日本語の文末表現に着目した英文作成支援システムの構築と実証

松山宏樹
岡田勇
江原暉将
宮崎瑞之
Dawn L. Miyazaki
宮澤信一郎

1．はじめに

英文作成を支援するシステムや英文を日本文に翻訳するシステムは実用化されてきている [1] とはいえまだ多くの課題を抱えている。広く実用化されているテンプレート型は、英文メール作成支援システムとして実用化されている [2] ものの、ビジネスレターといった定型化されている文書を扱うのには適しているが、そのような分野は限られているため対象が限定的である [3]。この欠点を解消すべくコーパス利用型翻訳システムが近年盛んに研究されてきている [4-9]。これは、キーワード及びキーセンテンスを入力し、コーパスから例文や類似文を抽出し提示する方式である。この方式は大量の適切なコーパスが入手できれば、類似文の抽出が可能となることから、いかに大量のコーパスを作成し、それからどのように類似文を見つけ出すかについて多くの研究が存在する。このシステムはあくまで類似文の抽出がシステムの出力となるため、最終的には人手で翻訳を行わなければならず、翻訳者の能力に依存してしまう。よって、より質の良い類似文を提示することが課題と

なる。このように、コーパスから適切な文を抽出するような翻訳メモリーは、その処理をどのようなアルゴリズムで行うかということが大きなポイントとなる。

本研究では、英文作成支援システムのうち、類似文提示型の翻訳メモリーの改良を行う。翻訳メモリーとは、翻訳したい文（キーセンテンス）を入力して、コーパス中から類似文を検索する手法を用いた翻訳支援ツールである。翻訳メモリーは、キーセンテンスから抽出されたキーワードあるいは入力されたキーワードを含む類似文を提示する方法である。キーワードを入力する方式の場合は辞書などで対訳を得るのは比較的容易である。一方で、キーセンテンスを入力する方式の場合は、文型が類似する文を提示することが可能ではあるが、その精度は十分ではない。よって、キーワードは類似しているが文型が異なるために参考にできないことが多い。例えば、「空港まで誰かに伊藤氏を迎えに行かせないといけません。」という文を、“I need to have someone pick Mr. Ito up at the airport.” のように自然で流暢な英語の文型に翻訳することは、英語のトレーニングを積んでいないと難しい。本研究では、翻訳したい文の文末表現に着目することで、文型を考慮した手法を提案する。

この文型や構文に注目した研究には、王軼諞、卜朝暉ら [10] による研究がある。この研究では、日本語文の構文特徴（文型、助詞、テンスなど）を考慮している。しかし、対象が「ある」「いる」といった存在文のみであり、かつ日本語と中国語における翻訳であるため、英語には対応できない。また、池原、阿部ら [11] による研究では、重文と複文を対象に文型パターン辞書の作成を行っているが、翻訳支援システムとしての機能はない。本研究では、文型に着目した新たな文間距離を定義し、その距離に基づいて類似文を提示するアルゴリズムを開発した。このアルゴリズムは、日本語文の文末に着目することを特徴としている。このアルゴリズムの有効性を評価するために、被験者実験を２度行い、それらの実施データに対して平均値の差の検定を行った。その結果、英文作成支援システムは辞書の使用と比べて Fluency

に関して 1% 有意で改善し、今回比較対照とした酒井らの SCOPE[12] と比べて Adequacy と Fluency について平均値は高いものの 5% の有意水準での差は得られなかった。

本論文の構成は以下の通りである。

２章にて、文型類似文の構築について記述する。３章にて、プロトタイプでの評価実験とそれに対する考察を記述する。４章にて、本システムの特徴と評価、結果について議論する。最後に５章にて、本稿のまとめと今後の課題を述べる。

２．文型類似文辞書の構築

我々は、日本語の文末表現に着目した文型類似文辞書を構築する。この辞書は類似文を抽出するために新たに定義された文間距離を用いている。従来の文間距離は対象となる日本文と使用されている単語が類似していたり、単語数の長さが類似していたりすることで定義されているのが一般的であるが、単語そのものは対訳が容易に手に入るのに比べて、構文や文型を特定化することは比較的困難である。また日本語の言い回しは文末で決まることが多い。たとえば一般的に日本語では疑問文や文のニュアンスは最終文節がどうであるかに大きく依存する。つまり、文末表現を重視した文間距離を定義することができれば、文型が類似する文を抽出することが容易になる。

文末表現に着目するために、文を係り受け解析し、根文節からの深さが深い文節同士の対応コストを小さくするように文間距離を定義する。根文節とは、係り受け解析の結果、最も上位の係り先となる文節であり、文で最も後ろに来る文節である。これは、日本語が典型的な主辞後置型言語であることに由来する。主辞後置型言語の特徴として、tense, aspect, modality を表す形態素が文末に置かれる。従って統語的、意味的特性が文末表現によって決定されやすい。このような方針で定義することで、倒置などの例外文を除けば、大枠である文型の類似性を捕らえることができる。この根文節からの深

さの情報に着目している点が、本システムの特徴である。

2.1　文型類似文辞書構築の流れ

文間の距離を用いて文型類似文辞書を作成する手順の概要を図１に示す。図１でクラスタリングとは、最遠隣法で距離の近い文を集める手法であり、これらの集合（クラスタ）を構築することで文型ごとの代表的な文を抽出し辞書を構築することができる。クラスタはすべての文と文との間の距離から作られる距離行列に基づいて行われるべきであるが、計算コストが大きいので、2段に分けてクラスタリングをしている。まず、我々は文末表現の一致する文を集めて仮の文集合を作り（文末でクラスタリング)、この仮の文集合の中で多くの文を含むものだけに対して文間の距離でクラスタリングを行うことで計算時間を現実的なものに抑える方針を採用する。クラスタリングによって作成されたクラスタが文型の候補となる。クラスタごとに距離行列を用いて最も中心となる文を選び、これを当該クラスタの代表的な文として選択する。これを集めることで文型類似文辞書の索引を作ることができる。それぞれの索引にはクラスタ内の文が表示される。以下に各処理について詳述する。

図１　文型類似分辞書作成過程

2.1.1　コーパス解析

ここでは、入力文の解析方法について述べる。NTCIR - 7 が提供したPATMT コーパスである特許文を対象にした約 180 万文のモデル訓練用日英対訳データ[1] [13] に対し、形態素解析と係り受け解析を行い、形態素単位

[1] 特許文には倒置や体言止めの文書はほとんど現れない。

の文節係り受け形式に変換する。形態素解析には ChaSen[14]、係り受け解析には CaboCha[15] を使用する。次に CaboCha の出力ファイルを形態素単位の係り受け形式ファイルに変換し、形態素単位の係り受け形式ファイルを、文節単位で処理できるように文節単位の文節係り受け形式ファイルに変換する（図２）。尚、PATMT コーパスを使用した理由は、特許文は科学技術英文に似ているということと、今回の目的に合った入手可能な大規模対訳コーパスがこのコーパスのみだったからである。

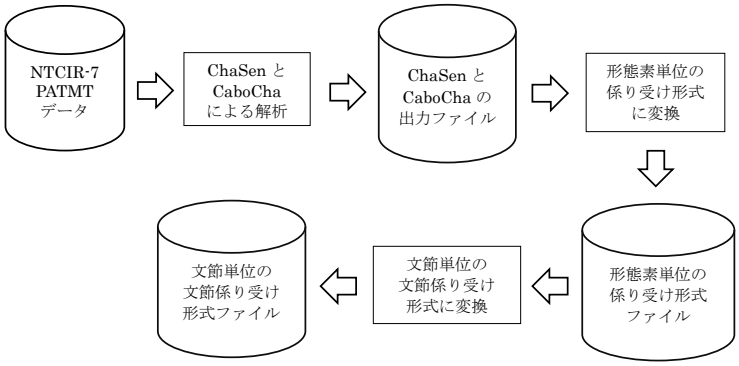

図２　コーパス解析

また後の処理で文節属性を付加する必要があるため、各文節に対して、我々が定義した受け種別と係り種別（表１）を付与し文節属性抽出ファイルを作成する（図３）。

表１　文節属性

受け種別

意味	記号
名詞	N
述語	V
サ変動詞	NV
述語名詞	NV
副詞・連体詞	E
接続詞	NV

係り種別

意味	記号
連用修飾	y
連体修飾	t
連用または連体修飾	ty
係助詞「は」	h
終止	s

図３　文節属性抽出ファイルの作成

前の処理で作成した文節単位の文節係り受け形式ファイルと文節属性抽出ファイルをマージし、係り受けと文節属性のマージファイルを作成する。このファイルから文末表現を抽出し文末表現ファイルを作成する。また根文節（文末文節）までの係り受けの数である係り受け深さデータを付加して係り受け深さのファイルを作成する（図４）。係り受け深さは文節ごとに根文節からの深さを係り属性をたどって根文節にたどり着くまでの経路長として定義する。

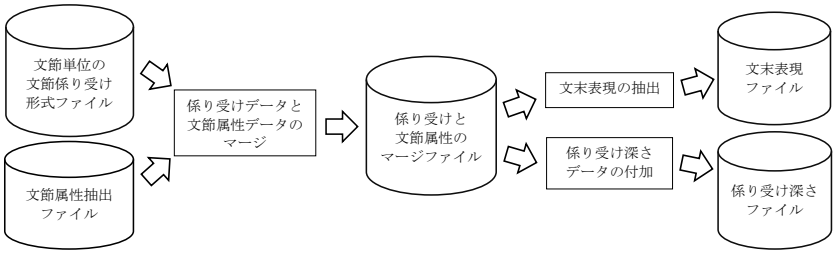

図４　文末表現ファイルと係り受け深さのファイルの作成

2.1.2　文末クラスタリングによるクラスタの作成

文末表現を集めたクラスタを作成する。このため先に作成した文末表現ファイルの文末表現を文節数単位で分割する。文節数は１文節、２文節、３文節単位に分割してファイルする。４文節以上は計算量の関係で無視する。その後､文節数ごとに文末表現を収集して文末表現のクラスタを作成する（図５）。このように文節数を３種類用いた理由は、文末の多くは１文節から３文節で構成され､それらに対処するためである。例えば｢得ることができる。｣という文末表現は３文節からなり､｢説明する｣という文末表現は１文節からなる。

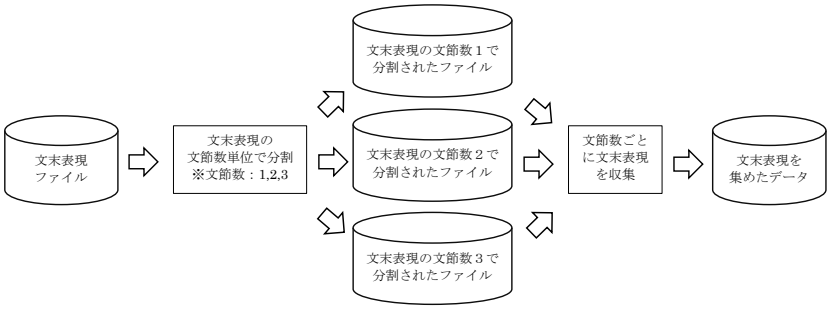

図５　文末クラスタリング

2.1.3　同一文末の文を距離によってクラスタリングしたファイルの作成

同一文末の文を距離計算してクラスタリングする。このため、以下の３つのステップを行う。

（１）距離でクラスタリングしたファイルの作成

係り受け深さファイルと文末表現を集めたデータから文末表現ごとのファイルを作成する。距離計算を現実的なものにするために、このうち深さ３以

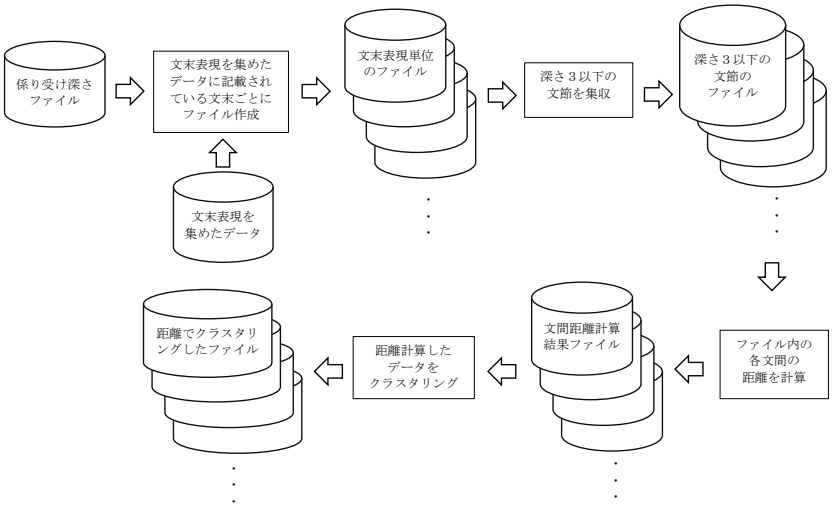

図６　距離でクラスタリングしたファイルの作成

下の文節のみを残し、ファイル内の各文間の距離を計算し文間距離計算結果ファイルを作成する（文間の距離の計算方法については 2.2 に示す）。さらに距離計算したデータをクラスタリングし、距離でクラスタリングしたファイルを作成する（図６）。また、深さの実例を図示する（図７）。

例文 （流体圧シリンダ３１の場合は流体が徐々に排出されることとなる）

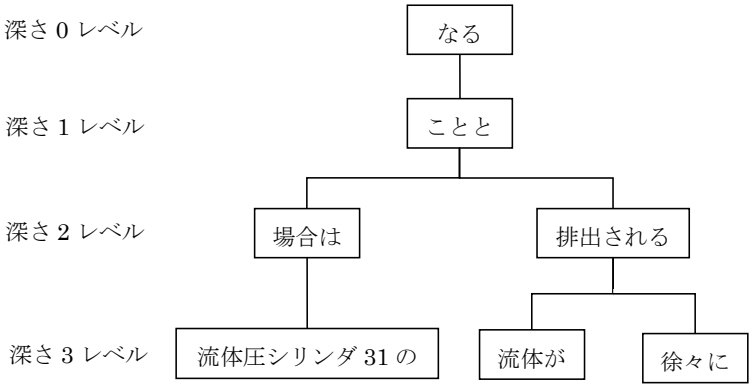

図７　深さの実例

（２）文末表現を含む文の ID を抽出したファイルの作成

作成された係り受け深さファイルと文末表現を集めたデータから、文末表現を集めたデータに記載されている文末を含む文の ID を文末ごとに抽出し、文末表現を含む文の ID を抽出したファイルを作成する（図８）。

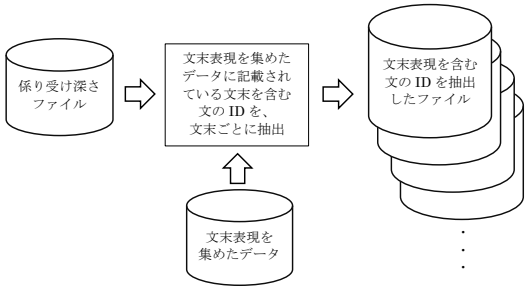

図８　文末表現を含む文の ID を抽出したファイルの作成

（3）同一文末の文を距離でクラスタリングしたファイルの作成

解析対象文のファイル（NTCIR-7 PATMT から解析対象を抽出した文集合）と距離でクラスタリングしたファイル、文末表現を含む文の ID を抽出したファイルの3種類のファイルをマージして、文型類似文辞書（同一文末を距離でクラスタリングしたファイル）を作成する。文末類似文辞書の各クラスタには、2.1 で述べたクラスタを代表する文が索引として付加されている（図9）。

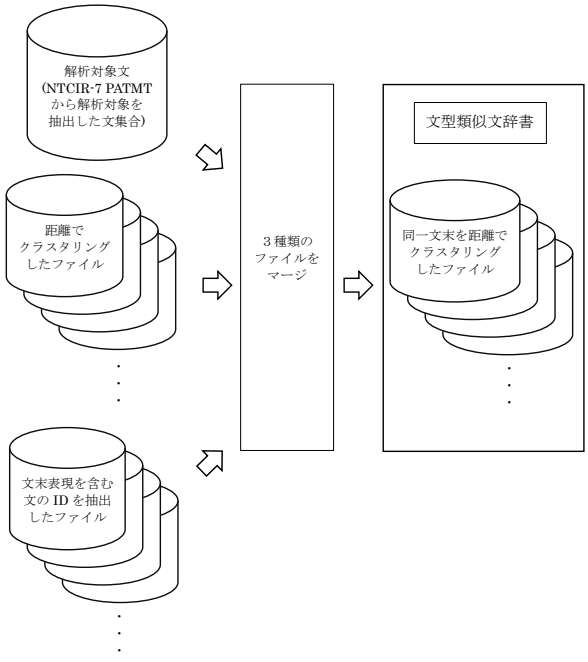

図9　同一文末の文を距離でクラスタリング

2.2　文間距離の計算アルゴリズム

本節で二つの文 a と b の文間距離を計算するアルゴリズムを定義する。ここで、英文作成支援システムの場合は、文 a は翻訳対象和文、文 b はデータベースから取り出した文となる。前処理として文 a、文 b ともに ChaSen

と CaboCha で形態素解析、文節解析、ならびに係り受け解析を行い、文 a、b ともに文節に分解する。その後、表 1 に示す文節の受け種別と係り種別をそれぞれの文節に自動付与し、そして係り受け深さデータを抽出する。これらの前処理を経て以下のアルゴリズムを行う。ここで各種パラメータの値は予備的検討において適切な用例が出力されるように調整したものであり、経験的に決定している。[2]

1）文 a の文節番号を i とし文節数を I とする$(i=1,\cdots,I)$。文 b の文節番号を k とし文節数をKとする$(k=1,\cdots,K)$。文 a に開始ダミーの文節 $i=0$ を考え、文 b にも開始ダミーの文節 $k=0$ を考える。

2）以下の条件を満たす場合、文間距離を上限値（ここでは 1.0）とする。

⑴文 a の文節数が 5 未満で、文 a と文 b の文節数の差が 2 以上。

⑵文 a の文節数が 10 未満で、文 a と文 b の文節数の差が 3 以上。

⑶文 a の文節数が 10 以上で、文 a と文 b の文節数の差が 4 以上。

3）上記に当てはまらない場合、すべての文節の組に対して文節対応コスト $C_{i,k}$ を次のように計算する。ここで $cost_1$ は文節間の語形や係り種別、受け種別の一致度合いによって決まり、文節 i と文節 k の根文節からの深さの小さいほうを$depth$ とし、$cost_2 = 2^{-depth}$ とする。

$$c_{i,k} = \cos t_1 \times \cos t_2$$

・文節 i と文節 k の内容語語形と機能語語形が一致 $\Rightarrow cost_1 = 0.0$

・文節 i と文節 k の受け種別と機能語語形が一致 $\Rightarrow cost_1 = 0.2$

・文節 i と文節 k の機能語語形が一致 $\Rightarrow cost_1 = 0.4$

・文節 i と文節 k の受け種別と係り種別が一致 $\Rightarrow cost_1 = 0.6$

・文節 i と文節 k の係り種別が一致 $\Rightarrow cost_1 = 0.8$

・文節 i と文節 k の受け種別が一致 $\Rightarrow cost_1 = 0.9$

・文節 i と文節 k の全てが不一致 $\Rightarrow cost_1 = 1.0$

[2] パラメータ値の妥当性の検証については今後の課題である。

4）すべての文節の組に対して累積文節対応コスト $d_{i,k}$ を次のように計算する。ここで $i=0$ と $k=1,\cdots,K$ に対し、累積文節対応コスト $d_{0,k}=k$ とし、$k=0$ と $i=1,\cdots,I$ に対する累積文節対応コスト $d_{i,0}=i$ として初期化する。

$$d_{i,k}=\min\left\{d_{i-1,k-1}+c_{i,k},\ d_{i,k-1}+c_{*,k},\ d_{i-1,k}+c_{i,*}\right\}$$

$$c_{*,k}=2^{-depth(k)}$$

$$c_{i,*}=2^{-depth(i)}$$

5）文間の距離を $d_{1,K}$ によって求める。

このように計算される文間距離は根文節（文末文節）の一致性を最重要視し、根文節からの深さが深くなるに従って重要度合いを減らした定義をしている。このように文末表現を重視した定義を行うことで文型の類似性を判断している。尚、この計算手順によって、文ａを「これにより、雌コンタクト長をさらに短くすることができる。」とし、文ｂを「溶存オゾン濃度が高ければ注入時間を短くすることができる。」とした場合の文間距離が 0.263 と計算されることが確認できる。ここで、文節対応コストと累積文節対応コストの計算結果をそれぞれ表２と表３に示す。

表２　文節対応コストの計算結果（例）

文節対応コスト							0	1	2	3	4	5	6	文bの文節番号k
								2	4	4	5	6	−1	係り先の文節番号
								4	3	3	2	1	0	文節kの根文節からの深さ
							ダミー	溶存オゾン濃度	高けれ	注入時間	短くする	こと	できる	自立語
								N	V	N	NV	N	NV	受け種別
								が	ば	を		が	。	付属語、句読点
								y	y	y	t	y	s	係り種別
0			ダミー				0.000	1.000	1.000	1.000	1.000	1.000	1.000	
1	2	4	これ	N	に	y	1.000	0.038	0.063	0.063	0.063	0.063	0.063	
2	5	3	より	V	、	y	1.000	0.063	0.075	0.100	0.125	0.125	0.125	
3	5	3	雌コンタクト長	N	を	y	1.000	0.063	0.100	0.025	0.125	0.125	0.125	
4	5	3	さらに	E		y	1.000	0.063	0.100	0.100	0.100	0.125	0.125	
5	6	2	短くする	NV		t	1.000	0.063	0.125	0.125	0.000	0.250	0.500	
6	7	1	こと	N	が	y	1.000	0.063	0.125	0.125	0.250	0.000	0.500	
7	−1	0	できる	NV	。	s	1.000	0.063	0.125	0.125	0.250	0.500	0.000	
文aの文節番号i	係り先の文節番号	文節iの根文節からの深さ	自立語	受け種別	付属語、句読点	係り種別								

表３　累積対応コストの計算結果（例）

累積対応コスト							0	1	2	3	4	5	6	文bの文節番号k
								2	4	4	5	6	−1	係り先の文節番号
								4	3	3	2	1	0	文節kの根文節からの深さ
							ダミー	溶存オゾン濃度	高けれ	注入時間	短くする	こと	できる	自立語
								N	V	N	NV	N	NV	受け種別
								が	ば	を		が	。	付属語、句読点
								y	y	y	t	y	s	係り種別
0			ダミー				0.000	1.000	2.000	3.000	4.000	5.000	6.000	
1	2	4	これ	N	に	y	1.000	0.038	0.100	0.163	0.225	0.288	0.350	
2	5	3	より	V	、	y	2.000	0.100	0.113	0.200	0.325	0.450	0.575	
3	5	3	雌コンタクト長	N	を	y	3.000	0.163	0.200	0.138	0.263	0.388	0.513	
4	5	3	さらに	E		y	4.000	0.225	0.263	0.263	0.238	0.363	0.488	
5	6	2	短くする	NV		t	5.000	0.288	0.388	0.388	0.263	0.513	0.763	
6	7	1	こと	N	が	y	6.000	0.350	0.513	0.513	0.513	0.263	0.763	
7	−1	0	できる	NV	。	s	7.000	0.413	0.638	0.638	0.763	0.763	0.263	
文aの文節番号i	係り先の文節番号	文節iの根文節からの深さ	自立語	受け種別	付属語、句読点	係り種別								

※網掛けは最短経路の文節対応である。

3．プロトタイプでの評価実験

前節の方法で構築した文型類似文辞書の有効性を評価するために、我々はこの辞書を組み込んだ英文作成支援システムのプロトタイプを作成し、評価実験を行った。

3.1　英文作成支援システムの概要

まず文型類似文辞書を用いた英文作成支援システムの概要を述べる。図10は英文作成支援システムのユーザインタフェースであり、日本語文入力エリアと文型類似文表示エリアからなる。日本語を入力して検索ボタンをクリックすると、該当する文型類似文が文間距離（類似度）順に表示される。図11は英文作成支援システムの構成であり、図10の処理の流れを表している。まず、ユーザインタフェースの日本語入力エリアに、英語に翻訳をしたい日本語文を入力する。次に、検索ボタンをクリックすると、入力文に対してChaSenで形態素解析、CaboChaで係り受け解析が行われる。その後、構築した文型類似文辞書の文型とパターンマッチングが行われる。この時に、文末に着目した距離計算（2.2節）がなされる。ここで全ての索引のうち、距離が近い順番に日英文のペアが文型類似文表示エリアに表示される。それぞれのペアから該当クラスタ内の全ての文を閲覧することができ、それを用いて翻訳者は英文作成を行う。

図10の例では、英語に翻訳をしたい日本語文として、「これにより、アンテナ本体1がルーフパネル20に確実に固着される。」を入力している。検索ボタンをクリックすると、最も日本語文末の距離が近い日本文として、1つ目に「この接着剤5によってセラミックスヒーター10がステー6に固着される。」が表示され、そのペアとなる英文として、“The adhesives 5 fix the ceramics heater 10 to the stay 6.”が表示される。以降、日本語文末の距離が近い順番に日英文のペアが文型類似文表示エリアに表示される。翻訳者はそれらを参照しながら、英語に翻訳をしたい日本語文を英訳していく。こ

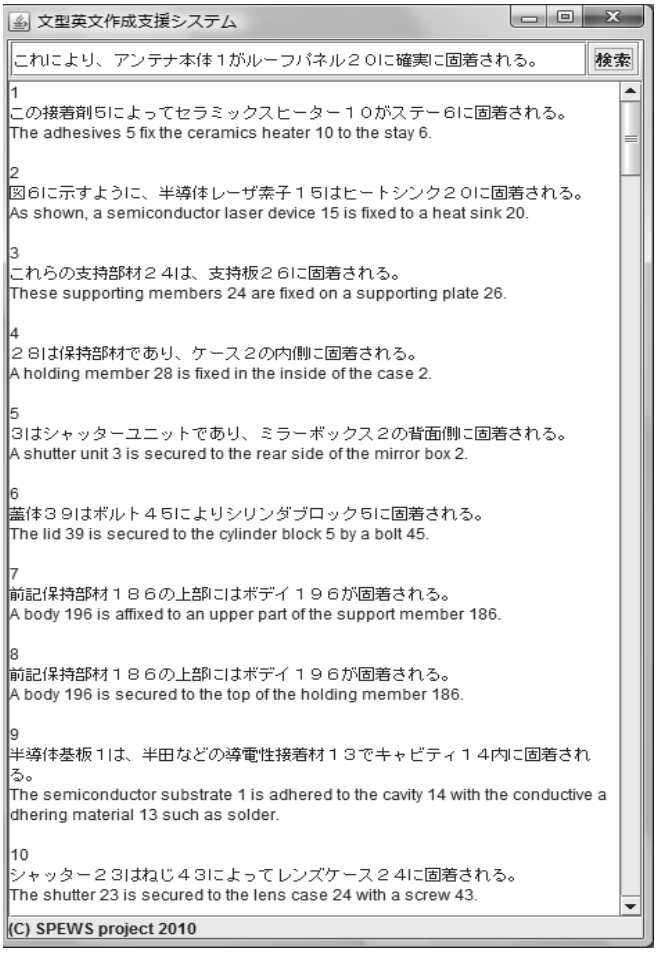

図 10　英文作成支援システムのユーザインタフェース

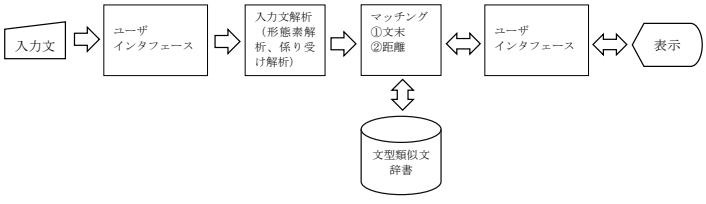

図 11　英文作成支援システムの構成

の例の「これにより、アンテナ本体１がルーフパネル 20 に確実に固着される。」という文を受動文で英訳すると、翻訳者は文型類似文表示エリアから受動文で書かれた文を探し、その文を参照して英文を作成する。この例では１の能動文（動詞が fix）ではなく２の受動文（動詞が be fixed to）を参照する。すると、翻訳者は “Therefore, a main body of antenna 1 is fixed to a roof panel 20, tightly.” のような自然な英語で文書を作成することができる可能性が上がる。

3.2　評価実験

今回の評価実験では、英文作成支援システムのプロトタイプを作成して行う。プロトタイプでは、ChaSen や CaboCha による入力文解析の処理は行わず、40 の日本語文に対して、あらかじめ距離計算された例文を類似度順に並べて文型類似文表示エリアに表示させる方法を採用した。

実験は 2011 年と 2013 年の２度実施した。１度目の実験では被験者は 12 名で、本システムを使って英文を作成する場合と、辞書を使って英文を作成する場合での比較実験を行った。被験者の TOEIC の点数の分布を表４に示す。被験者は学部生であり、科学技術文書（特許文）を対象に行った。

表４　実験１における被験者の TOEIC スコア

	被験者のTOEICスコア			
	500–599	600–699	700–799	800–899
人数	4	3	3	2

被験者の英文作成能力と長時間の解答作業（英文作成）のための疲労による影響をなくすため、以下のような交差検証法で実験を行った。

はじめに、40 ある問題を目検により難易度が同程度になるように２群に分割した。それぞれを問題 X、問題 Y と呼ぶ。次に英語の実力が平均的になるように被験者を４群に分け、順に１から４群とした。各群に対して、表５のように実験を行った。

表５　交差検証法によるグルーピングと実験の手順

	第１群	第２群	第３群	第４群
第１回	本システムを用いて問題 X を解答	本システムを用いて問題 Y を解答	辞書を用いて問題 X を解答	辞書を用いて問題 Y を解答
休憩	15 分	15 分	15 分	15 分
第２回	辞書を用いて問題 Y を解答	辞書を用いて問題 X を解答	本システムを用いて問題 Y を解答	本システムを用いて問題 X を解答

実験で被験者が作成した英作文に対して、Adequacy と Fluency を用いて人手で評価した [16][13]。この手法は機械翻訳の人手評価において広く用いられている。Adequacy に関しては日本人の大学英語教員が評価し、Fluency に関してはネイティブが評価を行った。それぞれ５点を最も高い評価、１点を最も低い評価とした。尚、英作文ができていないものに対しては評価対象外とした。

評価の事例を表６に示す。表６は問題グループ X の１番目の問題とその

表６　評価の事例

sequence No.	group	subject	system	Japanese sentence group	testing Japanese sentence No.	kind of the sentence	Adequacy	Fluency	sentence
1									
2						1 testing Japanese sentency			図3はマークに対応した監視画面を構成する場合の一例を示す図である。
3						2 reference tlanslation(Pro)			FIG. 3 is a diagram showing a typical case in which a monitor screen corresponding to a mark is formed.
4	1	1	A	X	1	3 made English sentence	4	2	Fig. 3 shows an example of constructing a watching monitor which corresponds to its mark.
5	1	2	A	X	1	3 made English sentence	2	2	FIG.3 shows an example of watching screen structure suited marks.
6	1	3	A	X	1	3 made English sentence	4	3	FIG. 3 shows an example of a situation which the monitoring display corresponding to the mark is provided.
7	2	4	B	X	1	3 made English sentence	2	3	Fig. 3 shows one example in the case of a watch screen corresponded to the mark.
8	2	5	B	X	1	3 made English sentence	2	3	Figure.3 shows as an example of mark structured a watch screen.
9	2	6	B	X	1	3 made English sentence	2	2	Fig3 shows case of instance that mark corresponded to watch screen.
10	3	7	B	X	1	3 made English sentence	2	2	FIG.3 is shown example structure to mark corresponding watch screen.
11	3	8	B	X	1	3 made English sentence	4	3	Figure 3 shows an example of the case of constructing a surveillance monitor corresponding the mark.
12	3	9	B	X	1	3 made English sentence	3	3	FIG. 3 shows an example which compose watching monitor correspond to marks.
13	4	10	A	X	1	3 made English sentence	3	2	Fig. 3 shows an example of the case of constructing the watching screen that correspondence to the mark.
14	4	11	A	X	1	3 made English sentence	5	3	Fig. 3 shows an example of the composition of a monitoring screen corresponded to the mark.
15	4	12	A	X	1	3 made English sentence	2	3	FIG.3 shows an example of the watch screen correspondence to mark.

翻訳結果及び評価結果である。「図３はマークに対応した監視画面を構成する場合の一例を示す図である。」という日本文に対して、12 名の被検者がそれぞれ本システムと対照システムを使用して翻訳を行った結果である。その結果に対して、評価者が Adequacy と Fluency を人手で付与している。尚、reference translation はプロの翻訳者が行った参照訳であり、もちろん被験者には開示しない。

解析手法として、実施データに対してそれぞれ Adequacy と Fluency の平均値の差の検定を行った。

評価の統計処理の結果を表７に示す。平均値は、１群から４群の間に多少の違いがあることが確認できるが、全群による検定では Adequacy と Fluency ともに、辞書のみの使用よりも本システムを使用した方が高く、また、全群による検定での２つの母平均の差の検定に関しては、それぞれ 95%、99% 以上で有意であった。この結果は、本システムの利用が、Adequacy を向上させることにつながり、また、作成する英文の Fluency を向上させることができるという点においては更に優れていることを示している。

表７　実験１における Adequacy と Fluency の検定結果

群	評価	システム	N	平均値	評価	等分散の仮定	等分散性のためのLeveneの検定		2つの母平均の差の検定		
							F 値	有意確率	t 値	自由度	有意確率（両側）
1群	Adequacy	本システム	58	2.793	Adequacy	等分散を仮定する。	0.115	0.735	−0.039	116	0.969
		辞書	60	2.800		等分散を仮定しない。			−0.039	115.879	0.969
	Fluency	本システム	58	3.241	Fluency	等分散を仮定する。	0.001	0.982	1.644	116	0.103
		辞書	60	2.950		等分散を仮定しない。			1.644	115.995	0.103
2群	Adequacy	本システム	60	2.883	Adequacy	等分散を仮定する。	2.934	0.089	3.171	118	0.002
		辞書	60	2.267		等分散を仮定しない。			3.171	115.802	0.002
	Fluency	本システム	60	3.067	Fluency	等分散を仮定する。	0.171	0.680	0.729	118	0.467
		辞書	60	2.950		等分散を仮定しない。			0.729	117.658	0.467
3群	Adequacy	本システム	60	2.717	Adequacy	等分散を仮定する。	0.883	0.349	2.104	118	0.038
		辞書	60	2.333		等分散を仮定しない。			2.104	117.003	0.038
	Fluency	本システム	60	3.100	Fluency	等分散を仮定する。	3.305	0.072	1.794	118	0.075
		辞書	60	2.800		等分散を仮定しない。			1.794	113.284	0.075
4群	Adequacy	本システム	60	2.333	Adequacy	等分散を仮定する。	0.708	0.402	−1.358	118	0.117
		辞書	60	2.583		等分散を仮定しない。			−1.358	116.577	0.117
	Fluency	本システム	60	2.933	Fluency	等分散を仮定する。	7.494	0.007	2.241	118	0.027
		辞書	60	2.583		等分散を仮定しない。			2.241	116.165	0.027
全群	Adequacy	本システム	238	2.681	Adequacy	等分散を仮定する。	0.500	0.480	1.968	476	0.050
		辞書	240	2.496		等分散を仮定しない。			1.968	475.241	0.050
	Fluency	本システム	238	3.084	Fluency	等分散を仮定する。	6.332	0.012	3.172	476	0.002
		辞書	240	2.821		等分散を仮定しない。			3.173	473.210	0.002

表８　実験２における被験者の TOEIC スコア

	被験者のTOEICスコア					
	200–299	300–399	400–499	500–599	600–699	700–799
人数	1	2	2	5	1	1

２度目の実験においては、１度目の実験と同じ手順で、辞書の代わりに SCOPE[12] を用いた。これは名古屋大学が提供しているフリーのフレーズ検索システムである。被験者は 12 名の情報工学を学んでいる大学院生であり、翻訳対象とする文も情報工学に関するものを選択して用いた。これによって、学生が専門分野の英文を作成する状況に近い状況を実現した。被験者の TOEIC のスコアの分布を表８に示す。尚、TOEFL のスコアのみの被験者が２名いたが、IELTS NAVI の IELTS/TOEFL/TOEIC/PTE/ 英検 スコア換算表目安 [17] に従って、TOEIC のスコアとして換算した。

実験１と同様に、作成された英作文に対して、Adequacy と Fluency を用いて人手で評価した。

解析手法として、実験１と同様の検定を行った。

評価の統計処理の結果を表９に示す。平均値は１群から４群の間には多少

表９　実験２における Adequacy と Fluency の検定結果

群	評価	システム	N	平均値	評価	等分散の仮定	等分散性のための Leveneの検定		2つの母平均の差の検定		
							F 値	有意確率	t 値	自由度	有意確率（両側）
1群	Adequacy	本システム	36	2.944	Adequacy	等分散を仮定する。	3.881	0.053	0.221	70	0.826
		SCOPE	36	2.889		等分散を仮定しない。			0.221	65.958	0.826
	Fluency	本システム	36	2.889	Fluency	等分散を仮定する。	2.388	0.127	−1.249	70	0.216
		SCOPE	36	3.222		等分散を仮定しない。			−1.249	67.099	0.216
2群	Adequacy	本システム	36	2.667	Adequacy	等分散を仮定する。	0.303	0.584	−0.132	70	0.895
		SCOPE	36	2.694		等分散を仮定しない。			−0.132	69.702	0.895
	Fluency	本システム	36	2.917	Fluency	等分散を仮定する。	0.640	0.427	0.219	70	0.827
		SCOPE	36	2.861		等分散を仮定しない。			0.219	69.248	0.827
3群	Adequacy	本システム	36	2.778	Adequacy	等分散を仮定する。	5.408	0.023	−0.113	70	0.910
		SCOPE	36	2.806		等分散を仮定しない。			−0.113	63.910	0.910
	Fluency	本システム	36	3.278	Fluency	等分散を仮定する。	0.075	0.785	2.543	70	0.013
		SCOPE	36	2.583		等分散を仮定しない。			2.543	69.998	0.013
4群	Adequacy	本システム	36	2.833	Adequacy	等分散を仮定する。	0.711	0.402	0.981	70	0.330
		SCOPE	36	2.583		等分散を仮定しない。			0.981	69.294	0.330
	Fluency	本システム	36	3.028	Fluency	等分散を仮定する。	0.735	0.394	0.117	70	0.907
		SCOPE	36	3.000		等分散を仮定しない。			0.117	69.985	0.907
全群	Adequacy	本システム	144	2.806	Adequacy	等分散を仮定する。	4.096	0.044	0.521	286	0.603
		SCOPE	144	2.743		等分散を仮定しない。			0.521	281.137	0.603
	Fluency	本システム	144	3.028	Fluency	等分散を仮定する。	0.022	0.883	0.856	286	0.393
		SCOPE	144	2.917		等分散を仮定しない。			0.856	285.458	0.393

の違いはあることが確認できる。全群による検定では Adequacy と Fluency ともに、本システムと対照システム（SCOPE）を比べた平均値は高い傾向が見られたが、5% の有意水準での差は認められなかった。

4．議論

本節では、本システムの特徴と評価、結果について議論する。

本システムは、文型類似文辞書を使用していることに大きな特徴がある。この辞書は、日本語文の文末表現を重視した文間距離を定義して構築された。辞書や本 [18-19] などを用いて英文作成をする場合、翻訳者自身で該当文型を探さなければならず、時間のロスなどコストが大きい。これに対し、提案したシステムでは入力文に対応する文型を自動的に計算し、類似する候補文を複数提示するため、ユーザの精神的及び作業的負荷が少ない。本システムは、クラスタによって類似文を集めているため、当該文に対する類似文を複数提示することが可能である。また、文型類似文を入力文に近い順に自動的に距離計算して提示することができる。このことによって、ユーザの精神的及び作業的負荷軽減に大きく貢献できる。

本システムは科学技術文（特に特許文）を対象としている。文型類似文辞書を構築するために使用したコーパスは NICIR-7 PATMT コーパスである。このコーパスは、特許翻訳テストコレクションである。テストコレクションとは、機械翻訳を評価するための文書データの集合である。今回、我々は特許翻訳テストコレクションの中の、日英の特許出願のまとまりである日英パテントファミリーから抽出された約 180 万文のモデル訓練用日英対訳データを使用している。そのため文型の網羅性が高く、また、日英両文とも人手で作成された文であるため文の完全性も高い。科学技術文の文書的特徴としては、複文や長文が多いことである。現段階では、これらの文を機械翻訳で正確に翻訳するのは困難である。この点は、既存のテンプレートによる英文作成支援や、キーワードを入力して検索された文を表示する英文作成支援シス

テムを使用しても解決は困難である。なぜなら、それらでは類似文には到達できず、文法構造的に異なる別の文を提示してしまう可能性が高いからである。それに対して、本システムでは、文末集合が日本語の構文決定の主要な要因であることに着目しているので、文法構造的に一致した類似文の提示が可能である。

本研究では、科学技術文書（特許文）を対象に最初に提案手法（本システム）を使用した場合と辞書を使用した場合での評価実験を行った。評価においては Adequacy と Fluency について人手で評価した。平均値の差の検定で評価した結果、辞書のみ使用時よりも Adequacy に優れ、Fluency に関しては更に優れていることが分かった。このことにより、意味が正しく伝達される英文の作成を支援することもさることながら、文法的・表現的にこなれた英文の作成を支援できるということが実験によって明らかになった。本研究の目的は、質の良い英文を作成することの支援であるため、その実現には近づくことができた。

2 回目の実験では SCOPE との比較を実施した。表 9 の全群での検定結果を見ると、Adequacy は「等分散性のための Levene の検定」の結果、有意確率が 0.044 となり 5% 水準で等分散の仮定を棄却した。これは被験者数が 12 名と少ないことが原因として考えられる。等分散性を仮定しない平均値の差の検定が示すように、平均値に関する帰無仮説を棄却できなかった。すなわち本実験は、提案するシステムが SCOPE と同水準であるという仮定を棄却できないことまでは示すことができた。人数を増やした追試によって、優位性を期待したい。

5．まとめと今後の課題

我々は、日英翻訳を行う際に重要となる文型の抽出を容易に行えるようにするため、コーパスを用いる方法のうち、キーワードの一致のみで候補文を抽出する従来型に代わって、文末に注目する方法を提案した。この方法では、

文末文節に着目した文間距離に基づき文集合をクラスタリング手法でグルーピングし、代表的な文を索引とする文型類似文辞書を構築する。この辞書を用いると、翻訳したい文と同じ文末構造を持つ文を優先的に複数文表示させることが可能となる。このような方法では、いわゆる機械翻訳のように直訳的に翻訳するのでは作れないような、こなれた英文を作成するときなどに効果を発揮すると予想される。それを確かめるために、文型類似文辞書を用いた英文作成支援システムのプロトタイプを構築した。このシステムは、日本語のキーワードではなく、日本語文そのものを入力して効率的に文型情報を考慮した英文を提示することができる。これを使用して、被験者を用いた評価実験を行った。結果として、辞書のみ使用した場合と比べて本システムを使用した場合では、作成された英文の Adequacy ならびに Fluency を改善でき、特に Fluency の改善に関して優れていた。類似システム（SCOPE）との比較においても同レベルの品質を維持できることは示せた。

日本語文末が類似している二つの文があるとき、それぞれに対応する英訳文も構造（統語）、意味ともに類似しているかどうかの実証は未了であるが、英文作成支援システムとしての有用性は実験的に示すことができた。上記の類似性の検証は今後の課題である。

また、定義した文間距離の妥当性の検証や、他分野のコーパスも用いてのシステムの実装も今後の課題として挙げられる。更に、今回は学生を対象に、２つの大学で実験を行ったが、より正確な実験結果を得るためにも、学生以外の異なった被験者で実験を行うことも考えられる。より多くの検証可能性に耐えうるシステムを構築するためにも、実験を繰り返す中で新たな知見を獲得し、アルゴリズムの改良とシステムの品質向上を目指していきたい。

その他の課題としては、文間距離を定義して作成された文型と、いわゆる言語学的な文型とが一致するかどうかという点である。現段階では文間距離は経験則によって定義している。そのため、必ずしも言語学的に正しい文型を抽出できているわけではない。この点においてはまだ議論の余地があるだろう。また、対象文が提示されたのちに、全ての文との文間距離を計算する

必要があり、コストがかかる。我々は、全ての文との文間距離を計算するのではなく、予めコーパスの類似文を一つのクラスタに集めて、そのクラスタの代表文との文間距離を計算する手法を提案しているが、この実装並びに効率性の向上などは今後の課題とする。更に、特許文だけに限らず、他分野のコーパスからの文型類似文の抽出と実装をすることで、実用性を高めていきたい。いずれにせよ、我々は、より性能の高い英文作成支援システムを構築するために、得られた知見と課題を考慮して実装を進めていきたいと考えている。

尚、本研究は文学作品の翻訳支援を対象にしたものではなく、あくまで科学技術文献の実務翻訳を対象にしたものである。科学技術の分野においては、言語の使用を制限し、曖昧性を排除する必要がある [20]。そのため、文章の美的価値や機智に富んだ表現よりも、曖昧性が排除された記述による文章構成に重きを置いている。科学技術文献における言語の平板化については、曖昧性を排除した記述をするためには必要なことであると考えている。しかしながら、大規模な文集合というデータから類似する文章を探し出す機械翻訳や英文作成支援システムの類は、人類の言語表現を画一化し、言語による独創的な表現方法を自らの力で探し出す機会を奪うという副作用も考えられ議論の余地が残されている。

参考文献

[1] 富永勲 , 佐藤雅之 : “機械翻訳システムの発展と実用化 (1) 国内外の開発の経緯と現状” , Journal of Information Processing and Management 33(7), 593-605 (1990)
[2] 東芝ソリューション株式会社 : “The 翻訳 2009 ビジネス” , http://pf.toshiba-sol.co.jp/prod/hon_yaku/business/index_j.htm（2015）
[3] 五十嵐健夫 : “文の構造を明示的に指定・表示することによる異言語間コミュニケーション” , 第 19 回インタラクティブシステムとソフトウェアに関するワークショップ , (2011)

[4] 堺武志：“日本語 - 外国語 文例翻訳ソフト : ことばさがし”，http://rose.ruru.ne.jp/multiplication/kotobaWin/jp.html (2010)
[5] 松原茂樹，江川誠二，加藤芳秀 : “英文用例検索システムＥＳＣＯＲＴ : 論文データベースを用いた図書館サービス”，情報プロフェッショナルシンポジウム予稿集，pp. 125-129 (2007)
[6] 大鹿広憲，佐藤学，安藤進，山名早人 : “Google を活用した英作文支援システムの構築”『DEWS2005』4B-i8 (2005)
[7] 成田真澄：“英文アブストラクト作成支援ツールのユーザ評価”，平成 12 年度 COE 形成基礎研究費研究成果報告 (先端的言語理論の構築とその多角的な実証 - ヒトの言語を組み立て演算する能力を語彙の意味概念から探る -)，神田外語大学，pp.309-318 (2001)
[8] 三好康夫 , 岡本竜：“用例に基づく対話的英作文支援システム”，JSiSE(教育システム情報学会)-W 第 3 回若手研究者フォーラム (2000)
[9] 小倉書店編：“CD-ROM 科学技術論文，報告書その他の文書に必要な英語文型・文例辞典”，小倉書店 (1997)
[10] 王軼謳 , 卜朝暉 , 宇野修一 , 浅井良信 , 池田尚志：“日中機械翻訳における存在文の翻訳処理について” , 言語処理大会 (2006)
[11] 池原 悟 , 阿部さつき , 徳久 雅人 , 村上 仁一：“非線形な表現構造に着目した日英文型パターン化” , 情報処理学会研究報告 . NL, 自然言語処理研究会報告 160, pp.49-56 (2004)
[12] 酒井佑太 , 小澤俊介 , 杉木健二 , 松原茂樹 :“英語論文からの表現集の自動生成”，言語処理学会　第 16 回年次大会（2010）
[13] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro. (2008): “Overview of the Patent Translation Task at the NTCIR-7 Workshop” , Proceedings of NTCIR-7 Workshop Meeting, pp.389-400, December 16-19, 2008, Tokyo, Japan.
[14] 松本裕治：“ChaSen -- 形態素解析器”，http://chasen-legacy.osdn.jp/ (2015)
[15] 松本裕治：“CaboCha/ 南瓜 : Yet Another Japanese Dependency Structure Analyzer”，http://taku910.github.io/cabocha/ (2015)
[16]Linguistic Data Annotation Specification: “Assessment of Fluency and

Adequacy in Chinese-English Translations Revision 1.0”, pp. 2-3, http://projects.ldc.upenn.edu/TIDES/Translation/TranAssessSpec.pdf (2002)
[17]IELTS NAVI – アイエルツ ナビ: “IELTS/TOEFL/TOEIC/PTE/ 英検 スコア換算表目安” ,http://ieltsnavi.com/score_conversion.html
[18] 佐藤元志著，田中宏明監修，古米弘明監修，鈴木穣監修：“英語論文表現例集 with CD-ROM　すぐに使える 5,800 の例文”，技報堂出版 (2009)
[19] 佐藤洋一編著 : “科学技術英語論文英借文用例辞典　英作文から英借文へ簡単！英語論文作成法”，オーム社 (2010)
[20] 深澤のぞみ：“科学技術論文作成を目指した作文指導 ―専門教員と日本語教師の視点の違いを中心に―”，公共社団法人　日本語教育学会 , 日本語教育 (1994)

まつやま　ひろき・秀明大学助教
おかだ　いさむ・創価大学准教授
えはら　てるまさ・江原自然言語処理研究室代表
みやざき　みつゆき・秀明大学准教授
ドーン　エル　ミヤザキ・早稲田大学講師
みやざわ　しんいちろう・秀明大学名誉教授