タイトル	対訳コーパスを用いた多言語間格フレーム対の自動獲得のための名詞句対の自動抽出手法
著者	三上，優；越前谷，博；桃内，佳雄；MIKAMI，Yuu；ECHIZEN－YA，Hiroshi；MOMOUCHI，Yoshio
引用	北海学園大学工学部研究報告(38)：131-139
発行日	2011-01-14

対訳コーパスを用いた多言語間格フレーム対の自動獲得のための名詞句対の自動抽出手法

三上　優[*]・越前谷　博[**]・桃内佳雄[**]

Automatic Extraction Method of Noun Phrase Pairs for Multi–lingual Case Frame using Parallel Corpus

Yuu Mikami[**], Hiroshi Echizen-ya[**] and Yoshio Momouchi[**]

要　旨

格フレームは機械翻訳における重要な知識の一つである．本研究では，言語間の格フレーム対を三言語間対訳コーパスより自動獲得する．その際には，言語資源が豊富な英語を中間言語に用いることで行う．例えば，日本語 - 英語 - スペイン語の三言語間対訳コーパスにおいて，英語 - 日本語対訳コーパスより，英語 - 日本語の格フレーム対を，英語 - スペイン語対訳コーパスより，英語 - スペイン語の格フレーム対を，獲得することにより，最終的に日本語 - スペイン語の格フレーム対を獲得する．本稿では，この第一段階として，自然言語文に最も多く出現する名詞句に着目し三言語間対訳コーパスより名詞句対を自動抽出するための手法およびその有効性について述べる．

1．はじめに

格フレームは機械翻訳において，意味的な情報が付与された重要な言語知識の一つである．格フレームの自動獲得の研究としては対訳コーパスを利用する手法[1)]やWeb上の大規模なテキストを利用する手法[2)]が提案されている．しかし，これらの手法は，構文解析ツールに強く依存するため，多言語への適用が困難であることが問題となる．本研究では，多言語間機械翻訳システムの構築に向けた様々な言語の格フレーム対の自動獲得を目的としている．本稿ではその第一段階として対訳コーパスより名詞句対を自動的に抽出するための手法を提案する．本手法は，構文解析ツールが豊富に存在し，かつ，最も広く利用されている英語を中間言語として利用することで，構文解析ツールが十分には得られない言語も含めた様々な言語間の名詞句対

[*] 北海学園大学大学院工学研究科電子情報工学専攻
[*] Division of Electronics and Information Engineering, Graduate School of Engineering, Hokkai–Gakuen University
[**] 北海学園大学工学部電子情報工学科
[**] Department of Electronics and Information Engineering, Faculty of Engineering, Hokkai–Gakuen University

を自動抽出する．**図 1** に日本語－英語－スペイン語の三言語間対訳コーパスにおける提案手法の概要を示す．日本語とスペイン語を構文解析ツールが不十分な言語と位置付け，日本語－英語－スペイン語間対訳コーパスを用いて，日本語－スペイン語間の名詞句対を自動抽出する．同時に，英語－日本語間と英語－スペイン語間の名詞句対を英語－日本語対訳コーパス，英語－スペイン語対訳コーパスよりそれぞれ抽出する．そして，最後に，英語の名詞句を介して，日本語－スペイン語の名詞句対を得る．また，本手法では，名詞句対を効率よく抽出するために名詞句ルールを自動獲得し，それを適用することで名詞句を自動抽出する．性能評価実験の結果，名詞句対の自動抽出におけるF値は0.271から0.311に向上した．

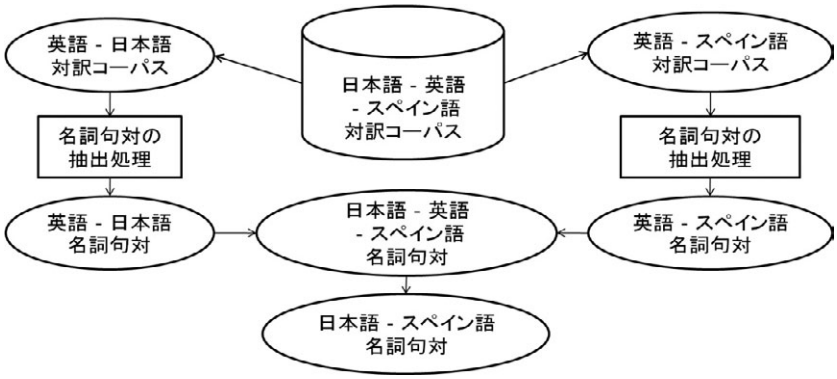

図 1　提案手法の概要

2．三言語間対訳コーパス

本研究では，英語を含む三言語間対訳コーパスを旅行用会話のテキストから得ることで作成した．**図 2** に日本語 - 英語 - スペイン語間対訳コーパスにおける対訳文の例を示す．

図 2 では，対訳関係にある英語，日本語，スペイン語の文に基づき，英語 - 日本語対訳文と英語 - スペイン語対訳文を得る．このように英語を中間言語とした対訳コーパスを用いること

英語：I'd like a single room with bath.
日本語：バス付きのシングルを 1 部屋お願いします。
スペイン語：¿Cuanto es la habitacion sencilla con bano?

⇩

英語－日本語対訳文
（I'd like a single room with bath.；バス/付き/の/シングル/を/ 1 /部屋/お/願い/し/ます/。）
英語－スペイン語対訳文
（I'd like a single room with bath. ; ¿Cuanto es la habitacion sencilla con bano?）

図 2　対訳コーパスの構成

により，日本語 - スペイン語間の名詞句対を抽出する．

3．名詞句対の自動抽出

3.1　対訳文間の共通部分に基づく名詞句対の抽出

本手法では，始めに二言語間の対訳コーパスを用いて，対訳文間の共通部分[3]に基づき名詞句対を自動抽出する．その際，構文解析ツールの豊富な英語を中間言語に用いることで，英文については構文解析ツールより名詞句を決定する．そして，その英語の名詞句に対応する部分を英語以外の言語文から抽出する．その処理過程を以下に示す．

（1）英語の名詞句が共通部分である２つの対訳文において，英語以外の言語文より名詞句の共通部分を決定する．

（2）英語以外の言語文より決定された共通部分が複数存在する場合には，英語の名詞句との類似度に基づき一意に決定する．

（3）決定された英語以外の言語文中の共通部分を英語の名詞句に対応する部分として，名詞句対を決定する．

図 3 に共通部分に基づく名詞句の抽出例を示す．**図 3** では対訳文１に対して構文解析ツール用いることで，“a tour”が名詞句となる．次いで，“a tour”に対応する名詞句を日本文から決定するために，“a tour”が出現する対訳文２を対訳コーパスから選択する．そして，日本文間中の共通部分を決定する[4]．対訳文１，対訳文２の日本文間では，“ツアー”と“か／。”が共通部分となる．そこで，“a tour”と“ツアー”，“a tour“と”か／。”の部分間類似度[5][6]をそれぞれ計算し，最も部分間類似度が高く，かつ，閾値以上のものを選択する．この場合，“a tour”と“ツアー”の部分間類似度が高くなることで，（a tour；ツアー）が名詞句対として抽

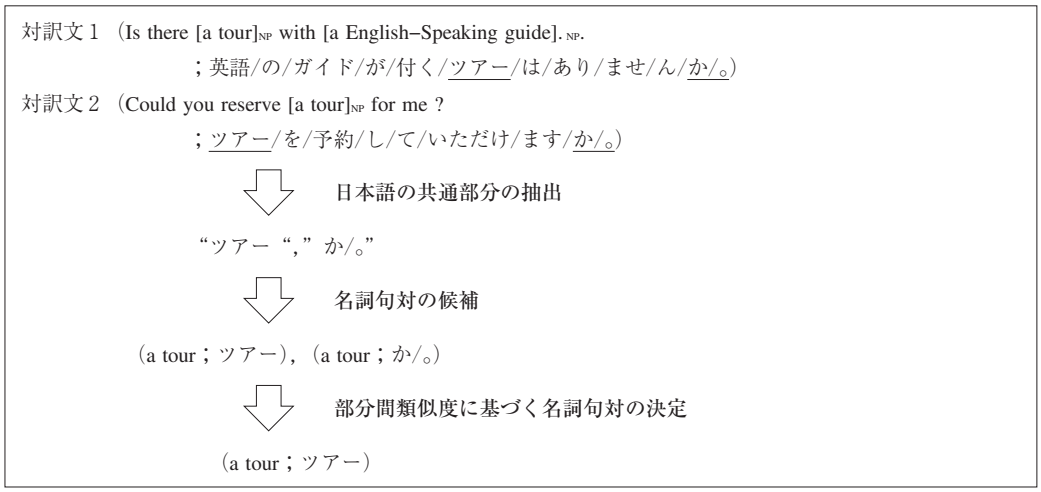

図 3　共通部分に基づく名詞句対の抽出例

出される．

部分間類似度については，3.3節で詳細を説明する．

3.2　名詞句ルールの自動獲得

本手法では，低出現頻度の名詞句を抽出するために有効となる，名詞句ルールを自動獲得する．その獲得処理の詳細を以下に記す．

（1）3.1節より抽出された名詞句対を含み，かつ，抽出された名詞句対の英語部分が名詞句の先頭と末尾に存在する対訳文を選択する．

（2）選択された対訳文の日本文において，bigram確率と部分間類似度に基づき，英文の名詞句に対応する英語以外の言語の名詞句を決定する．

（3）抽出された名詞句において共通部分以外の部分間を省略可能な部分[5)6)]として変数に置き換えることで名詞句ルールを獲得する．

（4）獲得された名詞句ルールにおいて差異部分をさらに変数に置き換えることで，より一般的な名詞句ルールを獲得する．

図 4 に英日の対訳文を用いた場合の名詞句ルール獲得の具体例を示す．

始めに，3.1節より抽出した名詞句対を含む対訳文を対訳コーパスより選択する．**図 4** で

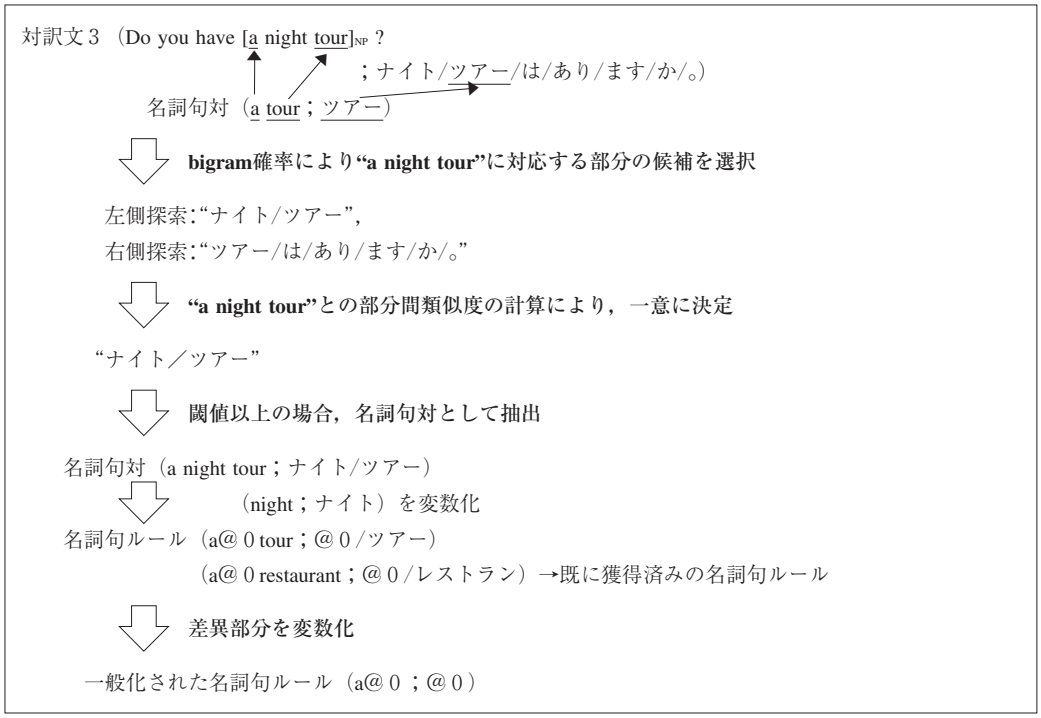

図 4　名詞句ルールの獲得の具体例

は，**図 3** より抽出された名詞句対（a tour；ツアー）を用いることで，対訳文 3 が選択される．すなわち，名詞句対の英語部分の“a”と“tour”が対訳文 3 の英文の名詞句“a night tour”の先頭と末尾の単語と一致し，かつ，名詞句対の日本語部分の“ツアー”が対訳文 3 の日本文中に存在するため，対訳文 3 が対訳コーパスより選択される．次いで，対訳文 3 の英文の名詞句“a night tour”に対応する部分を日本文中より決定する．日本文においては“ツアー”が共通部分であるため，“ツアー”を中心として探索を行う．左側探索においては，bi-gram確率を用いることで，“ナイト／ツアー”，右側探索においては，“ツアー/は/あり/ます/か/。”が選択され，“a night tour”の対応部分の候補となる．そして，“a night tour”と“ナイト/ツアー”，“ツアー/は/あり/ます/か/。”との間の部分間類似度を計算し，類似度が高く，閾値以上の組み合わせを選択する．その結果，“a night tour”と“ナイト／ツアー”の部分間類似度が高く，閾値0.600以上となるため，名詞句対として（a night tour；ナイト/ツアー）が抽出される．閾値として用いられる0.600は予備実験に基づき決定された．更に，この（a night tour；ナイト/ツアー）において名詞句対（a tour；ツアー）との共通部分以外の部分“night”と “ナイト”を省略可能な部分として，変数“@０”に置き換えることで名詞句ルール（a@０tour；@０/ツアー）を獲得する．

そして，この名詞句ルールと，他に獲得された名詞句ルール（a@０restaurant；@０／レストラン）との間で差異部分を変数化することで，より一般化された名詞句ルール（a@０；@０）が獲得される．

本手法では，更に，3.1節より抽出された名詞句対のペアに対しても差異部分を変数化することにより，名詞句ルールを獲得する．

3.3　名詞句ルールの適用

3.2節より獲得された名詞句ルールを用いることで，対訳コーパス中の低頻度の名詞句対を含め，より多くの名詞句対を効率よく抽出することが可能となる．名詞句ルールの適用の処理過程を以下に示す．

（1）名詞句ルールの変数以外の部分を含む対訳文を対訳コーパスより選択する
（2）名詞句ルールの変数部分に対応する部分の候補を日本文より抽出する．
（3）抽出された候補に対して，英語の名詞句との間の部分間類似度を計算し，閾値以上の場合，名詞句対とする．

図 5 に英日の対訳文における名詞句ルールの適用の具体例を示す．

図 5 では，構文解析ツールより“a barber’s shop”が名詞句として決定される．この名詞句の先頭の単語“a”と名詞句ルール（a@０；@０）中の変数以外の部分“a”が一致するため，対訳文 4 が選択される．次いで，対訳文 4 より“a barber’s shop”に対応する部分を日本

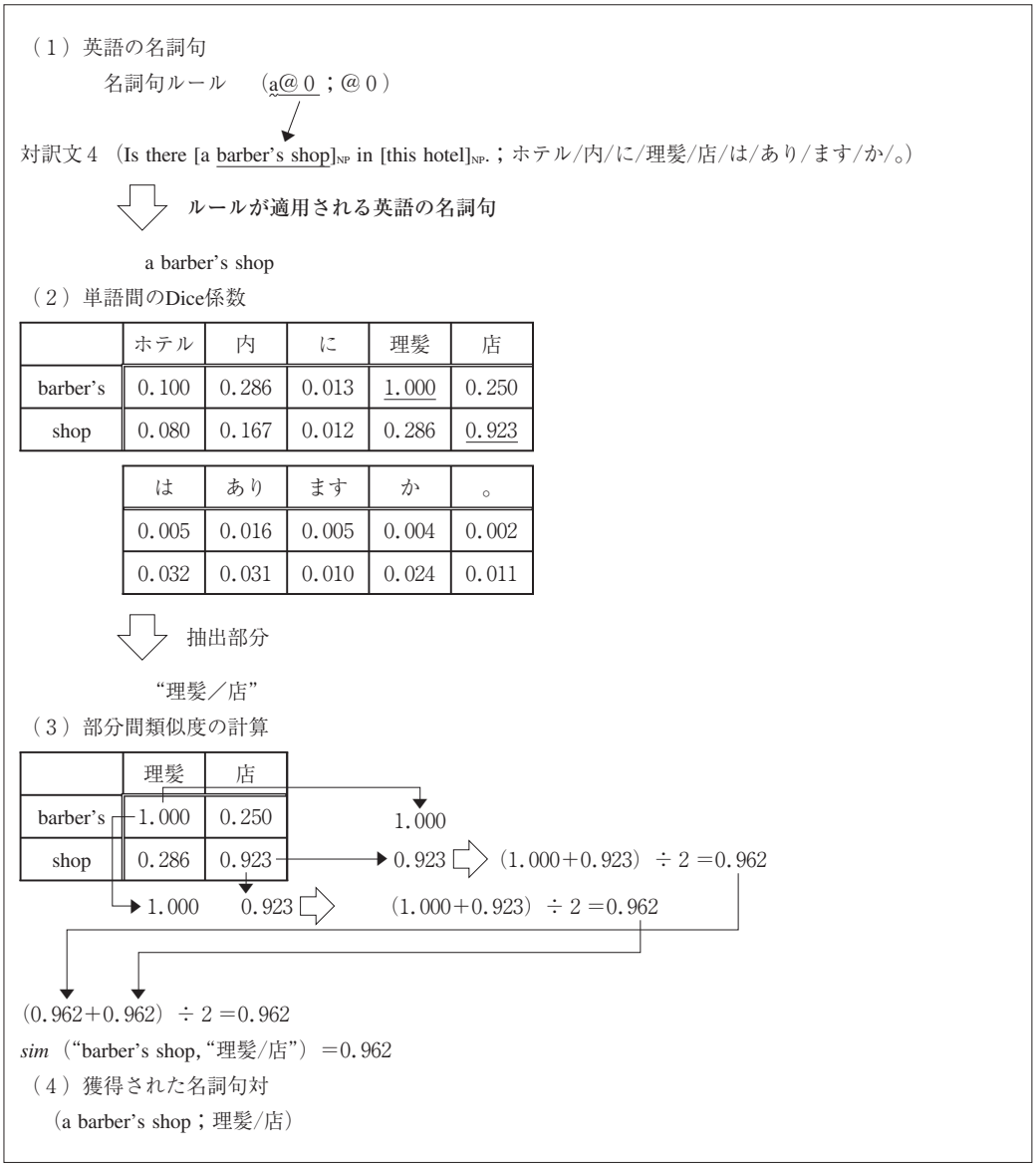

図5　名詞句ルールに基づく名詞句対の抽出例

文より決定することで名詞句対を抽出する．始めに，名詞句内の“barber's”と“shop”のそれぞれの単語と日本文中の全単語とのDice係数[7]を求める．ここでは適用される名詞句ルールの日本語部分は変数のみであるため，日本文中の全単語が対象となる．英文中においては名詞句“a barber's shop”において変数に対応する“barber's shop”が対象となる．**図5**の（2）より，“barber's”とのDice係数が最も高い値を示した単語は“理髪”，“shop”においては“店”となった．したがって，“理髪”から“店”までの部分“理髪／店”が対応する部分の候補と

なる．次いで“barber’s shop”と“理髪／店”との間の部分間類似度を求め，その値が閾値0.600以上の場合には，名詞句対と位置付ける．部分間類似度の計算例を**図5**の（3）に示す．表中の数値はDice係数の値を表している．“barber’s”との間で最もDice係数の値が大きい日本語の単語は“理髪”であるため，1.000が選択される．“shop”において“店”とのDice係数が最も高いため，その値0.923が選択される．選択されたDice係数の平均は0.962となる．同様に，“理髪”と“店”において，Dice係数の値が最も大きい英語の単語を選択し，平均値を求める．**図5**では0.962が得られる，これら2つの平均値に対して，更に，平均値を求めると，0.962となり，この値が部分間類似度となる．そして，この値は閾値0.600を上回るため名詞句対として（a barber’s shop；理髪／店）が得られる．このような名詞句対の抽出処理は他の言語の対訳コーパスにも適用可能である．例えば，英語－スペイン語の対訳コーパスからは名詞句対として（a barber’s shop；peluqueria）が抽出される．

4．性能評価実験

4.1　実験方法

本実験では実験データとして，旅行用会話文の文献[8]に掲載されている，日本語－英語－スペイン語の対訳コーパス1,042組を使用した．実験システムには名詞句ルールを適用したシステムと適用しないシステムの2つのシステムを用いて，それぞれ名詞句対の抽出を行ない，以下の式（1）～（3）による再現率，適合率，そしてF値を求めた．なお、抽出された名詞句の正誤判断は人手で行なった．

$$\text{再現率} = \frac{\text{正しく抽出された名詞句対の数}}{\text{対訳コーパス中の名詞句対の数}} \quad \text{——（1）}$$

$$\text{適合率} = \frac{\text{正しく抽出された名詞句対の数}}{\text{システムが出力した名詞句対の数}} \quad \text{——（2）}$$

$$\text{F値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \quad \text{——（3）}$$

4.2　実験結果

表1に名詞句対の抽出における再現率，適合率，そしてF値を示す．

英文に対する構文解析ツールには，Stanford Parser[9]を用いた．**表1**より名詞句ルールを使用したことにより，日本語－英語の再現率と適合率はともに変化しなかった．スペイン語－英語の再現率は4.1ポイント，適合率は4.8ポイント向上した．日本語－スペイン語の再現率は，3.4ポイント，適合率は，5.0ポイント向上した．その結果F値は0.04向上した．したがって，名詞句ルールの獲得とその適用が有効であることを確認した．

表 1　名詞句対の抽出の再現率と適合率，F値

	再現率		適合率		F値	
名詞句ルール	有り	無し	有り	無し	有り	無し
日本語－英語	33.8%	33.8%	46.8%	46.8%	0.393	0.393
スペイン語－英語	30.0%	25.9%	43.6%	38.8%	0.355	0.311
日本語－スペイン語	25.6%	22.2%	39.7%	34.7%	0.311	0.271

表 2　獲得された名詞句ルールの例

英語－日本語	英語－スペイン語
(a @ 0 restaurant；@ 0 /レストラン)	(the @ 0 city；@ 0 la ciudad)
(a @ 0 tour；@ 0 /ツアー)	(a @ 0 restaurant；un restaurant @ 0)
(the @ 0 telephone；@ 0 /電話)	(a @ 0 car；un coche @ 0)
(the @ 0 ；@ 0)	(a @ 0 ；un @ 0)
(a @ 0 ；@ 0)	

対訳文 5 (I'm looking for [a porter]$_{NP}$. ；Busco un mozo.)

名詞句ルール適用無しの場合

	un	mozo
a	0.36	0.01
porter	0.02	1.00

⇩ “a” と “un” のDice係数が閾値0.60より低く，“a” と “un” の対応は抽出できない

・抽出される名詞句対：(a porter；　mozo)

名詞句ルール適用有りの場合

・適用するルール：(a @ 0 ；un @ 0)

	mozo
porter	1.00

⇩ aとunの対応まで抽出できる

・抽出される名詞句対：(a porter；un mozo)

図 6　英語－スペイン語間の名詞句ルール適用

4.3　考察

表 1 より，名詞句ルールを適用することで，スペイン語－英語，日本語－スペイン語では，より正確に名詞句対の抽出が可能となったことが分かる．**表 2** に獲得された名詞句ルールの例を示す．また，**図 6** に名詞句ルールを適用することで日本語－スペイン語間の名詞句対が正しく抽出された例を示す．**図 6** では，名詞句ルールを適用しない場合に抽出される名詞句対は（a porter；mozo）となる．名詞句ルール（a@０；un@０）を適用した場合は，名詞句対として（a porter；un mozo）が得られ，冠詞の対応を正しく抽出できた．なお，日本語－英語では日本語に冠詞が存在しないため，冠詞の対応をとる必要がなく，名詞句ルールを使用した場合と使用していない場合では抽出される名詞句対に変化はなかった．

5．おわりに

本稿では対訳コーパスを用いた多言語間格フレーム対の自動獲得のための名詞句対の自動抽出手法について述べた．性能評価実験の結果，英語を中間言語とした日本語－スペイン語の名詞句対の抽出に対するF値が，提案手法により，0.271から0.311に向上した．今後は，対訳コーパスを増やすことにより精度向上を図る．また，他の言語による言語間対訳コーパスに適用することで，提案手法の有効性を確認する予定である．

参考文献

1）宇津呂武仁，松本裕治，長尾眞，“二言語対訳コーパスからの動詞の格フレーム獲得”情報処理学会論文誌，vol.34，No.５，pp.913－924，May．1993.

2）河原大輔，黒橋禎夫，“高性能計算機環境を用いたWebからの大規模格フレーム構築”情報処理学会研究報告（2006－NL－171（12）），pp.67－73，2006.

3）荒木健治，“自然言語処理ことはじめ―言葉を覚え会話のできるコンピュータ”，森北出版，東京，2004.

4）Hitoshi Echizen–ya, Kenji Araki, “Automatic Evaluation of Machine Translation based on Recursive Acquision of an Intuitive Common Parts Continuum, Proc. of the Eleventh Machine Translation Summit (MT SUMMIT XI), pp.151–158 (2007)

5）寺島涼，越前谷博，荒木健治，“学習型機械翻訳手法における省略可能性を用いた翻訳ルールの自動獲得とその有効性”情報処理学会研究報告（2008－NL－183（19）），pp.127－134，2008.

6）寺島涼，越前谷博，荒木健治，“対訳コーパスに基づく学習型機械翻訳における省略可能情報を用いた部分対応学習の有効性”　電子情報通信学会論文誌D, Vol.J93－D，No.３，pp.377－388，2010.

7）北村美穂子，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，vol.38，no.４，pp.727－736，1997.

8）JTBパブリッシングひとり歩きの４カ国語　会話ヨーロッパ編３　自由自在　英語・フランス語・イタリア語・スペイン語，2006.

9）D.Klein and C.D.Manning, “Accurate Unlexicalized Parsing”, Proc. of Meeting of the Association for Computational Linguistics (ACL2003), pp.423–430 (2003).