Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français

Gilles Adda[(1)], Joseph Mariani[(1)], Patrick Paroubek[(1)],
Martin Rajman[(2)], Josette Lecomte[(3)]

[(1)]LIMSI-CNRS
BP 133, F-91403 Orsay Cedex
{gadda,mariani,pap}@limsi.fr

[(2)]Laboratoire d'Intelligence Artificielle
Département Informatique
Ecole Polytechnique Fédérale de Lausanne
CH-1015 Lausanne, Suisse
martin.rajman@epfl.ch

[(3)]CNRS-INaLF
44, Avenue de la Libération, BP 30687
F-54063 Nancy-Cedex
josette.lecomte@inalf.cnrs.fr

Résumé

L'action GRACE est le premier exemple d'application du paradigme d'évaluation aux étiqueteurs morphosyntaxiques pour le français dans le cadre d'une campagne d'évaluation formelle, à participation ouverte et utilisant des données de grande taille. Après une rapide description de l'organisation et du déroulement de l'action ainsi que des problèmes posés par la nécessaire mise en place d'un référentiel commun pour l'évaluation, nous présenterons en détail la métrique *Précision-Décision* qui a été développée dans le cadre de GRACE pour la mesure quantitative des performances des systèmes d'étiquetage. Nous nous intéresserons ensuite aux résultats obtenus pour les participants à la phase de test de la campagne et indiquerons les aspects du protocole d'évaluation qui restent encore à valider sur les données recueillies. Enfin, nous conclurons en soulignant les incidences positives d'une campagne d'évaluation comme GRACE sur le domaine de l'ingénierie linguistique.

1. Action GRACE : organisation et déroulement

L'action GRACE d'évaluation des étiqueteurs morphosyntaxiques pour le français [Adda et al. 1997, Paroubek et al. 1998] a débuté en 1994 à l'initiative de Joseph Mariani (LIMSI) et de Robert Martin (INaLF). L'appel à participation a été lancé en novembre 1995 et a mené à la sélection de 21 équipes (y compris les 2 équipes organisatrices) venant de 5 pays différents (Canada, États-Unis, Allemagne, Suisse, et France). Les systèmes utilisés couvraient un ensemble d'approches très différentes, allant des systèmes stochastiques aux systèmes à base de règles. Certains participants ont même utilisé des analyseurs syntaxiques complets, dont les fonctionnalités ont été adaptées pour produire un marquage morpho-syntaxique. Les organisateurs ont également présenté[1] plusieurs approches minimalistes

[1] Voir note 3

(accès lexical simple et accès lexical complété par quelques règles heuristiques de désambiguïsation) et plusieurs versions de l'étiqueteur développé par E. Brill [Brill 1992], entraîné pour le français par l'INaLF et le Limsi.

La campagne GRACE proprement dite s'est déroulée en 3 phases successives : (1) *La phase d'apprentissage*, pendant laquelle a été distribué aux participants (en janvier 1996) un corpus (dit d'entraînement) d'environ 10 millions de mots, destiné à l'apprentissage et au calibrage des systèmes. (2) *La phase d'essais*, qui comprenait la distribution de données d'évaluation (environ 450 000 formes) qui ont été marquées par les participants durant l'automne 1996. L'objectif de cette phase était de tester le protocole en vraie grandeur et en collaboration avec les participants. Une journée atelier de présentation des résultats qui marquait l'ouverture de l'étape d'adjudication (révision des résultats et modification du protocole), réservée aux seuls participants, s'est déroulée en prélude aux JST'97 [Adda et al. 1997]. (3) *La phase de tests*, qui consistait en la réalisation de l'évaluation proprement dite, selon le protocole arrêté à la fin de la phase d'essais et avec de nouvelles données d'évaluation (environ 650 000 formes) qui ont été marquées par les participants à la fin de l'année 1997. Cette phase s'est déroulée pendant les trois premiers quarts de l'année 1998, et a été ponctuée par une journée d'ateliers (15 mai), au cours de laquelle les premiers résultats ont été présentés et discutés avec l'ensemble des participants, ainsi que par une présentation lors d'une journée de l'ATALA (24 octobre).

Les sources de corpus utilisées pour l'évaluation ont été, d'une part, la base textuelle (FRANTEXT) de l'INaLF (Institut National de la Langue Française) qui comporte des textes littéraires répartis sur 5 siècles, pour environ 180 millions de mots (plus de 3500 unités textuelles), et, d'autre part, le corpus du journal "Le Monde" qui a été distribué sur le CD-ROM ECI/MCI[2]. Le marquage de référence a été réalisé à l'INaLF par une linguiste experte.

La table 1 ci-après donne la liste des participants avec indication de leur pays d'origine, de leur implication (ou non) dans les différentes phases ainsi que le statut (public ou non) de leurs résultats (cette dernière information a été formellement demandée à tous les participants *avant* le début de la phase de tests). Dans la table, le signe « - » indique qu'une information n'est pas pertinente.

Les participants à GRACE	Pays	Essais	Tests	Res. Public
ATT Bell Laboratories	USA	oui	non	-
GREYC – URA 1526 – Université de Caen	FR	oui	oui	Oui
INGENIA-Langage Naturel S.A.	FR	oui	oui	Oui
CRISTAL – GRESEC – Université de Grenoble	FR	oui	oui	Oui
IAI Institüt für Angewandte Informationsforschung – Saarbrücken	D	oui	non	-
CNET, Dépt. LAA/EIA/AIA – France Telecom – Lannion	FR	oui	oui	Oui
XRCE Centre de Recherche Rank Xerox Grenoble – France	FR	oui	oui	Non
LATL – Laboratoire d'Analyse et de Technologie du Langage – Université de Genève	CH	oui	oui	Oui

[2] Le journal « Le Monde » distribue régulièrement ses textes sur CD-ROM. Sont actuellement disponibles l'ensemble des textes parus depuis 1987, soit plus de 200 millions de mots. Ces données peuvent être acquises, soit dans leur totalité (version commerciale « Le Monde sur CD-ROM »), soit en partie (maximum 5 années) auprès de l'ELRA.

LIA Laboratoire Informatique – Université d'Avignon *en collaboration avec* LPL Laboratoire Parole et Langage – Université d'Aix-en-Provence	FR	oui	oui	Non
Société T.GID	FR	oui	non	-
ISSCO – Université de Genève	CH	oui	oui	Oui
Société SYNAPSE – Toulouse	FR	oui	non	-
CLIPS Équipe TRILAN – IMAG –Grenoble	FR	oui	oui	Oui
ILR – Institüt für Linguistik/Romanistik - Universität Stuttgart *en collaboration avec* IMS – Institüt für maschinelle Sprachverarbeitung – Universität Stuttgart	D	oui	non	-
IBM Centre Scientifique – IBM-France	FR	oui	oui	Non
MEMODATA – Caen	FR	non	non	-
Société GSI-ERLI – Paris	FR	oui	non	-
CIT *puis* IRO –Québec	CA	oui	oui	Non
Les participants-organisateurs (évaluation séparée[3])				
INALF – Institut National de la Langue Française – CNRS	FR	oui	oui	Oui
Ecole Nationale Supérieure des télécommunications – Paris *puis* EPFL – École Polytechnique Fédérale de Lausanne	FR *puis* CH	non	non	-
Limsi – Laboratoire pour la Mécanique et les Sciences de l'Ingénieur – CNRS	FR	oui	oui	Oui

Table 1 : La liste des participants

2. La définition d'un référentiel commun

Une évaluation se fait nécessairement par rapport à un étalon. Dans le cas de GRACE, le choix a été fait d'évaluer les systèmes par comparaison, sur un corpus de référence, de l'étiquetage produit par les systèmes avec l'étiquetage (dit de référence) du même corpus, mais produit de façon manuelle par un expert humain. L'évaluation repose donc uniquement sur la comparaison des étiquettes (ou listes d'étiquettes) et ne tient aucun compte du fonctionnement interne des systèmes (approche de type « boîte noire »).

Toutefois, la mise en œuvre effective d'une telle procédure d'évaluation nécessite de définir, en accord avec les participants à l'évaluation, un *référentiel commun* permettant, en particulier, de comparer des systèmes qui n'utilisent pas nécessairement (1) ni les mêmes jeux d'étiquettes et (2) ni les mêmes procédures de segmentation pour définir les unités textuelles auxquelles sont affectées les étiquettes [Habert et al. 1998].

La possibilité de prendre en compte simultanément plusieurs jeux d'étiquettes distincts a été gérée dans GRACE par le biais de la notion de « tables de correspondances » : une description "pivot" (appelée jeu d'étiquettes GRACE) permettant de traduire les jeux

[3] Pour des raisons évidentes de respect de l'impartialité du protocole d'évaluation, les systèmes présentés par les organisateurs ont fait l'objet d'une évaluation séparée dont les résultats ne sont fournis qu'à titre indicatif.

d'étiquettes les uns dans les autres[4] a été définie à partir d'un jeu d'étiquettes initial dérivé du formalisme, proposé par les projets EAGLES [Leech & Wilson 1994] et MULTEXT [Ide & Veronis 1994] et adapté par itérations successives en interaction avec les participants. Le jeu d'étiquettes GRACE est constitué de 312 étiquettes (voir table 2 pour quelques exemples) dont une description complète est donnée dans [Rajman 1997] et est également accessible sur le site GRACE à l'URL http://www.limsi.fr/TLP/grace/.

Sur les 21 participants initialement inscrits pour les essais, 14 ont fourni une table de correspondances et 2 ont modifié leur système de façon à produire directement des étiquettes GRACE (leur table de correspondance est alors la fonction identité). Sur les 13 participants qui ont effectué les tests (organisateurs compris), 10 ont fourni une table de correspondances et 3 ont utilisé la fonction identité.

Pour résoudre le problème de l'alignement des unités lexicales du texte de référence avec celles du texte retourné par le participant, problème lié à l'utilisation de procédures de segmentation différentes, il a été décidé dans GRACE d'utiliser, pour le texte de référence, une segmentation minimale (toute séquence de caractères non séparateurs est une unité, les caractères séparateurs étant l'espace, le retour à la ligne et tous les caractères de ponctuation, y compris l'apostrophe et le trait d'union) associée à un algorithme de réalignement fondé sur l'utilitaire UNIX *diff* de comparaison de fichiers. Par cette méthode, les formes résiduelles qui résistent à tout réalignement sont peu nombreuses (de l'ordre de 2 % au maximum sur les données qui ont été traitées). En plus de ses bonnes performances, cette méthode présente également l'avantage d'identifier et d'extraire automatiquement les formes qui n'ont pu être correctement réalignées (et qui peuvent être alors analysées de façon spécifique).

Par ailleurs, afin de limiter autant que possible l'influence des différences de segmentation sur les résultats de l'évaluation, toutes les formes composées identifiées dans la référence ont fait l'objet d'un double marquage, en constituants et en locution, et, lors de l'évaluation, seules ont été prises en compte les étiquetages des participants qui présentaient une segmentation similaire à l'une des segmentations proposées par l'étiquetage de référence.

3. La mesure quantitative des performances : Précision et Décision

Une fois réalisées les phases de réalignement et de projection des étiquettes dans un jeu commun, la mesure effective des performances des systèmes d'étiquetage peut commencer.

Dans l'approche GRACE, on considère que, pour chaque unité lexicale traitée, un système d'étiquetage peut :

- soit effectuer un étiquetage « strict », c'est-à-dire associer à l'unité lexicale une étiquette unique (correcte ou erronée) ; dans la terminologie GRACE nous disons que le système a produit un étiquetage correct (cas « ok ») ou un étiquetage erroné (cas « erreur ») ;

- soit effectuer un étiquetage « ambigu », c'est-à-dire associer à l'unité lexicale une liste d'étiquettes possibles (liste d'alternatives) ; dans la terminologie GRACE, nous disons que le système produit alors un « silence »[5]. Ce silence sera dit « silence ok » si la liste produite ne contient que des étiquettes correctes, « silence erroné » si la liste ne contient que des étiquettes erronées et « silence vrai » dans les autres cas.

[4] Dans la terminologie GRACE, on parle de projection d'un jeu d'étiquettes dans un autre.

[5] Il est important de noter qu'un « silence » peut être dû à la projection du jeu d'étiquettes du participant vers le jeu d'étiquettes GRACE ; en effet, ce dernier a été construit de façon à être le plus fin possible et les étiquettes des participants correspondent de ce fait souvent à plusieurs étiquettes GRACE.

Sur la base de ces définitions, l'évaluation d'un système d'étiquetage peut alors être quantitativement caractérisée par les 7 grandeurs suivantes :

- nbCas : le nombre total d'occurrences d'unités lexicales ;
- *NONEVAL* : le nombre d'occurrences d'unités lexicales non retenues pour l'évaluation (en raison de problèmes d'étiquetage, de segmentation ou de réalignement) ;
- *OK* : le nombre d'occurrences d'unités lexicales correctement étiquetées (cas « ok ») ;
- *ERR* : le nombre d'occurrences d'unités lexicales étiquetées de façon erronée (cas « erreur ») ;
- SIL_{ok} : le nombre de silences pour lesquels toutes les alternatives sont correctes (« silences ok ») ;
- SIL_{err} : le nombre de silences pour lesquels toutes les alternatives sont erronées (« silences erronés ») ;
- $SILOK_{moy}$: le nombre moyen de silences qui pourraient être transformés en cas « ok » (étiquetage correct) par un choix aléatoire équiprobable effectué, pour chaque silence, parmi les alternatives évaluables proposées par le système d'étiquetage évalué.

Il est important de noter qu'à partir des 7 grandeurs indiquées ci-dessus, il est possible de dériver un certain nombre de grandeurs additionnelles également intéressantes, comme par exemple :

- *SIL* : le nombre total de silences, obtenu à partir de l'égalité
  $nbCas = NONEVAL + OK + ERR + SIL$;
- SIL_{sil} : le nombre total de silences contenant un mélange d'alternatives correctes et d'alternatives erronées (« silences vrais »), obtenu à partir de l'égalité
  $SIL = SIL_{ok} + SIL_{err} + SIL_{sil}$;
- $SILERR_{moy}$: le nombre moyen de silences qui pourraient être transformés en cas « erreur » par un choix aléatoire équiprobable, obtenu à partir de l'égalité
  $SIL = SILOK_{moy} + SILERR_{moy}$.

La caractérisation des résultats de l'évaluation d'un système d'étiquetage par le biais des 7 grandeurs (*nbCas*, *NONEVAL*, *OK*, *ERR*, SIL_{ok}, SIL_{err}, $SILOK_{moy}$) a l'avantage d'être très précise, mais également l'inconvénient d'être relativement peu maniable, en particulier dans le cas d'une démarche comparative, du fait de la difficulté de la visualisation des résultats dans un espace à 7 dimensions.
Pour cette raison, l'un des objectifs importants de l'action GRACE a été de proposer et de tester une représentation des résultats d'évaluation ayant la propriété d'être à la fois plus compacte et plus maniable tout en préservant l'information quantitative obtenue lors de l'évaluation.
Deux nouvelles grandeurs, la *précision* et la *décision*, librement inspirées des notions de « précision » et de « rappel » utilisées en recherche documentaire, ont été introduites à cet effet :

- la précision mesure la proportion d'étiquetages corrects parmi les étiquetages stricts (c.-à-d. non ambigus) ;
- la décision mesure la proportion d'étiquetages stricts parmi l'ensemble de tous les étiquetages (stricts et ambigus) évaluables.

De façon informelle, la décision indique donc dans quelle mesure un système d'étiquetage produit des résultats directement exploitables (c.-à-d. exploitables sans traitement additionnel pour lever les ambiguïtés d'étiquetage), alors que la précision quantifie la « justesse » de ces résultats.

De façon formelle, la précision P et la décision D d'un système d'étiquetage sont définies par :

- $P = OK / (OK + ERR)$
- $D = (OK + ERR) / (OK + ERR + SIL)$

Par définition, *P* et *D* varient entre 0 et 1 ; le couple (*P*,*D*), également appelé *score (P,D)*, peut donc être représenté sous la forme d'un point dans la portion de plan [0,1]×[0,1]. Le score (*P*,*D*) est utilisé dans GRACE pour caractériser le résultat de l'évaluation d'un système d'étiquetage.
Il est à noter cependant que l'indication d'un score (*P*,*D*) n'est pas équivalente à la donnée des 7 grandeurs *nbCas*, *NONEVAL*, *OK*, *ERR*, $nSIL_{ok}$, SIL_{err}, $SILOK_{moy}$. Toutefois, pour *nbCas* et *NONEVAL* fixés, la connaissance de *P* et de *D* permet de reconstruire les grandeurs *OK*, *ERR* et *SIL* par le biais des formules suivantes :

- $OK = (nbCas - NONEVAL) * P * D$
- $ERR = (nbCas - NONEVAL) * (1 - P) * D$
- $SIL = (nbCas - NONEVAL) * (1 - P)$

Pour permettre la reconstruction des autres grandeurs définies, il a été choisi dans GRACE d'indiquer 3 valeurs de précision supplémentaires :

- P_{min}, la précision qu'aurait le système si tous les silences étaient transformés en cas « erreurs » (sauf les « silences ok » pour lesquels, par définition, une telle transformation n'est pas possible) ; P_{min} est donc la précision minimale que pourrait avoir le système d'étiquetage évalué si l'on force sa décision à 1 ;
- P_{max}, la précision qu'aurait le système si tous les silences (sauf les « silences erronés ») étaient transformés en des cas « ok » ; P_{max} est donc la précision maximale que pourrait avoir le système d'étiquetage évalué si l'on force sa décision à 1 ;
- P_{moy}, la précision qu'aurait le système si les silences étaient transformés de façon aléatoire (équiprobable sur les alternatives) en cas « erreur » ou « ok » ; P_{moy} est donc la précision moyenne qu'aurait le système d'étiquetage s'il était complété, pour les cas de silence, par une procédure de choix aléatoire (équiprobable) parmi les alternatives proposées.

Formellement, P_{min}, P_{max} et P_{moy} sont définis par :

- $P_{min} = (Ok + SIL_{ok}) / (OK + ERR + SIL)$
- $P_{max} = (OK + SIL - SIL_{err}) / (OK + ERR + SIL)$
- $P_{moy} = (OK + SILOK_{moy}) / (OK + ERR + SIL)$

Pour *nbCas* et *NONEVAL* fixés, l'indication des 4 scores (*P*,*D*), (P_{min},*1*), (P_{moy},*1*) et (P_{max},*1*), autrement dit des 5 valeurs *P*, *D*, P_{min}, P_{moy}, et P_{max}, est alors bien équivalente à la donnée des 7 grandeurs *nbCas*, *NONEVAL*, *OK*, *ERR*, SIL_{ok}, SIL_{err}, $SILOK_{moy}$.

Ce résultat important a été systématiquement utilisé dans le cadre de GRACE, où le résultat complet de l'évaluation d'un système d'étiquetage est représenté sous la forme de 4 points (correspondant aux 4 scores (*P*,*D*) indiqués ci-dessus) dans la portion de plan [0,1]×[0,1]. Lorsque c'est nécessaire, les valeurs *nbCas* et *NONEVAL* sont quant à elles explicitement mentionnées dans les légendes des graphes produits (*NONEVAL* étant indiquée sous la forme du pourcentage $\%NONEVAL = 100 * NONEVAL / nbCas$).

La représentation des résultats de l'évaluation par le biais de scores (*P*,*D*) offre de nombreux avantages. Tout d'abord, cette représentation peut être réalisée (sans perte d'information) dans un plan, ce qui facilite grandement la visualisation des résultats. D'autre part, il est également important de noter que cette représentation se prête de façon assez naturelle à une démarche comparative. En effet, les 4 scores (*P*,*D*) caractérisant un système permettent de construire, de façon simple, dans le plan [0,1]×[0,1], la « zone de fonctionnement » du système, c'est-à-dire la région du plan dans laquelle se situera son score (*P*,*D*) selon la manière dont seront résolues les ambiguïtés d'étiquetage résiduelles correspondant aux silences.
Plus précisément, si l'on note *ok* (resp. *err*) le nombre de silences transformés en cas « ok » (resp. cas « erreur ») par le biais d'un post-traitement de désambiguïsation après étiquetage, le score (*P*,*D*) résultant pour le système est donné par les formules :

- $P(ok,err) = (OK+ok) / (OK+ERR+ok+err)$
- $D(ok,err) = (OK+ERR+ok+err) / (OK+ERR+SIL)$

avec les contraintes suivantes sur *ok* et *err* :
- $0 \leq ok \leq SIL$
- $0 \leq err \leq SIL$
- $0 \leq ok+err \leq SIL$

si l'on considère que le post-traitement des silences peut remettre en cause les listes d'alternatives associées à chacun des silences (c.-à-d. si l'alternative choisie après post-traitement n'est pas nécessairement dans la liste produite par le système),
ou encore :
- $0 \leq ok \leq SIL - SIL_{err}$
- $0 \leq err \leq SIL - SIL_{ok}$
- $0 \leq ok+err \leq SIL$

si l'on considère que le post-traitement des silences doit respecter les listes d'alternatives produites.

La région du plan [0,1]×[0,1] correspondant au second jeu, plus restrictif, de contraintes est indiquée à la figure 1. Les équations exactes des frontières des régions (portions de courbes *A*, *B*, *C*, *D* et *E*) peuvent bien sûr être dérivées des formules et inéquations indiquées ci-dessus, mais, dans une première approche du moins, la zone de fonctionnement d'un système d'étiquetage peut être approximée par un triangle dont les trois sommets sont les points P_1, P_2 et P_3 correspondant respectivement aux scores (P_{min},*1*), (*P*,*D*) et (P_{max},*1*).

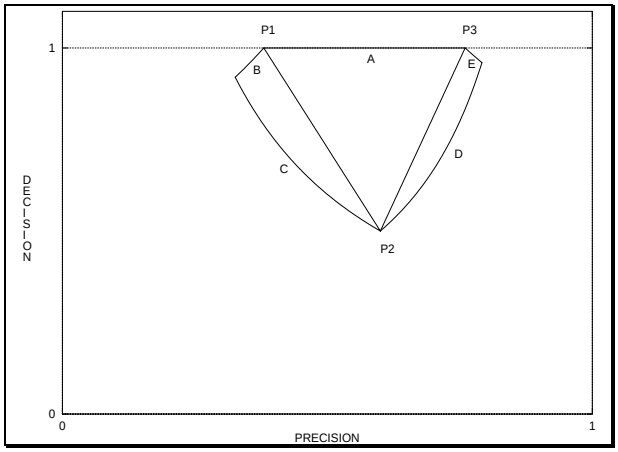


Figure 1: zone de fonctionnement

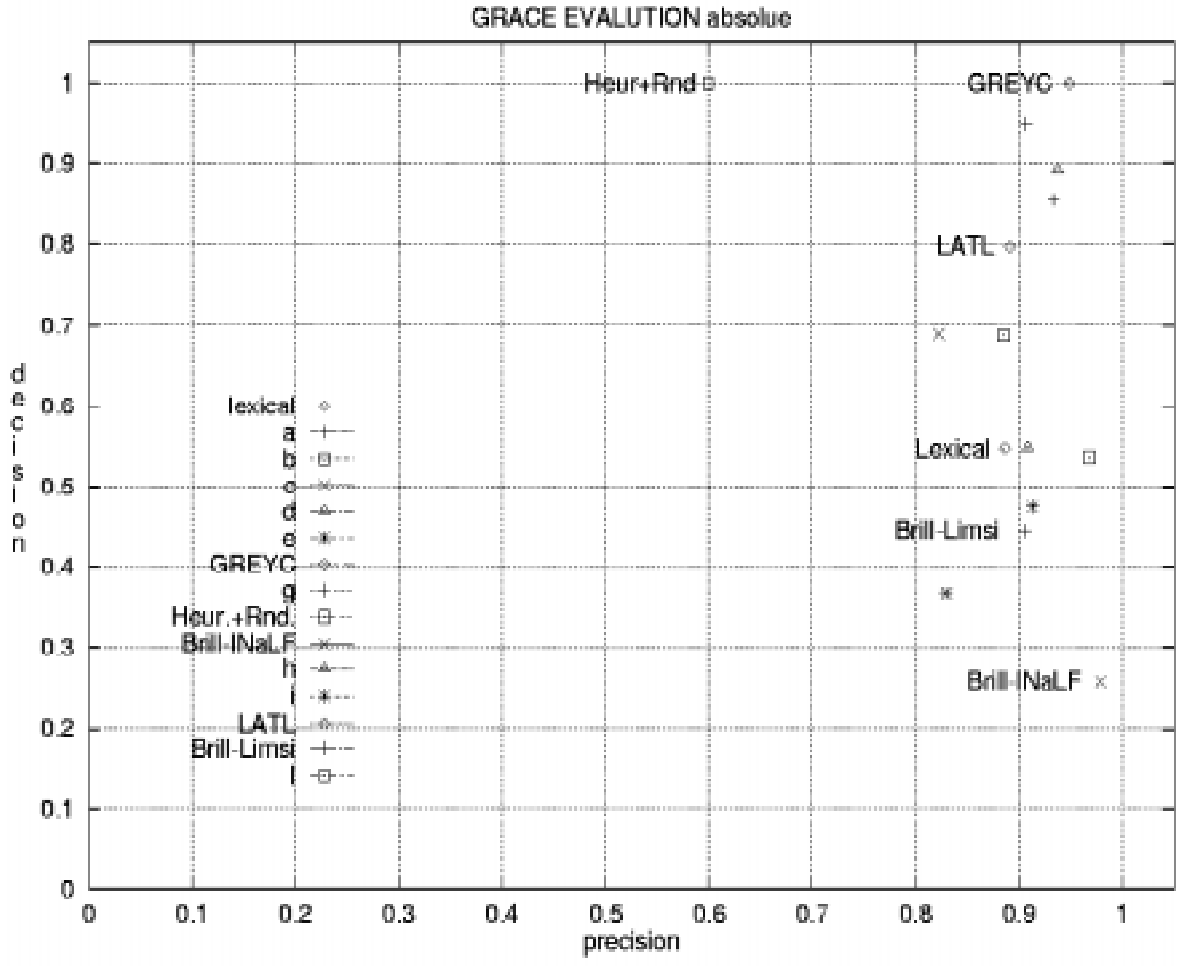
La définition de la notion de « zone de fonctionnement » est importante car elle permet de donner un support concret (analyse des positionnements respectifs et des intersections dans le plan [0,1]×[0,1]) à l'étude comparative des performances obtenues par plusieurs systèmes.

De façon complémentaire, les différentes précisions (P_{min}, P_{moy} et P_{max}) obtenues en forçant la décision à *1* (c'est-à-dire aucun silence résiduel) peuvent également servir de support pour une étude comparative des résultats, comme par exemple une analyse des variances permettant une meilleure évaluation des différences significatives entre les performances observées.

5. Les résultats

Les résultats de l'évaluation ont été présentés une première fois aux seuls participants en mai 1998 lors d'une journée atelier (début de l'étape d'adjudication), puis de manière publique (mais avec des résultats anonymes) lors de la journée ATALA du 24 Octobre 1998, intitulée « Le marquage morphosyntaxique: résultats de l'évaluation GRACE et perspectives ». Les résultats ont également été publiés sur le WEB (URL *http://www.limsi.fr/TLP/grace/*) en novembre 1998.

Dans le graphique ci-dessous sont représentés les scores (*P*,*D*) calculés pour les systèmes participants à la phase de test. Outre les systèmes présentés par les organisateurs (« Brill-INaLF » et « Brill-Limsi »), sont aussi identifiées les approches minimalistes (« Lexical », accès lexical simple, et «Heur+Rnd », accès lexical et règles heuristiques) ainsi que les systèmes des participants ayant accepté de rendre publics leurs résultats et qui les ont validés.



Il est à noter que les résultats présentés ci-dessus sont encore des résultats « bruts » qui nécessitent d'être analysés de façon détaillée en collaboration avec les participants.

En effet, outre le fait que les mesures décrites ci-dessus sont dépendantes de la qualité du marquage morphosyntaxique de référence, la projection des étiquettes des systèmes participants dans le jeu de référence introduit également un biais potentiel dans les mesures de performance. Une analyse critique de l'emploi des tables de correspondances et du biais introduit lors de la « projection » est présentée dans [Bertier & Lallich-Boidin 1998].

De plus, les opérations de réalignement peuvent aussi produire un biais lié à la granularité de l'étiquetage produit par le système du fait de la nature et de la taille des unités lexicales qu'il prend en compte. En particulier, il peut arriver que la segmentation de référence décompose une unité lexicale du participant en plusieurs unités lexicales de référence et il se produit alors un phénomène d'amplification du nombre d'unités qui peut accroître artificiellement les différences observées entre les systèmes.

La solution adoptée dans GRACE pour ces problèmes est de définir, en plus de l'évaluation mentionnée jusqu'à présent (appelée dans GRACE l'évaluation « absolue ») qui utilise les projections dans le jeu d'étiquettes de référence et la segmentation de référence, une autre forme d'évaluation, appelée évaluation « relative » qui utilise la segmentation et le jeu d'étiquettes du participant pour projeter les étiquettes de référence à l'aide de tables de correspondances « inverses » produites automatiquement par l'algorithme décrit dans [Adda et al. 1997]. L'objectif de l'évaluation relative est de compléter le protocole d'évaluation en y introduisant une approche de type « boîte transparente » (*glass box*) plus sensible au jeu d'étiquettes et à la segmentation du participant.

De même, l'évaluation de l'influence, sur les scores obtenus, de leur calcul sur un nombre limité de cas (les seules données de référence) est effectuée, dans le cadre de GRACE, par une analyse de la variance par la méthode Delta non paramétrique [Davison & Hinkley 1997] et par le calcul de coefficients d'accord comme le Kappa [Carletta 1996].

Jusqu'à présent, les efforts dans GRACE se sont concentrés sur la mise en œuvre de l'évaluation absolue. Cependant, le test et la validation des techniques d'évaluation relative sont en cours et donnent des résultats préliminaires encourageants. Toutefois, faute de temps, elles n'a pas encore pu être entièrement testée et validée sur les données des participants.

6. Conclusion

La première tentative d'application du paradigme d'évaluation à l'assignation de catégories morphosyntaxiques pour le français, dans une campagne ouverte et avec des données de grande taille, affiche un bilan très positif. Une communauté internationale de chercheurs et de développeurs s'est constituée autour de ce thème fédérateur. Parmi les 21 institutions participantes qui étaient présentes au début, 13 ont suivi la campagne jusqu'à son terme, et ce malgré sa durée (4 ans) et le manque de financement (seuls étaient pris en charge les déplacements) ! Bien que le problème de la comparaison des systèmes d'assignation n'ait pas été complètement résolu (les différences de segmentation entre systèmes peuvent par exemple augmenter artificiellement les différences de performances observées), des progrès certains ont été effectués, en particulier le test en vraie grandeur du paradigme d'évaluation, l'adoption d'un format standard d'annotation morphosyntaxique, la définition d'une nouvelle mesure (précision/décision), la construction d'une boîte à outils d'évaluation automatique, qui sera valorisée et rendue disponible dans le cadre du projet européen ELSE, et finalement avec la production d'une ressource linguistique qui n'existait pas auparavant : un corpus annoté, d'environ 1 million de formes. Ce corpus a été obtenu à moindre coût par le traitement des données qui ont été marquées par les participants dans le cadre de la campagne d'évaluation (projet MULTITAG du CNRS).

Description morphosyntaxique	**Etiquette**
Nom commun masculin singulier	`Ncms`
Nom propre féminin singulier	`Npfs`
Verbe 1ère personne sing. présent indicatif	`Vmip1s-`
Verbe auxiliaire participe présent	`Vmpp---`
Adjectif qualificatif masculin singulier	`Afpms`
Adjectif possessif féminin singulier	`Ds1fss-`
Pronom relatif masculin singulier	`Pr-ms--`
Article défini masculin singulier	`Da-ms-d`
Adverbe générique positif	`Rgp`
Préposition	`Sp`
Conjonction de coordination	`Cc`

Table 2: exemples d'étiquettes GRACE

Références

Adda G., Lecomte J., Mariani J., Paroubek P., Rajman M. (1997), Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de Parties du Discours pour le Français, *1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'Aupelf-Uref*, Avignon.

Bertier M. & Lallich-Boidin G. (1998), A Paradox Raised by the Evaluation of taggers, *1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

Brill E. (1992), A Simple Rule-Based Part of Speech Tagger, *3rd Conference on Applied Natural Language Processing*, Trento, Italy.

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

Davison A.C. & Hinkley D.V. (1997), *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, United Kingdom.

Habert B., Adda G., Adda-Decker M., Boula de Mareuil P., Ferrari S., Ferret O., Illouz G., Paroubek P. (1998), Towards Tokenization Evaluation, *1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

Ide N. & Véronis J. (1994), MULTEXT: Multilingual Text Tools and Corpora, *15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan.

Leech G. & Wilson A. (1994), EAGLES Morphosyntactic Annotation, Draft - Work in Progress, *Draft technical report EAG-CSG/IR-T3.1*, Lancaster.

Paroubek P., Adda G., Mariani J., Lecomte J., Rajman M. (1998), The GRACE French Part-Of-Speech Tagging Evaluation Task, *1st International Conference on Language Resources and Evaluation (LREC)*,Granada, Spain.

Rajman M. (1997), Format de description lexicale pour le français – Partie 2 : description morphosyntaxique, technical report GRACE, http://www.limsi.fr/grace/.