du corpus d'apprentissage et comprenant 60% des interventions, l'autre dédiée à la réalisation du corpus de test, composée des 40% d'interventions restantes. Du fait de l'utilisation d'un corpus parallèle, nous avons la garantie que les interventions utilisées pour l'apprentissage dans une langue, ne pourront pas être présentes dans le corpus de test d'une autre langue.

Enfin, nous créons les corpus d'apprentissage et de test pour chaque langue, en mélangeant l'ordre d'apparition de chacune des interventions. Ainsi, les corpus d'apprentissage de chaque langue comprennent tous trois les mêmes

[10]Les expressions anonymisées sont, par exemple pour la français, les suivantes : ELDR, GUE/NGL, PPE-DE, PSE, Verts/ALE, Verts/Alliance libre européenne, Alliance libre européenne, gauche unie, gauche unitaire européenne, gauche verte nordique, parti des socialistes européens, démocrates-chrétiens, démocrates européens, parti populaire européen, chrétien-démocrate, gauche unitaire européenne, groupe libéral, sociaux-démocrates, gauche chrétienne, démocrate chrétien, démocrate européen, démocrates chrétiens, social-démocrate, national-démocrate-chrétien, social-démocrate.

interventions, mais présentées dans un ordre différent, fixé de manière aléatoire[11], et ce, afin d'éviter tout biais (tel que la possibilité de dupliquer les résultats pour une langue sur les deux autres langues). Il en est de même pour les trois versions du corpus de test.

Dans la perspective de la tâche 3 d'identification du parti politique de chaque intervenant, nous avons décidé de limiter le corpus aux cinq partis les plus représentés en termes de nombre d'interventions, soit : 3 346 interventions attribuées au groupe ELDR, 4 482 au GUE/NGL, 11 429 au PPE-DE, 9 066 au PSE et 3 961 aux Verts/ALE. Le parti politique d'appartenance de chaque parlementaire renseigné sur le site Internet du Parlement européen nous a permis de constituer les données de référence de la troisième tâche, en associant à chaque intervention le parti politique de son orateur.

6.1.2 Exemples

ELDR. Monsieur le Président, l'UE est le principal donateur des territoires palestiniens. Selon l'ONU, depuis les bouclages, plus d'un million de Palestiniens vivent en dessous du seuil de pauvreté, soit deux dollars par jour. La conséquence des bouclages sur le plan humain est encore bien pire. Comme le commissaire Patten l'a dit, les malades ne peuvent se rendre à l'hôpital européen de Gaza car même les ambulances ne franchissent pas les barrages. Il est donc essentiel que nous continuions d'apporter notre soutien à cette région. Je suis heureuse de constater que le Conseil et la Commission partagent ce point de vue.

GUE-NGL. Madame la Présidente, avec la proposition de modification de l'OCM des fruits et légumes, la Commission aggrave les problèmes de l'organisation des marchés actuelle, ainsi que les injustices de la PAC, et elle occasionne plus de difficultés aux producteurs de fruits et légumes. Les mesures visant à éliminer le prix minimum sont particulièrement graves, comme dans le cas de la tomate destinée à l'industrie ; de la réduction de la limite maximale de l'aide pour le volume des fonds opérationnels de 4,5 % à 3 % de la valeur de la production commercialisée de chaque organisation de producteurs ; de la réduction de 9,1 % du montant des aides à la première campagne après la réforme de l'OCM ; et de la réduction de la quantité susceptible d'indemnité communautaire de retrait pour les agrumes.

PPE-DE. Madame la Présidente, à mon grand regret j'ai dû voter contre le budget parce que je trouve absolument insuffisants, et même inexistants, tous les articles destinés à chercher à améliorer les conditions de vie des personnes âgées et des retraités. J'ai vu, en outre, que nombre de ces fonds sont destinés aux célèbres programmes d'action communautaire. Je crois que ces programmes n'exercent pas la fonction utile qu'ils devraient avoir dans l'utilisation des fonds communautaires. Je crois que l'Union européenne doit modifier complètement la façon dont elle dépense l'argent des quinze états membres de l'union.

PSE. Monsieur le Président, Mesdames et Messieurs, permettez-moi de limiter mon intervention au problème du VIH/sida. Le rapport sur l'état de la population mondiale, qui vient juste de paraître, contient des chiffres effrayants. 14 000 hommes, femmes et enfants meurent chaque jour, en moyenne, de cette maladie. Elle est devenue la première cause de mortalité en Afrique subsaharienne. Dans le monde, plus de 60 millions de personnes ont été contaminées par le virus du sida, environ 22 millions d'entre elles sont décédées. Sur les 40 millions de personnes contaminées à l'heure actuelle, 95 % vivent dans des pays en développement et presque trois quarts en Afrique. Des 580 000 enfants de moins de 15 ans morts du sida, 500 000 – près de 90 % – vivaient en Afrique. Je pourrais poursuivre indéfiniment cette énumération de statistiques édifiantes.

Verts/ALE. Monsieur le Président, hier, une fois de plus, onze personnes provenant de pays africains ont perdu la vie près des côtes espagnoles lorsque leur embarcation a fait naufrage. Ce drame, qui se produit très fréquemment, ne peut nous laisser indifférents : je crois que, malgré la difficulté, il faut rechercher une solution afin que ces gens ne soient pas obligés de tenter d'atteindre les terres européennes d'une manière aussi dramatique et tragique. Je crois que, en pareilles circonstances, il convient de le rappeler ici.

[11] À titre d'exemple, l'intervention figurant en première position dans le corpus d'apprentissage français se retrouve en position 8 399 dans le corpus d'apprentissage anglais et en position 16 074 dans le corpus d'apprentissage italien.

6.2 Évaluation humaine de la tâche

Le test a été réalisé sur le corpus des débats parlementaires européens et reposait sur l'identification de quatre partis politiques (contre cinq dans la version finale de cette tâche) : ELDR, PPE-DE, PSE et Verts/ALE. La répartition des interventions par parti dans ce petit corpus d'évaluation était la suivante : 12 ELDR, 20 PPE-DE, 25 PSE, 6 Verts/ALE.

Rappel/précision. Pour chaque testeur humain, nous avons calculé le rappel et la précision qu'il a obtenu pour chacun des partis à identifier et avons ensuite calculé un macro-rappel ainsi qu'une macro-précision. Le tableau 8 montre les valeurs de macro-rappel et macro-précision obtenues par chaque testeur humain.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Rappel	0,41	0,35	0,39	0,23	0,47	0,35
Précision	0,42	0,34	0,37	0,27	0,42	0,34

TAB. 8 – Rappel et précision obtenus par les testeurs humains sur la tâche d'identification des partis politiques.

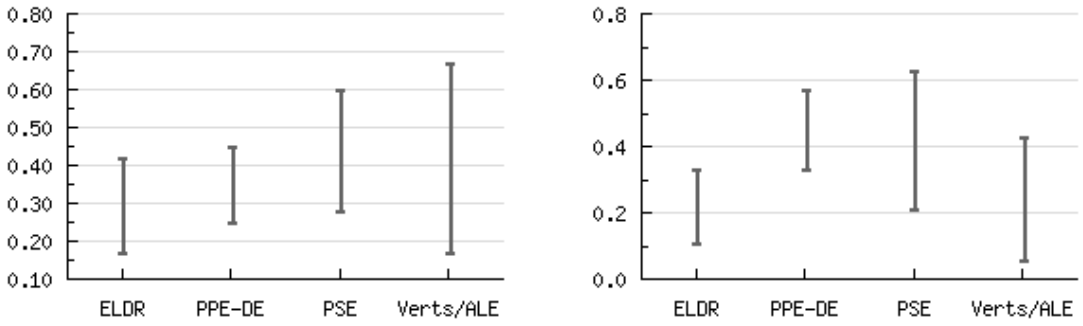

FIG. 4 – Valeurs minimales et maximales de rappel (graphique de gauche) et de précision (graphique de droite) obtenues par les testeurs humains pour chaque parti politique.

Ces résultats globaux masquent les différences qui peuvent exister entre les juges humains concernant la reconnaissance de chaque parti. Les graphiques de la figure 4 montrent ces différences en affichant pour chaque parti le minimum et le maximum des valeurs de rappel ou de précision obtenues par les juges humains. Les écarts entre les minima et les maxima montrent les désaccords entre juges, particulièrement accentués en ce qui concerne les valeurs de rappel pour le parti Verts/ALE. Il semble par ailleurs que les testeurs humains ont mieux réussi l'identification des partis PSE (précision maximale de 0,63) et PPE-DE (précision maximale de 0,57), deux partis correspondant au traditionnel clivage gauche/droite[12], que celle des autres partis. Les partis ELDR et Verts/ALE ont été moins bien identifiés, avec des valeurs minimum de précision tombant respectivement à 0,11 et 0,06). Par ailleurs, les juges humains semblent s'être mieux accordés sur l'identification du parti PPE-DE, l'écart entre minimum et maximum étant plus réduit que pour les autres partis.

Coefficient Kappa. Nous avons également confronté les résultats des différents testeurs au moyen du coefficient κ, défini par (Cohen, 1960) et repris par (Carletta, 1996). Ce coefficient permet de mesurer le taux d'accord entre deux juges. Sur cette tâche, le coefficient κ a varié entre -0,14 et 0,24, soit des accords qualifiés de très mauvais à médiocre, avec une majorité d'accords qualifiés de mauvais (pour des valeurs de κ comprises entre 0,02 et 0,17). Les accords entre juges sont donc mauvais sur cette tâche et leurs performances plutôt médiocres dans l'ensemble comme nous l'avons vu au paragraphe précédent (voir tableau 8).

Matrice de confusion. Enfin, à partir des matrices de confusion pour chaque juge humain, nous avons voulu retrouver quelles étaient les confusions entre partis les plus fréquentes. Pour cela, nous avons rassemblé dans le

[12]Le parti PSE – Parti Socialiste Européen étant à gauche tandis que le parti PPE-DE – Parti Populaire Européen (démocrates chrétiens)-Démocrates Européens est un parti de droite.

tableau 9 les répartitions, en pourcentage, de l'effectif de chaque parti parmi les autres partis, suivant les attributions effectuées par les juges humains. Ainsi, pour le parti ELDR (première ligne), les réponses effectivement attribuées à ce parti par les testeurs humains ont été au minimum de 17% et au plus de 42% de l'effectif réel des interventions de ELDR. Toujours concernant l'effectif réel des interventions du parti ELDR, les testeurs humains en ont attribué – à tort – entre 17% et 50%, suivant le juge, au parti PPE-DE, entre 0% et 33% au parti PSE et entre 0% et 42% au parti Verts/ALE. Les différences importantes qu'on observe entre minima et maxima d'attributions rendent évidents les désaccords entre juges. Tous ces éléments montrent la difficulté de cette tâche.

	Partis attribués par les juges humains aux interventions d'un même parti			
Parti	**ELDR**	**PPE-DE**	**PSE**	**Verts/ALE**
ELDR	17% < 42%	17% < 50%	0% < 33%	0% < 42%
PPE-DE	10% < 40%	25% < 45%	5% < 50%	5% < 30%
PSE	16% < 28%	8% < 32%	28% < 60%	4% < 28%
Verts/ALE	0% < 17%	0% < 5%	4% < 29%	17% < 67%

TAB. 9 – Valeurs minimales et maximales d'attribution de l'effectif de chaque parti, dans ce parti et les autres, résultant du test humain.

Une lecture globale de ce tableau permet néanmoins de mettre en évidence des « couples » de partis politiques pour lesquels certains juges ont rencontré des difficultés d'identification. Les erreurs d'identification qui ont été les plus importantes sont les suivantes : PPE-DE au lieu de ELDR (jusqu'à 50% d'interventions ELDR attribuées à PPE-DE), PSE au lieu de PPE-DE (jusqu'à 50% de mauvaises attributions), PPE-DE au lieu de PSE (jusqu'à 32% de mauvaises attributions), et enfin PSE au lieu des Verts/ALE (jusqu'à 29% de mauvaises attributions).

Si l'on établit une échelle des partis politiques, en classant les quatre partis testés de gauche à droite sur l'échiquier politique, nous obtenons la représentation suivante : Verts/ALE (gauche écologique) – PSE (gauche) – ELDR (centre-droit) – PPE-DE (droite). En se référant à cette échelle, il apparaît qu'une partie des erreurs concernent des partis proches sur cette échelle. Mais on observe également des confusions entre les deux « gros » partis que sont le PSE et le PPE-DE, plus dans le sens PPE-DE pris pour des PSE que l'inverse. Enfin, les confusions entre partis situés aux extrémités de notre échelle sont fortement asymétriques : les interventions des Verts/ALE ont rarement été prises pour des interventions des PPE-DE (maximum 5%), en revanche les interventions des PPE-DE ont plus souvent été prises pour des interventions des Verts/ALE (30% maximum).

6.3 Résultats

Cette tâche de détermination du parti politique auquel appartient l'orateur d'une intervention s'est révélée difficile, tant pour les testeurs humains que pour les participants au défi.

Trois équipes ont initialement fait part de leur intention de participer à cette tâche. Seule l'équipe de l'Université de Montréal (D. Forest et al.) a finalement soumis des fichiers de résultats, au nombre de trois. Cette participation repose uniquement sur le corpus en français. Le tableau 10 met en évidence les valeurs de rappel et précision obtenues par cette équipe pour chaque parti politique ainsi que la F-mesure de chacune des soumissions.

Soumission	**Type de valeurs**	**ELDR**	**GUE-NGL**	**PPE-DE**	**PSE**	**Verts/ALE**	**F-mesure**
Nombre de documents attendus		1338	1794	4571	3626	1585	
1	Rappel	0,189	0,393	0,437	0,360	0,233	0,320
	Précision	0,210	0,345	0,447	0,365	0,226	
2	Rappel	0,231	0,332	0,498	0,394	0,207	0,339
	Précision	0,236	0,422	0,452	0,370	0,252	
3	Rappel	0,202	0,376	0,462	0,383	0,243	0,334
	Précision	0,205	0,384	0,462	0,369	0,255	

TAB. 10 – Rappels et précisions obtenus par parti politique pour chacune des trois soumissions de la tâche 3 par l'équipe de Montréal.

Des partis mieux identifiés que d'autres. Nous pouvons constater que les valeurs de rappel et précision augmentent avec le nombre de documents attendus. Plus un parti aura eu de documents disponibles, plus l'apprentissage aura été efficace, et en conséquence meilleurs auront été les résultats du corpus de test. Cette tendance se vérifie sur les trois soumissions pour lesquelles les valeurs de rappel et précision les plus élevées se rapportent aux partis PPE-DE (4571 documents), PSE (3626 documents) et GUE-NGL (1794 documents).

Nous avions observé chez les évaluateurs humains de cette tâche des taux de rappel et précision plus élevés pour les deux principaux partis que sont le PPE-DE et le PSE que pour les autres partis (voir tableau 11).

Juges / Parti	1	2	3	4	5	6
ELDR	0,33/0,42	0,27/0,25	0,11/0,17	0,31/0,42	0,33/0,33	0,18/0,17
PPE-DE	0,57/0,40	0,37/0,35	0,33/0,25	0,56/0,45	0,39/0,45	0,40/0,30
PSE	0,63/0,48	0,42/0,44	0,57/0,32	0,53/0,36	0,41/0,28	0,50/0,60
Verts/ALE	0,14/0,33	0,43/0,50	0,06/0,17	0,29/0,67	0,22/0,33	0,29/0,33

TAB. 11 – Valeurs de rappel et de précision (rappel/précision) obtenues par les six évaluateurs humains pour chaque parti politique.

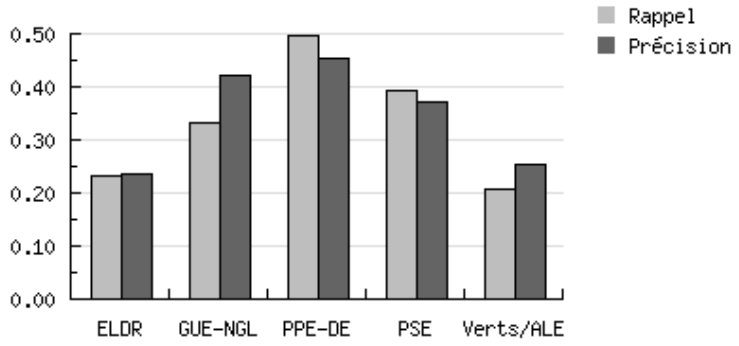

FIG. 5 – Valeurs de rappel et précision obtenues par parti sur la meilleure soumission (n° 2).

La figure 5 montre que les valeurs de rappel dépassent celles de la précision pour les deux principaux partis (PPE-DE et PSE), autrement dit ceux pour lesquels il y a eu plus de documents disponibles (tant dans le corpus d'apprentissage que dans celui de test), alors que c'est l'inverse pour les autres partis. Cela montre que l'attribution d'une intervention à ces partis a été préférentielle par rapport aux partis à plus faible effectif d'interventions.

7 Conclusion

L'édition 2009 du Défi Fouille de Textes a été pleine d'enseignements pour les organisateurs. L'évaluation humaine de la tâche a permis, comme d'habitude, de mieux cerner les problèmes, à défaut d'y trouver toujours des solutions.

À l'exception de la tâche 1 de classification de documents en *objectif* ou *subjectif*, les deux autres tâches se sont révélées difficiles et ont suscité peu de participations. Néanmoins, l'ensemble des données de référence sur les corpus de la tâche 2 de détection des passages subjectifs d'un texte nous semble devoir être intéressante pour la communauté concernée par ce type de problématique. Cet ensemble constitue une pré-annotation de corpus qui nous semble utile. À ce titre, les participants qui ont joué le jeu doivent en être remerciés.

Un autre aspect qui nous semble intéressant est le lien entre le nombre de passages subjectifs d'un texte et l'orientation subjective de ce texte. En effet, si l'abondance de passages clairement subjectifs suffit à donner l'impression que le texte est subjectif, cette abondance n'est cependant pas nécessaire d'après nos tests humains de la section 5.2.

Par ailleurs, les résultats médiocres de la tache 3 d'identification du parti politique d'un orateur montrent que, même dans un contexte parlementaire où les opinions sont supposées s'exprimer clairement, il n'est pas facile d'attribuer le bon parti à l'orateur. En soi, c'est un résultat intéressant.

8 Remerciements

Cet atelier bénéficie du soutien financier du projet CapDigital DoXa (traitement automatique des opinions et sentiments[13], convention DGE n° 08 2 93 0888). Nous exprimons notre gratitude envers le LIP6 (*Laboratoire d'Informatique de l'Université Paris 6*[14]) pour son soutien logistique.

Nous exprimons nos remerciements à la société ELDA (*Evaluations and Language resources Distribution Agency*[15]) pour son implication dans cette campagne d'évaluation au travers de la mise à disposition de ses corpus.

Nous remercions également les testeurs humains *(Arnaud, Béatrice, Cyril, Isabelle, Jean-Baptiste, Martine et Sarra)* qui ont bien voulu prendre un peu de leur temps pour tester les différentes tâches. Merci à Anne « la p'tite » pour la version italienne du site et à Jean-Baptiste pour la version anglaise.

Enfin, nous tenons à souligner combien nous avons apprécié la participation des différentes équipes de cette nouvelle édition, alors que cela représente une surcharge de travail conséquente et non financée dans le cadre de projets.

Références

Berthelin J.-B., Grouin C., Hurault-Plantet M. et Paroubek P. (2008). Human judgement as a parameter in evaluation campaigns. In *Coling 2008 : Proceedings of the workshop on Human Judgements in Computational Linguistics*, p. 17–23, Manchester, UK : Coling 2008 Organizing Committee.

Carletta J. (1996). Assessing agreement on classification tasks : the kappa statistics. *Computational Linguistics*, **2**(22), 249–254.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Nakache D. et Métais E. (2005). Evaluation : nouvelle approche avec juges. In *INFORSID*, p. 555–570, Grenoble.

Paroubek P., Robba I., Vilnat A. et Ayache C. (2008). EASY, Evaluation of Parsers of French : what are the Results ? In European Language Resources Association (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

[13] `http://www.projet-doxa.fr/`
[14] `http://www.lip6.fr`
[15] `http://www.elda.org/`