社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－FI－80 (2)
2005－NL－169 (2)
2005／9／29

外部手段非依存型アライメントシステム『あられ』

デロワ中村 弥生†

本稿では、パラレルコーパスアライメントシステム『あられ』(AlALeR - Système d'Alignement Autonome, Léger et Robuste) を紹介する。本システムは、従来の手法と異なり、形態素解析器や辞書等の外部手段を用いることなく、日本語テキストと仏語あるいは英語テキストとの文レベル対応付けを行う。本稿の提案するアライメント手法は、日本語表記法の特徴を十分に利用した部分形態素解析とトランスデューサを用いた書き換えによる外来語の効果的な利用に特徴づけられる。

Autonomous Alignment System : AlALeR

YAYOI NAKAMURA-DELLOYE†

The present paper describes the AlALeR System, an Autonomous, Robust and Light Alignment System. Capable of aligning at the sentence level a French text and a Japanese one, the AlALeR System doesn't use any external tool, such as morphological analysers or dictionaries, contrary to existing methods. This system is characterized by a partial morphological analysis taking advantage of some peculiarities of japanese writing system, and by the transcription of loan words with a transducer.

1. はじめに

これまで提案された大多数の文レベルアライメント手法は、その実装および計算の簡素性に特徴づけられる。この簡素性は、主に、文字列の分布情報や長さといった内部情報のみを用いいることによるものである。

これに対し、日本語を扱うアライメントシステムにおいては、単語分割のために形態素解析器、そして日本語とは様々な面でその性質を異にする英語等の言語との対応付けを効果的に行うために対訳辞書が用いられてきた。

しかしながら、特に文レベルのアライメントは、他の処理の準備段階として利用されるきわめて基本的な作業であり、日本語を扱うアライメントにおいても外部手段に依存することのない手法が望まれる状況も多々存在するものと考えられる。そこで、日本語表記法の特徴を十分に利用し、形態素解析器や辞書等の外部手段を用いることなく日本語テキストのアライメントを行うシステム『あられ』(AlALeR) の開発を試みることとなった。

本稿では、外部手段非依存型アライメントシステム『あられ』を紹介する。まず従来手法を概観した後、システムの概要といくつかの過程の詳細を述べ、最後に実験結果について報告する。

2. 従来手法とシステム『あられ』

アライメント技術の研究は、機械翻訳の研究の枠組みの中で生まれた。そのため、先駆者の研究はすべてその実装および計算の簡素性を探求しており、これらの研究が生み出した手法は、文字列の分布情報[4] や長さ[1)3)] といった内部情報のみを用いいることを特徴としている。

これら先駆的研究の改良においても、欧米の研究者たちは、これらの方向性を追随し、外部情報を必要としない、同一語源語 (cognates) といった新たな概念を導入することによる改良が進められた[10)7)5)]。

しかし、日本語には、語境界を示す文字が存在しないことから、日本語を扱った研究においては、かなり早い時期からアライメントシステムに形態素解析器が導入された[8)]。また、日本語は、自然言語処理の分野で主に扱われている英、仏、独語等の言語と、語彙レベルにおいても構文レベルにおいてもその性質を異にすることから、これらの言語を扱った内部情報のみを用いる手法の単純な応用は不可能であった。そのため、日本語を扱った研究においては、対訳辞書が利用され、簡素性よりも精度が重視された[13)]。

システム『あられ』においては、まず語分割の問題は、字種切り法に改良を加えた部分形態素解析の利用により解決された。また、トランスデューサを用いた書き換えにより外来語を効果的に利用することで、対訳辞書を用いることなく語レベルの効率的な対応付けも実現した。

3. 『あられ』のアライメント手法

3.1 システム概要

システムの入力は、日本語テキストと仏語あるいは英語テキストで構成されたパラレルコーパスである。多言

† パリ第７大学大学院（言語科学）
フランス国立科学研究センター（CNRS）Laboratoire LaTTiCe (UMR 8094)

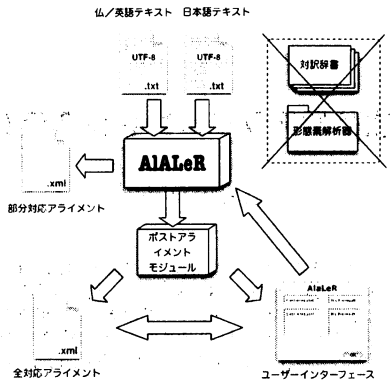

図 1 システム『あられ』

語処理において常に障害となるエンコーディングの問題に対処するため、入力には UTF8 を用いたテキスト形式のファイルを用いる。

『あられ』は、アライメントの結果として、極めて信頼性の高い部分対応付けテキスト、あるいは、オプションが選択された場合には、全対応付けテキストを出力する。全対応付けオプションが選択されると、ポストアライメントモジュールがメインプロセスで対応付けされなかった文の対応付けを行う。

これらの結果は、XML 形式のファイルで出力されるか、あるいはユーザーインターフェースに送られる。ユーザーインターフェースを利用すると、出力結果を読みやすい形態で表示できるだけでなく、必要な場合には出力結果を容易に修正できる☆。

3.2 アライメントの流れ

本システムは、文献[4] で提案された語彙分布情報に基づく手法を採用している。この手法は、対応する文は、それぞれが対応する要素で構成されるという仮定に基づいたもので、まず語の大まかな対応付けを行い、得られた対応語をもとに文の対応付けを行うというものである。

本システムのアライメント工程は、大きくわけて 2 つの過程で構成される。

(1) **語リスト構築過程**：トークンを、同一語源語（cognates）、不変化語（transfuges）、カタカナ語、その他の 4 種に選別し、それぞれリストを作成する。

(2) **アライメント過程**：
- **前処理過程（プレアライメント）**：アライメントの可能性を制限するために、同一語源語、不変化語、カタカナ語を用いてアンカーの設置を行う。
- **メインプロセス**：文に含まれる語の分布類似度の計算により、文の対応付けを行う。

☆ システムは C++ で、ユーザーインターフェースは Carbon（Apple API）で実装されている。

以下、これら各過程について述べる。システムは取り扱う言語によって多少動作が変わるが、本稿ではシステム独自の特徴がより明らかとなるよう、特に和仏テキストのアライメントについて話を進める。

3.3 語リスト構築過程

語リスト構築過程は、4 つの過程から構成される。

- 文リスト（LPH）構築過程
- トークンリスト（LMOT）構築過程
- トークンの選別による 4 種の語リスト作成過程：不変化語（transfuges）リスト（LTRNS）、同一語源語（cognates）リスト（LCOG）、カタカナ語リスト（LKTKN）、内容語リスト（LEX）
- トークンの原形化（Lemmatization）と内容語インデックス（ILX）作成過程

3.3.1 文リスト構築過程

すでに文献[11] で述べられているように、英または仏語テキストの文分割は、それ自体が一つの課題となっている。これらのテキストの文分割は、代表的な文分割文字であるピリオドの多義性から、決して単純な作業とは言えない。本システムにおいても、略語（U.S.A.）、記号列（abc@cdf.fr）、数（1.5）等の例外を的確に処理できるいくつかの詳細な規則が定義されている。

日本語の句点は、そのような多義的な性格が薄く、（日本語本来の句点を用いた文章では）文分割は容易であると言える。

3.3.2 トークン抽出過程

仏語テキスト側のリストは、事前に定義された語分割文字に挟まれた文字列を抽出することにより、構築される。

語分割文字を持たない日本語テキスト側のリスト構築には、語分割のための特別な作業が必要であり、通常この作業は形態素解析によって行われる。しかしながら、字種切り法を用いて、完全な形態素解析を行うことなく大部分の内容語を抽出することは可能である。字種切り法のみを用いて、完璧に文を語に分割することはもちろ

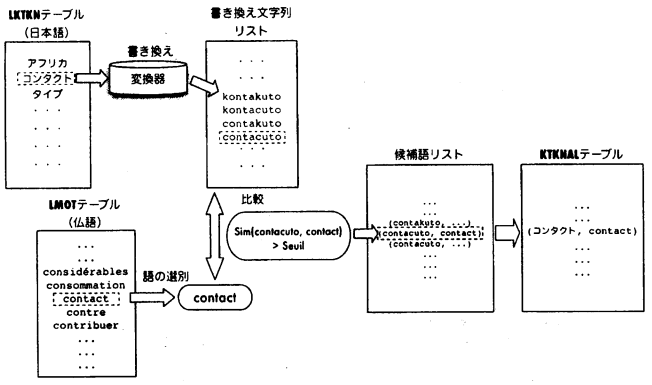

図 2 カタカナ語対応付けの流れ

ん不可能であるが、漢字とカタカナの文字列を抽出することで多くの内容語は認識される。

しかしこの方法で得られたリストには、機能語は含まれないことになる。そこで、システムは事前に定義された機能語リストを用いて、仏語テキストのトークンリストから機能語を削除する。

3.3.3 語の選別

次に、仏語および日本語テキストから抽出されたトークンリストに含まれる語の選別が行われ、4つのリスト、不変化語リスト、同一語源語リスト、カタカナ語リスト、内容語リストが新たに構築される。

選別が終わると、内容語リスト以外の3つのリストに含まれる語の対応付けが行われ、対応不変化語テーブル (TRAL)、対応同一語源語テーブル (COGAL)、対応カタカナ語テーブル (KTKNAL) が構築される。これらの対応はその形態から計算が可能であり、しかも、こうして得られた対応付けは、分布類似度から計算されるものよりも信頼性が高い。

同一語源語 (cognates) とは、同一または形態の類似した文字列で、歴史的に関係の深い言語間の語彙によく見られるものであり、例えば英語と仏語の「generation/génération」や「error/erreur」といったものを指す。同一語源語の概念の導入により、語彙情報を全く用いない統計的なアライメント手法を、簡単にそして経済的に改良することが可能となるが、その効果は同一語族に属する言語間にほぼ限られる。しかし、文献[2]では、表記にローマ字を用いることが可能である日本語においても、同一語源語が有効であることが報告されている。

『あられ』においては、2つの入力テキスト双方に現れる全く同一のローマ字列のみを同一語源語として扱う。システムは、まず、日本語テキストからローマ字語を抽出し同一語源語リストを作成する。次にこのリストを用いて、仏語テキストから同一文字列を抽出し、仏語側のリストを作成する。

不変化語 (transfuges) は、数字や記号といった翻訳に際して形態が変化しない語である。不変化語は、アライメント研究の分野でも初めは同一語源語に含まれていたが、文献[6]において新たなカテゴリーとして定義された。不変化語リストは、単に、日本語、仏語テキストからそれぞれ記号、数字列を抽出し作成される。

3つ目のリストは、カタカナ語を集めたものである。カタカナ語対応付けの流れを図2に示した。日本語テキストから抽出されたカタカナ語は、まずカタカナ書き換え用に開発されたトランスデューサを用いてシステムにより1つまたは複数のローマ字列に書き換えられる。次に、仏語トークンリストの選別過程において、処理中の仏単語とあるカタカナ語の書き換えローマ字列の1つの類似度が既定の閾値を満たした場合、その仏単語は候補語リストに保存される。仏語トークンリストの選別が終了すると、それぞれのカタカナ語について、最も対応する可能性の高い仏単語が、抽出された候補語の中から選ばれ、対応ペアが形成され、対応カタカナ語テーブルに加えられる。対応仏単語の見つからなかったカタカナ語は、分布類似度の計算によって対応付けされる可能性を残しておくために、内容語リストに移される。

カタカナ語の書き換えローマ字列と仏語単語間の類似度の計算は、文献[5]において英仏語間の同一語源語認識に用いられた最大並行部分文字列 (sous-chaine maximale parallèle) の手法に近いものである。本システムで用いられた式は、カタカナ語の書き換え文字列の持つ特殊な条件に対応するものであり、共通子音数を考慮したことが最大の特徴である。これは、共通の子音をより多く含む文字列のペアの類似度が、母音が偶然的に一致した文字列ペアのそれよりも高くなるようにするために加えられた修正である。

3.3.4 トークンの原形化 (Lemmatization)

3.3.4.1 仏語トークンの原形化

本システムは、文献[4]の提案手法を採用している。この手法は、複数のトークンに共通な前接あるいは後接の部分文字列を検出し、意味を含む語幹を決定するという

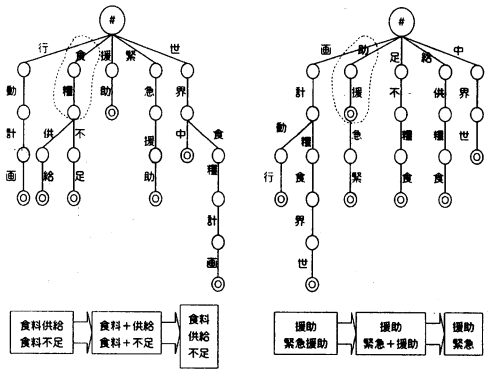

図 3 前接 (図左)・後接 (図右) 部分文字列検出用トライ

ものである。この作業は、トライと呼ばれるデータ構造を利用することで効率的に実装される。

3.3.4.2 日本語トークンの原形化

字種切り法により得られたトークンリストには、さらなる分割を必要とする、複数の語で構成された同一種文字列が含まれている。このような文字列における構成語間の境界の検出は、仏語の原形化処理同様、前接あるいは後接の部分文字列検出としてとらえることができる。そこで、字種切り法では行えなかった単語分割に、トライに基づく原形化の手法を応用した。

日本語処理における特徴は、語幹以外の部分が接辞ではなく、それ自身が他の1つまたは複数の独立した語であるという点にある。例えば、「ab」という文字列からは、「b」という接辞を除いた「a」という1つの語幹が得られるのではなく、「a」「b」という2つの単語が得られることになる。図3は、7つの文字列を入力として作られた、前接、後接部分文字列検出用のトライである。この図は、どのようにトライを用いて日本語の同一文字種でまとめられたトークンが分割されるかを示している。

以上の処理が終了すると、文レベルの対応付けに必要な、語レベルの対応付けを行うためのデータが揃い、アライメント過程へと移行する。

3.4 アライメント過程

2つの入力テキストは、各行が仏語テキストの各文に対応し、各列が日本語テキストの各文に対応した2次元配列で表現される。

まず、前処理であるプレアライメントが行われ、その後、アライメントのメインプロセスへと移行する。メインプロセスは、3つの作業を反復する構成で、これらの過程においては、「対応候補文ペア (Candidats des PaiRes de phrases à aligner)」テーブル、「対応語 (Mots ALignés)」テーブル、「アライメント結果 (Résultat d'ALignement)」テーブルといったデータ構造がそれぞれ構築される。

3.4.1 プレアライメント

プレアライメントは、処理範囲を制限する信頼性の高いアンカーを検出することを目的としている。本システムのプレアライメントは、文献[5]の提案手法を応用したもので、TRAL、COGAL、KTKNAL テーブルを用いて行われる。

3.4.2 「対応候補文ペア (CPR)」テーブル

CPR テーブルは、対応候補文のペアを示す行列である。アライメントの対角線性の仮定に基づいた本手法において、候補ペアに対応する要素は、行列の対角線を軸とし楕円形を形成する。

3.4.3 「対応語 (MAL)」テーブル

MAL テーブルは、それぞれが対訳であると考えられる語ペアの集合である。

語の対応付けは、各語の分布類似度に基づいて行われる。同一対応候補文ペアに含まれるすべての語をそれぞれ比較し、それらの分布情報から類似度を算出する。すでにこれまで2つの単語間の分布類似度を計算するための多くの定式が提案されてきた。本システムでは、それらの中から文献[12]によって提案された Dice 係数の改良式を応用した。この改良式は、従来の手法においては、類似度の算出とは別の過程で扱われていた出現頻度自体を重みとして用いることにより、単語対の同時出現頻度の割合と同時に出現頻度自体を考慮することを可能とした。本システムでは、単語対の同時出現文数の情報をも付加し、この改良式にさらなる修正を加えた。本手法の改良により、前述の手法では類似度が同一となっていた単語対間の比較精度が改善された。

3.4.4 「アライメント結果 (RAL)」テーブルとポストアライメントモジュール

RAL テーブルは、相互に対訳であると考えられる文のペアの集合である。

文の対応付けを行うために、TRAL、COGAL、KTKNAL に加え、前述の方法で構築された MAL テーブルを用いて、CPR テーブルに含まれるそれぞれの候補文ペアがこれら

	日		仏／英		翻訳モデル										
	字数	文数	語数	文数	0-1	1-0	1-1	1-2	1-3+	2-1	2-2	2-3+	3-1	3-2	3+ -3+
Bio	3615	75	1418	69	0	0	55	7	1	3	0	0	0	0	0
FIV	2597	52	1176	54	0	0	43	3	0	2	0	0	0	0	1
G8	3077	47	1398	53	0	0	38	1	0	7	0	0	0	0	0
EU	14308	238	3881	252	0	4	208	5	1	17	0	0	0	0	0
Unicode	14155	268	4224	274	1	0	195	22	1	19	2	0	1	1	1
Balth	11491	423	4835	321	1	2	185	68	20	9	13	1	0	0	0
Zadig	69475	2198	26271	1900	7	6	1190	300	64	103	20	5	18	4	4

表 1 実験使用テキスト

		Bio	FIV	G8	Unicode	EUJP	Balth	Zadig
プレアライメント	再現率	0.57	0.53	0.42	0.62	0.81	0.23	0.14
	適合率	0.98	0.93	1	0.96	0.98	0.99	0.91
メインプロセス	再現率	0.81	0.66	0.95	0.87	0.90	0.49	0.66
	適合率	1	1	1	0.98	0.99	0.97	0.95
ポストアライメント	適合率	0.98	0.92	0.98	0.92	0.96	0.86	0.86

表 2 アライメント結果

のテーブルに含まれる対応語をいくつ有しているかを計算する。ある対応候補文ペアが閾値以上の対応語ペアを含む場合、それらの文は対訳文であると判断される。こうして検出された対訳文をアンカーとし、新たに CPR テーブルの構築を行い、以上のアライメント作業を繰り返す。

2 度目のターンが終了すると、一定のカバー率を満たす信頼度の高い部分対応付け結果が得られる。この部分対応付け結果から、ポストアライメントモジュールによって全対応付け結果を得ることが可能である。ポストアライメントモジュールは、メインプロセスで対応付けされなかった文を抽出し、それらの文から得られるすべての組み合わせの対応可能性をそれぞれの文の長さに基づいて計算する。この対応可能性の値をもとに動的計画法を用いて、すべての文がもう一方のテキストの少なくとも 1 文と対応付けされるようアライメントを行う。

4. 評　　価

本システムの精度を測るため、5 つの和仏コーパス、2 つの和英コーパスに対して対応付け実験を行った（PowerMac G5 使用）。使用テキストは、雑誌「*Label France*」の記事 2 つ☆1（以下「Bio」「FIV」と呼ぶ）、G8 文書（以下「G8」）、ユニコードマニュアル「*How to Unicode*」☆2（以下「Unicode」）、EU 文書（以下「EU」）、文学作品 2 つ（ヴォルテール「*Zadig*」、アナトール・フランス「*Balthasar*」、以下それぞれ「Zadig」「Balth」）である。使用テキストの性質は表 1 にまとめた。この表から文学作品は他のテキストに比べ翻訳モデルの種類が豊富であることがわかる。複合モデル（1-3 といった複数の文で構成された翻訳モデル）は、語彙情報を全く用いない統計手法を用いた際に結果を大きく誤らせる要因であり、自動アライメントの分野において大きな課題の一つとなっている。

今回の実験では、それぞれのテキストについて、プレアライメント、メインプロセス（部分対応）、ポストアライメント（全対応）から得られた 3 つの結果の分析を行った。対応付けの精度は、表 2 に示されたとおりである☆3。

表 2 に示された結果から、本システムが複合モデルに対しても柔軟に対処していることがわかる。この頑健性は、信頼性の極めて高い部分対応付けを導入したことによるものである。

部分対応付けの段階で信頼性を追求した結果、いくつかのテキストではメインプロセスの結果の再現率が低くなっている。しかしながら、本システムにおいては、文字列長に基づくポストアライメントがこの欠点を十分補っていると考える。

本手法の最大の欠点は、行列を使用した場合の必要メモリの膨大さにある。対応文を STL を用いたリスト表現にした場合、使用メモリは 1 割以下に減少するが、計算時間は 20 倍に増大する。本システムではこの問題に対し、アライメントに用いられるすべての行列がスパースマトリックスであることに注目し、その特性を生かしたデータ構造を実装し用いることで対応した。

上記の実験に加え、形態素解析器☆4を用いた実験も行った。この実験では、形態素解析器を用いて行った対応付けと本手法から得られた結果の間にほとんど差異が認められなかった。しかも、カタカナ語対応付けの結果に関しては、本手法が形態素解析器を用いた場合よりも有効であることを示す興味深い結果が得られた。カタカ

☆1 フランス外務省発行の雑誌

☆2 インターネットサイト：（仏語版） http://www.freenix.fr、（日本語版） http://www.linux.or.jp

☆3 再現率は、入力テキスト全体の文のうち、もう一方のテキストの少なくとも 1 つの文と対応付けされた文の割合を示す。適合率は、実際に対応付けが行われた文のうち、もう一方のテキストの少なくとも 1 つの文と的確に対応付けされた文の割合を示す。

☆4 形態素解析器には、ChaSen を使用した。

	Bio		FIV		G8		Unicode		EU		Balth		Zadig	
形態素解析器の使用	無	有	無	有	無	有	無	有	無	有	無	有	無	有
抽出語数	50	57	43	45	21	23	163	162	62	63	34	37	152	166
対応付け結果	23	24	19	14	10	8	50	44	29	30	17	17	68	62
エラー	3	4	1	0	1	1	2	2	1	1	0	1	2	2
適合対応付け数	20	20	18	14	9	7	48	42	28	29	17	16	66	60

表 3 カタカナ語の対応付け結果

ナ語対応付けの実験結果を表 3 に示す。形態素解析器を用いた場合が本手法の精度を下回った要因は、主に、形態素解析器がカタカナ語を過分割したことによる。カタカナ語は、多くの場合、固有名詞または新語であるため、辞書に登録されていないことが多い。そのため、辞書に依存した形態素解析器の手法よりも同一テキスト内語彙の比較に基づく本手法がより論理的な分割を実現できたためだと考えられる。

5. 結論と今後の課題

本システム『あられ』によって得られたアライメント結果から、辞書や形態素解析器等の外部手段に依存しない日本語テキストのアライメントシステムの開発が可能であることが立証されたといえる。これらの結果は、まず第一にカタカナ語の対応付けを導入したことによるものであると言える。カタカナ語は、特に翻訳文書には数多く見られるが、固有名詞、新語であることが多く、その場合、辞書に登録されていないため、書き換えによって原語を検出するという戦略は、きわめて効果的であった。このカタカナ書き換えによる対応付けの手法は、日仏語以外の多言語間における利用はもちろん、Web からの対訳語自動獲得等の他分野への応用も考えられる、まだまだ多くの可能性を備えた新たなアライメント手法であると言える。

本稿で紹介した実験のほかに、特許文書に対しても対応付けを行った。特許文書には、不変化語が多く含まれるため、大変満足のゆく結果が得られた。しかしながら、これらの文書は、一文が長く、文レベルのアライメントは段落の対応付けに似たものとなってしまう。翻訳メモリの構築等において、文よりも小さな単位におけるアライメントがより有効であることは、すでに文献[9] でも指摘されたとおりである。

今後は、対象単位を文から節に変更し、和仏パラレルコーパスの対応付けを行っていく予定である。これらの研究は、翻訳メモリの構築だけでなく、対照言語学等の分野においても有効なデータを提供できるものであると考える。

参 考 文 献

1) PeterF. BROWN, JenniferC. LAI, and RobertL. MERCER. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169 - 176, 1991.

2) K.CHURCH, I.DAGAN, W.GALE, P.FUNG, and B.SATISH J.HELFMAN. Aligning parallel texts : do methodes developed for english-french generalize to asian languages ? In *Proceedings of the Pascific Asia Conference on Formal and Computational Linguistics*, 1993.

3) WilliamA. GALE and KennethW. CHURCH. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(3):75–102, 1993.

4) Martin KAY and Martin RÖSCHEISEN. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, March 1993.

5) Olivier KRAIF. Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation. *TAL*, 42(3), 2001.

6) Jean-Marc LANGÉ and Éric GAUSSIER. Alignement de corpus multilingues au niveau des phrases. *TAL*, 36(1–2), 1995.

7) Philippe LANGLAIS. Alignement de corpus bilingues : intérêt, algorithmes et évaluation. In *Bulletin de Linguistique Appliquée et Générale, numéro Hors Série*, pages 245–254. Université de Franche-Comté, 1997.

8) H MURAO. Studies on bilingual text alignment. Bachelor thesis, Kyoto University, 1991. in Japanese.

9) M. SIMARD. *Mémoire de traduction sous-phrastique*. Thèse de doctorat en informatique, Université de Montréal, 2003.

10) M.SIMARD, G.FOSTER, and P.ISABALLE. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Thoretical and Methodological Issues in Machine Translation (TMI)*, pages 67 - 81, juin 1992.

11) M.SIMARD and Pierre PLAMONDON. Bilingual sentence alignment : Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80, 1998.

12) 北村美穂子 and 松本裕治. 対訳コーパスを利用した対訳表現の自動抽出. 情報処理学会論文誌, 38(4):727- 736, 1997.

13) 春野雅彦 and 山崎毅文. 辞書と統計を用いた対訳アライメント. 情報処理学会研究報告, 1996-NL-112:23-30, 1996.