Traitement automatique de textes parallèles : le cas de l'occitan moderne

Gérard Ligozat

LIMSI, Université Paris-Sud

et

Université Adam Mickiewicz, Poznań

1. Textes parallèles et bitextes

On parle de deux textes parallèles, lorsqu'on a affaire à un texte dans une langue et à sa traduction dans une autre langue. Par exemple, le texte de « Mirèio » de Mistral, en occitan, et sa traduction en français, par Mistral lui-même, constituent deux textes parallèles (Mistral, 1859).

Une constatation toute simple est à l'origine du projet que nous présentons ici, à savoir : une grande partie de la littérature moderne en occitan est en fait constituée de textes parallèles !

Exemple 1 : *Mirèio, Cant segound* (traduction et notes de F. Mistral)

Cantas, cantas, magnanarello ! Que la culido es cantarello ! Galant son li magnan e s'endormon di tres ; Li amourié soun plen de fiho Que lou bèu tèms escarrabiho, Coume un vòu de blóundis abiho Que raubon sa melico i roumanin dóu gres.	Chantez, chantez, *magnanarelles* ! (1) Car la cueillette aime les chants. Beaux sont les vers à soie, et ils s'endorment de leur troisième somme ; (2) Les mûriers sont pleins de jeunes filles Que le beau temps rend alertes et gaies, Telles qu'un essaim de blondes

	abeilles Qui dérobent leur miel au romarin des champs pierreux.

(1) Magnanarelles (*magnanarello*). On désigne par ce mot les femmes préposées à l'éducation des vers à soie (*magnan*).

(2) Ils s'endorment de leur troisième somme (*s'endormon di tres*). Les vers à soie vivent à l'état de larve trente-quatre jours environ, et dans cet intervalle changent quatre fois de peau. A l'approche de chaque mue, ils s'engourdissent et cessent de manger (*dormon*). On dit *dourmi de la proumiero, di dos, di tres, di quatre*, ce qui signifie littéralement : *dormir de la première mue, des deux mues, des trois mues*, etc.

Avant d'aller plus loin, citons deux autres exemples de textes parallèles occitans-français ; l'un est tiré de « Le Troubadour », de Fabre d'Olivet (Fabre d'Olivet, 1803) ; l'autre correspond à un passage de « Portulan », de Roland Pécout (Pécout, 1978), et à sa traduction par Marie-Jeanne Verny (Verny, 2004).

Exemple 2 : Fabre d'Olivet, *Le Troubadour*

Chaûmès pas mai ; hoû ! vendemiaires, hoppo ! dépés ! léû al trabal ; vous espéroun lous bouteilhaires, acò's proù chourat, avén sal prenéz banastas é desquetas, levadoùs, semals e tinetas. L'un, qu'acampe ; l'autre qu'estuja ; resto plus-rès dins lou tinel, que la rapuga tout'essuja ; lou trouilhaire espéro un faissel.	Ne tardez pas d'avantage ; holà ! vendangeurs, debout ! allons, vite, à l'ouvrage ; les pauvres grappilleurs sont là qui vous attendent ; allons, c'est assez prendre du repos ; le travail vous rappelle. Et tôt, prenez ces larges marines d'osier, ces petites corbeilles, ces cornues propres à conserver le jus de raisin. Que l'un cueille le fruit de la vigne, tandis que l'autre l'entassera dans les cornues ; il ne reste bientôt plus rien, dans la cuve à fouler, que la grappe dépouillée, et le marc aride ; le fouleur attend une nouvelle cuvée.

Exemple 3 : Roland Pécout, *Portulan*

De ton ajocador auses de bruchs, de rises. Davalas. Mostrar que siás aquí. Reintegrar ton tu. Cubrir teis uelhs d'un agach espinchable. Una còla de pichons, bralhas mascaradas, morets. Te fan la saludada. Respondes a cadun, que cadun es cadun. Un pichonet te pren la man e compta lei dets, puei s'escacalassa. Lei grands volon de cigarretas.	Perché, tu entends des bruits, des rires. Tu descends. Montrer que tu es là, te réintégrer. Te vêtir de regard regardable. Une bande de gamins, aux vêtements déchirés, te salue. Tu réponds à chacun, car chacun est chacun. Un petit te prend la main et compte les doigts, puis éclate de rire. Les grands veulent des cigarettes.

On parle d'alignement de textes parallèles (au niveau des phrases) lorsque les phrases de chacun des deux textes sont mises en correspondance.

Nous avons vu ci-dessus trois exemples de textes parallèles. Un alignement au niveau des phrases donnerait pour le début de l’exemple 3 le résultat suivant :

De ton ajocador auses de bruchs, de rises.	Perché, tu entends des bruits, des rires.
Davalas.	Tu descends.
Mostrar que siás aquí.	Montrer que tu es là, te réintégrer.
Reintegrar ton tu.	
Cubrir teis uelhs d'un agach espinchable.	Te vêtir de regard regardable.
Una còla de pichons, bralhas mascaradas, morets.	Une bande de gamins, aux vêtements déchirés, te salue.
Te fan la saludada.	

On constate en particulier qu'à deux reprises deux phrases de l'original donnent lieu à une seule phase dans la traduction.

La notion de *bitexte* a été introduite par Harris (Harris, 1988) pour désigner des textes parallèles alignés. On considère également, en particulier dans le domaine de la traduction automatique, des alignements à d'autres niveaux qu'à celui des phrases, en particulier au niveau des mots. Plus généralement, on parle de *multitexte* lorsqu'on est en présence de plusieurs traductions alignées d'un même texte dans plusieurs langues.

2. Un corpus considérable

Bien qu'il ne soit pas lieu ici de faire un bilan et d'essayer de déterminer quel pourcentage de l'écrit en occitan moderne possède une traduction en français, il ne fait pas de doute que cette proportion est élevée. Pour ce qui relève de l'École du Félibrige, la traduction française en regard semble être très largement majoritaire. Bien qu'une tendance à la publication monolingue ait été très nette autour de l'IEO dans la seconde moitié du XX^{ème} siècle, il n'en reste pas moins

que nombre d'auteurs contemporains ont été traduits, s'ils ne se sont pas traduits eux-mêmes comme cela est très souvent arrivé (Max Rouquette et Bernard Manciet étant deux des exemples les plus marquants).

Les textes parallèles occitans-français étant nombreux, il nous paraît intéressant et important de les mettre à profit dans le cadre du traitement automatique des langues, et cela dans un contexte de développement de ressources linguistiques pour l'occitan moderne. Pour quelles raisons ? La réponse découlera de la réponse à une question plus générale : à quoi peuvent servir les textes parallèles ? Avant de traiter cette question, nous allons voir rapidement quels ont été jusqu'à présent les travaux consacrés aux textes parallèles occitans-français.

3. Corpus parallèles, utilisations

L'intérêt pour la recherche des corpus parallèles (ensemble de textes parallèles) et, plus généralement, les corpus comparables (ensembles de textes en diverses langues consacrés à des thèmes analogues) est bien identifié par Ajmer et Altenberg (Aijmer et Altenberg, 1996) :

« Les corpus parallèles et comparables :

– mettent en relief les différences entre textes sources et traductions, et entre textes provenant de locuteurs natifs et non-natifs ;

– amènent à des idées nouvelles sur les langues comparées, idées qui auraient eu peu de chances de résulter de la simple étude de corpus monolingues ;

– peuvent être utilisés pour toute une gamme de tâches comparatives, et augmentent notre connaissance des caractéristiques typologiques et culturelles propres à chaque langue, ainsi que des caractéristiques universelles ;

– peuvent être utilisés pour de nombreuses applications pratiques, par exemple pour la lexicographie, l'enseignement des langues, et la traduction. »

Pour mieux caractériser la nature de ces applications, il peut être utile de revenir brièvement sur la nature et les méthodes du traitement automatique des langues (TAL) tel qu'il se pratique actuellement.

4. Le Traitement Automatique des Langues

L'idée même de traitement automatique de traductions peut paraître bien problématique pour un linguiste : la machine n'est-elle pas condamnée, lorsqu'elle va prendre un texte, à en faire une traduction mot-à-mot, au pire une traduction mécanique comme on en trouve maint exemple dans certaines notices d'utilisation, alors que la traduction implique au contraire la connaissance profonde du « génie de la langue », et suppose la créativité, l'intelligence ? Traitement automatique n'est-il pas nécessairement synonyme de traitement bêtement et irrémédiablement mécanique ?

Eh bien, c'est précisément le côté systématique et « bêtement » mécanique du traitement automatique qui présente un intérêt : si le linguiste échafaude des hypothèses, développe des théories, leur automatisation permet de les mettre à l'épreuve sans complaisance. L'utilisation de corpus met en relief, par ailleurs, la portée et la valeur réelle de ces hypothèses ou théories. Et éventuellement leurs failles et leurs insuffisances.

S'agissant de traductions, le traitement automatique permet d'envisager trois phases :

– l'étude automatique de la traduction permet de formuler des hypothèses ;

– ces hypothèses peuvent alors être testées automatiquement ;

– après le passage au filtre du traitement automatique apparaissent les phénomènes qui ne se réduisent pas à ces principes, et qui peuvent constituer l'intérêt principal des textes étudiés.

Dernier point, les outils de traitement automatique (aide à la traduction) sont effectivement d'ores et déjà bien présents dans la communauté des traducteurs professionnels, qui sont à la fois

demandeurs envers la discipline du TAL, et inspirateurs par les problèmes qu'ils soulèvent. Ne serait-ce que de ce point de vue, un examen de ce que ces outils peuvent apporter pour les études occitanes serait justifié.

5. Bref résumé de l'histoire du TAL

Les activités relevant du TAL peuvent être rattachées à deux types d'approches : l'approche symbolique d'une part, et l'approche statistique d'autre part.

L'approche symbolique est la plus ancienne. Elle consiste essentiellement à fonder le traitement automatique sur une représentation explicite des règles à l'oeuvre dans la langue. Par exemple, s'il s'agit de faire une analyse automatique de la structure des phrases, on utilisera une grammaire de type Chomsky, qui énonce des règles explicites de bonne formation, ou l'un des formalismes grammaticaux qui ont été développés dans les années 70 et 80 sur la base de grammaires algébriques (ou « context-free ») : Lexical Functional Grammar, tree grammars, HPSG, grammaires d'unification diverses.

Cette approche se caractérise donc par l'utilisation de connaissances linguistiques explicites dans les domaines de la morphologie, de la syntaxe, de la sémantique, qui peuvent être accompagnées de connaissances sur « le monde » (utilisation de schémas, de scénarios, etc.).

L'approche statistique est plus récente, et ne se développe de manière massive que depuis les années 80, avec des réussites inattendues et impressionnantes. Son succès est lié au changements intervenus à la fin de cette décennie, qui a vu le développement d'Internet et ouvert la possibilité de travailler sur des corpus de très

grande taille.

Il est remarquable que l'approche statistique se fonde en général sur un modèle qui est celui le modèle du canal bruité. Par exemple, la traduction d'un texte T1 d'une langue L1 vers un texte T2 d'une langue L2 :

traduction : T1 --> T2

est vu comme une tâche de décodage : le texte T1 est considéré comme le résultat du passage d'un texte hypothétique T2 au travers d'un canal de transmission bruité. Il s'agit donc pour traduire de décoder ce produit bruité T1 pour reconstituer le texte T2 d'origine en surmontant les difficultés induites par le brouillage qu'a réalisé le canal de transmission..

transmission : T2 (entrée) --> CANAL --> T1 (sortie)

Par exemple, si l'on doit traduire le terme occitan *ajocador* en français, on considère que ce terme résulte après brouillage d'un terme français qu'il s'agit de retrouver.

De manière analogue, la reconnaissance de la parole sera vue comme le problème de décodage d'un signal sonore qui est la version « brouillée » d'une suite de mots présents en entrée.

On utilisera donc pour ce faire des modèles probabilistes (de la langue L2, donc, dans notre exemple, de l'occitan) : un tel modèle détermine la probabilité d'apparition de chacune des chaînes de mots possibles (dans notre exemple, en occitan) et du canal de transmission dont on devra estimer les paramètres (dans notre exemple, la probabilité de passage de l'occitan au français). Typiquement, cela se fait en deux phases :

– utilisation d'un corpus d'apprentissage pour la détermination des paramètres ;

– évaluation du modèle obtenu sur des corpus de test.

La constatation, surprenante à premier abord, qui a pu être faite au cours de la dernière décennie, est que ces méthodes peuvent être remarquablement efficaces alors même qu'elles n'utilisent que peu ou pas de connaissances sur la langue. On a ainsi construit et développé des outils très efficaces tels que des analyseurs morphologiques (POS taggers), des analyseurs syntaxiques, des logiciels de désambiguïsation, etc. Et, pour ce qui concerne les textes parallèles, des outils d'alignement automatique.

6. Alignement de textes parallèles

Le traitement automatique des textes parallèles est en effet centré sur l'opération *d'alignement***.** Comme nous l'avons mentionné plus haut, cet alignement peut être réalisé à divers niveaux : alignement de phrases, de mots, d'expressions, l'idée générale étant que, en présence de deux textes T1 et T2 dans deux langues respectives L1 et L2, il s'agit de déterminer quels sont les éléments du texte T2 qui correspondent à ceux du texte T1.

6.1. Techniques d'alignement

Il n'est pas lieu ici de décrire les diverses techniques utilisées pour réaliser l'alignement de deux textes. On pourra se reporter pour une présentation générale à l'article de Véronis (Véronis, 2000). Contentons-nous de mentionner l'existence de techniques descendantes, dans lesquelles on commence par aligner au niveau global du texte (par exemple, en supposant que la première phrase de T1 correspond à la première phrase de T2), pour ensuite raffiner progressivement, et des techniques ascendantes, dans lesquelles on part de mots utilisés comme points d'ancrage, pour utiliser ensuite ces points d'ancrage pour l'alignement à un niveau supérieur. Les mots

utilisés comme points d'ancrage peuvent être ce que l'on appelle des « cognates », par exemple des mots possédant un préfixe commun, ou des mots proches en terme de « distance d'édition », comme *government* et *gouvernement* (Simard et al., 1992 ; Ribeiro et al., 2001).

Répétons ici encore que, le plus souvent, très peu d'informations linguistiques sont utilisées. Par exemple, les travaux pionniers (mais efficaces) que constituent ceux de Brown et al. (Brown et al., 1990) et de Gale et Church (Gale et Church, 1993) n'utilisent que la connaissance de longueur des phrases (évaluée en nombre de mots ou de caractères) pour la mise en correspondance.

S'agissant d'alignement de phrases, on est amené à considérer la situation où, dans un bitexte, *m* phrases dans T1 correspondent à *n* phrases dans T2 : on parle alors de correspondance (m : n).

Par exemple, considérons à nouveau un extrait de l'exemple 3.

Mostrar que siás aquí.	Montrer que tu es là, te réintégrer.
Reintegrar ton tu.	
Cubrir teis uelhs d'un agach espinchable.	Te vêtir de regard regardable.

Les deux premières phrases en occitan correspondent à une seule en français : la correspondance est donc (2 : 1). Pour la troisème phrase, la correspondance est (1 : 1).

On peut ainsi caractériser les omissions comme cas (1 : 0) et les ajouts comme cas (0 : 1).

Dans la pratique, la plupart des systèmes partent de l'hypothèse que le cas (1 : 1) est majoritaire dans les textes parallèles qu'ils traitent, et ne considèrent que rarement les cas où *m* et *n* sont supérieurs à 2 (fusions complexes).

7. Textes parallèles occitans-français : travaux antérieurs

L'idée de travailler sur des textes parallèles occitans-français a donné lieu il y a quelques années au Projet MULTEXT-Cataloc (1997), qui s'inscrivait dans le cadre plus général du Projet MULTEXT (1994). Ce dernier avait pour ambition de développer des ressources linguistiques (lexiques, corpus, etc.) et de procéder à une adaptation des outils et des standards, en particulier pour des langues pauvres en ressources linguistiques. Le projet MULTEXT-East (Multext-East, 2007) concernait un certain nombre de langues de l'Europe de l'Est et de la Baltique (Bulgare, Croate, Estonien, Hongrois, Lituanien, Résien, Roumain, Russe, Slovène, Serbe, Tchèque). MULTEXT-Cataloc était un premier pas vers la création de telles ressources pour l'occitan.

A notre connaissance, le projet en question a abouti à la création de bitextes pour deux textes parallèles : le premier étant « Lo vesitaire dis estello » , de Rémi Blancon, et le second « Mirèio ». Les textes utilisés dans cette expérience sont tirés du corpus numérisé que représente la base de textes du « CIEL d'OC ».

Dans le premier cas, une traduction en français a été réalisée dans le cadre du projet. Les auteurs de ce travail accompagnent la description de leur travail du commentaire suivant :

> « La production d'un corpus aligné dans le contexte du traitement automatique de langues minoritaires peut paraître à première vue paradoxale dans la mesure où l'existence même d'un corpus pour une langue minoritaire est fragile (ces langues étant plus orales qu'écrites), rendant a fortiori leur traduction dans une langue non-minoritaire hasardeuse (puisque l'intérêt majeur est la production de textes dans la langue même) d'autant que la langue non-minoritaire alignée est souvent perçue comme historiquement concurrente de la langue minoritaire. »

Les doutes des auteurs ne nous paraissent pas justifiés : d'une part, l'opposition entre langue parlée et langue écrite ne nous paraît pas pertinente ici, outre que la tradition écrite de l'occitan est ancienne et

bien établie ; d'autre part, et c'est à nos yeux l'aspect le plus important, la traduction en français est bien souvent très révélatrice du point de vue de son auteur, du statut qu'il accorde à l'une et à l'autre langue, des publics qu'il vise, pour ne citer que quelques éléments. Pour leur part, les auteurs tirent de leur travail la conclusion suivante :

« On peut donc conclure de l'analyse des données de l'alignement que les changements de mise en page (versification -> prose) sont assez délicats à gérer automatiquement mais que dans les autres cas, l'alignement donne des résultats tout à fait satisfaisants ».

Raison de plus, à nos yeux, pour ne pas en rester là. L'étude des bitextes devrait donc se révéler particulièrement riche et fructueuse lorsqu'on aura la possibilité d'exploiter la richesse du corpus potentiel que constitue l'ensemble des textes de la littérature moderne et contemporaine en déployant pour cela toute la panoplie des techniques qu'offre à l'heure actuelle la recherche sur les textes parallèles.

8. Un nouveau projet : PARALOC (textes PARALlèles en OCcitan)

Il s'agit de mette à profit la richesse que constitue l'existence d'un corpus de textes parallèles de la littérature occitane moderne, et d'appliquer sur ce corpus les techniques de traitement automatique.

La première étape consistera donc à constituer un (des) corpus bilingue(s) numérisé(s) de la littérature occitane moderne.
En premier lieu, un état des lieux devra être réalisé, afin de déterminer précisément quels sont les textes possédant une ou plusieurs traductions, quels sont ceux d'entre eux qui sont d'ores et déjà disponibles sous forme électronique, et de définir une politique de numérisation, doublée le cas échéant d'un programme de traduction.

Un choix des priorités à appliquer devra être défini.

Cette première étape devrait pouvoir être réalisée avec la collaboration des programmes de numérisation déjà existants – nous pensons en particulier au projet de la base TELÒC (Bras, 2008 ; Thomas, 2008), et à celui du CIEL d'OC (Ciel d'Oc, 2000).

Il s'agira ensuite d'appliquer à ces corpus les méthodes de traitement automatique des textes parallèles. Cette partie du projet est beaucoup plus ouverte, l'idée étant que l'existence du corpus permettra aux chercheurs de définir des projets plus spécifiques en fonction de leurs propres intérêts.

Les nombreuses applications envisageables relèvent principalement de trois domaines : la lexicographie, la traduction et l'enseignement.

8.1. Applications des bitextes à la lexicographie

– constitution d'outils de concordances : on peut ainsi disposer aisément de toutes les traductions d'un mot donné, avec le contexte correspondant ;

– développement d'outils d'aide pour les dictionnaires en ligne ;

– extraction automatique de dictionnaires de mots simples ou d'unités complexes.

Citons un exemple d'utilisation des concordances pour la formation des traducteurs. Il est donné par Zanettin (Zanettin, 1998). Il s'agit de la traduction de l'italien vers l'anglais. Une recherche très élémentaire (occurrences successives, en italien, des termes *podio* et *gradino*, à une distance de cinq mots au plus, couplée avec une recherche similaire pour les termes *podium* et *step* permet aux auteurs de mettre en évidence le fait que la traduction la plus naturelle en anglais de l'expression *salire il gradino piu alto del podio* doit être *win the gold medal* plutôt qu'une traduction mot-à-mot.

8.2. Applications des bitextes à la traduction et à

l'enseignement de la langue

– résolution de problèmes de traduction, par la constitution de bases de solutions ;

– création et maintenance de textes parallèles eux-mêmes ;

– vérification de traductions, levée d'ambiguïtés ;

– correction d'erreurs (faux amis, omissions, rajouts, cohérence terminologique) ;

– anticipation de la frappe et amélioration de la reconnaissance vocale lors de dictée automatisée.

– constitution de banques d'exemples ;

– réalisation de méthodes d'enseignement basées sur les suites préfabriquées.

9. Conclusions et perspectives

Le point de départ de notre recherche tient en deux constatations : la première est l'existence potentielle d'un riche corpus bilingue pour la littérature occitane moderne. La seconde est que le traitement des textes parallèles est intéressant pour de multiples applications.

Notre projet consistera donc dans un premier temps à numériser ce corpus, puis à l'exploiter, et à fournir aux chercheurs des outils pour cette exploitation, tout en développant des applications particulières. Les applications les plus immédiates relèvent de la lexicologie, de la traduction, et de l'enseignement de la langue, mais il est clair que de nombreux domaines sont concernés. Par exemple, la stylistique, mais aussi la sociolinguistique pourront tirer profit d'un examen automatisé qui mettrait en évidence la façon dont telle ou telle traduction procède à des changements de registre systématiques, à des effacements des termes techniques : dans notre exemple 2, on pourra ainsi étudier de

manière systématique comment (et pourquoi) les termes techniques *banastas é desquetas, levadoùs, semals e tinetas* se réduisent dans la traduction à des *petites corbeilles,* des *cornues propres à conserver le jus de raisin* ; on pourra de même étudier dans la traduction de Mistral par lui-même (exemple 1) la glose systématique des termes perçus comme « typiques » sinon « exotiques », comme le sont les termes *magnanarelle* ou *cantarello*.

Nous pensons également que le traitement automatique contribuera à suggérer aux chercheurs des traitements nouveaux de questions anciennes, et peut-être à en soulever de nouvelles. Nous pensons par exemple à deux questions qui viennent à l'esprit lorsque l'on discute de traduction de l'occitan :

– qu'est-ce qu'un locuteur natif lorsqu'il est question de langues minorisées ?

– dans le même contexte, qu'est-ce qu'une traduction de qualité ?

Un sujet de réflexion sur ces deux points nous est fourni par l'extrait suivant :

> « Je pourrais vous les donner ici dans leur belle langue originale, mais j'aime mieux vous les traduire en m'aidant de la naïve traduction en pur français classique faite par le poète lui même. Nul ne sait mieux ce qu'il a voulu dire ; notre français à nous serait un miroir terne de son œuvre ; le sien à lui est un miroir vivant. À nous deux, nous répondrons mieux aux nécessités des deux langues. Lisons donc. C'est moi qui parle, mais c'est lui qui chante. »

Il s'agit du *Quarantième Entretien* du *Cours familier de littérature* d'Alphonse de Lamartine, daté de 1859, et qui est consacré à *Mirèio*.

Bibliographie

Aijmer K., Altenberg B. (1996). Introduction. In Aijmer K., Altenberg B. and Johansson M. (eds.), *Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies in Lund*, 4-5 March 1994. Lund, Lund University Press, pp. 11-16.

Bras M. (2008), Batelòc : Cap a una basa informatizada de tèxtes occitans, *dans ce volume*.

Brown P., Cocke J., Della Pietra S., Jelinek F., Lafferty J., Mercer R., & Roossin P. (1990), A statistical approach to machine translation. *Computational Linguistics, vol. 16*.

Ciel d'Oc (2000) : http://sites.univ-provence.fr/tresoc/

Fabre d'Olivet (1803), *Le Troubadour, poésies occitaniques*, réédition Lacour, Nîmes, 1997.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora, *Computational Linguistics, 19(3*), pp. 75-102.

Harris, B. (1988), Bi-texts: A new concept in translation theory, *Language Monthly*, 54, pp. 8-10.

Mistral, F. (1859), *Mirèio*, Garnier-Flammarion, 1978.

Multext-Cataloc : http://aune.lpl.univ-aix.fr/projects/multext-cataloc/reports/CORP-oc.html

Multext-East Home Page (2007) : http://nl.ijs.si/ME/

Pécout, R. (1978) *Portulan, Itinerari en orient*, Vent Terral.

Ribeiro, A., Dias, G., Lopes, G., Mexia, J. (2001). Cognates Alignment. In: Bente Maegaard (Ed.), *Proceedings of the Machine Translation Summit VIII (MT Summit VIII) - Machine Translation in the Information Age,* Santiago de Compostela, Spain, pp. 287-292.

Simard, M., Foster, G., Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In : *Proceedings of TMI92*, Montreal, Canada, pp. 67-81.

Thomas J. (2008), Cap a una basa informatizada de tèxtes occitans, *dans ce volume*.

Verny M.-J. (2004), Enrasigament o nomadisme : trajectoire d'un écrivain occitan de la fin du XXe siècle, Roland Pécout / Marie-Jeanne Verny, Puèglaurenç IEO.

Véronis, J. (2000). From the Rosetta stone to the information society: a survey of Parallel Text Processing. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora* (pp. 1-24). Dordrecht: Kluwer Academic Publishers.

Zanettin F. (1998), Bilingual comparable corpora and the training of translators. In Laviosa, Sara. (ed.) *META, 43:4, Special Issue*. The corpus-based approach: a new paradigm in translation studies, pp. 616-630.