Trois expériences d'évaluation dans le cadre du développement d'un système d'alignement sous-phrastique

Sylwia Ozdowska

NCLT, School of Computing, Dublin City University
Dublin 9, Irlande
sozdowska@computing.dcu.ie

RÉSUMÉ. *Nous présentons la démarche que nous avons adoptée pour mener à bien une évaluation système dans le contexte du développement d'un système d'alignement sous-phrastique,* ALIBI. *À cet égard, nous examinons trois procédures d'évaluation qui correspondent à des aspects fondamentaux de la mise au point de systèmes de traitement automatique des langues : une évaluation par annotation des sorties du système qui permet d'observer le comportement de chaque composante prise isolément, une évaluation avec des références multicorpus qui permet d'observer le comportement d'un système selon le type de corpus qu'il prend en entrée et une évaluation avec une référence standard disponible publiquement qui permet d'observer son comportement par rapport à des outils de même famille. Nous décrivons chaque expérience d'évaluation et faisons le point sur la nature des résultats qu'elle fournit ainsi que leurs apports.*

ABSTRACT. *We present the approach we adopted to carry out a system oriented evaluation in the context of the development of a word alignment system,* ALIBI. *We look at three evaluation procedures corresponding to fundamental aspects in the development of natural language processing systems: an evaluation resulting from the annotation performed on output alignments to accurately examine the behaviour of each module under study, an evaluation with multicorpus references allowing to observe the behaviour of a system depending on the type of input corpus and an evaluation with a standard reference allowing to observe its behaviour in comparison with other systems of the same type. We describe each procedure and discuss the measures particular to each of them as well as their contribution to system oriented evaluation.*

MOTS-CLÉS : *évaluation système, évaluation multicorpus, évaluation standard, annotation multi-juge, accord interjuge, alignement sous-phrastique.*

KEYWORDS: *system oriented evaluation, cross-corpus evaluation, gold standard evaluation, multiple annotation, inter-annotator agreement, word alignment.*

1. Introduction

En traitement automatique des langues (TAL), il existe différents types d'évaluation — évaluation système, évaluation par la tâche et évaluation utilisateur — dont le choix est conditionné entre autres par la maturité du système testé (Hirschman *et al.*, 2003 ; Paroubek, 2004). Dans cet article, nous nous intéressons à une évaluation système dans le contexte du développement d'un système d'alignement sous-phrastique, ALIBI (Ozdowska, 2006). Ce type d'évaluation permet d'estimer les performances intrinsèques d'un système en tant que technologie et non en tant qu'application opérationnelle en condition réelle (Chaudiron, 2004). L'estimation s'effectue par comparaison des sorties proposées par le système avec des données de référence et sur la base de mesures communes, à savoir précision, rappel et f-mesure. La constitution d'une référence passe par une phase d'annotation manuelle au cours de laquelle un humain attribue aux données de test les informations que le système à évaluer est censé produire. Dans le cas de l'alignement sous-phrastique, ces informations sont des liens de correspondance traductionnelle entre des mots ou des groupes de mots dans un échantillon de biphrases, *i. e.* de phrases alignées.

Le projet Blinker a été l'un des premiers à proposer un cadre pour la constitution de données de référence pour cette tâche (Melamed, 1998a ; Melamed, 1998b). Deux principes ont été adoptés pour garantir la fiablité des données produites : la mise au point préalable d'un guide d'annotation et le recours à plusieurs juges pour l'annotation des mêmes données[1]. Cette démarche a par la suite été reproduite de manière plus ou moins fidèle dans le cadre de campagnes d'évaluation de systèmes d'alignement telles que ARCADE I (Véronis *et al.*, 2000) ou HLT-NAACL'03 (Mihalcea *et al.*, 2003) ou encore à titre individuel (Merkel, 1999 ; Kraif, 2001). Dans les campagnes d'évaluation, différentes solutions ont été retenues concernant notamment : l'estimation de l'accord interjuge avant l'évaluation, la prise en compte des références multiples résultant de l'annotation multijuge et la prise en compte des alignements avec des segments de plusieurs mots dans le calcul des mesures de performances lors de l'évaluation[2].

Ainsi, dans ARCADE I :

– les données de test sont annotées par deux juges sur la base de conventions d'annotation ;

– l'accord entre deux juges X et Y résulte de la moyenne géométrique de l'accord de X vis-à-vis de Y et de l'accord de Y vis-à-vis de X ;

– l'estimation des performances s'effectue avec la référence pour laquelle le recouvrement avec les alignements de chaque système est maximal ;

1. Les données de référence produites sont disponibles à l'adresse `http://www.cs.nyu.edu/~melamed/ftp/data/`.
2. Pour un panorama détaillé sur l'état des évaluations en alignement sous-phrastique voir (Ozdowska, 2006).

– les alignements avec des segments de plusieurs mots sont évalués en fonction du degré de recouvrement avec l'alignement de référence.

Dans HLT-NAACL'03, qui reprend le protocole d'évaluation proposé dans (Och *et al.*, 2003) :

– les données de test sont annotées par deux juges qui caractérisent les alignements soit comme sûrs (attribut S) soit comme probables (attribut P) ;

– l'accord entre les juges n'est pas calculé ;

– l'estimation des performances est effectuée avec une référence finale unique dans laquelle les alignements pour lesquels il y a accord interjuge sur S conservent cet attribut, les autres alignements sont considérés comme P[3] ;

– les alignements avec des segments de plusieurs mots sont décomposés en liens individuels de type P entre chaque mot du segment source et chaque mot du segment cible.

Nous retenons de ces travaux que, de façon standard, l'évaluation des systèmes d'alignement sous-phrastique s'appuie sur une référence correspondant à un seul type de corpus, construite indépendamment des sorties des systèmes et dont la fiabilité est maximisée grâce à l'intervention de plusieurs annotateurs sur les mêmes données. Ce type d'évaluation pemet de comparer différents systèmes dédiés à la même tâche sur la base de leurs performances.

En dehors des campagnes d'évaluation, le développement d'un système implique des phases d'évaluation individuelle qui lui sont spécifiques dans la mesure où elles visent à donner des indications précises sur la validité des principes que le système met en œuvre ainsi que sur ses perspectives d'évolution. Dans le contexte du développement d'ALIBI, qui est un système d'alignement sous-phrastique à base de règles syntaxiques (cf. section 2.1), l'objectif des évaluations individuelles était double. D'une part, il s'agissait de rendre compte de manière fine du fonctionnement de chaque règle d'alignement syntaxique et de cibler au mieux les modifications possibles pour chacune d'elles. Pour ce faire, il fallait que chaque règle soit représentée par un nombre de cas d'application suffisamment élevé pour que l'estimation de ses performances soit significative. C'est pourquoi nous avons choisi d'annoter directement les sorties fournies par le système pour construire une importante base de cas validés représentant toutes les règles qui composent ALIBI et permettant d'évaluer chacune d'elles de manière individuelle. D'autre part, il s'agissait d'apprécier la généricité des principes mis en œuvre sur différents types de corpus tout en tenant compte de l'état des évaluations en alignement. Pour ce faire, nous avons choisi de construire notre propre référence sur la base de l'examen des campagnes d'évaluation standard.

Si le recours à une évaluation par annotation directe des sorties du système n'est pas nouveau (Ahrenberg *et al.*, 2000), il semble tout de même relativement rare dans le domaine de l'alignement. Quant à l'évaluation multicorpus, elle relève d'une démarche inédite dans ce domaine. Combinées ensemble, les trois expériences d'évalua-

3. Les performances sont mesurées séparément sur chaque ensemble d'alignements S ou P.

tion que nous avons menées — avec une référence obtenue par annotation des sorties du système, avec une référence multicorpus et avec une référence standard — nous ont permis de tester le système sous différents aspects, tous justifiés. Le point commun de ces trois expériences est la construction d'une base de cas annotés manuellement et le calcul de mesures de performances sur la base de ces cas.

Dans un premier temps, nous présenterons brièvement le principe de fonctionnement d'ALIBI (section 2). Puis nous décrirons nos trois expériences d'évaluation :

– évaluation avec une référence construite à partir des sorties du système (désormais *évaluation par annotation des sorties*) permettant de vérifier le fonctionnement de chaque règle d'alignement (section 3) ;

– évaluation avec une référence multicorpus (désormais *évaluation multicorpus*) permettant d'apprécier l'influence du type de corpus sur les performances (section 4) ;

– évaluation avec une référence standard disponible publiquement (désormais *évaluation standard*) permettant de comparer les performances entre systèmes dédiés à la même tâche (section 5).

Nous motiverons les choix effectués à chaque étape et ferons le point sur les apports de chaque expérience d'évaluation et la nature des résultats fournis pour mettre en avant leur complémentarité. Enfin, nous discuterons d'un problème spécifique à l'évaluation de l'alignement sous-phrastique, à savoir la prise en compte des alignements avec des segments de plusieurs mots (section 6).

2. Contexte

2.1. *Système* ALIBI

ALIBI est un système d'alignement sous-phrastique qui vise à mettre en correspondance des unités textuelles de taille inférieure à la phrase qui sont potentiellement en relation de traduction (Ozdowska, 2006). Le principe d'alignement qu'il met en œuvre découle de l'hypothèse formulée par F. Debili *et al.* (1996) sur la base du principe d'analogie (Lepage, 2006). Ce principe est le suivant : il s'agit de partir d'un couple amorce, c'est-à-dire de deux mots en relation de traduction dans une biphrase (*Community* et *Communauté* dans la figure 1), et d'aligner les mots qui sont en relation syntaxique avec ce couple (*ban* et *interdire*).

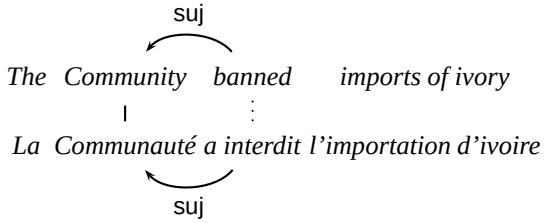

Figure 1. *Alignement syntaxique*

Dans le système ALIBI, les couples amorces sont repérés automatiquement par une technique de type statistique. Les relations de dépendance sont identifiées par les analyseurs SYNTEX pour le français et l'anglais (Bourigault *et al.*, 2005 ; Bourigault, 2007). Le principe de propagation est exprimé sous forme de règles qui s'appuient sur des patrons de la forme :

RecteurE-relationE-RégiE/RecteurF-relationF-RégiF

Les patrons tiennent compte de divergences catégorielles et syntaxiques classiques entre l'anglais et le français. Lorsqu'il s'appuie sur les mêmes relations syntaxiques (relationE = relationF), un patron d'alignement est dit identique. Lorsqu'il s'appuie sur des relations syntaxiques différentes (relationE ≠ relationF), un patron d'alignement est dit compatible. Chaque patron d'alignement peut s'appliquer dès lors que l'un des couples de recteurs (RecteurE/RecteurF) ou de régis (RégiE/RégiF) est un couple amorce (CA), le tout — patron avec couple amorce — constitue une règle d'alignement. Ainsi, qu'il soit identique ou compatible, chaque patron correspond à deux règles d'alignement, une où c'est le couple de régis qui est le CA et une où c'est le couple de recteurs qui est le CA :

1) RecteurE-relationE-RégiE/RecteurF-relationF-RégiF ;

2) RecteurE-relationE-RégiE/RecteurF-relationF-RégiF.

Ainsi, dans l'exemple (1) le CA de régis *Community*/*Communauté* permet d'aligner les verbes recteurs *ban*/*interdire* grâce à la relation sujet. La règle correspondant à ce cas d'alignement s'écrit : N-suj-V/N-suj-V.

(1) *The Community* ***banned*** *imports of ivory*
La Communauté ***a interdit*** *l'importation d'ivoire*

Inversement, dans l'exemple (2) le CA de recteurs *haul*/*haler* permet d'aligner les noms régis *net*/*filet* grâce à la relation sujet. La règle correspondant à ce cas d'alignement s'écrit : N-suj-V/N-suj-V.

(2) *The* ***net*** *is hauled to the shore*
Le ***filet*** *est halé à terre*

2.2. *Corpus*

Le principe d'alignement par propagation syntaxique est testé sur trois corpus anglais/français alignés au niveau des phrases : INRA, JOC et HANSARD. Le corpus INRA comprend des articles de recherche et de vulgarisation en agronomie. Il a été constitué à l'Institut National de la Recherche Agronomique dans le cadre d'une expérience sur l'enrichissement de la base de données terminologiques alimentée et exploitée par les

traducteurs du service linguistique[4]. Le corpus JOC se compose de questions-réponses traitées à la Commission Européenne. Initialement élaboré dans le cadre des projets MLCC et MULTEXT (Ide *et al.*, 1994) et utilisé pour la campagne d'évaluation de systèmes d'alignement ARCADE I[5], il est distribué par l'ELRA/ELDA[6]. Enfin, le corpus HANSARD rassemble des débats parlementaires canadiens[7]. Largement exploité dans le domaine de l'alignement et de la traduction statistique, une version étendue de ce corpus a été utilisée pour la campagne d'évaluation HLT-NAACL'03 (Mihalcea *et al.*, 2003) dont les données sont disponibles publiquement[8].

Le tableau 1 résume les principales caractéristiques chiffrées concernant les trois corpus : le nombre de biphrases (**biphrases**), le nombre de mots-occurrences total (**m-o**$_e$+**m-o**$_f$) et par langue (**m-o**$_e$/**m-o**$_f$).

	INRA		**JOC**		**HANSARD**	
biphrases	7137		8759		8000	
m-o$_e$+**m-o**$_f$	289785		439722		252419	
m-o$_e$/**m-o**$_f$	137221	152564	201362	238360	119804	132615

Tableau 1. *Caractéristiques des corpus* INRA, JOC *et* HANSARD

Dans les sections qui suivent, nous décrivons les trois expériences d'évaluation menées — évaluation par annotation des sorties, évaluation multicorpus et évaluation standard —, rendant chacune compte d'un aspect particulier du système.

3. Évaluation par annotation des sorties pour la vérification du fonctionnement des composantes du système

3.1. *Principe de l'évaluation par annotation des sorties*

En phase de développement d'un système modulaire, deux types d'analyses nous semblent particulièrement importants :

– une analyse quantitative détaillée offrant la possibilité d'observer de manière précise le fonctionnement de chaque composante du système prise isolément ; dans le cas d'ALIBI, il s'agit de chaque règle d'alignement ;

– une analyse qualitative mettant au jour les failles du système par une catégorisation des causes d'erreurs à partir de l'analyse des cas d'erreurs.

Bien qu'il ne s'agisse pas nécessairement d'une démarche standard, nous avons opté pour une évaluation consistant à annoter directement les sorties fournies par le

4. Nous remercions A. Lacombe de l'INRA de nous avoir autorisée à utiliser ce corpus.
5. http://www.up.univ-mrs.fr/veronis/arcade/arcade1/index.html
6. Evaluations and Language resources Distribution Agency (http://www.elda.org/)
7. http://parl.gc.ca/common/Chamber_House_Debates.asp?Language=F
8. http://www.cse.unt.edu/~rada/wpt/

système pour répondre à ces deux objectifs. Il s'agit d'une démarche progressive par essai-erreur. Elle est menée la plupart du temps par le concepteur de l'application qui est à la fois juge et partie et lui permet d'alterner des phases d'intégration ou d'ajustement de composantes avec des phases de vérification de leurs performances. À cet égard, ce type d'évaluation offre un avantage important, à savoir la possibilité de sélectionner les cas à évaluer et donc de s'assurer que, pour chaque composante ou sous-tâche testée, le nombre de cas validés est suffisant pour que l'évaluation soit représentative.

Pour nous assurer que le principe d'analogie mis en œuvre dans ALIBI s'appliquait dans le contexte de l'alignement sous-phrastique dans un nombre de cas suffisamment important et dans des configurations syntaxiques suffisamment diversifiées, nous avons adopté la démarche suivante : définition de règles d'alignement, évaluation de chaque règle prise isolément et ajustement en fonction de la précision mesurée. L'estimation de la précision individuelle de chaque règle d'alignement définie suppose une évaluation à grande échelle. En effet, chaque règle doit être représentée par un nombre de cas d'application suffisamment élevé pour que l'évaluation soit représentative. Afin de répondre à ce critère de représentativité, nous nous sommes imposé comme contrainte de valider l'ensemble des cas d'application d'une règle si leur nombre était inférieur ou égal à cinquante et au moins cinquante cas d'application si leur nombre était supérieur à cinquante. Le choix des cas évalués s'est fait de manière séquentielle, au fur et à mesure du parcours des alignements produits, avec possibilité de passer certains cas soit parce qu'une décision s'avérait trop difficile à prendre, soit parce qu'il s'agissait de cas redondants. Cette manière de procéder ne permet sans doute pas de s'assurer que l'évaluation ne repose pas sur des bases trop subjectives du fait de l'intervention d'un seul juge et de la sélection des cas annotés. Par ailleurs, cette expérience d'évaluation nous a montré qu'il existe un seuil au-delà duquel on constate une stabilité quantitative (taux de précision) et qualitative (type de cas évalués) des résultats et qu'on pourrait envisager de déterminer en fonction du nombre de cas d'application d'une règle. En effet, à cet égard le nombre de cas évalués pour certaines règles d'alignement est parfois excessif.

Pour ce qui est du versant qualitatif de l'évaluation, l'évaluation par annotation des sorties du système offre la possibilité de mettre à la disposition de l'évaluateur l'ensemble des informations qui ont conduit à produire un résultat donné et qui sont essentielles à son interprétation. Dans ce cas d'ALIBI, les alignements sont produits grâce à des relations syntaxiques et des couples amorces, et ce dans des contextes biphrastiques donnés. Ces informations permettent notamment de catégoriser les cas d'erreurs en fonction de leur origine.

3.2. *Apports de l'évaluation par annotation des sorties*

Du point de vue quantitatif, le tableau 2 montre un extrait des résultats de l'évaluation d'un ensemble de règles d'alignement, chacune d'elles étant évaluée de manière individuelle, et ce sur deux corpus. À chaque règle d'alignement évaluée est associé

le nombre de cas d'application de la règle (**Cas**), le nombre de cas évalués (**Éval.**) et celui de cas corrects (+) ainsi que la précision obtenue (**P**) définie par :

$$P = \frac{\text{nb d'alignements corrects évalués}}{\text{nb d'alignements évalués}}$$

	INRA				JOC			
	Cas	**Éval.**	+	**P**	**Cas**	**Éval.**	+	**P**
	Règles identiques							
V-suj-N/V-suj-N	2425	591	533	0,90	2581	588	538	0,91
V-suj-N/V-suj-N	1557	286	256	0,89	997	213	202	0,95
V-obj-N/V-obj-N	1528	595	567	0,95	2177	437	423	0,97
V-obj-N/V-obj-N	543	136	131	0,96	514	125	121	0,98
V-att-N|Adj/V-att-N|Adj	545	52	50	0,96	253	50	50	1,00
V-att-N|Adj/V-att-N|Adj	817	135	117	0,87	493	113	92	0,81
N-prep-V/N-prep-V	21	21	20	0,95	58	50	41	0,82
N-prep-V/N-prep-V	24	24	24	1,00	55	50	50	1,00
N-prep-N/N-prep-N	3554	755	598	0,79	4826	534	436	0,82
N-prep-N/N-prep-N	1740	414	379	0,91	2396	161	146	0,91
...	...							
	Règles compatibles							
N-nn-N/N-prep-N	3121	547	450	0,82	3341	291	243	0,83
N-nn-N/N-prep-N	1766	378	367	0,97	1496	65	63	0,97
N-prep-N/N-nn-N	132	76	18	0,24	147	50	16	0,32
N-prep-N/N-nn-N	35	35	26	0,74	43	43	19	0,44
N-nn-N/N-adj-Adj	619	62	55	0,89	2199	134	110	0,82
N-nn-N/N-adj-Adj	587	75	66	0,88	1137	53	44	0,83
N-adj-Adj/N-nn-N	280	51	49	0,96	207	57	39	0,68
N-adj-Adj/N-nn-N	256	51	45	0,88	171	53	43	0,81
N-prep-N/N-adj-Adj	59	50	8	0,16	92	50	13	0,26
N-prep-N/N-adj-Adj	35	35	14	0,40	129	50	11	0,22
N-adj-Adj/N-prep-N	243	53	29	0,55	452	50	32	0,64
N-adj-Adj/N-prep-N	335	51	22	0,43	470	52	37	0,71
...	...							

Tableau 2. *Extrait des résultats de l'évaluation d'un ensemble de règles d'alignement par annotation des sorties*

Ces données permettent de comparer la précision des règles sous différents angles — comparaison deux à deux pour les règles issues d'un même patron, comparaison globale des règles entre elles et comparaison en fonction du corpus — et de déterminer quelles sont les configurations syntaxiques où le principe d'analogie est vérifié de manière massive. On peut ainsi juger de la pertinence de chaque règle, retenir celles qui

ont une bonne précision, écarter celles qui ne sont pas performantes en terme de précision ou de productivité (nombre de cas couverts par une règle), comme par exemple la règle N-prep-N/N-adj-Adj dont la précision est de 0,40 sur le corpus INRA et de 0,22 sur le corpus JOC et la productivité faible sur les deux corpus. Plus globalement, on voit que les règles identiques offrent une meilleure précision que les règles compatibles, que certaines règles sont plus productives que d'autres, que la compatibilité entre deux relations syntaxiques différentes est orientée et non pas symétrique, etc.

Du point de vue qualitatif, on peut distinguer différents types d'erreurs : celles liées aux traitements effectués en amont de l'application des règles d'alignement, à savoir l'analyse syntaxique et l'amorçage, et celles liées à des phénomènes propres à la traduction tels que les différences de degré d'explicitation, les différences de densité lexicale ou encore les reformulations. Si on reprend l'exemple de la règle N-prep-N/N-adj-Adj, les configurations syntaxiques qu'elle met en jeu sont souvent corrélées à une différence d'explicitation dans les deux langues. Comme le montre la biphrase (3), la préposition anglaise, *of* dans ce cas, est rendue par une construction plus étoffée en français, à savoir un participe passé et une préposition. L'application de la règle à partir de *treatment*/*traitement* conduit à un alignement erroné. La régularité du phénomène, et donc de l'erreur, explique le manque de précision de la règle.

(3) *the monitoring of the treatment of the **refugees** in these camps*
*le contrôle du traitement **réservé** aux réfugiés dans ces camps*

L'évaluation par annotation des sorties se justifie pleinement dans le cadre de l'élaboration d'un système même s'il est vrai qu'elle est sans doute liée à un type d'architecture particulier, à savoir l'architecture modulaire. Elle permet au concepteur de l'application de vérifier la pertinence des différentes composantes et de maîtriser par là même l'enrichissement progressif du système. La possibilité de sélectionner les cas à évaluer rend justice à l'outil parce qu'elle permet de prendre en compte l'ensemble des situations où il est en mesure d'intervenir. Cette possibilité de sélection amène en revanche la question du caractère potentiellement subjectif de l'évaluation : si l'échantillon de données à annoter n'est pas fixé au départ, la possibilité de mettre à l'écart des cas difficiles peut biaiser les résultats de par la liberté laissée à l'évaluateur.

4. Évaluation multicorpus pour la comparaison des performances en fonction du corpus

4.1. *Principe de l'évaluation multicorpus*

Comme nous venons de le montrer, l'évaluation par annotation des sorties constitue un cadre privilégié pour vérifier quantitativement et qualitativement le bon fonctionnement des composantes d'un système ou encore son efficacité sur les sous-tâches qui concourent à la réalisation d'une tâche donnée. En revanche, elle ne semble pas constituer une démarche standard pour ce qui est de l'évaluation des systèmes d'ali-

gnement sous-phrastique, cette dernière se faisant en général à l'aide d'une référence construite indépendamment des sorties d'un système. D'autre part, cette méthode apparaît comme trop limitée pour évaluer un système dans son intégralité, c'est-à-dire estimer l'apport de la stratégie testée de manière relative par rapport à une stratégie de base aussi bien que ses performances absolues, notamment parce qu'elle rend difficile la prise en compte du rappel. Alors que la précision est obtenue à partir des réponses que le système propose et que l'on juge comme étant correctes ou non, le rappel ne peut être calculé qu'à partir d'informations additionnelles marquées par l'évaluateur. Enfin, elle ne permet pas d'apprécier de manière suffisamment fine la variabilité intercorpus et donc la généricité du principe d'alignement implémenté. Pour ces raisons, nous nous sommes tournée vers une évaluation avec une référence multicorpus construite indépendamment des sorties de notre système et en adéquation avec les pratiques en vigueur dans le domaine de l'alignement.

La prise en compte de la dimension intercorpus devrait constituer l'un des aspects fondamentaux de l'évaluation dans le sens où elle met en évidence l'influence du type de corpus sur la stratégie testée. Cependant, comme le notent A. Kilgarriff *et al.* (2003), force est de constater qu'elle reste encore très marginale, sans doute parce que difficile à mettre en œuvre. Nous avons pris le parti de considérer l'observation de la variabilité intercorpus comme faisant partie intégrante de notre protocole d'évaluation. Pour ce qui est de l'alignement sous-phrastique, il existe des références pour la Bible (Melamed, 1998a), le Hansard (Och *et al.*, 2003) et le JOC (Véronis *et al.*, 2000 ; Kraif, 2001). Cependant, étant donné que ces références n'ont pas été construites selon les mêmes conventions d'annotation, il est impossible d'y recourir pour mesurer des variations intercorpus. C'est pourquoi, dans notre cas, ce type de comparaison a nécessité la construction d'une référence multicorpus obtenue à partir d'échantillons de chacun des trois corpus de test. Pour ce faire, nous avons procédé conformément aux standards évoqués en introduction, à savoir en utilisant un guide d'annotation et en faisant appel à plusieurs annotateurs pour vérifier la fiabilité des références produites par le biais du calcul de l'accord interjuge.

Le calcul de l'accord interjuge permet de s'assurer que les mêmes décisions sont prises dans la plupart des cas par les différents annotateurs. En effet, il y a inévitablement une part de subjectivité dans tout processus d'annotation manuelle dans la mesure où il est soumis à des interprétations personnelles pouvant varier d'un individu à l'autre. Cependant, selon la tâche considérée, cette part de subjectivité et les différences d'annotation qui en découlent peuvent être plus ou moins importantes. Ainsi, compte tenu de la complexité du processus de traduction, l'annotation manuelle de correspondances traductionnelles présente un degré de difficulté plus élevé que celle de relations syntaxiques ou d'étiquettes morphosyntaxiques par exemple. Les correspondances sont souvent floues et ne s'établissent pas nécessairement mot à mot ce qui conduit notamment à des choix de segmentation différents. Ainsi, dans la biphrase (4), l'annotation choisie est plutôt synthétique — elle concerne des segments comme dans le cas de *allis shad*/*grande alose* — alors que dans la biphrase (5), elle est plutôt analytique — elle concerne des mots comme dans le cas de *shad*/*alose*.

(4) *The **[allis shad]**$_1$ **[is_considered to be]**$_2$ a vulnerable species*
*La **[grande alose]**$_1$ **[est considérée comme]**$_2$ une espèce vulnérable*

(5) *The **allis**$_1$ **shad**$_2$ **is_considered**$_3$ **[to be]**$_4$ a vulnerable species*
*La **grande**$_1$ **alose**$_2$ **[est considérée]**$_3$ **comme**$_4$ une espèce vulnérable*

Le calcul de l'accord interjuge repose sur le principe suivant. Quel que soit le nombre d'annotateurs sollicités, les références construites sont confrontées deux à deux. Soient X et Y, deux références construites à partir des mêmes données par des annotateurs distincts qui contiennent respectivement un ensemble d'alignements, autrement dit des couples de mots sources et cibles $(u, v) \in X$ et $(u, v) \in Y$. Estimer l'accord interjuge revient à comparer les couples $(u, v) \in X$ et les couples $(u, v) \in Y$ pour déterminer la part de ceux qui sont communs aux deux références. Pour cela, il faut mesurer :

– l'accord de X vis-à-vis de Y : $Accord_{X/Y} = \frac{\text{nb alignements communs}}{\text{nb alignements de Y}}$;

– l'accord de Y vis-à-vis de X : $Accord_{Y/X} = \frac{\text{nb alignements communs}}{\text{nb alignements de X}}$.

L'accord interjuge combine les valeurs $Accord_{Y/X}$ et $Accord_{X/Y}$, par exemple au moyen d'une moyenne géométrique : $Accord = 2 \times \frac{Accord_{Y/X} \times Accord_{X/Y}}{Accord_{Y/X} + Accord_{X/Y}}$

Des références fiables construites selon les mêmes conventions d'annotation permettent de suivre l'évolution des performances en fonction du corpus.

4.2. *Apports de l'évaluation multicorpus*

Trois personnes (J1, J2 et J3) ont participé à la construction des références multijuges pour les trois corpus : le concepteur de l'application ainsi que deux non-spécialistes de l'alignement bilingues. Chaque référence contient 120 biphrases dont 60 sont annotées par deux juges différents — 20 biphrases annotées par J1 et J2, 20 par J2 et J3 et 20 par J1 et J3 — et 60 par un seul juge. Chaque lot de 20 biphrases a fait l'objet d'une évaluation du point de vue de l'accord interjuge au moyen d'une moyenne géométrique comme mentionné ci-dessus. Le tableau 3 présente les taux d'accord obtenus pour chaque corpus sur l'ensemble des couples annotés en commun par deux juges.

	AJ1J2	**AJ1J3**	**AJ2J3**
INRA	0,90	0,89	0,88
JOC	0,87	0,86	0,85
HANSARD	0,76	0,82	0,72

Tableau 3. *Accord interjuge sur les trois références*

Globalement, l'accord interjuge est élevé : son taux est supérieur à 0,7. Les chiffres sont relativement homogènes d'un couple d'annoteurs à l'autre. En revanche, pour un couple d'annotateurs donné, l'accord est nettement inférieur sur le corpus HANSARD, 0,77 en moyenne, que sur INRA et sur JOC, 0,89 et 0,86 en moyenne.

Par ailleurs, comme le montre le tableau 4, les annotations manuelles permettent de caractériser les corpus en terme de répartition des types de correspondances : un-un (*disease*/*maladie* dans la biphrase (6)), avec élément zéro (cas où un mot source ou cible n'a pas de correspondant, comme par exemple *d'abord*), avec segment de plusieurs mots (cas où plusieurs mots sources et/ou cibles interviennent dans la correspondance, comme dans *about*/*au sujet de*). Les annotations effectuées reflètent également les pratiques des annotateurs.

(6) *A few words about this disease might help to put things into perspective*
D'abord quelques mots au sujet de cette maladie pour vous situer

	un-un	**avec zéro**	**segments**
	INRA		
J1	58%	18%	24%
J2	64%	15%	21%
J3	57%	13%	30%
	JOC		
J1	55%	25%	20%
J2	51%	22%	27%
J3	53%	21%	26%
	HANSARD		
J1	39%	19%	42%
J2	43%	21%	36%
J3	45%	25%	30%

Tableau 4. *Caractéristiques des trois références*

Cette caractérisation est intéressante étant donné que, comme nous venons de le mentionner, moins il y a de correspondances de type un-un et plus il y a de chances que les variations entre les juges soient importantes. Dans le corpus INRA, la traduction est proche du mot à mot et les correspondances sont principalement de type un-un. Dans le corpus HANSARD, la traduction est beaucoup plus libre, ce qui entraîne une augmentation de la taille des segments mis en correspondance et s'accompagne de variations dans la délimitation de ces segments. Les alignements entre des segments de plusieurs mots étant décomposés en liens individuels entre chaque mot source et chaque mot cible[9], plus la taille des segments augmente, plus les différences pèsent lourd dans

9. Nous reviendrons sur cette question dans la section 6.

l'estimation de l'accord interjuge. Si on reprend les deux alignements proposés pour *is considered to be*/*est considéré comme* dans les biphrases (4) et (5) données ci-dessus, le premier correspond à 9 liens et le second à 4, ce qui fait une différence de 5 liens.

Dans la mesure où l'appréciation de l'accord entre les juges sur chaque corpus a donné des résultats satisfaisants, nous avons unifié les annotations produites par chacun d'entre eux sur un corpus donné. Ainsi, nous disposons de trois références uniques de 180 biphrases (qui correspondent à 120 biphrases différentes), une pour chaque corpus, auxquelles nous confrontons les sorties du système pour estimer ses performances en terme de précision (P), rappel (R) et f-mesure (F). Dans notre cas, ces mesures se définissent de la manière suivante :

$$P = \frac{\text{nb d'alignements corrects trouvés}}{\text{nb d'alignements trouvés}}$$

$$R = \frac{\text{nb d'alignements corrects trouvés}}{\text{nb d'alignements attendus}}$$

$$F = \frac{2PR}{P+R}$$

Le tableau 5 montre le compotement d'ALIBI par rapport à une stratégie de base et en fonction du corpus. Dans notre cas, la stratégie de base correspond à un alignement obtenu avec un outil statistique, à savoir GIZA++ (Och *et al.*, 2003), entraîné sur nos trois corpus[10]. La confrontation des performances du système avec cette stratégie permet de mesurer l'apport (positif ou négatif) de la propagation syntaxique à l'alignement.

	INRA		**JOC**		**HANSARD**	
	Base	**ALIBI**	**Base**	**ALIBI**	**Base**	**ALIBI**
P	0,95	0,91 $(-0,04)$	0,93	0,87 $(-0,06)$	0,89	0,82 $(-0,07)$
R	0,66	0,75 $(+0,09)$	0,58	0,67 $(+0,09)$	0,43	0,53 $(+0,10)$
F	0,78	0,82 $(+0,04)$	0,71	0,75 $(+0,04)$	0,58	0,64 $(+0,06)$

Tableau 5. *Évaluation en fonction des stratégies de base et des corpus*

Plusieurs choses peuvent être notées à partir des chiffres présentés. Premièrement, par rapport à la stratégie de base, la propagation permet de gagner en f-mesure, ce résultat étant relativement homogène sur les trois corpus. ALIBI couvre plus de cas d'alignement que la stratégie de base mais au détriment de quelques points de précision. Par ailleurs, on constate des variations significatives selon les corpus. Ainsi, sur le corpus de textes scientifiques INRA, la précision est de 0,91 le rappel de 0,75. Sur le corpus de textes institutionnels JOC, les chiffres sont respectivement de 0,87 et 0,67. Enfin, sur le corpus de débats parlementaires HANSARD, ils sont de 0,82 et 0,53.

10. GIZA++ est disponible à l'adresse `http://www.jfoch.com/GIZA++.html`.

Du point de vue qualitatif, l'origine exacte des variations intercorpus reste difficile à identifier. Néanmoins, on peut tout d'abord supposer que, compte tenu de la caractérisation des références en fonction des types de correspondances, la granularité des alignements constitue l'un de ces facteurs, notamment pour ce qui est de l'estimation du rappel. En effet, l'analyse de la répartition des correspondances en fonction de leur type montre que le corpus INRA présente le taux le plus élévé de correspondances un-un. Par conséquent, dans le cas de HANSARD, il y a une surgénération beaucoup plus importante d'alignements de référence correspondant à des équivalences de type un-plusieurs, plusieurs-un ou plusieurs-plusieurs que dans les deux autres corpus. La prise en compte des alignements plusieurs-plusieurs en tant que liens individuels de chaque mot source vers chaque mot cible pose le problème de la représentativité du rappel estimé par rapport à la couverture effective de l'alignement. Ensuite, la qualité de l'analyse syntaxique peut également entrer en ligne de compte pour ce qui est des variations observées aussi bien au niveau de la précision que du rappel. Enfin, on peut supposer que selon les corpus la correspondance au niveau des structures syntaxiques reflète plus ou moins bien une correspondance au niveau sémantique.

En l'absence de références disponibles pour différents types de corpus, l'évaluation multicorpus n'est pas une chose aisée. En effet, l'élaboration de références est une tâche coûteuse et ce d'autant plus que plusieurs annotateurs sont sollicités. De ce fait, les références produites ont une taille limitée. L'évaluation n'étant pas menée à grande échelle, se pose alors le problème de la fiabilité des résultats. Bien évidemment, il n'est pas concevable, ni même souhaitable d'ailleurs, de construire des références gigantesques ou exhaustives mais de trouver un point d'équilibre garantissant que les performances mesurées sont stables. Dans notre cas, il est probable qu'avec 180 biphrases, ce point d'équilibre est loin d'être atteint.

5. Évaluation avec une référence standard pour la comparaison de différents systèmes dédiés à la même tâche

5.1. *Principe de l'évaluation avec une référence standard*

Quand il s'agit de comparer les performances obtenues à celles de systèmes dédiés à la même tâche, les meilleures conditions sont celles de campagnes d'évaluation où l'ensemble des systèmes est soumis aux mêmes contraintes pour ce qui est des données mises à disposition et du temps imparti pour les exploiter. À défaut, il est au minimum nécessaire de recourir à des données de référence standard mises à disposition publiquement. Les systèmes peuvent alors être comparés sur la base des taux de précision, de rappel et de f-mesure. En effet, lorsque les évaluations sont faites à titre individuel, avec des données non standard, les comparaisons s'avèrent délicates, et ce malgré l'existance de mesures de performances communes. La principale raison à cela est que les systèmes ne sont pas évalués dans les mêmes conditions lorsqu'ils le sont de manière individuelle : en général, les paramètres de l'évaluation — type de référence, paire de langues, type de corpus, etc. — varient d'un système à l'autre (Ahrenberg *et al.*, 2000 ; Véronis, 2000). Pour mettre en parallèle les performances d'ALIBI avec

celles des systèmes existants dédiés à la même tâche, nous avons donc choisi d'utiliser la référence mise à disposition dans le cadre de la campagne d'évaluation HLT-NAACL'03 (Mihalcea *et al.*, 2003). Comme nous l'avons mentionné en introduction, cette référence contient les alignements effectués par deux annotateurs sur 447 biphrases du Hansard. La nature de chaque lien est spécifiée par un attribut, à savoir S pour les alignements considérés comme sûrs par les deux annotateurs et P pour tous les autres cas, avec S $\subset$ P. La figure 2 présente un exemple de phrase annotée. Les alignements S sont en traits pleins et les P en pointillés.

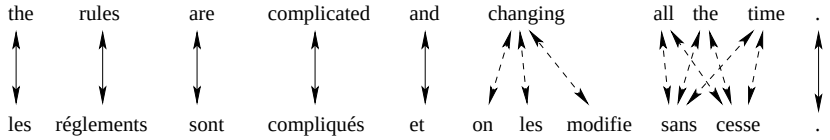

Figure 2. *Annotation de référence pour la campagne d'alignement HLT-NAACL'03*

La réutilisation de données de référence existantes pose le problème de leur éventuelle adaptation à un contexte d'exploitation nouveau dans le cas où celui-ci n'est pas identique au contexte d'origine. En effet, ALIBI travaille sur du texte présentant une segmentation en mots qui ne correspond pas nécessairement à la segmentation standard prenant l'espace comme séparateur[11]. Or, c'est sur la base de cette dernière qu'a été construite la référence HLT-NAACL'03. Dans le cas de la biphrase de la figure 2, par exemple, les séquences *all the time* et *sans cesse* sont regroupées et donc considérées comme un seul mot dans chaque langue et non pas comme trois et deux mots respectivement. Pour être en mesure d'évaluer les alignements produits par ALIBI, nous avons donc dû adapter la référence originale à cette nouvelle segmentation. Si une partie de ce travail d'adaptation a pu être effectuée de manière automatique, nous avons tout de même été obligée de procéder à une vérification et des corrections manuelles. La figure 3 montre le résultat de cette conversion sur la même biphrase que précédemment. Les différences par rapport à la segmentation standard, autrement dit les regroupements faits dans SYNTEX, sont signalées à l'aide d'un tiret bas (_).

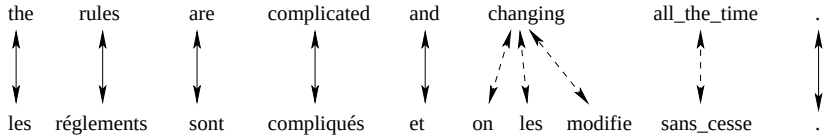

Figure 3. *Annotation de référence après conversion*

Il en ressort que, dans certains cas, la récupération d'une référence existante ne peut se faire qu'au prix d'une expertise humaine conséquente de sorte qu'il peut arriver

11. Il s'agit d'une segmentation propre à l'analyseur SYNTEX utilisé en amont d'ALIBI.

que l'investissement nécessaire à cette récupération soit identique, voire supérieur, à celui requis pour la mise au point d'une nouvelle référence.

5.2. *Apports de l'évaluation avec une référence standard*

Les tableaux 6 et 7 résument les performances d'ALIBI et celles des meilleurs systèmes en terme de f-mesure qui ont pris part à la campagne HLT-NAACL'03 — Ralign (Simard *et al.*, 2003), XRCE (Déjean *et al.*, 2003), BiBr (Zhao *et al.*, 2003) et ProAlign (Lin *et al.*, 2003) — telles qu'elles ont été estimées à partir de la même référence. Les systèmes Ralign, XRCE, BiBr et ProAlign auxquels nous nous comparons utilisent principalement des approches statistiques. Ils exploitent éventuellement différents types de contraintes : contrainte de compositionnalité pour Ralign et contraintes syntaxiques pour BiBr et ProAlign. Les performances affichées pour ces systèmes sont tirées de (Mihalcea *et al.*, 2003). Pour l'ensemble des systèmes, les chiffres du tableau 6 sont ceux obtenus sur les alignements de type S. Les chiffres du tableau 7 sont ceux obtenus sur les alignements de type P. De plus, dans le cas d'ALIBI, seuls ont été pris en compte lors de l'évaluation les alignements ne portant pas sur les ponctuations.

	ALIBI	**Ralign**	**XRCE**	**BiBr**	**ProAlign**
P	88,78%	72,54%	55,54%	63,03%	71,94%
R	66,86%	80,61%	93,46%	74,59%	91,48%
F	76,28%	76,36%	69,68%	68,32%	80,54%

Tableau 6. *Mise en parallèle des performances du système* ALIBI *par rapport à celles des participants à la campagne d'évaluation HLT-NAACL'03 (alignements S)*

	ALIBI	**Ralign**	**XRCE**	**BiBr**	**ProAlign**
P	90,81%	77,56%	89,65%	66,11%	96,49%
R	22,49%	38,19%	34,92%	30,06%	28,41%
F	36,05%	51,18%	50,27%	41,33%	43,89%

Tableau 7. *Mise en parallèle des performances du système* ALIBI *par rapport à celles des participants à la campagne d'évaluation HLT-NAACL'03 (alignements P)*

À l'examen du tableau 6, on remarque que les performances globales de notre système sont de niveau comparable à celles des autres. En revanche, les chiffres du tableau 7 sont plus contrastés. Tout comme pour les alignements S, la précision du système ALIBI sur les alignements P reste élevée. Cependant, le rappel est trop faible pour concurrencer les autres systèmes. Il en ressort que les règles d'alignement syntaxique mises en œuvre dans ALIBI ne sont pas conçues pour produire des alignements ambigus de type P où des déterminants peuvent se trouver alignés avec des verbes, par

exemple. Le système ProAlign, utilisant des contraintes syntaxiques, souffre d'ailleurs du même problème comparativement aux trois autres systèmes à fonctionnement plus purement statistique. Néanmoins, il faut souligner que pour arriver aux résultats affichés, les systèmes statistiques nécessitent d'importantes quantités de données pour s'entraîner, 1,3 million de biphrases dans le cadre de la campagne HLT-NAACL'03, alors que les performances d'ALIBI ne sont pas conditionnées par la taille du corpus. De plus, contrairement à bon nombre de techniques statistiques, les résultats obtenus avec ALIBI sont parfaitement interprétables au niveau individuel, c'est-à-dire qu'il est possible de déterminer avec précision les raisons pour lesquelles un alignement a réussi ou échoué. Le recours aux données syntaxiques et à leur encodage explicite sous forme de règles d'alignement rend les traitements transparents là où la complexité mathématique des modèles statistiques en fait souvent des boîtes noires. Enfin, comme nous l'avons déjà signalé, la décomposition des correspondances impliquant des segments de plusieurs mots en alignements individuels de type P pose problème. Il n'est pas certain qu'il s'agisse d'une solution adéquate pour rendre compte des performances, plus particulièrement pour fournir un rappel qui soit représentatif. En effet, cette décomposition introduit une surgénération importante des alignements de référence — une correspondance telle que *all the time*/*sans cesse* est représentée par 6 alignements — par rapport aux alignements qui peuvent être effectivement produits par les systèmes. Ce mode de jugement par rapport à la référence semble donc problématique dans le sens où le caractère « segmental » de cette dernière contraste avec le caractère « non segmental » des alignements produits. Nous discuterons plus en détail cette question dans la section suivante.

Le recours à une référence commune ne suffit pas toujours à garantir exactement les mêmes conditions d'évaluation en dehors du cadre des campagnes d'évaluation. Comme on vient de le voir, la réutilisation d'une référence standard en dehors du contexte d'origine peut nécessiter certaines adaptations changeant par là même quelque peu les conditions d'évaluation. Si tel est le cas, les performances mesurées ne sont sans doute pas directement comparables avec celles mesurées dans le contexte d'origine. Il n'en reste pas moins que l'utilisation d'une référence standard autorise une mise en parallèle des performances qui est sans aucun doute plus légitime que dans le cas de l'utilisation d'une référence individuelle.

6. Prise en compte des alignements avec des segments de plusieurs mots

La prise en compte des alignements avec des segments de plusieurs mots constitue une difficulté propre à la tâche d'alignement sous-phrastique. Comme nous l'avons évoqué plus haut, si on considère le nombre de mots de part et d'autre d'une biphrase qui composent un alignement, on peut distinguer différents types de correspondances : un-zéro, zéro-un, un-un, un-plusieurs, plusieurs-un, plusieurs-plusieurs. Les correspondances qui font intervenir des segments de plusieurs mots dans la partie source ou cible d'une biphrase peuvent être représentées et prises en compte lors de l'évaluation de différentes manières. Soient m et n, le nombre de mots dans les segments source

et cible respectivement, avec $m \geq 1$ et $n > 1$, ou $m > 1$ et $n \geq 1$. Si les segments ne sont pas considérés comme des unités, leur mise en correspondance est faite par l'énumération de tous les liens individuels entre chaque mot du segment source et chaque mot du segment cible (Melamed, 1998a ; Mihalcea *et al.*, 2003). Dans ce cas, le nombre de liens est égal à $m \times n$. C'est ce que nous appelons une référence mot à mot. En revanche, s'ils sont considérés comme des unités, les segments sont mis en correspondance à la manière de liens un-un standard lorsque ceux-ci portent sur un seul mot source et un seul mot cible (Ahrenberg *et al.*, 2000 ; Carl *et al.*, 2003 ; Tiedemann, 2003). Dans ce cas, un seul lien est généré entre le segment source et le segment cible et ce indépendamment des valeurs de m et de n. C'est ce que nous appelons une référence segment à segment. Prenons l'exemple de la biphrase (7).

(7) *The rules are complicated and changing all the time*
Les règlements sont compliqués et on les modifie sans cesse

Le tableau 8 présente les alignements de référence correspondant à cette biphrase selon que ceux-ci sont construits mot à mot (C_{mot}), ou segment à segment ($C_{segment}$).

Référence mot à mot C_{mot} (14 couples)	
the	les
rules	règlements
are	sont
complicated	compliqués
and	et
changing	on
changing	les
changing	modifie
all	sans
all	cesse
the	sans
the	cesse
time	sans
time	cesse
Référence segment à segment $C_{segment}$ (7 couples)	
the	les
rules	règlements
are	sont
complicated	compliqués
and	et
changing	on les modifie
all the time	sans cesse

Tableau 8. *Référence mot à mot versus référence segment à segment*

Dans le cas de C_{mot}, le nombre d'alignements de référence s'élève à 14 dont 6 relèvent de la correspondance traductionnelle entre les segments *all the time* et *sans cesse* Dans le cas de $C_{segment}$, le nombre d'alignements de référence est de 7. Les correspondances traductionnelles entre *changing* et *on les modifie* d'une part, et *all the time* et *sans cesse* d'autre part, sont établies segment à segment grâce à un lien unique, comme des correspondances un-un standard.

Le choix de l'une ou l'autre des deux stratégies pour la prise en compte des correspondances autres qu'un mot pour un mot entraîne des variations dans l'estimation des taux de précision et de rappel (Carl *et al.*, 2003 ; Tiedemann, 2003). Reprenons l'exemple de la biphrase (7). Supposons que pour cette biphrase le système retourne les 8 alignements présentés dans le tableau 9.

Alignements système (8 couples)	
the	les
rules	règlements
are	sont
complicated	compliqués
and	et
changing	modifie
the	cesse
time	cesse
Alignements corrects par rapport à C_{mot} (8 couples)	
the	les
rules	règlements
are	sont
complicated	compliqués
and	et
changing	modifie
the	cesse
time	cesse
Performances mot à mot	
P	**R**
1 (8/8)	0,57 (8/14)
Alignements corrects par rapport à $C_{segment}$ (5 couples)	
the	les
rules	règlements
are	sont
complicated	compliqués
and	et
Performances segment à segment	
P	**R**
0,62 (5/8)	0,71 (5/7)

Tableau 9. *Performances par mots versus performances par segments*

On voit que les performances varient considérablement en fonction de la référence utilisée en terme de précision et de rappel. La précision passe de 1 pour la référence C_{mot} à 0,62 pour la référence $C_{segment}$. En effet, alors que les alignements *changing*/*modifie*, *the*/*cesse* et *time*/*cesse* sont répertoriés comme corrects dans la première, ce sont les alignements complets *changing*/*on les modifie* et *all the time*/*sans cesse* que l'on s'attend à trouver selon la seconde. Les alignements *changing*/*modifie*, *the*/*cesse* et *time*/*cesse* sont donc considérés comme incorrects. Quant au rappel, il suit le mouvement inverse : 0,57 par rapport à la référence mot à mot C_{mot} et 0,71 par rapport à la référence segment à segment $C_{segment}$[12]. Comme on l'a vu plus haut, le nombre de liens est deux fois supérieur dans C_{mot} par rapport à $C_{segment}$ du fait de l'annotation des liens un-plusieurs et plusieurs-plusieurs par l'ensemble des liens un-un qui les composent, ce qui augmente les chances d'avoir les bons alignements parmi ceux qui sont retournés mais diminue celles de les avoir tous. Les variations observées peuvent donc être conséquentes et cependant les stratégies adoptées lors d'évaluations individuelles ne sont pas toujours clairement explicitées dans les travaux qui en rapportent les résultats.

Compte tenu des chiffres de précision et de rappel mesurés à l'aide de chaque référence, il semble qu'une référence mot à mot a tendance à surévaluer la précision alors qu'une référence segment à segment surévalue le rappel. Dans le premier cas, on peut se poser la question de savoir quel est l'intérêt de trouver un lien tel que *the*/*cesse* si l'on ne trouve pas les autres liens permettant de mettre en correspondance les deux expressions complètes *all the time*/*sans cesse*. Dans le second cas, le fait de mettre un lien direct entre les deux segments *all the time* et *sans cesse* ne suppose-il pas que l'on ait été capable de les isoler à un moment donné du processus d'alignement, chose dont les systèmes actuels ne sont pas véritablement capables ? Cet aspect peut être pris en compte lors de l'évaluation en incluant les alignements partiels dans les alignements corrects avec pénalisation éventuelle en fonction du recouvrement entre l'alignement de référence et l'alignement proposé (Véronis *et al.*, 2000). L'écart par rapport à une précision calculée avec une référence mot à mot tend ainsi à se réduire.

7. Conclusion

L'évaluation par annotation des sorties d'un système particulier apparaît adéquate pour vérifier quantitativement et qualitativement la pertinence des composantes d'un système prises individuellement et pour comparer les performances de leurs versions successives en terme de précision dans le cadre d'une démarche par essai-erreur. La possibilité de sélectionner les cas à évaluer parmi les sorties du système permet de s'assurer que l'ensemble des composantes est bien pris en compte et ce à une échelle suffisante pour que les résultats puissent être considérés comme représentatifs. En revanche, dès lors que différents corpus, d'une part, et différents systèmes dédiés à la même tâche, d'autre part, sont concernés, l'évaluation par rapport à une référence construite indépendamment des sorties du système est la seule valable. Les difficultés

12. À titre indicatif, F est de 0,73 pour C_{mot} et de 0,66 pour $C_{segment}$.

liées à la prise en compte de la dimension intercorpus ou encore à la comparaison entre systèmes, deux aspects fondamentaux de l'évaluation des systèmes de TAL, peuvent alors être d'autant mieux résolues que les performances sont estimées en accord avec les standards qui prévalent dans le domaine concerné. À cet égard, bien qu'elle puisse à première vue paraître modeste, l'expérience d'évaluation sur trois corpus de nature différente relève d'une démarche originale dans le domaine de l'alignement sous-phrastique. Elle met clairement au jour des différences notables entre les corpus ainsi que leurs répercussions sur les performances du système testé. Pour ce qui est de la comparaison entre systèmes, il est indéniable que les meilleures conditions sont offertes par les campagnes d'évaluation où exactement les mêmes données sont mises à disposition de tous les participants, dans des limites de temps identiques pour tous. À défaut de participation à une campagne d'évaluation, la réutilisation d'une référence mise à disposition dans de tels contextes constitue la condition minimale d'une mise en parallèle des performances. Les paramètres des évaluations effectuées à titre individuel, sans données standard, variant généralement d'un système à l'autre, les performances mesurées ne sont pas comparables.

8. Bibliographie

Ahrenberg L., Andersson M., Merkel M., « A knowledge-lite approach to word alignment », *in* J. Véronis (ed.), *Parallel Text Processing : Alignment and Use of Translatin Corpora*, Kluwer Academic Publishers, Dordrecht, chapter 5, p. 97-138, 2000.

Bourigault D., Un analyseur syntaxique opérationnel : SYNTEX, Thèse d'habilitation à diriger les recherches, Université Toulouse 2, France, 2007.

Bourigault D., Fabre C., Frérot C., Jacques M.-P., Ozdowska S., « SYNTEX, analyseur syntaxique de corpus », *Atelier EASy (Évaluation des Analyseurs Syntaxiques), Actes de la Conférence Traitement Automatique des Langues Naturelles*, Dourdan, France, 2005.

Carl M., Fissaha S., « Phrase-based Evaluation of Word-to-Word Alignments », *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Canada, p. 31-35, 2003.

Chaudiron S., « La place de l'usager dans l'évaluation des systèmes de recherche d'informations », *in* S. Chaudiron (ed.), *Évaluation des systèmes de traitement de l'information*, Hermès, Paris, p. 287-310, 2004.

Debili F., Zribi A., « Les dépendances syntaxiques au service de l'appariement de mots », *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artifi cielle*, Rennes, France, p. 81-90, 1996.

Déjean H., Éric Gaussier, Goutte C., Yamada K., « Reducing parameter space for word alignment », *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Canada, p. 23-26, mai, 2003.

Hirschman L., Mani I., « Evaluation », *in* R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, p. 414-429, 2003.

Ide N., Véronis J., « MULTEXT (Multilingual Tools and Corpora) », *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japon, p. 588-592, 1994.

Kilgarriff A., Grefenstette G., « Introduction to the special issue of Web as a Corpus », *Computational Linguistics*, vol. 29, n° 3, p. 333-338, 2003.

Kraif O., Constitution et exploitation de bi-textes pour l'aide à la traduction, Thèse de doctorat, Université de Nice Sophia Antipolis, France, 2001.

Lepage Y., « Analogie et traitement automatique des langues », *Actes de la Conférence Traitement Automatique des Langues Naturelles*, Leuven, Belgique, p. 781-791, 2006.

Lin D., Cherry C., « ProAlign : Shared Task System Description », *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Canada, p. 11-14, 2003.

Melamed I. D., Annotation Style Guide for the Blinker Project, Rapport technique, Institute for Research in Congnitive Science, University of Pennsylvania, Philadelphie, États-Unis, 1998a.

Melamed I. D., Manual Annotation of Translational Equivalence : The Blinker Project, Rapport technique, Institute for Research in Congnitive Science, University of Pennsylvania, Philadelphie, États-Unis, 1998b.

Merkel M., Annotation style guide for the PLUG link annotator, Rapport technique, Université de Linköping, Suède, 1999.

Mihalcea R., Pedersen T., « An Evaluation Exercise for Word Alignment », *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Canada, p. 1-10, 2003.

Och F. J., Ney H., « A Systematic Comparison of Various Statistical Alignment Models », *Computational Linguistics*, vol. 1, n° 29, p. 19-51, 2003.

Ozdowska S., ALIBI, un système d'ALIgnement BIlingue à base de règles de propagation syntaxique, Thèse de doctorat, Université Toulouse 2, France, 2006.

Paroubek P., « L'évaluation des systèmes d'analyse morphosyntaxique et syntaxique », *in* S. Chaudiron (ed.), *Évaluation des systèmes de traitement de l'information*, Hermès, Paris, p. 101-125, 2004.

Simard M., Langlais P., « Statistical Translation Alignment with Compositionality Constraints », *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Canada, p. 19-22, 2003.

Tiedemann J., Recycling Translations. Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing, Thèse de doctorat, Uppsala Universitet, Suède, 2003.

Véronis J., « Alignement de corpus multilingues », *in* J.-M. Pierrel (ed.), *Ingénierie des langues*, Hermes, Paris, chapter 6, p. 151-171, 2000.

Véronis J., Langlais P., « Evaluation of Parallel Text Alignment Systems. The ARCADE Project », *in* J. Véronis (ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Kluwer Academic Publishers, Dordrecht, chapter 19, p. 369-388, 2000.

Zhao B., Vogel S., « Word Alignment Based on Bilingual Bracketing », *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Canada, p. 15-18, 2003.