ÉVALUATION NUMÉRIQUE DE LA RÉSOLUTION DE LA RÉFÉRENCE : CRITIQUES ET PROPOSITIONS

ANDREI POPESCU-BELIS*

Résumé - Abstract

Le regroupement des expressions référentielles d'un texte qui désignent une même entité commence à être à la portée des systèmes informatiques. L'évaluation de cette capacité est vue ici comme la comparaison de la réponse du système avec une réponse attendue par l'évaluateur. Après avoir défini le cadre théorique, nous examinons trois mesures de qualité existantes, dont une ayant servi dans une campagne d'évaluation (MUC). Nous proposons ensuite de nouvelles mesures. Puis, nous étudions différentes généralisations. À la suite de ces arguments théoriques, des exemples concrets sont donnés à la fin.

It has become possible for a computer program to group together some of the referring expressions which denote the same entity in a given discourse. The evaluation of this capacity is viewed here as the comparison between the system's response and the one expected by the evaluator. After defining the theoretical framework, three existing quality measures are analyzed, one of which has been used in an evaluation campaign (MUC). New measures are then proposed, and various generalizations are examined. Following the theoretical arguments, examples are given in the end.

Mots Clefs - Keywords

Résolution de la référence, Évaluation, Mesures formelles de qualité

Reference resolution, Evaluation, Formal quality measures

---

* LIMSI–CNRS, Groupe Langage et Cognition, BP 133, 91403 ORSAY Cedex. Mel : popescu@limsi.fr .

Je remercie Benoît Habert, Christian Jacquemin, Patrick Paroubek et deux évaluateurs anonymes, qui ont grandement contribué à l'amélioration de cet article. Je remercie également Anne Reboul, Isabelle Robba et Gérard Sabah pour notre collaboration sur la référence (projet Cervical du GIS - Sciences de la Cognition).

INTRODUCTION

La compréhension d'un texte par un humain ou un programme informatique passe par l'identification des entités ou objets dont traite le texte, et qui sont dénotés par des expressions du texte. Cette faculté d'identification est tellement élémentaire qu'elle est évaluée chez un individu par des tests généraux de compréhension, portant plutôt sur les propriétés des entités. Les programmes informatiques n'ont pas encore atteint ce niveau, mais il existe déjà de nombreux systèmes cherchant à identifier les entités d'un texte, en vue de tâches plus complexes. Se pose alors le problème de l'évaluation en soi de cette capacité, par l'examen de la réponse du système. Le but du présent travail est d'examiner différentes solutions algorithmiques apportées au problème de l'évaluation de la résolution de la référence, de montrer leurs limites, puis de proposer de nouvelles méthodes et des généralisations.

Après une brève description du problème de la référence (§1.1) et un exemple (§1.2), nous posons le problème de l'évaluation (§2). Des critères de cohérence pour l'évaluation numérique, utilisés par la suite, sont énoncés (§2.3). Un cadre théorique pour l'évaluation de la référence est ensuite proposé (§3.1) et illustré sur l'exemple donné (§3.2). Vient ensuite l'analyse proprement dite des mesures d'évaluation existantes (§4.1 à §4.3), parmi lesquelles nous détaillons la mesure des conférences MUC. Afin de dépasser les limites de ces mesures, nous proposons quatre nouvelles mesures (§4.4 à §4.7), parmi lesquelles nous détaillons une mesure dite « à noyaux » et une mesure entropique. Plusieurs généralisations sont ensuite analysées : le cas où l'on ne compte pas les expressions non coréférentielles dans le score final (§5.2), le cas où le système n'a pas identifié toutes les expressions attendues (§5.3) et le cas où l'on considère plusieurs types de coréférences (§5.4). À la fin, nous donnons des exemples numériques sur plusieurs textes, certains construits pour la circonstance (§6.1), d'autres réels et plus longs, soumis à notre système de résolution de la référence (§6.2).

1. APPROCHE COMPUTATIONNELLE DE LA REFERENCE

1.1. Position du problème

Le concept de référence est classique en philosophie, sans qu'un consensus semble se dégager sur sa nature ou son fonctionnement. Dans une optique plus applicative, nous entendrons par référence le lien entre une *expression référentielle* (ER) d'un texte et une *entité* du monde évoqué par le texte (cf. Evans G. 1985, Reboul A. 1998, Récanati F. 1993). Les *entités* sont ici, de façon générale, les perceptions « stables dans le temps » (Givón T. 1979:320-321) – objets ou personnages – dont on prédique des propriétés ou des relations (thème *vs.* rhème, groupe nominal *vs.* verbe).

Le modèle que nous défendons (Popescu-Belis A. *et al.* 1998) représente un compromis entre la pertinence linguistique et la faisabilité informatique. Notre système aborde (pour commencer) la résolution de la référence comme construction des ensembles d'ER désignant une même entité. Ces ensembles ou *classes* peuvent ensuite constituer la base de traitements plus sophistiqués, comme dans les systèmes de R. Gaizauskas *et al.* (1995) et

S. Luperfoy (1992), notamment par l'introduction d'une structure complexe modélisant chaque entité, que nous nommons *représentation mentale*.

Il existe une approche encore plus élémentaire, destinée au traitement automatique, qui vise seulement à établir des *liens* dits de *coréférence* entre paires d'ER (en général consécutives) qui désignent la même entité (Hirschman L. 1997). Cette conception sous-tend fréquemment la résolution de l'anaphore pronominale (Sidner C.L. 1983, Mitkov R. 1998). Les approches par classes et par liens sont équivalentes puisque la connaissance des liens permet de déduire les classes, et une linéarisation des classes conduit à des liens. En revanche, en tant que technique de résolution, l'approche par classes possède une plus grande pertinence cognitive, comme cela ressort des analyses pragmatiques de la référence (Reboul A. 1998) ; nous défendons son efficacité dans (Popescu-Belis A. *et al.* 1998).

1.2. Étude d'un exemple

Afin de préciser d'emblée ces notions, nous avons noté et numéroté les ER du texte suivant (extrait d'un topo-guide d'alpinisme) :

> *Le sommet Ouest$_{(1)}$ se trouve à 3854m. Pour l'$_{(2)}$atteindre, emprunter sur 150m un petit couloir$_{(3)}$ qui$_{(4)}$ est souvent glacé. Ce couloir$_{(5)}$ démarre derrière le sommet Sud$_{(6)}$ (3742m), qui$_{(7)}$ est, lui$_{(8)}$, facile à atteindre. Ce deuxième sommet$_{(9)}$ est bien visible, car il$_{(10)}$ est très saillant. Pour rejoindre ce petit bastion$_{(11)}$, on doit le$_{(12)}$ viser depuis le grand couloir inférieur$_{(13)}$, assez facile à gravir. Bien qu'il$_{(14)}$ soit initialement large, celui-ci$_{(15)}$ se$_{(16)}$ resserre peu à peu. Attention, ce rassurant entonnoir$_{(17)}$ reste très longtemps enneigé.*

Ici, la « compréhension » du texte devrait aboutir par exemple à la construction d'un schéma faisant figurer les deux sommets et les deux couloirs. La résolution de la référence sera la construction d'ensembles d'ER désignant la même entité (ER dites coréférentes) ; par opposition, la résolution de la coréférence détermine des liens deux à deux entre ER coréférentes. Le résultat correct ou *clé*, consiste ici en quatre ensembles ou classes :

> K_1) Le sommet Ouest$_{(1)}$ + l'$_{(2)}$
> K_2) un petit couloir$_{(3)}$ + qui$_{(4)}$ + Ce couloir$_{(5)}$
> K_3) le sommet Sud$_{(6)}$ + qui$_{(7)}$ + lui$_{(8)}$ + Ce deuxième sommet$_{(9)}$ + il$_{(10)}$ + ce petit bastion$_{(11)}$ + le$_{(12)}$
> K_4) le grand couloir inférieur$_{(13)}$ + il$_{(14)}$ + celui-ci$_{(15)}$ + se$_{(16)}$ + ce rassurant entonnoir$_{(17)}$

Supposons qu'un système produise avec ces ER la *réponse* suivante :

> R_1) Le sommet Ouest$_{(1)}$ + l'$_{(2)}$ + le sommet Sud$_{(6)}$ + qui$_{(7)}$ + lui$_{(8)}$ + Ce deuxième sommet$_{(9)}$ + il$_{(10)}$
> R_2) un petit couloir$_{(3)}$ + qui$_{(4)}$ + Ce couloir$_{(5)}$ + ce petit bastion$_{(11)}$ + le$_{(12)}$

+ *le grand couloir inférieur*$_{(13)}$ + *il*$_{(14)}$ + *celui-ci*$_{(15)}$ + *se*$_{(16)}$
R_3) *ce rassurant entonnoir*$_{(17)}$

Comment estimer ici la qualité du système ? Comment, en d'autres termes, comparer la clé et la réponse ? On peut par exemple remarquer que si on élimine deux appariements erronés mais « explicables » (*sommet Ouest / sommet Sud* et *petit couloir / petit bastion*), les autres expressions paraissent être traitées correctement. Le problème d'une mesure de la proximité entre clé et réponse se pose donc avec acuité.

2. LA QUESTION DE L'EVALUATION

2.1. Évaluation et mesure

Avant d'entrer dans des considérations plus techniques, il est utile de situer notre problème dans la perspective plus large de l'évaluation des systèmes, perspective que nous avons développée ailleurs (Popescu-Belis A. *à paraître*). On distingue généralement plusieurs approches de l'évaluation : outre l'opposition « boîte noire » / « boîte de verre », on peut citer l'évaluation centrée sur l'utilisateur, ou l'évaluation de l'impact scientifique ou économique.

Pour des analyses et des exemples concernant ces différentes approches, on pourra se reporter au livre de K. Sparck Jones et J. Galliers (1996), et aux actes de conférences spécialisées, comme la Première Conférence Internationale sur les Ressources et l'Évaluation Linguistiques, LREC'98 (Rubio A. *et al.* 1998), ou les Journées Scientifiques et Techniques FRANCIL, bilan d'actions francophones d'évaluation (FRANCIL 1997). Pour l'évaluation fondée sur l'appréciation des utilisateurs, on peut évoquer la méthodologie proposée dans le projet européen EAGLES (EWG 1996)

Alors que l'évaluation de type « boîte de verre » étudie plutôt la structure, les connaissances et les techniques d'un système, l'évaluation de type « boîte noire » se limite aux résultats produits par le système. Ceux-ci sont mesurés sur une tâche et des données précises et sont convertis en scores numériques ou en appréciations discrètes. On effectue souvent des évaluations compétitives, regroupées en campagnes collectives, par exemple MUC, TREC, TDT[1] (cf. Hirschman L. 1998). Leur intérêt est de comparer concrètement l'aptitude de différentes techniques à résoudre un même problème, afin de faire avancer les recherches et donner une idée concrète du niveau atteint.

Pour évaluer la capacité d'un programme à effectuer certaines tâches, on effectue en général une ou plusieurs *mesures* de ses performances sur un ensemble de données, suivies d'un *bilan*. Nous concevons (cf. Popescu-Belis A. *à paraître*) une mesure de qualité comme une correspondance entre un ensemble de valeurs de qualité objectives, rendues en langue par « parfait », « bon », « moyen », « mauvais », « nul » et un ensemble

[1] MUC : Message Understanding Conferences (cf. §2.2), TDT : Topic Detection and Tracking Project (segmentation, détection et suivi de sujets dans un flot d'information), TREC : Text Retrieval Conferences (recherche d'information).

d'appréciations ou notes, par convention ici l'intervalle [0%; 100%] ou un ensemble discret. Les valeurs de qualité objectives, selon les buts de l'évaluation, peuvent être théoriquement déterminées en demandant à un grand nombre d'experts leur avis sur le système évalué, puis en calculant la moyenne des réponses. Une mesure de qualité calculable automatiquement vise précisément à se substituer à ce processus de consultation.

2.2. Évaluation modulaire dans les conférences MUC

Les *Message Understanding Conferences* (Conférences sur la compréhension de messages) ont proposé une évaluation compétitive de la capacité d'un système à « comprendre » des articles courts sur un thème donné[2]. La tâche consistait à remplir un schéma ou formulaire donné à l'avance avec des données extraites du texte examiné (cf. Grishman R. et Sundheim B. 1996, MUC-6 1995). Plusieurs sous-tâches ou étapes de traitement étaient distinguées, par exemple l'étiquetage des noms d'entités, la détection des coréférences, le remplissage des nombreux champs de formulaire contenant les acteurs et attributs d'une situation. Les organisateurs avaient constitué des jeux de données et de réponses attendues (clés) à chaque étape.

Lorsque cela est possible, l'évaluation des étapes intermédiaires, par module, permet d'obtenir des scores plus informatifs. Une telle évaluation est possible lorsque les systèmes possèdent le même découpage en sous-tâches, mais elle est coûteuse, nécessitant la préparation de clés pour chacune des étapes. D'autre part, si l'on souhaite évaluer chaque module, il faut à notre avis évaluer les modules homologues sur les mêmes données, et non sur celles produites par les modules antérieurs de son système. Dans ce dernier cas, on risque de pénaliser injustement les modules qui sont précédés par des modules peu efficaces.

La résolution de la (co)référence était une des sous-tâches de MUC-6 et 7. Celle-ci dépend fortement de la bonne identification des expressions référentielles du texte traité, et conditionne à son tour le remplissage correct des formulaires MUC. Son évaluation fine devrait utiliser le même jeu d'ER pour tous les participants, contrairement aux tests MUC. L'identification des ER ne fait pas partie *stricto sensu* de la tâche de résolution de la coréférence, et l'on devrait par conséquent évaluer seulement le groupement correct, au sens de la validité des coréférences. Toutefois, nous étudions l'option contraire au §5.3.

2.3. Critères de cohérence des mesures d'évaluation

Des critères généraux de cohérence pour les mesures de qualité, résumés ci-après, permettent de départager plusieurs mesures, afin d'en choisir une pour une session d'évaluation. Ces critères de *méta-évaluation* sont développés dans (Popescu-Belis A. *à paraître*).

---

[2] Il y a eu sept campagnes d'évaluation MUC, sur l'anglais, chacune suivie d'une conférence (actes publiés, *http://www.muc.saic.com*). La dernière, MUC-7 en 1998, a regroupé dix-huit participants, en majorité américains (aucune équipe française).

Afin de mesurer une certaine capacité d'un système, on lui soumet des données D appartenant à un certain domaine Δ, et on apprécie sa réponse $rep(D)$ à l'aide de la mesure m. Nos critères portent sur ces mesures m, qui visent à faire correspondre à une « qualité objective » du système une valeur dans [0%; 100%] (cf. §2.1). Nous nous limitons au cas où $m(rep(D))$ mesure la proximité entre la réponse du système $rep(D)$ et l'ensemble des réponses correctes ou acceptables $REP_{CORR}(D)$ que les humains savent construire. Les critères doivent être vérifiés quelle que soit la donnée $D \in \Delta$, et sont sujets à preuves, contre-exemples ou arguments.

2.3.1. Borne supérieure de la mesure (BS)

Lorsqu'une réponse $rep(D)$ est parfaite, elle doit recevoir le score maximal. La condition réciproque doit aussi être satisfaite : le score maximal doit être atteint seulement par les réponses parfaites. La clé étant par hypothèse connue exactement, (BS) est un critère qui peut être prouvé pour une mesure m donnée :

(BS) $\quad m(rep(D)) = 100\% \Leftrightarrow rep(D) \in REP_{CORR}(D)$

2.3.2. Borne inférieure de la mesure (BI)

Inversement, une mesure cohérente doit attribuer le score minimal (0% ou la plus mauvaise note) aux réponses que les humains considèrent aussi comme « les plus mauvaises », et réciproquement. Si $REP_M(D)$ est l'ensemble des réponses « les plus mauvaises », on demande que :

(BI) $\quad m(rep(D)) = 0\% \Leftrightarrow rep(D) \in REP_M(D)$

Tel quel, le critère (BI) est seulement sujet à argumentation, puisqu'il dépend de la définition de $REP_M(D)$, que l'on peut rarement préciser, du fait qu'on s'y intéresse peu en général. Si l'on observe que

$$m(rep(D)) = 0\% \Leftrightarrow rep(D) \in m^{-1}(0\%),$$

le critère (BI) revient à comparer $m^{-1}(0\%)$, i.e. les réponses notées avec 0%, avec $REP_M(D)$. On peut détailler ce critère. D'abord on demande que :

(BI-1) $\quad m^{-1}(0\%) \subset REP_M(D)$

c'est-à-dire que toutes les réponses notées 0% soient vraiment « très mauvaises ». $REP_M(D)$ étant difficile à préciser, ce critère sera rarement examiné, tout au plus sur quelques exemples. Mais on doit aussi avoir :

(BI-2) $\quad m^{-1}(0\%) \supset REP_M(D)$
(BI-2′) $\quad$ *les mauvaises réponses doivent recevoir des scores faibles*

à savoir toutes les réponses « très mauvaises » sont bien notées 0%. Ce critère est plus facile à étudier à l'aide d'exemples concrets de mauvaises réponses, par exemple : pas de traitement, traitement aléatoire ou trivial, etc. Le critère (BI-2′) est une approximation du critère (BI-2).

Une question qui conditionne (BI-2) est la suivante : existe-t-il réellement des réponses qui sont notées 0% ? Si la réponse est négative, il est clair que (BI-2) ne peut avoir lieu, à moins que toute réponse soit bonne ($REP_M(D) = \varnothing$). D'où le critère suivant, démontrable, et son approximation :

(BI-3) $m^{-1}(0\%) \neq \varnothing$

(BI-3′) *les scores minimaux doivent être faibles*

Lorsque la mesure de qualité est une moyenne de plusieurs mesures sur *rep*(*D*) (par exemple une moyenne arithmétique), pour qu'elle puisse atteindre 0%, il faut vérifier que les mesures qui la composent peuvent atteindre 0% simultanément – i.e. que les ensembles de réponses qui les annulent ne sont pas disjoints.

2.3.3. Indulgence / sévérité de deux mesures

Le choix d'une mesure de qualité se fait souvent par rapport à d'autres mesures. Nous appelons *indulgence relative sur le domaine* Δ la propriété d'une mesure de produire des scores toujours supérieurs à ceux d'une autre pour les mêmes réponses aux données $D \in \Delta$ (Δ sera omis s'il est évident). Le cas contraire est appelé *sévérité relative*.

(IS) m_1 est plus indulgente que m_2 sur Δ si et seulement si
$\forall D \in \Delta$, $\forall rep(D)$, $m_1(rep(D)) \geq m_2(rep(D))$, et $m_1 \neq m_2$

La vérification n'en est pas facile, puisqu'il faut la réaliser pour toute réponse rep(*D*). Aussi, m_1 peut être plus indulgente que m_2 sur le domaine $\Delta_1 \subset \Delta$, et le contraire sur $\Delta_2 \subset \Delta$, et $\Delta_1 \cap \Delta_2 = \varnothing$. L'intérêt de comparer deux mesures est de pouvoir choisir la plus sensible aux performances attendues, augmentant ainsi les variations des scores. Si ceux-ci sont plutôt faibles, on prendra une mesure plus indulgente, et inversement.

3. CADRE D'EVALUATION DE LA RESOLUTION DE LA REFERENCE

Le problème de la résolution de la référence se pose en des termes particulièrement clairs : étant donné un ensemble d'ER, il faut regrouper les ensembles d'ER coréférentes, qui constituent des classes d'équivalence pour la relation de coréférence. Cette définition exclut de la tâche l'identification des ER (cf. §2.2 et §5.3) ; ainsi, les ER utilisées pour la réponse sont exactement les mêmes que celles de la clé. En outre, la définition se limite aux coréférences du type identité, i.e. à la dénotation d'un *même* objet par deux ER d'un texte (nous y reviendrons au §5.5). Toute ER appartient à une et une seule classe d'équivalence (éventuellement réduite à un seul élément), puisqu'elle possède par définition un seul référent déterminé ; les classes forment une *partition* de l'ensemble des ER.

Le problème de l'évaluation revient donc à mesurer le degré de ressemblance entre deux partitions en classes du même ensemble d'ER. Toutefois, il n'existe pas de mesure intrinsèque de cette similitude. En mathématiques, on se limite en général aux cas « partition plus fine / moins fine / non comparable », mais des théories plus récentes abordent l'aspect quantitatif

du problème (cf. §4.7). Ce cadre formel est suffisamment précis pour permettre en tout cas une analyse des différentes mesures de qualité suivant les critères précédents.

3.1. Définitions et notations

Soit E l'ensemble des ER à grouper en classes d'ER coréférentes. La clé est une partition de E notée P_K, c'est-à-dire un ensemble de parties de E, $P_K = \{K_1, K_2, \ldots, K_n\}$, disjointes, non vides, et recouvrant E (classes d'équivalence). De même, la réponse est une autre partition de E notée $P_R = \{R_1, R_2, \ldots, R_m\}$. Ces classes peuvent être des singletons (pas de lien de coréférence) – le cas où ceux-ci ne sont comptés est discuté au §5.2 (les résultats sont invariants pour les mesures MUC et noyaux).

La réponse parfaite est celle où pour chaque K_i il existe R_j telle que $R_j = K_i$ (donc $P_R = P_K$). Lorsque ce n'est pas le cas, il est utile de considérer toutes les classes réponse contenant des éléments d'une classe clé K donnée. On définit ainsi la projection de K sur P_R comme l'ensemble des fragments en lesquels K est partagée dans la réponse :

(DÉF.1) $\pi(K) = \{A \mid \exists\, R_j \in P_R \text{ telle que } A = K \cap R_j \text{ et } A \neq \varnothing\}$

On introduit aussi l'ensemble des classes réponse qui contiennent les fragments en question :

(DÉF.2) $\pi^*(K) = \{R_j \mid R_j \in P_R \text{ et } R_j \cap K \neq \varnothing\}$

La projection d'une classe réponse R sur P_K est définie réciproquement par :

(DÉF.3) $\sigma(R) = \{B \mid \exists\, K_i \in P_K \text{ telle que } B = R \cap K_i \text{ et } B \neq \varnothing\}$

(DÉF.4) $\sigma^*(R) = \{K_i \mid K_i \in P_K \text{ et } K_i \cap R \neq \varnothing\}$

Par conséquent, $\pi(K) \subset \mathit{Parties}(K)$, $\pi^*(K) \subset P_R$, $\sigma(R) \subset \mathit{Parties}(R)$ et $\sigma^*(R) \subset P_K$.

Si on note $|\ldots|$ le nombre d'éléments, alors $|P_K|$ est le nombre de classes clé, et $|P_R|$ le nombre de classes réponse. Puisqu'il y a au moins une projection (la classe K entière) et au plus $|K|$ (fragmentation complète en singletons), on a aussi :

(PROP.1) $1 \leq |\pi(K)| \leq |K|$ et $1 \leq |\sigma(R)| \leq |R|$, où $K \in P_K$ et $R \in P_R$

Le *taux de coréférence* du texte est donné par $|E| / |P_K|$, i.e. le nombre moyen d'ER par classe d'équivalence clé (voir exemples au §6.2 pour des textes réels).

3.2. Application à un exemple

L'exemple initial (§1.2) contient 17 ER, avec quatre entités ou classes clé. Le système est supposé avoir construit trois entités ou classes réponse (table 1 et figure 1). Le taux de coréférence de la clé est de 17/4 = 4.25 ER/classe, celui de la réponse de 17/3 = 5.67 ER/classe.

Clé : P_K	Réponse : P_R
K_1 : 1, 2	R_1 : 1, 2, 6, 7, 8, 9, 10
K_2 : 3, 4, 5	R_2 : 3, 4, 5, 11, 12, 13, 14, 15, 16
K_3 : 6, 7, 8, 9, 10, 11, 12	R_3 : 17
K_4 : 13, 14, 15, 16, 17	

Table 1. Classes clé et réponse pour le texte du §1.2

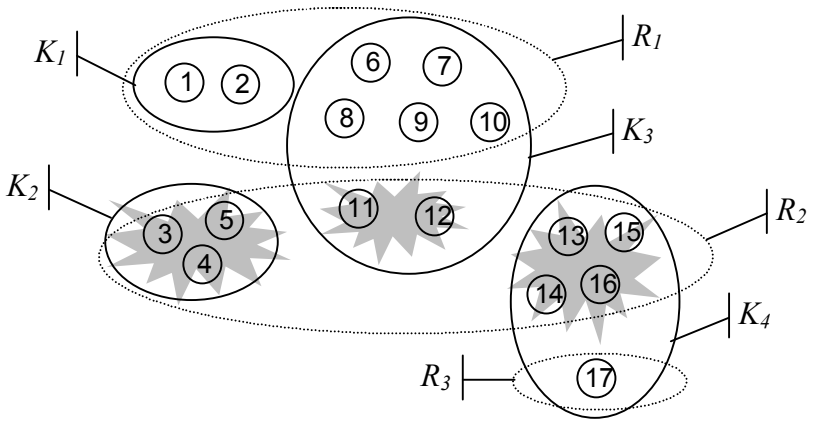

Figure 1. Représentation des classes clé (trait plein) et réponse (pointillé) pour le texte du §1.2. Les tâches grises sont les projections de R_2 sur P_K

La visualisation des projections est aisée sur la figure 1. K_1 et K_2 se projettent sur P_R en un seul fragment, alors que K_3 et K_4 sont divisées en deux : $\pi(K_3)$ = {{6, 7, 8, 9, 10}, {11, 12}}, et $\pi(K_4)$ = {{13, 14, 15, 16}, {17}}. Par conséquent, $\pi^*(K_1) = \{R_1\}$, $\pi^*(K_2) = \{R_2\}$, $\pi^*(K_3) = \{R_1, R_2\}$, $\pi^*(K_4) = \{R_2, R_3\}$. De façon analogue, R_1 se projette en deux fragments, R_2 en trois fragments (grisés sur la figure 1) et R_3 en un seul. On a donc $\sigma^*(R_1) = \{K_1, K_3\}$, $\sigma^*(R_2) = \{K_2, K_3, K_4\}$, et $\sigma^*(R_3) = \{K_4\}$.

3.3. Rappel, précision, *f-mesure*

Deux concepts fondamentaux utilisés pour les mesures de qualité sont le *rappel* et la *précision*. Introduits à l'origine en recherche documentaire par Van Rijsbergen (1979), ils s'appliquent aussi à d'autres tâches. De façon générale, lorsqu'un système doit, à partir d'un texte ou d'un ensemble de textes, produire un ensemble résultat, deux types d'erreurs peuvent se produire : le système peut omettre dans sa réponse des éléments qui auraient dû y figurer ou bien ajouter des éléments qui ne devraient pas y figurer. Les premières erreurs sont dites de *rappel* et les secondes de *précision*.

On peut alors afficher les scores dans un plan <*rappel*, *précision*>. Pour obtenir un seul score, en vue d'une évaluation comparative stricte, on considère souvent la *f-mesure*, à savoir la moyenne harmonique du rappel et de la précision :

(DÉF.5) $f - mesure\,(r,p) = \dfrac{2}{1/r + 1/p}$ ou bien 0 si $r = 0$ ou $p = 0$

Son avantage par rapport à la moyenne arithmétique est d'être plus proche de la valeur la plus faible, et ce d'autant que celle-ci est proche de 0. Elle pénalise donc les trop grandes inégalités entre rappel et précision, et les valeurs proches de 0.

Si l'on conçoit la résolution de la référence comme tâche de recherche des liens de coréférence corrects (MUC-6 1995), alors les erreurs de rappel sont les liens omis par le système, et celles de précision les liens inexacts. D'apparence claire, cet énoncé soulève de nombreuses difficultés : (1) les mêmes classes d'équivalence peuvent être définies par des configurations de liens différentes ; (2) le nombre de liens doit être converti à une proportion comprise entre 0 et 1, afin de permettre la comparaison des performances sur des textes différents ; (3) le statut des singletons n'est pas clair. En réalité, l'équivalence logique entre l'approche par liens et celle par classes est seulement qualitative, et non quantitative.

4. ANALYSE DES MESURES DE QUALITE

M. Vilain *et al.* (1995) ont apporté une amélioration notable aux définitions du rappel et de la précision pour la coréférence proposées à MUC-5. Les auteurs ont trouvé une méthode pour calculer la proportion de liens manquants et de liens incorrects sans faire appel à une identification précise de ces liens, puisque plusieurs configurations peuvent conduire aux mêmes classes d'équivalence (terme proposé par M. Vilain). La mesure proposée a été jugée satisfaisante, et utilisée pour MUC-6 et 7.

Des études plus récentes mettent en évidence des situations où les scores de la *mesure MUC* ne reflètent pas correctement la qualité d'une réponse. R. Passonneau (1997) propose une extension prenant en compte l'accord aléatoire (*mesure* κ). Partant d'un contre-exemple, A. Bagga et B. Baldwin (1998a, 1998b) proposent leur *mesure B*[3]. Suite à l'analyse de plusieurs cas réels ou simulés, nous proposons quatre nouvelles mesures : deux fondées sur les *classes-noyaux*, une mesure distributionnelle et une mesure entropique[3]. Enfin, R. Mitkov (1998) défend l'idée d'une mesure comparative pour la résolution des anaphores, mais celle-ci ne remplace pas une mesure de qualité (cf. §5.1).

Nous allons décrire le principe de chaque mesure, et l'étudier sur l'exemple du §1.2. La formalisation que nous donnerons nous permettra d'examiner notamment la satisfaction ou non des critères de cohérence (BS) et (BI) (démonstrations en annexe). Les scores sont notés avec trois lettres : la première désigne la mesure, la deuxième est *R* ou *P* (rappel ou précision) et la troisième est *S* ou *E* (succès ou erreur, avec *S* + *E* = 100%).

[3] Ces mesures ont été élaborées pour le système réalisé avec I. Robba. Une description préliminaire des trois premières se trouve dans (Popescu-Belis A. et Robba I. 1998b).

4.1. Mesure MUC (M. Vilain *et al.*)

Pour compter les erreurs de rappel, cette mesure calcule pour chaque classe clé K le nombre minimal de liens de coréférence qui manquent pour reconstituer K à partir de ses projections sur la réponse (éléments de $\pi(K)$). Il faut ensuite sommer sur toutes les classes K de P_K et normaliser ; le taux de succès sera le complémentaire à 100% du taux d'erreur.

Ainsi, sur la figure 1, les classes K_1 et K_2 ne sont pas scindées ($|\pi(K_1)|=|\pi(K_2)|=1$), mais K_3 est scindée en deux. La mesure MUC estime, de façon indulgente, qu'un seul lien a été omis pour K_3 parmi six ($|K_3|=7$), p.ex. celui entre ER_{10} et ER_{11}. De même, pour K_4, un lien a été omis parmi quatre, p.ex. celui entre ER_{16} et ER_{17}. On obtient $MRE = (0+0+1+1) / (1+2+6+4) \approx$ $\approx 15\%$, donc $MRS \approx 85\%$

Formellement, nous avons établi pour le succès de rappel la formule suivante, plus synthétique que celle de M. Vilain *et al.* (1995) :

(DÉF.6) $$MRS(P_R, P_K) = \frac{|E| - \sum_{K \in P_K} |\pi(K)|}{|E| - |P_K|}$$ et $MRS = 1$ si $|E| = |P_K|$

Bien que cette formule dérive par calcul d'une expression plus intuitive et algorithmique du score, nous lui attribuons tout de même le statut de définition. On remarque qu'il faut $|E| - |P_K|$ liens pour former les classes de P_K. Pour la précision, en partant du nombre de liens erronés dans une classe réponse R, on obtient la formule symétrique :

(DÉF.7) $$MPS(P_R, P_K) = \frac{|E| - \sum_{R \in P_R} |\sigma(R)|}{|E| - |P_R|}$$ et $MPS = 1$ si $|E| = |P_R|$

Lorsque les dénominateurs sont nuls (cas non précisé par les auteurs), nous avons choisi des conventions cohérentes : si $|P_K|=|E|$, cela signifie qu'il n'y a aucune coréférence dans la clé, donc il ne peut y avoir d'erreur de rappel[4]. Si $|P_R|=|E|$, le système n'a effectué aucune résolution, et n'ayant postulé aucun lien de coréférence, il ne peut avoir fait d'erreur de précision. Remarquons toutefois que dans ces deux cas la *f-mesure* est nulle (sauf si $|P_K|=|P_R|=|E|$).

On démontre (en annexe) le résultat suivant, qui apparaît expérimentalement sur les résultats MUC, et se retrouve dans les définitions utilisées en recherche d'information :

(PROP.2) les numérateurs de *MRS* et *MPS* sont égaux

Qu'en est-il des critères de cohérence ? Le critère (BS) est satisfait grâce aux résultats suivants (le symbole $\exists!$ signifie « il existe un unique ») :

[4] Il peut être intéressant de tester des systèmes aussi sur ces textes.

(PROP.3) $MRS = 100\% \Leftrightarrow \forall K \in P_K\ ,\ \exists!\ R \in P_R$ tel que $K \subset R$
$MPS = 100\% \Leftrightarrow \forall R \in P_R\ ,\ \exists!\ K \in P_K$ tel que $R \subset K$
f-mesure $= 100\% \Leftrightarrow MRS = MPS = 100\% \Leftrightarrow P_K = P_R$

Les critères (BI) sont plus délicats à étudier, car on ne peut trouver facilement les réponses notées 0%. Le critère (BI-3) – le score 0% peut être atteint – est certes satisfait grâce au cas particulier où aucune résolution n'est faite (toutes les réponses R sont des singletons), puisque alors $MRS = 0\%$, $MPS = 100\%$ (par convention), et la *f-mesure* est nulle. Nous pouvons cependant argumenter que le critère (BI-2′) – les mauvaises réponses doivent recevoir des scores faibles – n'est pas respecté, à l'aide d'un exemple de mauvaise réponse. En effet, pour la résolution triviale par groupement de toutes les ER nous obtenons facilement :

(PROP.4) $P_R = \{E\} \Rightarrow MRS = 100\%$ et $MPS = \dfrac{|E| - |P_K|}{|E| - 1}$

Ainsi, ce qui d'un point de vue informationnel est une « très mauvaise réponse » peut recevoir une note non nulle, et d'autant meilleure que le taux de coréférence de la clé est important. Par ailleurs, on prouve (en annexe) les minorations suivantes pour *MRS* et *MPS* :

(PROP.5) $MRS \geq \dfrac{|E| - |P_K| \cdot |P_R|}{|E| - |P_K|}$ et $MPS \geq \dfrac{|E| - |P_K| \cdot |P_R|}{|E| - |P_R|}$

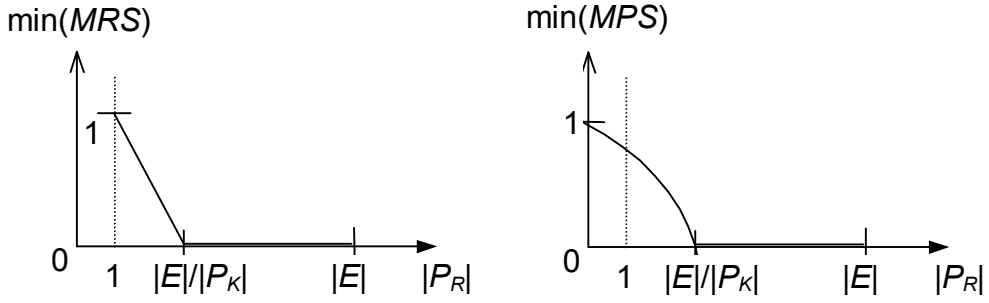

Figure 2. Borne inférieure établie pour la mesure MUC (rappel et précision) en fonction du nombre de classes réponse

En représentant graphiquement ces bornes inférieures en fonction de $|P_R|$ (figure 2) on s'aperçoit que si $|E| / |P_K| >> 2$ (taux de coréférence important), alors il suffit d'avoir une réponse telle que $|P_R| < |E| / |P_K|$ (peu de classes réponse, fort groupement des ER) pour être sûr d'obtenir un score

strictement positif. Ainsi, pour des textes à fort taux de coréférence, la mesure MUC ne satisfait pas le critère (BI-3′), donc ni (BI-2′) ni (BI)[5].

4.2. Extension de la mesure MUC à l'aide du κ (R. Passonneau)

Le but de cette étude est de mesurer l'accord entre deux humains chargés d'annoter la clé sur un texte. Passonneau montre que l'accord n'est pas total lors de deux annotations indépendantes (bien qu'il s'améliore après concertation) ; la mesure MUC se révèle trop indulgente pour mesurer ce désaccord. La mesure du *kappa* (Krippendorff 1980), est alors utilisée pour estimer la possibilité d'accords par hasard. Cela s'applique aussi à la comparaison d'une réponse et d'une clé.

Brièvement, l'idée est d'estimer, pour deux partitions différentes P_{R1} et P_{R2}, le nombre de liens sur lesquels il y a accord (i.e. présents ou absents en même temps dans P_{R1} et P_{R2}), le nombre de liens absents dans P_{R1} mais présents dans P_{R2} et inversement. Ces quantités permettent de mesurer la quantité d'accords qui dépasse le hasard, à l'aide du coefficient κ. Celui-ci peut varier de 1, accord parfait, jusqu'à -1, corrélation opposée, 0 marquant l'indépendance statistique. Pour notre exemple, on trouve κ = -0.18, ce qui rend compte de la « transversalité » de la clé et de la réponse (figure 1).

Au-delà de la pertinence de sa visée, cette mesure soulève à nos yeux trois difficultés : (1) l'appel aux liens de coréférence est critiquable puisque ceux-ci semblent avoir une pertinence cognitive plus réduite que les classes d'équivalence ; (2) le remplacement du rappel et de la précision par un seul coefficient est certainement moins informatif ; (3) plus grave, κ est obtenu par simple calcul à partir de *MRS* et *MPS* (car il est autrement impossible d'identifier les liens précis) – il peut donc bien sembler moins indulgent en valeur que la mesure MUC, il n'en sera pas plus informatif.

4.3. Mesure B^3 (A. Bagga et B. Baldwin)

Cette mesure cherche à pénaliser les fusions de classes volumineuses, qui peuvent cacher en réalité une tactique triviale. Le rappel et la précision sont d'abord calculés pour chaque ER. Le rappel B^3 pour une ER d'une classe *K* mesure la proportion de la classe *K* qui est contenue dans la classe réponse *R* contenant l'ER. Plus formellement :

(DÉF.8) $$BRS(ER_i) = \frac{|R \cap K|}{|K|} \quad \text{et} \quad BPS(ER_i) = \frac{|R \cap K|}{|R|}$$

où $ER_i \in R$ et $ER_i \in K$

Dans notre exemple, $BRS(ER_1)$=2/2, $BRS(ER_3)$=3/3 et cela est en fait valable pour toute ER de K_1 resp. K_2. En revanche, $BRS(ER_6)$=5/7, $BRS(ER_{11})$=2/7 – ce sont les deux valeurs possibles pour des ER de K_3 – et $BRS(ER_{13})$=4/5, $BRS(ER_{17})$=1/5.

On réalise ensuite la moyenne sur l'ensemble des ER[6] :

---

[5] Une deuxième minoration pour *MRS* pourra être déduite de (PROP.7) – minoration pour *CRS* – et (PROP.8) – *CRS* moins indulgente que *MRS*.

(DÉF.9) $$BRS = \frac{1}{|E|} \cdot \sum_{\substack{K \in P_K \\ R \in P_R}} \frac{|R \cap K|^2}{|K|} \quad \text{et} \quad BPS = \frac{1}{|E|} \cdot \sum_{\substack{K \in P_K \\ R \in P_R}} \frac{|R \cap K|^2}{|R|}$$

On peut facilement prouver que la limite de 100% est atteinte seulement pour $P_R = P_K$, ce qui vérifie le critère (BI). En revanche, les critères (BI-2) et (BI-3) ne sont pas satisfaits puisque le score de 0% n'est jamais atteint, les scores pour chaque ER n'étant jamais nuls. Nous avons prouvé (annexe) la minoration suivante, qui montre que la mesure B^3 est au voisinage des scores nuls plus discutable que la mesure MUC.

(PROP.6) $$\frac{|P_K|}{|E|} \leq BRS \leq 1 \quad \text{et} \quad \frac{|P_R|}{|E|} \leq BPS \leq 1$$

4.4. Mesure classes-noyaux ou mesure C

Nous avons élaboré cette mesure pour calculer de façon plus réaliste que la mesure MUC les nombres de liens manquants et erronés.

4.4.1. Principe

La mesure C est fondée sur l'idée de *noyau*, qui est *la classe réponse c*(K) qui « correspond le mieux » à la classe clé K*, à savoir la classe réponse qui contient le plus grand nombre d'ER de *K*. Au lieu de calculer le nombre minimal de liens manquants entre les projections de *K* (MUC), on compte comme erreurs de rappel *toutes* les ER hors du noyau $c^*(K)$, et on normalise. Pour la précision, il faut considérer symétriquement le noyau $c^*(R)$ de chaque classe réponse *R*.

Sur la figure 1, les projections de R_2 sur P_K sont en grisé ; la plus grande étant celle sur K_4, on a $c^*(R_2) = K_4$ – et $c(R_2) = \{ER_{13}, ER_{14}, ER_{15}, ER_{16}\}$. De même, $c^*(R_1) = K_3$, et $c^*(R_3) = K_4$.

Dans l'autre sens, K_1 et K_2 étant entièrement contenues, respectivement, dans R_1 et R_2, on a $c^*(K_1) = R_1$ et $c^*(K_2) = R_2$ – et $c(K_1) = K_1 \cap R_1 = K_1 = \ = \{ER_1, ER_2\}$, et $c(K_2) = K_2 \cap R_2 = K_2 = \{ER_3, ER_4, ER_5\}$. La plus grande projection de K_3 sur P_R est celle sur R_1 (cinq éléments) donc $c^*(K_3) = R_1$ et $c^*(K_4) = R_2$. Le fait que les noyaux de K_1 et K_3, resp. K_2 et K_4, soient identiques reflète l'intuition (cf. figure1) que la réponse ne parvient pas à distinguer ces entités.

Les scores se calculent facilement. Pour le rappel, il n'y a pas d'erreur pour K_1 et K_2. En revanche, on compte une erreur pour K_4, et deux erreurs pour K_3, car ER_{11} et ER_{12} ne sont pas rattachées au noyau $c^*(K_3) = R_1$ (la méthode MUC ne compte ici qu'une seule erreur). Il y a donc trois erreurs sur un total de 13 possibles (au maximum 1 pour K_1, 2 pour K_2, 6 pour K_3 et

---

[6] Les auteurs parlent d'une moyenne pondérée, dans laquelle soit les ER, soit leurs classes ont un poids égal, et semblent privilégier la première variante. Cependant, aucune formule analytique n'est donnée.

4 pour K_4), donc un succès de rappel *CRS* = 10/13 ≈ 77% (le rappel MUC étant *MRS* = 11/13 ≈ 85%).

Pour la précision, on souhaite que les R_j soient peu fragmentées en projection sur P_K, puisque le nombre de ces projections correspond aux liens erronés dans la réponse. Ainsi, R_1 donne lieu à deux erreurs (ER_1 et ER_2 hors de $c^*(R_1) = K_3$), et R_2 donne lieu à cinq erreurs (ER_3, ER_4, ER_5, ER_{11} et ER_{12}), alors que la méthode MUC en compte deux. Il y a donc 7 erreurs sur 6+8+0 possibles, d'où une précision *CPS* = 7/14 ≈ 50% (la précision MUC étant bien plus élevée, *MPS* = 11/14 ≈ 79%).

4.4.2. Résultats formels

La définition formelle des « pré-noyaux » $c(K_i)$ et $c(R_j)$ est :

(DÉF.10) $$c(K) = \underset{A \in \pi(K)}{ArgMax} |A| \quad \text{et} \quad c(R) = \underset{B \in \sigma(R)}{ArgMax} |B|$$

à savoir la plus grande projection de K (resp. R) sur P_R (resp. P_K). Lorsque plusieurs projections ont la taille maximale, on effectue un tirage au sort, ce qui ne change pas le nombre d'ER n'appartenant pas au pré-noyau. Le noyau de chaque classe est défini par :

(DÉF.11) $c^*(K) = R$ où $R \supset c(K)$ et $R \in P_R$
$c^*(R) = K$ où $K \supset c(R)$ et $K \in P_K$

Les scores sont symétriques en K et R, et s'expriment ainsi :

(DÉF.12) $$CRS = \frac{\left(\sum_{K \in P_K} |c(K)|\right) - |P_K|}{|E| - |P_K|} \quad \text{et} \quad CRS = 1 \; si \; |P_K| = |E|$$

(DÉF.13) $$CPS = \frac{\left(\sum_{R \in P_R} |c(R)|\right) - |P_R|}{|E| - |P_R|} \quad \text{et} \quad CPS = 1 \; si \; |P_R| = |E|$$

Le prolongement *CPS* = 100% dans le cas où $|P_R| = |E|$ (aucune résolution), analogue à celui de *MPS*, signale qu'il n'y a aucun lien erroné construit[7]. Le prolongement *CRS* = 100% s'applique au cas particulier $|P_K| = |E|$, où il n'y a aucune coréférence à trouver dans la clé, donc aucune erreur de rappel possible, seulement des erreurs de précision.

Il est facile de voir que la mesure C vérifie le critère (BS) ($\forall i$, $|c(K_i)| = |K_i|$, et de même pour R_j). Pour le critère (BI), il est à nouveau diffi-

[7] On peut se demander si ce prolongement est un « prolongement par continuité » : en fait, un tel prolongement n'existe pas. Si, au lieu de la réponse « nulle » $|P_R| = |E|$ (aucune résolution), le programme regroupe seulement deux ER (donc $|P_R| = |E| - 1$), on a *CPS* = 100% si celles-ci sont réellement coréférentes, et *CPS* = 0% sinon. Le rappel sera en général nul ou très faible.

cile de caractériser les solutions qui conduisent à une *f-mesure* nulle, mais il en existe : ainsi, le cas « aucune résolution » ($|P_R| = |E|$) implique un rappel nul, donc une *f-mesure* nulle (sauf dans le cas particulier où $|P_K| = |E|$).

On peut prouver (cf. annexe) les résultats suivants :

(PROP.7) $CRS \geq \frac{|R_m| - |P_K|}{|E| - |P_K|}$ et $CPS \geq \frac{|K_m| - |P_R|}{|E| - |P_R|}$

où K_m et R_m désignent les plus grandes classes clé et réponse

Ainsi, (BI-2′) n'est pas satisfait à cause de *CPS* si l'entité principale est fréquente. Lorsqu'un un système « sait » que tel est le cas, il lui suffit de fusionner toutes les ER (solution triviale $P_R = \{E\}$) pour avoir un score non nul. Cet inconvénient est analogue à celui de la mesure MUC, toutefois moins aigu, car MUC est plus indulgente (PROP.8).

4.4.3. Sévérité relative

La mesure C est plus sévère que la mesure MUC, comme nous l'avons déjà vu sur l'exemple du §4.4.1 :

(PROP.8)
- Pour E, P_K, P_R fixés, $CPS \leq MPS$ et $CRS \leq MRS$
- $CRS = MRS \Leftrightarrow$
  $\forall K \in P_K$, $\pi(K) \backslash \{c(K)\}$ est un ensemble de singletons
- $CPS = MPS \Leftrightarrow$
  $\forall R \in P_R$, $\sigma(R) \backslash \{c(R)\}$ est un ensemble de singletons

Cela démontre (cf. annexe) que la mesure C satisfait le but visé, qui était de proposer une mesure moins indulgente que la mesure MUC. Cette sévérité relative est visible aussi sur les exemples du §6. Les autres mesures ne se classent pas par ordre d'indulgence, comme le montrent les résultats numériques du §6 (scores parfois plus grands, parfois plus petits). De toute façon, certaines mesures se prêtent mal aux calculs exigés par une comparaison théorique.

4.5. Mesure classes-noyaux-exclusifs ou mesure XC

Nous avons aussi étudié une mesure fondée sur des *noyaux exclusifs* (notés *xc**). Ceux-ci sont analogues aux noyaux, sauf qu'ils sont construits de façon à être distincts : si $K_i \neq K_j$, alors $xc^*(K_i) \neq xc^*(K_j)$. Le but est de pénaliser les situations où plusieurs noyaux *c** sont confondus, en obligeant le système à avoir une unique « représentation » par entité.

L'algorithme de construction des *xc** commence par la classe clé la plus importante. Sur la figure 1, le noyau exclusif de K_3 correspond à sa plus grande projection, $xc^*(K_3) = R_1$. Désormais, R_1 ne peut plus être le noyau exclusif d'une autre classe clé. La deuxième classe clé, dans l'ordre, est K_4, et son noyau exclusif sera $xc^*(K_4) = R_2$ (plus grande projection). Par conséquent, pour K_2, comme R_2 est indisponible, $xc^*(K_2) = \varnothing$, et de même $xc^*(K_1) = \varnothing$. La construction symétrique pour P_R est dépourvue de sens dans notre interprétation.

On compte ensuite le nombre d'ER correctes (rappel) et le nombre d'ER incorrectes (précision) dans chaque *xc**(*K*) (cf. table 2, ligne 1, pour les valeurs obtenues sur l'exemple du §1.2). La place manque pour donner les formules, mais la définition algorithmique rend celles-ci difficilement exploitables pour des calculs. Outre le critère (BI), évident, on ne peut donner qu'une idée expérimentale (§6) des performances de cette mesure. Conçue à l'origine pour être plus sévère que la mesure par noyaux « non-exclusifs », l'expérience montre qu'elle ne l'est pas toujours.

4.6. Mesures de recouvrement quantitatif

La comparaison du nombre de classes clé et réponse est en pratique une méthode rapide pour évaluer la réponse d'un système. Bien sûr, l'égalité n'assure pas que la réponse soit parfaite. Mais si le nombre de classes réponse dépasse le nombre correct, alors le système fait probablement plus d'erreurs de rappel que d'erreurs de précision, et inversement (la démonstration de ce point est à l'étude, pour différentes définitions du rappel et de la précision).

Pour notre exemple, les tailles des classes clé sont, par ordre décroissant, (7, 5, 3, 2), et celles des classes réponse (9, 7, 1, '0'). Une distance entre ces deux vecteurs est (|9–7|, |7–5|, |1–3|, |0–2|) = (2, 2, 2, 2), donc une différence de 8 ER sur un maximum de 32 (valeur atteinte pour (17, 0, …, 0) et (1, …, 1)). D'autre part, R_1 et R_2 sont plus grandes que K_1 et K_2, et l'inverse pour R_3 et 'R_4' ; la réponse contient donc trop de regroupements.

Nous avons défini deux mesures *distributionnelles* calculées à partir du réarrangement des classes de P_K et de P_R par taille décroissante (par exemple $|K_1| \geq |K_2| \geq \ldots \geq |K_n|$ et $|R_1| \geq |R_2| \geq \ldots \geq |R_m|$). La somme des valeurs absolues des différences de taille pour chaque couple (R_i, K_i) fournit une première mesure de la distance entre clé et réponse, dite de recouvrement distributionnel (*RCVT*).

Séparons maintenant tous les indices pour lesquels $|K_i| \geq |R_i|$ des indices pour lesquels $|K_i| < |R_i|$. Une deuxième mesure compare les moyennes de ces indices (pondérés par les $||K_i| - |R_i||$). Si les premiers sont globalement plus petits que les seconds, alors en moyenne les classes réponse sont plus petites que les classes clé (et plus nombreuses) ; cela constitue *une D-erreur de rappel*, le cas contraire étant *une D-erreur de précision*. Faute de place, nous omettrons les formules mathématiques.

Ces deux mesures ne sauraient être employées seules, car ce n'est pas parce que les classes réponse et clé ont même taille (*RCVT*), ou que leurs différences sont uniformément distribuées (*D-erreur*), que ces classes sont identiques (le critère (BS) n'est pas satisfait). Ces mesures vérifient le critère (BI-1) étendu, i.e. une réponse mal notée est certainement une « mauvaise réponse ». De plus, la *D-erreur* indique le sens dans lequel il faut modifier un système, à savoir améliorer sa précision ou son rappel.

4.7. La mesure informationnelle ou entropique (mesure H)

Cette dernière mesure est plus solidement ancrée que les précédentes dans un modèle théorique. Elle est fondée sur la notion d'entropie, développée dans la théorie de l'information (Shannon C. et Weaver W. 1949), et sur

les études de la transmission d'information par un canal. Le concept qui fonde notre analogie est celui d'*information référentielle.* L'analogie permet de mieux comprendre le sens de la *mesure entropique*, et ses propriétés théoriques se démontrent à l'aide de résultats déjà établis (Ash R.B. 1965).

Le modèle du canal de communication comprend une *source*, i.e. une variable aléatoire (v.a.) pouvant prendre certaines valeurs, et un *récepteur*, une autre v.a. Une *transmission* consiste en l'émission d'un élément (valeur) par la source, et la lecture d'un autre élément sur le récepteur. Ces éléments ne sont pas comparables (car ils appartiennent à des ensembles différents), mais la fiabilité du canal de transmission se manifeste par une correspondance stable entre certains éléments de la source et du récepteur. La théorie définit une information moyenne émise par la source à chaque transmission, qui est l'entropie de la distribution de probabilité de la v.a. source, et de même une information moyenne produite par le récepteur. Puis, la théorie définit une information (ou entropie) conditionnelle de la source connaissant la valeur du récepteur, qui mesure les *pertes* dans le canal. Le concept symétrique est celui de *gains non pertinents* du canal. Ces deux valeurs sont calculées à l'aide de la loi du couple < v.a. source, v.a. récepteur >.

Nous identifions dans ce cadre la source avec l'ensemble des entités ou référents clé P_K, et le récepteur avec P_R. Les lois des variables aléatoires correspondent en fait à des partitions de l'espace de probabilité Ω sur lequel elles sont définies, qui est ici l'ensemble des ER. Autrement dit, chaque « transmission » est en fait la production d'une ER, à laquelle correspondent un certain élément de la source (référent émis) et un élément du récepteur (référent compris). La compréhension est bonne (clé proche de la réponse) lorsque les éléments du récepteur correspondent de façon stable au cours de la transmission à des éléments de la source. La loi du couple < v.a. source, v.a. récepteur > est calculée à l'aide des intersections entre classes de P_K et P_R.

Plus formellement, les quantités d'information référentielle de la source et du récepteur sont :

(DÉF.14) $$H(P_K) = -\sum_{K_i \in P_K} \frac{|K_i|}{|E|} \cdot \log \frac{|K_i|}{|E|} \quad \text{et} \quad H(P_R) = -\sum_{R_j \in P_R} \frac{|R_j|}{|E|} \cdot \log \frac{|R_j|}{|E|}$$

La quantité d'information véhiculée par le récepteur à propos de la source est notée $H(P_K|P_R)$, et mesure la connaissance apportée par la *partition réponse* P_R à propos de la *partition source (clé)* P_K. Ainsi, $H(P_K|P_R)$ et la quantité symétrique $H(P_R|P_K)$ sont définies par :

(DÉF.15) $$H(P_K|P_R) = -\sum_{(K_i,R_j)\in P_K\times P_R} \frac{|K_i \cap R_j|}{|E|} \cdot \log \frac{|K_i \cap R_j|}{|R_j|}$$

$$H(P_R|P_K) = -\sum_{(K_i,R_j)\in P_K\times P_R} \frac{|K_i \cap R_j|}{|E|} \cdot \log \frac{|K_i \cap R_j|}{|K_i|}$$

(avec le prolongement $0\cdot\log(0) = 0$).

On démontre (cf. Ash R.B. 1965) les relations suivantes :

(PROP.9)
- $H(P_R) - H(P_K) = H(P_R|P_K) - H(P_K|P_R)$
- $0 \le H(P_R|P_K) \le H(P_R)$
- $0 \le H(P_K|P_R) \le H(P_K)$

Nous interprétons l'égalité de la (PROP.9) comme *l'égalité fondamentale de l'information référentielle* : l'information reçue est égale à l'information émise moins les pertes (erreurs de précision ou confusions) plus les gains injustifiés (erreurs de rappel ou distinctions). Le rappel et la précision sont donc :

(DÉF.16) $$HRS = \frac{H(P_R) - H(P_R|P_K)}{H(P_R)} \quad \text{et} \quad HPS = \frac{H(P_K) - H(P_K|P_R)}{H(P_K)}$$

avec $HRS = 1$ si $H(P_R) = 0$, et $HPS = 1$ si $H(P_K) = 0$.

Les résultats existants (Ash R.B. 1965) conduisent aux cas d'égalité qui correspondent aux critères de cohérence (BS) et (BI). (BS) est satisfait, car :

(PROP.10) *f-mesure* = 100% $\Leftrightarrow$ $H(P_R|P_K) = H(P_K|P_R) = 0 \Leftrightarrow P_R = P_K$

Contrairement aux autres mesures, on peut préciser exactement pour la mesure H les cas où la *f-mesure* est nulle, grâce aux résultats :

(PROP.11) « *f-mesure* = 0 » équivaut à l'une des trois conditions :
- $H(P_R) = 0$ et $H(P_K) \neq 0$ – (fusion des ER dans la réponse)
- $H(P_k) = 0$ et $H(P_R) \neq 0$ – (fusion des ER dans la clé)
- $H(P_k) \neq 0$ et $H(P_R) \neq 0$ et « P_K et P_R sont indépendantes »

Le dernier critère, l'indépendance des v.a. ou des partitions, est le cas où la connaissance de la réponse n'apporte rien sur la connaissance de la clé :

(PROP.12) Les assertions suivantes sont équivalentes :
- « P_K et P_R sont indépendantes »
- $H(P_K) = H(P_K|P_R)$
- $H(P_R) = H(P_R|P_K)$
- les vecteurs $(|K_1 \cap R_j|, \ldots, |K_i \cap R_j|, \ldots, |K_n \cap R_j|)_{1 \le i \le n}$ sont proportionnels pour $1 \le j \le m$
- les vecteurs $(|K_i \cap R_1|, \ldots, |K_i \cap R_j|, \ldots, |K_i \cap R_m|)_{1 \le j \le m}$ sont proportionnels pour $1 \le i \le n$

En d'autres termes, chaque K_i doit se projeter sur P_R selon les mêmes proportions, et *vice versa*. Ces conditions montrent que le score nul ne peut pas toujours être atteint lorsque le nombre de classes réponse $|P_R|$ est fixé, puisque les classes K_i contiennent un nombre entier d'ER, qui n'est pas divisible à volonté. Il peut donc être impossible de réaliser la condition de proportionnalité précédente, et donc d'atteindre le score nul.

La mesure H présente un fondement théorique solide, et apparaît à ce titre comme « meilleure » que les autres mesures. Elle n'en est toutefois pas plus indulgente ou plus sévère, comme l'indiqueront les résultats numériques, mais s'avère plus analysable au voisinage de ses bornes.

5. ÉTUDE DE GENERALISATIONS

5.1. Évaluation par rapport aux tactiques/systèmes élémentaires

Pour la résolution de l'anaphore pronominale, R. Mitkov (1998) analyse les performances de son système en termes d'amélioration par rapport à un système simpliste (*baseline*). Cette proposition suppose l'existence préalable d'une mesure numérique, et suggère de l'utiliser *différentiellement*, en fixant par convention le score minimal non à 0% mais à celui d'une tactique simpliste. On adapte donc la mesure pour qu'elle atteigne 0% (critère BI-3).

L'une quelconque des mesures utilisées peut *a priori* convenir ici, mais quelle tactique simpliste choisir ? Il existe une tactique simpliste conduisant à 100% de rappel (fusionner toutes les ER), une autre donnant une bonne précision (fusionner le moins possible d'ER), mais elles sont par trop triviales, et conduisent aux extrêmes opposés. Il serait plus pragmatique de définir le résolveur de coréférences élémentaire comme celui qui utilise seulement l'accord en genre et en nombre, et l'identité de la tête du groupe nominal constituant l'ER[8]. Une proposition plus simple et calculable est de comparer la clé à une réponse consistant en une distribution aléatoire des ER en classes (par exemple de même profil distributionnel que la clé). En principe, les mesures H et *kappa* s'annulent sur une telle réponse, et les considérations sur H (PROP.12) montrent comment la construire.

Nous avons utilisé l'évaluation différentielle pour déterminer la contribution de chaque connaissance de notre système au score final (Popescu-Belis A. et Robba I. 1998a). Nous avons fait trois groupes de règles, et produit les réponses lorsque l'un des groupes alternativement, puis deux, étaient désactivés. Ceci a permis l'identification du groupe le plus important (l'accord sémantique). Pour augmenter la certitude des scores différentiels, nous avons pris la moyenne de plusieurs mesures de qualité.

5.2. Considération ou non des classes singleton

Du point de vue des *liens de coréférence*, les ER de la clé qui ne coréfèrent avec aucune autre ER ne devraient pas être prises en compte dans le calcul. De fait, dans les résultats détaillés de MUC-6, seules apparaissent

[8] Ou seulement l'identité du premier nom, puisque la détermination de la tête d'un GN peut ne pas être considérée comme une opération simple (*baseline*).

les classes de taille supérieure ou égale à 2. Pourtant, dans le modèle théorique MUC, les classes d'équivalence peuvent être aussi des singletons.

Pour départager ces deux options, il faut calculer deux variantes du rappel *MRS* : d'abord avec toutes les $K \in P_K$ puis seulement avec les K ayant deux ER ou plus (P^2_K) (idem pour *MPS*). Toutefois, dans la valeur de chaque $|\pi(K)|$ il faut faire intervenir aussi les $R_j \in P_R$ singletons, puisqu'elles correspondent à des ER non attachées. Il est donc inexact d'affirmer que l'on ne tient absolument pas compte des singletons. En utilisant les formules, nous avons démontré (en annexe) que les singletons ne posaient pas problème :

(PROP.13)
- Les rappels *MRS* et *CRS* sont invariants, qu'ils soient calculés sur P_K ou sur $P^2_K = \{K \mid K \in P_K \wedge |K| \geq 2\}$.
- Les précisions *MPS* et *CPS* sont invariantes, qu'elles soient calculées sur P_R ou sur $P^2_R = \{R \mid R \in P_R \wedge |R| \geq 2\}$.

Ces invariances ne s'appliquent pas à la mesure B^3. La mesure κ utilise en même temps *MRS* et *MPS*, donc on doit considérer les classes singletons. Il en est de même pour la mesure entropique H, puisque les classes singletons interviennent dans l'entropie de la source et du récepteur. La mesure XC se prête plus difficilement aux calculs.

En conclusion, il est certainement plus cohérent et homogène de prendre en compte les classes singletons dans les calculs ; de toute façon, *il faut* toujours le faire pour les singletons de R dans le calcul du rappel et ceux de K dans le calcul de la précision. Dans les deux cas où on a le choix de les considérer ou non (mesures MUC et C), cela ne change rien au résultat.

5.3. Ensembles d'ER différents pour la clé et la réponse

Lorsque le système est chargé aussi de l'identification des ER, la clé P_K et la réponse P_R ne sont plus des partitions du même ensemble E d'ER. Nous avons argumenté (§2.2) qu'une évaluation modulaire devrait fournir aux systèmes le même ensemble d'ER clé, et leur demander seulement de partitionner celui-ci. Si on désire néanmoins évaluer le résultat combiné des deux tâches, il faut adapter les mesures précédentes.

Dans la description de la mesure MUC (Vilain M. *et al.*1995), les exemples sont toujours donnés avec les mêmes ensembles d'ER, mais l'implémentation réalisée prévoit aussi le cas contraire. L'étude de ce programme montre qu'il obéit à la formulation que nous donnerons (DÉF.17). Les auteurs de la mesure B^3 ne traitent pas ce cas : leur mesure, conçue pour la coréférence inter-documents, suppose que le système possède l'ensemble des entités correctes pour chaque document.

Nous proposons, lorsque les ensembles d'ER de la clé E_K et de la réponse E_R sont différents, de les enrichir chacun avec les ER manquantes pour aboutir au même $E = E_K \cup E_R$. Les partitions sont enrichies chacune avec les singletons correspondants.

(DÉF.17) Si $E_K \neq E_R$ alors poser $E = E_K \cup E_R$,
$P'_K = P_K \cup \{ \{er\} \mid er \in E_R \setminus E_K \}$ et

$P'_R = P_R \cup \{ \{er\} \mid er \in E_K \setminus E_R \}$
puis utiliser les mesures précédentes avec E, P'_K et P'_R .

Ces modifications ne changent pas les scores MUC et noyaux (cf. PROP.13) – qui intègrent donc de façon particulièrement aisée le cas où $E_K \neq E_R$. On peut imaginer d'autres façons de procéder, mais cette méthode nous paraît de loin la plus naturelle.

5.4. Limitation à un sous-ensemble d'ER. Cas des pronoms

Afin d'évaluer la résolution de la référence pour une catégorie particulière d'ER, par exemple les noms propres, la méthode la plus simple est de reconstruire l'ensemble E et les partitions P_K et P_R restreints à cette catégorie. Une autre idée serait de calculer à l'aide des mesures précédentes un score pour chaque ER. Seule la mesure B^3 semblerait capable de calculer des scores par ER, et non par classe, mais on voit sur la (DÉF.8) que le score B^3 est le même pour toutes les ER d'une intersection $K_i \cap R_j$ donnée, donc ne caractérise pas spécialement *une* ER. Il semble donc impossible de contourner la première méthode.

L'évaluation de la résolution des anaphores pronominales *n'est pas* la restriction de l'évaluation de la référence à la classe des pronoms, comme on pourrait le croire. En effet, la résolution des anaphores ne consiste pas seulement à regrouper entre eux les pronoms coréférents, mais nécessite des rattachements à des ER non pronominales (antécédents). Cela montre déjà que l'ensemble des clés est nécessaire pour l'évaluation.

Le couplage pronom/antécédent rend impossible, à nos yeux, l'évaluation restreinte aux anaphores dans une réponse P_K . Une tentative élaborée indépendamment par S. Azzam *et al.* (1998), ainsi que par I. Robba et l'auteur (travail non publié), est la suivante. Pour chaque ER à évaluer (pronom), on examine s'il y a dans sa classe réponse au moins une ER non pronominale de sa classe clé ; s'il n'y en a pas, on compte une erreur de rappel (nous omettrons les formules, peu parlantes). Cette définition est trop indulgente, car la présence d'un antécédent et du pronom dans une même classe réponse n'est pas un gage de correction. Pour la précision, on compte une erreur chaque fois que dans la classe réponse de l'ER pronom il y a un antécédent non coréférentiel. Le résultat est trop sévère, comme l'ont montré nos expérimentations.

Ces difficultés font écho à la critique pragmatique du dualisme pronom/antécédent (Reboul A. 1994) : ces deux catégories doivent être traitées comme des instances d'expressions référentielles, en remplaçant les liens anaphoriques par des liens référentiels entre expressions et représentations. Si on désire garder les classes d'équivalence, cognitivement pertinentes, on doit renoncer à évaluer à part les liens anaphoriques, puisque ceux-ci ne se distinguent pas des autres liens de coréférence.

Afin d'évaluer spécifiquement la résolution des pronoms, il faut demeurer dans le paradigme pronom/antécédent, et demander au système de fournir un antécédent précis pour chaque pronom. Dans ce cas, si un couple < pronom, antécédent > n'appartient à la même classe clé de P_K , on compte une erreur (sinon un succès). On peut convenir, avec R. Mitkov et L. Belguith

(1998) d'appeler « rappel » le nombre de pronoms bien résolus parmi le total de pronoms existants et « précision » le nombre de pronoms bien résolus parmi le total de pronoms traités par le programme. Les mesures précédentes cèdent donc la place à un simple comptage.

5.5. Différents types de coréférence

D'un point de vue cognitif ou linguistique, les relations référentielles ne se limitent pas à la simple identité du référent. Les liens du type tout/partie, contenant/contenu, type/instance, individu/fonction, variable/valeur, etc., font en effet partie du phénomène de la référence, sans se ramener à la désignation d'un même objet invariable. Formellement, on doit donc considérer des liens *typés* entre ER, faisant l'objet d'un encodage spécifique dans la clé (Bruneseaux F. 1998). Deux solutions existent pour leur évaluation.

La première solution, réductrice, consiste à remplacer ces liens par des liens de type « identité » là où cela peut avoir un sens, et négliger les autres cas – solution de MUC-7 (Hirschman L. 1997). Ainsi, on regroupe une variable et sa valeur, un individu et sa fonction lors de la mention, etc. Ceci permet de se ramener à un seul type de coréférence sans toutefois négliger certains cas particuliers fréquents.

La deuxième solution consiste à limiter la relation de coréférence à la désignation stricte de la *même* entité (identité), et de concevoir les autres relations comme des relations non pas entre ER mais entre classes, ou bien *représentations mentales*, structures associées aux entités (Popescu-Belis A. *et al.* 1998). Les mesures étudiées s'appliquent alors à deux niveaux. D'abord, comme avant, on évalue la réponse P_R pour les liens « identité ». Puis, indépendamment pour chaque type de lien, on réapplique les mesures à la partition de P_R en super-classes correspondant aux différents types de liens référentiels hormis l'identité. Il reste bien sûr à étudier la spécificité de chaque type de lien : par exemple le lien tout/partie est transitif, mais pas le lien individu/fonction. Cette étude, qui fournirait la possibilité d'évaluer séparément chaque type de lien, dépasse le cadre du présent travail.

6. ÉTUDE D'EXEMPLES CONCRETS

Nous avons programmé les mesures précédentes dans l'interface d'évaluation de l'atelier de traitement de la référence développé avec I. Robba (Popescu-Belis A. et Robba I. 1998a, Popescu-Belis A. *et al.* 1998). Le système consiste en un résolveur construisant les classes réponse, accompagné d'un ensemble d'outils permettant de baliser les ER et les classes clé dans des textes. Il est possible d'importer/exporter un texte où les classes sont données à l'aide de balises SGML (Bruneseaux F. 1998, Popescu-Belis A. 1998), et surtout de comparer deux partitions des ER.

6.1. Clés et réponses artificielles

Afin d'illustrer les mesures sur certains cas particuliers, il est utile de construire une clé et une réponse artificielles. Concrètement, on construit un schéma de texte au format SGML avec des ER fictives, et on saisit ensuite une clé, puis une réponse. Notre programme fournit automatiquement les résultats de l'évaluation à l'écran. Nous avons testé les exemples suivants :

(1) Le premier exemple est le texte du §1.2 (figure 1).

(2a) On considère ensuite un texte avec 10 ER et deux classes clé, notées K_1 = {1, 2, 3, 4, 5} et K_2 = {6, 7, 8, 9, 10}, pour lequel on suppose d'abord que le système ne fait aucune résolution, i.e. P_R = {{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}}.

(2b) Sur le même texte, on suppose que le système a fusionné toutes les ER en une seule classe réponse R_1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, et donc P_R = {R_1}.

(3a) Nous étudions ensuite le texte donné en exemple à MUC-6. Ce texte contient 147 ER formant 15 classes clé, sans considérer les singletons. Il y a en tout 50 pronoms, mais seules 5 classes clé en contiennent. On suppose que le système ne résout rien, $|P_R|$ = 147.

(3b) Sur ce même texte, on suppose que le système fusionne toutes les ER, $|P_R|$ = 1.

(3c) Sur ce même texte, on suppose que le système, incapable de résoudre les 50 ER pronominales, les regroupe toutes en une seule classe, et résout correctement les 97 autres ER ($|P_R|$ = 15 + 1).

Les résultats sont donnés (en %) dans les tables 2 (rappel et précision) et 3 (*f-mesure*) ; il suffit de donner deux chiffres significatifs, car les exemples sont assez extrêmes, et de plus l'accord des humains sur la clé n'est pas non plus parfait.

	MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*XRS*	*XPS*	*D-err*	*RVT*	*HRS*	*HPS*
1	85	79	74	49	77	50	53	59	34-p	77	55	37
2a	0	100	20	100	0	100	20	100	41-r	20	30	100
2b	100	89	100	50	100	44	50	50	51-p	50	100	0
3a	0	100	10	100	0	100	10	100	39-r	10	40	100
3b	100	90	100	19	100	31	31	31	58-p	31	100	0
3c	96	97	65	79	67	82	69	84	7-r	86	76	81

Table 2. Rappels et précisions (%) pour les exemples artificiels

	MUC	B^3	K	C	XC	H
1	81	59	–18	61	56	44
2a	0	33	0	0	33	46
2b	94	67	0	62	50	0
3a	0	19	0	0	19	57
3b	95	33	0	47	31	0
3c	97	71	66	73	76	78

Table 3. *F-mesures* (%) et κ pour les exemples artificiels

Le premier exemple (1) affiche des valeurs relativement élevées des scores pour la réponse assez confuse vue au §1.2, sans illustrer une tendance particulière. Les cas (2a) et (3a) donnent les scores d'un système qui « ne fait rien ». Les cas (2b) et (3b) correspondent au système qui « fusionne tout », tactique simpliste qui doit aussi être notée comme « très mauvaise ». On constate une baisse de la précision, et un rappel à 100% (sauf pour la mesure XC). La *f-mesure* est loin d'être nulle, surtout pour la mesure MUC – les autres étant moins indulgentes envers cette tactique triviale. Le cas (3c), encore plus réaliste, montre l'indulgence de certains scores (en particulier MUC) alors que 30% des ER sont mal résolues.

Quant à l'indulgence relative, ces résultats confirment que les seules mesures comparables sont MUC et C (la seconde étant plus sévère). Pour tout autre couple, on ne peut déduire d'inégalité valable sur tous les exemples (la mesure κ mise à part). En outre, ces scores sont covariants, c'est-à-dire augmentent ou diminuent ensemble selon les exemples : ainsi, (3c) est meilleur en *f-mesure* que (3b) et (3a) pour tous les scores.

6.2. Performances de notre système sur des textes réels

Notre résolveur de coréférences cherche à regrouper toutes les ER dénotant un même référent. L'algorithme est proche de celui décrit dans (Lappin S. et Leass H.J. 1994) et nous l'avons décrit ailleurs (Popescu-Belis A. *et al.* 1998). Les ER sont examinées de façon séquentielle ; le résolveur doit déterminer pour chacune si elle se rapporte à une entité déjà introduite (auquel cas il l'ajoute à la classe correspondante) ou au contraire si elle concerne une entité nouvelle (auquel cas il crée une nouvelle classe). Pour chaque ER, le résolveur sélectionne les classes candidates en étudiant la compatibilité « moyenne » de l'ER courante avec chaque classe, à partir de l'accord en genre, nombre et « sémantique » avec chacun des éléments de cette classe. S'il n'y a pas de classe candidate, le programme crée une nouvelle classe, sinon il choisit la classe la plus active, selon un coefficient numérique d'activation périodiquement mis à jour.

	VA	LPG.eq	LPG
Mots	2630	7405	28576
ER ($\|E\|$)	638	686	3359
Classes clé ($\|P_K\|$)	372	216	480
$\|E\|$ / $\|P_K\|$	1.72	3.18	7.00
ER nominales	510	390	1864
ER pronoms	102	262	1398
ER non analysées	26	34	97

Table 4. Propriétés quantitatives des textes utilisés

L'expérimentation des mesures sur des textes réels se heurte à la difficulté de constituer les ressources correspondantes (Bruneseaux F. 1998,

Popescu-Belis A. 1998), car, en effet, on doit construire manuellement la clé. Nous disposons pour l'heure d'une nouvelle de Stendhal – *Vittoria Accoramboni* (VA), balisée au LIMSI – et d'un fragment du *Père Goriot* de Balzac (LPG), balisé au LORIA, Nancy, mais seulement avec les personnages principaux. Nous avons extrait de LPG un fragment LPG.eq d'une taille en ER comparable à VA. Des données statistiques sur les trois textes figurent dans la table 4. La taille de ces textes est assez importante (environ 100 pages pour LPG), mais il faut surtout noter le taux de coréférence $|E| / |P_K|$ important de chaque texte. Ce taux rend l'algorithme MUC particulièrement indulgent, comme observé au §4.1.

	MRS	*MPS*	*BRS*	*BPS*	*CRS*	*CPS*	*XRS*	*XPS*	*D-err*	*RVT*	*HRS*	*HPS*
VA	70	78	75	75	53	47	70	79	15-R	85	89	89
.eq	62	77	50	57	43	36	41	65	17-R	73	71	71
LPG	70	88	37	52	43	44	35	61	14-R	66	59	64

Table 5. Résultats du système sur les textes réels (%)

	MUC	B^3	K	C	XC	H
VA	74	75	57	50	74	89
LPG.eq	69	53	20	39	50	71
LPG	78	43	9	43	44	61

Table 6. *F-mesures* (%) et κ pour les résultats donnés dans la table 5

Les résultats de notre système sont donnés dans les tables 5 (rappel et précision) et 6 (*f-mesure*). Les scores paraissent élevés, si on les compare à ceux des participants aux tests MUC (dans les 60%). Il ne faut pas oublier deux différences. D'abord, nous n'évaluons pas la reconnaissance des ER, et donnons au système les ER correctes, contrairement aux tests MUC, pour les raisons exposées au §2.2. Puis, nos textes étant bien plus longs que ceux utilisés pour MUC, nos scores bénéficient d'un biais inhérent à la mesure MUC pour des classes volumineuses.

On observe que le système obtient des scores variables, malgré la nature assez semblable des trois textes. Les mesures MUC et C (C à un degré moindre) sont plus indulgentes lorsque le nombre d'ER augmente (LPG par rapport à VA et LPG.eq), alors que la « qualité » du système reste la même. Les mesures H, XC et B^3 varient dans le sens inverse. Les *f-mesures* ne sont donc pas covariantes de LPG.eq à LPG, elles augmentent pour MUC et C et diminuent pour H, XC et B^3. En revanche, entre les textes de longueur analogue (LPG.eq et VA), les différentes mesures (y compris κ) sont en accord pour désigner la réponse sur VA comme objectivement « meilleure » que celle sur LPG.eq – cela étant dû aux spécificités de chaque texte par rapport aux connaissances du système.

CONCLUSION

Nous avons commencé par situer notre approche computationnelle de la résolution de la référence, puis nous avons montré l'intérêt de l'évaluation numérique, et donné un cadre formel et des critères de cohérence pour les mesures de qualité. Après avoir posé les définitions nécessaires, nous avons décrit trois mesures existantes, et établi des formules précises et parfois originales, ainsi que des propriétés nouvelles. Ensuite, pour répondre aux imperfections de ces mesures, nous avons proposé quatre nouvelles mesures, et fourni pour deux d'entre elles une analyse détaillée de leur comportement et intérêt. Nous avons aussi examiné plusieurs extensions importantes, et illustré nos résultats par des exemples concrets.

Il reste qu'au terme de cette recherche, aucune des mesures envisagées ne semble s'imposer par des qualités intrinsèques. La mesure entropique semble certes la mieux justifiée théoriquement, mais cela relève de l'argumentation, non de la preuve. On peut suggérer que chaque mesure est adaptée à un certain domaine de données, ou à une plage de qualité des programmes. Plus généralement, il semble à ce stade qu'il n'existe pas de mesure unique capable de saisir objectivement la qualité d'un programme de résolution de la référence. Il est alors encourageant de considérer que chacune des mesures précédentes, fruit de la réflexion d'experts différents, fournit un point de vue différent sur la qualité d'une réponse. Nous pouvons dans ce cas affirmer que, lorsque ces mesures appliquées à deux programmes produisent des réponses covariantes, c'est-à-dire uniformément meilleures pour l'un des programmes, nous devons avec d'autant plus de confiance estimer ce programme meilleur.

RÉFÉRENCES

ASH Robert B. (1965) : *Information Theory*, New York, Interscience Publishers.

AZZAM Saliha, HUMPHREYS Kevin et GAIZAUSKAS Robert (1998) : "Evaluating a Focus-Based Approach to Anaphora Resolution", *Actes COLING-ACL '98*, Montréal, Canada, vol. I/II, pp. 74-78.

BAGGA Amit et BALDWIN Breck (1998a) : "Algorithms for Scoring Coreference Chains", *Actes LREC'98 Workshop on Linguistic Coreference*, Grenade, Espagne.

BAGGA Amit et BALDWIN Breck (1998b) : "Entity-Based Cross-Document Coreferencing Using the Vector Space Model", *Actes COLING-ACL '98*, Montréal, Canada, vol. I/II, pp. 79-85.

BRUNESEAUX Florence (1998) : "Noms propres, syntagmes nominaux, expressions référentielles", *Langues : cahier d'études et de recherches francophones*, vol. 1, n° 1, pp. 46-59.

EVANS Gareth (1985) : *The Varieties of Reference*, Oxford, Oxford University Press.

EWG (1996) : *EAGLES Evaluation Group. Final Report*, Center for Sprogteknologi, Copenhague, Danemark, EAG-EWG-PR.2.

FRANCIL (1997) : *Actes des Premières Journées Scientifiques et Techniques du réseau FRANCIL de l'AUPELF-UREF*, Avignon, AUPELF-UREF.

GAIZAUSKAS Robert, WAKAO T., HUMPHREYS Kevin, CUNNINGHAM Hamish et WILKS Yorick (1995) : "University of Sheffield: Description of the LaSIE System as used for MUC-6", *Actes MUC-6*, Columbia, MD, pp. 207-220.

GIVÓN Talmy (1979) : *On understanding grammar*, New York, Academic Press.

GRISHMAN Ralph et SUNDHEIM Beth (1996) : "Message Understanding Conference-6: A Brief History", *Actes COLING-96*, Copenhague, Danemark, pp. 466-471.

HIRSCHMAN Lynette (1997) : *MUC-7 Coreference Task Definition*, MITRE Corp.

HIRSCHMAN Lynette (1998) : "Language Understanding Evaluations: Lessons Learned from MUC and ATIS", *Actes First International Conference on Language Resources and Evaluation (LREC '98)*, Grenade, Espagne, vol. 1/2, pp. 117-122.

KRIPPENDORFF Klaus (1980) : *Content Analysis: An Introduction to its Methodology*, Beverly Hills, CA, Sage Publications.

LAPPIN Shalom et LEASS Herbert J. (1994) : "An Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics*, vol. 20, n° 4, pp. 535-561.

LUPERFOY Susann (1992) : "The Representation of Multimodal User Interface Dialogues Using Discourse Pegs", *Actes 30e ACL*, University of Delaware, Newark, DE, pp. 22-31.

MITKOV Ruslan (1998) : "Robust pronoun resolution with limited knowledge", *Actes COLING-ACL '98*, Montréal, Canada, vol. II/II, pp. 869-875.

MITKOV Ruslan et BELGUITH Lamia (1998) : "Pronoun resolution made simple: a robust, knowledge-poor approach", *Actes TALN '98*, Paris, pp. 42-51.

MUC-6 (1995) : *Proceedings of the 6th Message Understanding Conference (DARPA MUC-6 '95)*, San Francisco, CA, Morgan Kaufman.

PASSONNEAU Rebecca J. (1997) : *Applying Reliability Metrics to Co-Reference Annotation*, Technical Report Columbia University - Department of Computer Science, CUCS-017-97.

POPESCU-BELIS Andrei (1998) : "How Corpora with Annotated Coreference Links Improve Anaphora and Reference Resolution", *Actes First International Conference on Language Resources and Evaluation (LREC'98)*, Grenade, Espagne, vol. 1/2, pp. 567-572.

POPESCU-BELIS Andrei (*à paraître*) : "L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures", *Langues : cahiers d'études et de recherches francophones*.

POPESCU-BELIS Andrei et ROBBA Isabelle (1998a) : "Evaluation of Coreference Rules on Complex Narrative Texts", *Actes Second Colloquim on Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2)*, Lancaster, UK, pp. 178-185.

POPESCU-BELIS Andrei et ROBBA Isabelle (1998b) : "Three New Methods for Evaluating Reference Resolution", *Actes LREC'98 Workshop on Linguistic Coreference*, Grenade, Espagne.

POPESCU-BELIS Andrei, ROBBA Isabelle et SABAH Gérard (1998) : "Reference Resolution Beyond Coreference: a Conceptual Frame and its Application", *Actes COLING-ACL '98*, Montréal, Canada, vol. II/II, pp. 1046-1052.

REBOUL Anne (1994) : "L'anaphore pronominale : le problème de l'attribution des référents", in J. Moeschler, A. Reboul, J.-M. Luscher et J. Jayez (eds.), *Langage et pertinence*, Nancy, Presses Universitaires de Nancy, pp. 105-173.

REBOUL Anne (1998) : "A relevance theoretic approach to reference", *Actes Relevance Theory Workshop*, Luton, UK.

RÉCANATI François (1993) : *Direct Reference: from Language to Thought*, Oxford, UK, Basil Blackwell.

RUBIO Antonio, GALLARDO Natividad, CASTRO Rosa et TEJADA Antonio, eds. (1998) : *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Grenade, Espagne, ELRA.

SHANNON Claude E. et WEAVER Warren (1949) : *The Mathematical Theory of Communication*, Urbana, IL, University of Illinois Press.

SIDNER Candace L. (1983) : "Focusing in the Comprehension of Definite Anaphora", in M. Brady et R. Berwick (ed.), *Computational Models of Discourse*, Cambridge, MA, MIT Press, pp. 267-330.

SPARCK JONES Karen et GALLIERS Julia (1996) : *Evaluating Natural Language Processing Systems: An Analysis and Review*, Berlin, Springer-Verlag.

VAN RIJSBERGEN Cornelis J. (1979) : *Information Retrieval*, London, Butterworth.

VILAIN Mark, BURGER John, ABERDEEN John, CONNOLLY Dennis et HIRSCHMAN Lynette (1995) : "A Model-Theoretic Coreference Scoring Scheme", *Actes MUC-6*, Columbia, MD, pp. 45-52.

ANNEXE 1

Les démonstrations des théorèmes énoncés sont données ci-dessous.

(PROP.2) – Dans le numérateur de *MRS*, on exprime $|\pi(K_i)|$ comme $\Sigma_{R_j \in P_R} 1_{K_i \cap R_j}$, et dans le numérateur de *MPS* on exprime $|\sigma(R_j)|$ comme $\Sigma_{K_i \in P_K} 1_{K_i \cap R_j}$, où $1_{K \cap R}$ vaut 1 si $K \cap R \neq \varnothing$ et 0 sinon. On voit alors que l'expression résultante est la même pour *MRS* et *MPS*.

(PROP.3) – Le sens '$\Leftarrow$' est évident : si chaque K est incluse dans une R, c'est que sa projection n'est pas scindée, $|\pi(K)|=1$, et *MRS*=1. Pour le sens '$\Rightarrow$', puisque l'erreur est une somme de termes positifs, il faut que chaque $|\pi(K)|=1$, à savoir chaque K n'intersecte qu'une seule classe réponse, dans laquelle elle est donc contenue (cqfd). La démonstration est analogue pour la précision, puis le résultat sur la *f-mesure* s'en déduit facilement.

(PROP.5) – La somme au numérateur de *MRS* contient exactement $|P_K|$ termes, chacun inférieur à $|P_R|$; en effet, une classe clé K se projette en au maximum $|P_R|$ fragments, un par classe réponse. De même pour *MPS*.

(PROP.6) – Dans la double somme de *BRS*, on regroupe les termes de même classe clé K_i. Montrons d'abord que $|K_i|^2 \geq \Sigma_{R_j \in P_R} |K_i \cap R_j|^2 \geq |K_i|$. En effet, $\Sigma_{R_j \in P_R} |K_i \cap R_j| = |K_i|$, et on élève au carré pour la première majoration ; la seconde utilise $|K_i \cap R_j|^2 \geq |K_i \cap R_j|$. Pour finir, on divise les inégalités précédentes par $|K_i|$ et on somme sur tous les K_i. De même pour *BPS*.

(PROP.7) – Pour *CPS* : soit une classe clé K_i. Alors, $\forall R_j \in P_R$, on a par définition $|c(R_j)| \geq |K_i \cap R_j|$ (le pré-noyau est la plus grande des projections),

donc $\Sigma_{R_j \in P_R} |c(R_j)| \geq \Sigma_{R_j \in P_R} |K_i \cap R_j| = |K_i|$. On choisit alors $K_i = K_m$, la plus grande classe clé (choix intéressant). De même pour *CRS*.

(PROP.8) – Si le dénominateur est nul, l'inégalité $MRS \geq CRS$ est vraie (par convention). Sinon, elle revient à l'inégalité des numérateurs, et se transforme facilement en $\Sigma_{K_i \in P_K}(|K_i| - |\pi(K_i)|) \geq \Sigma_{K_i \in P_K}(|c(K_i)| - 1)$. Or, l'inégalité $|K_i| \geq |c(K_i)| + |\pi(K_i)| - 1$ exprime que si on ajoute les ER du pré-noyau de K_i et une ER pour chaque *autre* projection de K_i, on trouve moins d'ER que le total de K_i. Cela est évident, et l'égalité a lieu si et seulement si toutes les projections de K_i, sauf peut-être le pré-noyau, sont des singletons. Cela doit être vrai $\forall K_i \in P_K$ pour que $MRS = CRS$. De même pour $MPS \geq CPS$.

(PROP.13) – Pour le rappel, on compare les formules pour E et P_K avec celles pour E^2 et $P^2{}_K$ (on élimine les K_i singletons et les ER correspondantes). Le dénominateur de *MRS* et *CRS*, $|E| - |P_K|$, est invariant car on retranche le même nombre d'éléments à E et à P_K pour arriver à E^2 et $P^2{}_K$. Les numérateurs s'écrivent comme $\Sigma_{K_i \in P_K}(|K_i| - |\pi(K_i)|)$ et $\Sigma_{K_i \in P_K}(|c(K_i)| - 1)$, où l'on voit que les classes singleton ne comptent pas, car elles ont une seule projection, et leur pré-noyau est un singleton.