日英新聞の記事および文を対応付けるための高信頼性尺度

内山　将夫[†]　井佐原　均[†]

大規模な日英対訳コーパスを作ることを目的として，1989 年から 2001 年までの読売新聞と The Daily Yomiuri とから日英記事対応と文対応とを得た．そのときの方法は，まず，内容が対応する日本語記事と英語記事とを言語横断検索により得て，次に，その対応付けられた日英記事中にある日本語文と英語文とを DP マッチングにより対応付けるというものである．しかし，それにより対応付けられた記事対応や文対応には，間違った対応 (ノイズ) が多く含まれる．そのため，我々は，本稿において，そのようなノイズを避けて，正しい対応のみを得るための信頼性の高い尺度を提案し，その信頼性の評価をした．実験の結果，我々の提案した尺度を用いることにより，良質な記事対応や文対応が得られることがわかった．また，その数は，良質な記事対応は約 4 万 7 千であり，文対応は，1 対 1 対応が約 15 万，1 対 1 対応以外が約 3 万 8 千であった．これらは，現時点で一般に利用できる日英 2 言語コーパスとしては最大のものである．

キーワード：　日英対訳コーパス，記事アライメント，文アライメント

Reliable Measures for Aligning Japanese-English News Articles and Sentences

Masao Utiyama[†] and Hitoshi Isahara[†]

We have aligned Japanese and English news articles and sentences, extracted from the Yomiuri and the Daily Yomiuri newspapers, to make a large parallel corpus. We first used a method based on cross-lingual information retrieval to align the Japanese and English articles and then used a method based on dynamic programming (DP) matching to align the Japanese and English sentences in these articles. However, the articles and sentences included many incorrect alignments. To remove these, we propose two measures that evaluate the validity of the alignments. Using these measures, we successfully extracted a valid correspondence of about 47 thousands article pairs, 150 thousands 1-to-1 sentence pairs, and 38 thousands 1-to-many sentence pairs. We were therefore able to build the largest Japanese-English parallel corpus available to the public.

KeyWords:　*Japanese-English parallel corpus, article alignment, sentence alignment*

1　はじめに

日英対訳コーパスは，機械翻訳などの自然言語処理において必要であるばかりでなく，英語学や比較言語学，あるいは，英語教育や日本語教育などにとっても非常に有用な言語資源であ

† 通信総合研究所, Communications Research Laboratory

る．しかしながら，これまで，一般に利用可能で，かつ，大規模な日英対訳コーパスは存在していなかった．

そのような背景の中で，我々は，比較的大規模な日本語新聞記事集合およびそれと内容的に一部対応している英語新聞記事集合とから，大規模な日英対訳コーパスを作ることを試みた．

そのための方法は，まず，内容が対応する日本語記事と英語記事とを得て，次に，その対応付けられた日英記事中にある日本語文と英語文とを対応付けるというものである．

ここで，我々が対象とする日本語記事と英語記事においては，英語記事の内容が日本語記事の内容に対応している場合には，その英語記事は，日本語記事を元にして書かれている場合が多いのであるが，その場合であっても，日本語記事を直訳しているわけではなく，意訳が含まれていることが多く，更に，日本語記事の内容の一部が英語記事においては欠落していたり，日本語記事にない内容が英語記事に書かれている場合もある．また，記事対応付けを得るための日本語記事集合と英語記事集合についても，英語記事集合の大きさは日本語記事集合の大きさの6%未満であるので，日本語記事の中で，対応する英語記事があるものは極く少数である．

そのため，記事対応付けおよび文対応付けにあたっては，非常にノイズが多い状況のなかから，適切な対応付けのみを抽出しなくてはならないので，対応の良さを判断するための尺度は信頼性の高いものでなくてはならない．

本稿では，そのような信頼性の高い尺度を，記事対応付けと文対応付けの双方について提案し，その信頼性の程度を評価する．また，作成した対応付けデータを試験的に公開したときの状況についても述べ，そのようなデータが潜在的に有用な分野について考察する．

以下では，まず，対応付けに用いた日英新聞記事について概要を述べ，次に，記事対応付けの方法と文対応付けの方法を述べたあとで，それぞれの対応付けの精度を評価する．最後に考察と結論を述べる．また，付録には，実際に得られた文対応の例を示す．

2　対応付けに用いた日英新聞記事

対応付けの元データは，日本語記事は「読売新聞」，英語記事は「The Daily Yomiuri」であり，それぞれ「読売新聞記事データ」における1989年9月から2001年12月までの記事を利用した．この期間における年間の記事数は，日本語記事は10万から35万程度であり，英語記事は4千から1万3千程度である．また，総記事数は，日本語記事は約200万であり，英語記事は約11万である．このように，英語記事の方が少ないので，対応付けにおいては，各英語記事に対応する日本語記事を求めることにした．

記事のメタ情報として，The Daily Yomiuriには，1996年7月中旬から,「本紙翻訳 = Y/N」という情報が各記事に付いている．これは，その英語記事を書くにあたって，読売新聞の記事を元にしたかどうかという意味であるので，1996年7月中旬からは,「本紙翻訳=Y」である英語記事についてのみ，対応する日本語記事を求めることにした．このときの英語記事の数は 35318

である．一方，1996年7月中旬以前には，そのような情報はないので，全ての英語記事について対応する日本語記事を求めることにした．このときの英語記事の数は 59086 である．なお，以下では，1996年7月中旬以前の記事集合を「1989-1996」と書き，1996年7月中旬以降の記事集合を「1996-2001」と書くことにする．

1989-1996については，全英語記事を利用するため，1996-2001と違って，そもそも，各英語記事について対応する日本語記事がない場合がある．そのため，どのくらいの英語記事に，対応する日本語記事があるかを推測するために,「本紙翻訳=Y」の割合を，1997年から2001年の記事について調べたところ，67.9％であった．

対応を求めるにあたって，各英語記事に対応する日本語記事は，互いに近い日付であると考えられる．そのため，各英語記事について，その日付の前後2日の範囲の日本語記事の中から対応する記事を見付けることにした．このとき，1日分の英語記事について，日本語記事は5日分があるが，このときの平均記事数は，1989-1996については，英語記事が24，日本語記事が1532，1996-2001については，英語記事が18，日本語記事が2885である．

このように，非常に曖昧性があり，かつ，対応記事も場合によっては存在しないという，ノイズの多い状況のなかから対応記事を見つける必要があるので，信頼性の高い記事対応(評価)尺度が必要である．

また，文対応についていえば，たとえ記事同士が対応していたとしても，その対応は，直訳関係にあるものは少なく，どちらかというと，日本語記事を材料として英語記事を書いたというような状況である．たとえば，以下の例では，英語と日本語とで，e1，e3，e4とj1，j2，j3，j4とによる3対4に複雑に絡みあう対応があり，その間にe2とj5による対応がある．

> <e1> Two bullet holes were found at the home of Kengo Tanaka, 65, president of Bungei Shunju, in Akabane, Tokyo, by his wife Kimiko, 64, at around 9 a.m. Monday. </e1> <e2> Police suspect right-wing activists, who have mounted criticism against articles about the Imperial family appearing in the Shukan Bunshun, the publisher's weekly magazine, were responsible for the shooting. </e2> <e3> Police received an anonymous phone call shortly after 1 a.m. Monday by a caller who reported hearing gunfire near Tanaka's residence. </e3> <e4> Police found nothing after investigating the report, but later found a bullet in the Tanakas' bedroom, where they were sleeping at the time of the shooting. </e4>
> <j1> 二十九日午前八時五十五分ごろ、東京都北区赤羽西四文芸春秋社長、田中健五さん（６５）方の二階東側外壁に、短銃で撃たれた跡があるのを、妻喜美子さん（６４）が見つけた。 </j1> <j2> 赤羽署で調べたところ、寝室の外壁に二か所の穴が確認され、銃弾一発が寝室内から発見された。 </j2> <j3> これに先立ち、午前一時すぎ、田中さん方周辺で「短銃の発射音のような音が二、三発聞こえた」という匿名の通報が同署にあり、署員が確認に向かったが、この時点で銃痕は発見できなかった。 </j3> <j4> 発射音がしたころ、田中夫妻は寝室で就寝中だったという。 </j4> <j5> 同社が発行している週刊誌「週刊文春」が、最近、皇室批判記事を掲載していたことから、同署では、皇室批判に反発する右翼の犯行の可能性があるとみて、捜査をしている。 </j5>

このような文対応は，人間の観察者(たとえば，日英記事のスタイルを比較研究しているような人)にとっては価値があるが，文対応の結果を自然言語処理，たとえば，機械翻訳に利用しようとしている場合には，今のところは，有用性は限定されている．そのため，なるべく直訳同士にあるような文対応を抽出したいのであるが，このような状況から直訳に近い文対応を抽出するためには，信頼性の高い文対応(評価)尺度が必要である．

3　対応付けの方針

これまで，1節で，日英対訳コーパスが必要とされてていることを述べ，また，2節において，対応付けの元となる日英新聞記事に付いて述べた．本節では，これらの節に基づいて，本稿における，記事対応付けおよび文対応付けの方針について述べる．それは以下の2点である．

(1)　まず，日英記事対応付けと文対応付けとは，本稿における目的ではあるが，そのような対応付けをすること自体の目的は，その対応付けの結果を利用して，機械翻訳なり英語教育なりに役立てることである．そのため，対応付けについては，もし，既存の言語資源および手法を利用することにより，ある程度の量と精度の対応付けが得られるなら，あえて新しい言語資源や手法を開発することなく，既存の言語資源や手法を有効に利用する．

(2)　しかし，対象とするコーパスには，多くのノイズがあるため，既存の言語資源や手法をそのまま利用した場合に得られる対応付けには，間違った対応付けも多く含まれる．そのため，その対応付けのなかから，良さそうな対応付けのみを抽出するための信頼性の高い尺度を考える．

こうした場合には，対象とするコーパスに潜在的に存在する対応付けのうちで，既存の言語資源や手法により抽出されなかったものは利用できない，そのため，対応付けの再現率は低い可能性がある．しかし，良さそうな対応付けとして抽出されたものの精度は高いことが期待できる．

つまり，上記の方針は，再現率よりも精度を重視するということである．以下，この方針に基づき，4節と5節では，既存の言語資源や手法に基づいて記事対応と文対応とを取る方法について述べ，6節では，得られた対応付けの中から良さそうな対応付けを得る尺度について述べる．

4　記事対応付けの方法

記事対応付けは，言語横断検索の枠組で行なう．つまり，英語記事を質問とし，それに関連する記事を日本語記事データベースから検索することにより，与えられた英語記事と対応する日本語記事を見付ける．

このとき，一般に，質問である英語記事を日本語に変換するか，あるいは，データベースである日本語記事を英語に変換する必要がある．本研究では，データベースである日本語記事を英語(の単語集合)に変換した．そうした主な理由は，手元にある言語資源が日英方向の変換に便利だったからである．

4.1　日本語記事の英単語集合への変換

我々は，辞書引きに基づいて日本語記事を英単語集合に変換することにした．利用した日英辞書は，EDR 日英対訳辞書，EDICT (一般的な日英対訳辞書)，ENAMDICT (固有名詞の日英対訳辞書) である[1]．これらの辞書の見出し語に対して，IPADIC (version 2.4.4) の品詞体系を付与し，茶筌[2] (version 2.2.8) の追加辞書として利用した．追加したエントリ数は，EDR 日英対訳辞書が約 18 万，EDICT が約 6 万，ENAMDICT が約 22 万である[3]．

こうすることにより，茶筌の解析結果から容易に日英対訳辞書のエントリがアクセスできるようになる．たとえば,「あおぎ見た月」は，追加辞書なしの状態では

あおぎ	あおぐ	動詞-自立
見	見る	動詞-自立
た	た	助動詞
月	月	名詞-一般

と形態素解析される (形態素情報の一部を省略) が，追加辞書ありの状態では

あおぎ見	あおぎ見る	動詞-自立
た	た	助動詞
月	月	名詞-一般

のように解析され，特に工夫をせずとも，複合語である「あおぎ見る」の訳語として「look up」「face upwards」「look up to」「respcet」「admire」などが得られる．また，この方法によると，「くすの木台に行く」を形態素解析した場合のように，辞書にない単語に起因する解析誤りである「くす/の/木/台/に/行く」のようなものも「くすの木台/に/行く」として解析でき，かつ，「くすの木台」の訳語として「Kusunokidai」も容易に得られる．このように，IPADIC を増強することにより，解析誤りを避けながら，容易に日英辞書の辞書引きができると共に，複合語や固有名詞の翻訳という言語横断検索において重要な作業も同時にできるため，この方法は有用である．

このようにして日本語の各単語 (もしくは複合語) において，その品詞が内容語 (主に名詞) に相当するものから英訳語を得て，そこから簡単なヒューリスティクスにより主辞を抽出し当該日本語単語の変換結果としたが，このとき，各単語についてその全ての訳語の主辞全てを変換結果として採用するとすると，訳語の主辞のなかには当該文脈の訳として適当でないものもあるため，検索結果に悪影響を与えると考えられる．そのため，なるべく，訳語として適当なものだけを変換結果として利用したい．そうするためには，訳語の曖昧性を解消すれば良いのだが，それを正確にするのは困難である．そのため，ここでは，ヒューリスティクスとして，まず，訳語の主辞の中から，より多くの訳語に含まれているようなものを優先し，次に，同順位のものについては，その訳語の主辞に対応する日本語単語を含む日本語記事の年と同年の英語記事において，その訳語の主辞を含む英語記事数 (document frequency, df) が多いような訳語

1 http://www.csse.monash.edu.au/~jwb/edict.html

2 http://chasen.aist-nara.ac.jp/index.html.ja

3 ここで追加したエントリは，IPADIC に含まれていないもののみである．なお，たとえば，EDR 日英対訳辞書と EDICT となどで，個別に追加した辞書間における重複があったとしても，それらの重複を除去することはせずに追加している．

の主辞を優先する[4]ことにした．そして，このヒューリスティクスにより優先付けられた上位2個のみを変換結果として利用した．なお，dfが0であるような訳語の主辞は，最初から，候補に含めない．

4.2　英語記事からの日本語記事の検索

一旦，日本語記事が英単語集合に変換されてしまえば，あとは，通常の情報検索と同様にして，質問として与えられた英語記事に最も類似するような日本語記事(の英単語集合への変換結果)を検索することができる．そして，その日本語記事をもって対応記事とする．このときの英語記事と日本語記事の類似度としては，BM25 (Robertson and Walker 1994)を利用した．BM25は，情報検索に有用な尺度として知られており，TREC[5] (Text REtrieval Conference) やNTCIR[6] (NII-NACSIS Test Collection for IR Systems) でも，その有効性は実証されている．

ここで，質問である英語記事Qと日本語記事の変換結果Dとの類似度BM25(D, Q)は以下である．

$$\mathrm{BM25}(D,Q) = \sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

ただし，
TはQに含まれる単語(ターム)である．
$w^{(1)}$はTの重みであり，$w^{(1)} = \log \frac{(N-n+0.5)}{(n+0.5)}$．
Nは，検索対象の文書集合における全文書数である．ただし，検索対象は，質問である英語記事の日付の前後2日の範囲の日本語記事(の英単語への変換結果)である．
nは，Tを含む文書の数である．
$K = k_1((1-b) + b\frac{dl}{avdl})$である．ただし，$k_1$，$b$，$k_3$は経験的に定める定数であり，本研究では，$k_1 = 1$, $b = 1$, $k_3 = 1000$である．また，dlは，Dの長さであり，$avdl$は，文書集合における文書の長さの平均値である．ただし，文書の長さとは，その文書に含まれる単語の延べ数のことである．
tfは，Dに含まれるTの数である．
qtfは，Qに含まれるTの数である．

以上をまとめると，記事対応付けにおいては，日本語記事を英単語集合に変換し，その変換結果に対して，英語記事を質問として情報検索をし，その結果のBM25による類似度が1位の日本語記事を，英語記事の対応記事とする．この対応付けられた日英記事中にある日本語文と英語文との対応付けは，次節で述べる方法で行なう．

5　文対応付けの方法

日英記事における文間の対応はDPマッチングで求めた(Gale and Church 1993; Utsuro, Ikeda, Yamane, Matsumoto, and Nagao 1994)．DPマッチングで文対応を得るアルゴリズムの簡潔な記述には(Utsuro et al. 1994)を参照せよ．ここでは，日本語文(集合)から得られた内

4 たとえば,「今日」には「today」「nowadays」「this day」「present day」などが訳としてあるが，このうち，主辞だけをみると,「day」が一番多いので，まず，これを取る．次に，「nowadays」「today」の中から，dfが高いものを取る．
5 http://trec.nist.gov/
6 http://research.nii.ac.jp/~ntcadm/index-en.html

容語集合 J と英語文 (集合) から得られた内容語集合 E との類似度，$\mathrm{SIM}(J,E)$ についてのみ述べる．

$$\mathrm{SIM}(J,E) = \frac{\mathrm{co}(J \cap E) + 1}{|J| + |E| - 2\mathrm{co}(J \cap E) + 2}$$

である[7]．ただし，$\mathrm{f}(x)$ を文 X における x の頻度とすると $|X| = \sum_{x \in X} \mathrm{f}(x)$ である．また，$\mathrm{co}(J \cap E)$ は，J 中の単語と E 中の単語との 1 対 1 対応を，日英および英日対訳辞書に基づき求めた場合の集合を $J \cap E = \{(j,e) | j \in J, e \in E\}$ とすると，$\mathrm{co}(J \cap E) = \sum_{(j,e) \in J \cap E} \min(\mathrm{f}(j), \mathrm{f}(e))$ である．

J と E と $J \cap E$ とは，以下のようにして求めた．まず，辞書引きにあたって，日本語文については，茶筌により形態素解析をした結果から，内容語および複合語を抽出した．これが J である．また，英語文については，Brill's Tagger (Brill 1992) により品詞付けをし，基本形を WordNet [8] のライブラリを利用して求め，その結果から，内容語と複合語を抽出した．これが E である．次に，$J \cap E$ については，ある (j,e) の組 $(j \in J \wedge e \in E)$ について，もし，j の訳語に e があるか，e の訳語に j がある場合には，(j,e) には対応の可能性があるとし，そのような全ての対応の可能性のなかから，訳語の曖昧性の低いほうから 1 対 1 に対応付けていった．すなわち，(j,e) の曖昧性として，j の訳語の数を採用し，それの小さいものから対応付けをしていくのだが，既に，(j,e) のどちらかでもが選ばれている対応はスキップする，という方法を採用した．なお，このときの訳語の対応付けに用いた対訳辞書は，EDR 日英対訳辞書と EDR 英日対訳辞書を統合して生成した日英および英日対訳辞書である．これらの辞書において，日英方向のエントリ数は約 32 万，英日方向のエントリ数は約 37 万である．

以上のように定義された類似度を用いて，文対応を付けたが，このとき，文対応付けに用いたプログラムでは，DP マッチングにおける文間の対応としては，1 対 n もしくは n 対 1，ただし，$1 \leq n \leq 6$ しか許していない．

この条件下で，文対応プログラムの精度を，人手により文対応が付けられている，白書データ (日本電子工業振興協会 2000) に適用することにより求めた．白書データには，18 対の日英ファイルがあるが，そのうち，訳抜け (0 対 n もしくは n 対 0 の文対応) の数が 3 以下の 12 ファイルを対象とした．これらのファイル対について，日本語文の平均数は 413，英語文の平均数は 495 である．

このとき，再現率の平均は 0.982，適合率の平均は 0.986 である．これより，このプログラムの精度は十分に高いと言える．なお，

$$\text{再現率} = \frac{\text{プログラムの得た文対の中で正しい対の数}}{\text{正しい対の総数}}$$

7 単語集合同士の類似度については，その他にも様々なものが考えられるが，それらについて詳細な比較検討はしていないが，本節で述べる文対応付けの実験結果の精度からは，$\mathrm{SIM}(J,E)$ が文対応付けの類似度として妥当なものであると言える．

8 http://www.cogsci.princeton.edu/~wn/

$$\text{適合率} = \frac{\text{プログラムの得た文対の中で正しい対の数}}{\text{プログラムが推定した対の総数}}$$

ただし，1対nの文対応からは，n個の対が得られる．たとえば，文J_1と文E_1, E_2, E_3が対応しているとすると，得られる対は$(J_1, E_1), (J_1, E_2), (J_1, E_3)$の3個である．

我々は，辞書のみに基づいて文対応付けをした．それに対して，(Utsuro et al. 1994)は，辞書情報に統計情報を組合せることにより，文対応の精度が向上すると述べている．しかし，我々のプログラムの精度は既に十分に高いので，統計情報は利用しなかった．

6　記事対応尺度と文対応尺度

4節と5節とにおいて，記事対応の類似度BM25と文対応の類似度SIMとを導入した．しかしながら，これらの類似度のみを利用して記事対応や文対応を付けた場合には，7節や9節で実験で示すように，十分に精度の高い記事対応や文対応を得ることはできない．そのため，本節では，記事対応と文対応の双方について，新たな尺度を定義する．本研究の主要な貢献は，以下で述べる二つの尺度AVSIMとSntScoreとを提案し，その性能を実験により詳細に検討すると同時に，大規模な日英対応付けコーパスを構築し，それを一般に利用可能にした点である．

まず，記事対応についてであるが，我々は，4節において，日本語記事Jと英語記事Eの類似度としてBM25(J, E)を導入した．この類似度は，単語集合間の類似度であるので，文の順序などは考慮できない．そのため，文の順序を考慮できる記事対応尺度として，AVSIM(J, E)を定義する．これは，JとEとの文対応[9]を$\{(J_1, E_1), ..., (J_m, E_m)\}$としたとき，以下の式である．

$$\text{AVSIM}(J, E) = \frac{\sum_{k=1}^{m} \text{SIM}(J_k, E_k)}{m}$$

AVSIMが高い値となるのは，個々の文対応の類似度SIMが高い場合であるので，そのような場合には，記事としての対応も良いと考えた．

次に文対応の良さの尺度について述べる．5節で述べたように，我々の文対応付けプログラムの精度は，白書データのように日本語文と英語文とが原文と訳文という関係にあるようなものを対応付ける限りにおいては，高精度である．しかし，2節で述べたように，日本語記事と英語記事との関係は，一般には，原文と訳文という関係ではない．そのため，5節の方法で文対応付けをした場合には，適切な対応と共に不適切な対応も多く得られる．そのようにノイズの多い状況から，適切な対応のみを抽出するためには，文対応の尺度として，文類似度だけでなく，記事対応の尺度も利用すれば良いと考えた．そのため，日本語記事Jと英語記事Eとの記事対応における，文J_kとE_kとの文対応尺度として，

$$\text{SntScore}(J_k, E_k) = \text{AVSIM}(J, E) \times \text{SIM}(J_k, E_k)$$

9 これらは，もし対応が1対nである場合には，文と文集合との対応となるが，そのような場合も含めて文対応と呼ぶ．

を定義した．この尺度は，同一記事対応内で文対応を比べる場合には文類似度SIMと同じ順位を与えるが，異なる記事間での文対応の比較では，文類似度だけでなく，記事対応の尺度値も高いような文対応を優先する．

7　記事対応付けの精度

7.1　無作為抽出による精度評価

記事対応付けは，各英語記事との類似度BM25が高い日本語記事を検索することによりなされる．このとき，類似度1位の日本語記事についての記事対応付けの精度を1996-2001と1989-1996とについて表1に示す[10]．

表 1　類似度1位の記事対応の精度

評価値	1996-2001			1989-1996		
	下限	割合	上限	下限	割合	上限
A	0.49	0.59	0.69	0.20	0.29	0.38
B	0.06	0.12	0.18	0.08	0.15	0.22
C	0.03	0.08	0.13	0.03	0.08	0.13
D	0.13	0.21	0.29	0.38	0.48	0.58

表1において,「評価値」とは，記事対応の良さの人手による判定の評価値であり，その基準は，Aは「記事全体の記述の5〜6割程度以上について意味の対応がとれる」，Bは「2〜3割程度以上5〜6割程度以下について意味の対応がとれる」，Dは「全然違う」，Cは「A,B,D以外」である[11].「割合」とは，1996-2001 と 1989-1996 のそれぞれから，100記事対応ずつを一様無作為抽出したときに，その評価値であった記事対応の割合である.「下限」「上限」とは，割合の95%信頼区間の下限と上限である．

2節で述べたように，1996-2001については,「本紙翻訳=Y」なる英語記事のみを対象したが，1989-1996については，全英語記事を対象とした．そのため，1989-1996の精度は，1996-2001よりも低い．また，1996-2001の精度が1989-1996の精度よりも高いといっても，それでも，評価値Aが約60%，AもしくはBが約70%であるので，BM25による記事対応付けの結果をそのまま利用した場合には，ノイズとなる記事対応が多すぎる．

我々の観察によれば，評価値がAもしくはBの記事対応は，そこから日英言語表現間の対応

10 評価を記述する際には1996-2001をメインとする．その理由は，今後とも The Daily Yomiuri には「本紙翻訳=Y/N」の情報が付くと考えられるので，1996-2001の精度評価の方が相対的に重要と考えられるからである．

11 A,B,C,Dの判定は第1著者がした．判定については，1996-2001については，ダブルチェックをした．1996-2001についての初回の判定における各評価値の割合は，A=0.62, B=0.09, C=0.09, D=0.20である．したがって，同一評価者内においては判定結果は安定していると言える．なお，1996-2001については，更に，類似度1 位の評価値がCかDの場合には，10位以内までを見て，A,Bがないかを探した場合の各評価値の割合は，A=0.62, B=0.15, C=0.05, D=0.18であるので，類似度1位のものとそれほど違わない．そのため，1989-1996については，類似度1位のもののみしか判定しなかった．

が抽出できそうという意味において，有用な記事対応である．このような記事対応のみを抽出するには，BM25による記事対応付けの結果をそのまま全て利用するのではなく，対応の良さにより対応付けの結果をソートし，その上位のみを抽出すれば良い．

7.2　ソートした場合の記事対応の精度

記事対応の良さの指標として，AVSIMとBM25のどちらが適当かを比較した．表1と同じデータに対して，それぞれの値の降順により記事対応をソートし，評価値がAもしくはBの場合を正解とし，各順位までにおける正解の個数とその割合とを調べた．それを表2に示す．表2から，我々は，AVSIMの方がBM25よりも，記事対応の良さとして適切な尺度であると判断した．

表 2　順位と精度

順位	1996-2001				1989-1996			
	AVSIM		BM25		AVSIM		BM25	
	数	割合	数	割合	数	割合	数	割合
5	5	1.00	5	1.00	5	1.00	2	0.40
10	10	1.00	8	0.80	10	1.00	4	0.40
20	20	1.00	16	0.80	19	0.95	9	0.45
30	30	1.00	25	0.83	28	0.93	16	0.53
40	40	1.00	34	0.85	34	0.85	24	0.60
50	50	1.00	39	0.78	37	0.74	28	0.56
60	60	1.00	47	0.78	42	0.70	30	0.50
70	66	0.94	55	0.79	42	0.60	35	0.50
80	70	0.88	62	0.78	43	0.54	38	0.47
90	71	0.79	68	0.76	43	0.48	40	0.44
100	71	0.71	71	0.71	44	0.44	44	0.44

AVSIMの精度の方がBM25の精度よりも高い理由は，6節で述べたように，AVSIMが，BM25と違って，個々の文対応の良さまでも考慮した尺度であるからと考える．

7.3　評価値とAVSIM

人手により判定された評価値A,B,C,DとAVSIMとの対応の程度を調べることを目的とし，表1と同じデータに対して，各評価値となった記事対応について，AVSIMの統計量を求めた．それらを，1996-2001については表3に，1989-1996については表4に示す．

表 3　AVSIMの統計量(1996-2001)

評価値	数	下限	平均	上限	閾値	有意差
A	59	0.176	0.193	0.209	0.168	**
B	12	0.122	0.151	0.179	0.111	**
C	8	0.077	0.094	0.110	0.085	*
D	21	0.065	0.075	0.086		

表 4　AVSIMの統計量(1989-1996)

評価値	数	下限	平均	上限	閾値	有意差
A	29	0.153	0.175	0.197	0.157	*
B	15	0.113	0.141	0.169	0.131	
C	8	0.092	0.123	0.154	0.097	**
D	48	0.076	0.082	0.088		

これらの表において,「数」とは，その評価値であった記事対応の数である．また,「平均」とは，そのような記事対応のAVSIM の平均値であり,「下限」および「上限」は，平均値の95%信頼区間の下限と上限である.「閾値」は，その評価値であるような記事対応と，次の評価値であるような記事対応とを分けるときに，どのAVSIMで区切れば良いかを示す．たとえば，表3では，A判定とB判定とはAVSIMの値が0.168により分かれる．この閾値は，線形判別分析により求めた値である．また,「有意差」の欄にある「**」と「*」は，それぞれ，その評価値と次の評価値とで平均値に差があるかを，Welch検定により片側検定したときに，その差が，1% と5%水準で有意であることを示す.

二つの表において，1989-1996のBとCとの区分を除いては，全ての評価値において，各評価値と次の評価値とでは，平均値に有意な差があることがわかる．このことから，AVSIMは，各評価値を十分に明確に区切ることができると言える．なお，1989-1996では，BとCが分かれていないことについて，その理由を調べた．そうすると，実際，1989-1996 では，C だといっても，記述の重複が，1996-2001 の C と比べて，多いものが多かった．定性的には，1996-2001 の C は,「D ではない(全然違うわけではない)」という意味で C であり，1989-1996 の C は,「B かもしれない」という意味で C であった.

次に，1996-2001と1989-1996とで，同じ評価値を与えられた記事対応のAVSIMの平均値に統計的に有意な差があるかを調べた．つまり，たとえば，表3では，評価値Aの平均値は0.193であり，表4では，0.175であるが，この二つの平均値の差が統計的に有意かどうかを両側検定によるWelch検定により調べたところ，有意水準5%においては，A,B,C,Dいずれの評価値においても有意差はみられなかった．そのため，1996-2001と1989-1996とで，同じ評価値の記事対応は，同じ程度のAVSIMであると判断した．そのため，AVSIMは，異なる記事集合を利用した場合であっても，安定して，記事対応の良さを示す指標であると考える.

表 5　評価値と記事数の推定

	1996-2001	1989-1996	計
A	15491	16004	31495
B	9244	5999	15243
C	4944	10258	15202
D	5639	26825	32464
計	35318	59086	94404

最後に，表3と表4にある閾値[12]に基づいて，1989-1996と1996-2001とについて，A,B,C,Dであるような記事数を推定した結果を表5に示す．表より，評価値がAもしくはBと推定される記事対応は，全体では，46738 (= 31495 + 15243)だけある．我々は，約4万7千という記事対応は，訳語抽出などの自然言語処理への応用や，英語教育などへの応用にとって，十分有効に利用できる量であると考える．

以上より，AVSIMは，人手による評価値A,B,C,Dに良く対応した尺度であり，かつ，異なる記事集合においても同一評価値については安定した数値をとる尺度であることがわかった．また，AVSIMに基づいて記事対応を抽出することにより，約4万7千の良質な記事対応が抽出できることが期待できることがわかった．

8　記事対応付けの精度向上の可能性

7節で述べたように，AVSIMは，記事対応の良さを示す信頼性の高い尺度である．そのため，BM25の代り(もしくは重みつき和などによる組み合わせで)，最初からAVSIMを利用して記事対応を求めれば，7.1節で述べた全体的な精度も向上すると考えられる．

しかし，我々は，現時点では，BM25による類似度1位の記事対応についてのみしか，AVSIMを求めていない．その理由は，10位以内などの比較的少しの記事をみただけでは記事対応精度に顕著な向上がないからであり，かつ，現時点での文対応プログラムの実行速度が遅いからである．今の文対応プログラムでは，一記事あたりの対応を取るために，数秒は掛る[13]．そのため，一位同士の対応についてAVSIMを得るだけでも，9万4千記事程度なので，数日間は掛かる．したがって，たとえば，100位以内をみるだけでも，数100日間掛かることになる．これは非現実的である．

しかし，今後，もっと高速の文対応プログラムを作り，それを利用することにより，より高精度な記事対応が得られるものと考えている．

また，今は，各英語記事について，その記事の日付の前後2日の範囲しか調べていないが，記事によっては，5日前のものが翻訳されているものがあった．このようなものまでカバーするためには，もっと広い範囲から対応候補記事を集める必要がある．

この2点は，システム全体を効率化しスケールアップすることにより達成可能なので，将来的には実現したい．

12 表3での閾値の丸めていない値は，0.1681076526, 0.111106681, 0.08531399165であり，表4では， 0.1566618237, 0.130510963, 0.09692189387である．これらの値を実際には利用した．
13 プログラムの動作環境は，CPUは Pentium-4 1500MHz，OSはRed Hat Linux 7.1である．

9　文対応付けの精度

2節で述べたように，たとえ，日英記事間に内容上の対応があったとしても，文間対応があるとは限らないので，対応付けられた記事から得られる文対応はノイズが多いものとなる．そのため，BM25による類似度1位の記事対応全てから得られる文対応全てをSntScoreにより降順にソートし，その上位のみを利用することにより対応の良いものを抽出することにした[14]．

このような文対応の数は，1989-1996と1996-2001を合せた全体で，約130万だけある．なお，ここでの文とは，日本語文については，簡単なプログラムにより，句点などで日本語記事を分割した結果であり，英語文については，MXTERMINATOR (Reynar and Ratnaparkhi 1997)に対して前処理と後処理を適用して英語記事を分割した結果である．

文対応のなかでは，1対1対応が最も重要である．また，文対応といっても，新聞記事には，中見出しなどの，必ずしも文でないものもある．そのため，1対1対応のなかで，文末が句点やピリオドなどで終っているもののみを取り出し，これを特に「1:1」と呼び，その他の対応を「1:n」と呼ぶことにする．1:1の数は，約64万ある．1:nの数は，約66 万ある．

1:1の精度を求めるために，SntScoreにより降順にソートされた上位30万対応について，3万対応ごとに100ずつを一様無作為抽出した．この各対応について，x/o の2値評価をした[15]．ここで，x は「意味が全然違う」であり，oは「意味が全然違うことはない」である．その結果の x/o の数を表6に示す．

表 6　順位と1:1の精度

範囲	o数	x数
1 -	100	0
30001 -	99	1
60001 -	99	1
90001 -	97	3
120001 -	96	4
150001 -	92	8
180001 -	82	18
210001 -	74	26
240001 -	47	53
270001 -	30	70

表から分かるように，順位が下っていくにつれて，x の数が指数的に増加している．このこ

14 文対応精度評価は，1989-1996 と 1996-2001 とを分けずに行なう．その理由は以下の2 点である．(1) まず，作成するコーパスでは，1989-2001 全体から選んだ文対応のなかから良く対応していそうなもののみを抽出したい．そのためには，全体を評価した方が良い．(2) 記事対応の精度評価の結果から，同程度のAVSIMは，1989-1996 と 1996-2001 とで同じ評価値に対応するので，SntScoreも1989-1996と1996-2001とで分ける必要はないと考えられる．

15 評価は第1著者がした．ダブルチェックによると，初回の判定と2回目の判定とで100個あたり多くて2,3個程度のo/xの違いがあった．したがって，同一評価者内においては判定結果は安定していると言える．なお，1:n についてはダブルチェックはしていない．

とは，SntScoreが，効率良く，適切な1:1を上位に順位付けていることを示している．

表6から，15万対までは十分に信頼できる対応であると言える．なお，15万対までの o の累積の割合は 0.982 である．

表 7　順位と 1:n の精度

範囲	1:n の数	o 数	x 数
1 -	38090	98	2
90001 -	59228	87	13
180001 -	71711	61	39

次に，1:nの精度を求めるために，SntScoreにより降順にソートされた上位について，表6の「1-90000」「90001-180000」「180001-270000」の各範囲について，それらの1:1 のSntScoreの範囲に収まるような1:nの精度を求めた．精度を求めるときには，1:1のときと同様に，各範囲から100対を一様無作為抽出し，x/o の2値評価をした．その結果を表7に示す．表より，「1-90000」の範囲の38090個の1:nについては，精度の良い対応であると言える．

以上述べたように，SntScoreにより文対応をソートすることにより，1:1と1:nの双方について，上位には，十分に精度の高い文対応が得られる．次に，SIMについて，SntScoreとの比較のため，その精度を述べる．比較にあたっては，SIMの降順でソートした上位における精度を調べ，その精度により比較する．

まず，1:1についてであるが，SIMの降順により1:1をソートした場合の上位15万対から100対を一様無作為抽出してo/xの判定をした結果は，o数=93・x数=7であった．これを，表6に示される，SntScoreにおける上位15万対における無作為抽出500対でのo数=491・x数=9と比べると，比率の差の検定を片側検定ですると，有意水準1% (実際には0.16%)でSntScoreの方が有意にoの比率が高い．

次に，1:nについてであるが，1:1のときと同様に，1:nをSIMの降順にソートし，上位38090対から100対を一様無作為抽出してo/xの判定をした結果は，o数=89・x数=11であった．これを，表7に示されるSntScoreにおける上位38090対における無作為抽出100対でのo数=98・x数=2と比べると，比率の差の検定を片側検定ですると，有意水準1% (実際には0.49%)でSntScoreの方が有意にoの比率が高い．

これらより，1:1と1:nの双方について，SntScoreの方が，有用な尺度であると言える．SntScoreの精度の方がSIMの精度よりも高い理由は，6節で述べたように，SntScoreが，SIMと違って，記事対応の良さまでも考慮した尺度であるからと考える．

10　関連研究

自動的に記事対応を得ることを目的とする研究はいくつかある．そのうち，(Collier, Hirakawa, and Kumano 1998)は，言語横断検索に機械翻訳を利用した場合と辞書引きを利用した場合とを比較しており，再現率が高いとき(多くの記事対応を得たいとき)には，辞書引きの方が有利だとしている．我々も，表1のデータの1996-2001についてのみ，シャープ株式会社の機械翻訳支援システムを利用して精度評価をしてみたが，その結果は，統計的に有意ではないが，辞書引きの結果の精度の方が高かった[16]．これらのことから，辞書引きの方が記事対応を得るには適しているのではないかと考えられる．

また，(Matsumoto and Tanaka 2002)は，日経産業新聞について，英語記事と日本語記事との対応付けをしていて，その精度は，97%と非常に高精度である．しかし，彼らは，同じ方法を，NHKの報道記事の対応付けに対しても適用しているが，その場合の精度は69.8%であり，彼らの方法が，全ての場合で高精度であるわけではないということも示している．そのため，彼らの方法を読売新聞の記事対応付けに利用した場合にも同様に高い精度が得られるかは明かではない．

これらの従来の記事対応を得る研究と我々の研究との主要な違いは次の2点である．

(1)　まず，記事対応の評価尺度について，我々は，DPマッチングによる文対応付けの結果を利用した信頼性の高い尺度を提案した．それに対して従来の研究は bag-of-words に基づいた尺度を利用している．なお，情報検索において，質問文と文書との類似度を求める際にDPマッチングを利用する方法が(Yamamoto, Yamamoto, Umemura, and Church 2000)により提案されているが，彼らの研究対象と我々の研究対象とは異なるし，かつ，DPマッチングの方法や，評価尺度の定義も異なる．

(2)　次に，我々は，記事対応の結果から文対応までを，実際に，大規模に得た．(高橋, 松尾, 古瀬 1999)は，記事対応の結果から文対応を得ることを構想してはいるが，実際に文対応を得ているわけではない．加えて，我々は，対応付けの結果が一般に研究および教育目的に利用できるようにしているが，これは日英対応付けコーパスとしては初めての試みである．

11　データ公開

我々は，本稿で述べた日英新聞記事対応付けの結果を数値情報としてエンコーディングすることにより，読売新聞と The Daily Yomiuri の記事データを持っている場合には，対応付けの結果が復元できるデータを作った．また，9節で述べた文対応について，日英それぞれの文末が句

16 ただし，このときには，英語記事を日本語に翻訳し，その翻訳結果を質問として日本語記事からなるデータベースを検索した．これは，本稿でこれまで説明してきた方法である，日本語記事を英語単語集合に変換する方法の逆であるので，厳密な比較ではない．

点やピリオドなどで終了しているものについて，1:1の上位15万対と1:nの上位3万対とを，読売新聞社からの許可を得て，生の文として，上記数値データに追加したデータを試験的に公開した．

公開した期間は2002年10月23日から2002年11月22日の1ヶ月間であり，公開の情報は，言語処理学会のメイリングリストを通じて流した．その結果，31の機関からデータ入手の申し込みを受けた[17]．

それら機関の内訳[18]は，国内が28，国外が3であった．また，企業からの申し込みは4件あり，そのほかの27件は中学・高校・大学もしくは研究機関であった．また，自然言語処理関係の研究機関から15件，その他の機関からは16件であった．それら16件は，言語学関係が13件，中学・高校が2件，また，民間企業で翻訳業務をしている企業からの申し込みが1件あった．

これらの内訳から，このような日英対応付けデータが，自然言語処理の研究機関だけでなく，言語学や中学・高校の英語教育などに関わる人にとっても関心の高いものであることがわかる．

12　今後の課題

本稿で述べた対応付けは，3節で述べた方針に基づいている．すなわち，既存の言語資源や手法を用いて対応付けをして，その結果から，なるべく対応の良さそうなもののみを抽出するというものである．

その結果，7節や9節で述べたように，上位にソートされた記事対応や文対応については，十分に精度の高い対応が得られた．しかし，ソートの適用対象は，4節や5節の方法で求められた記事対応や文対応であるので，どんなに精度良くソートしたとしても，最初に求められた記事対応や文対応に含まれているよりも多くの正解対応を得ることはできない．

たとえば，記事対応では，BM25により検索された記事対応しか対象としていないため，BM25の検索精度により，抽出できる記事対応の精度は制限される．この記事対応付けについては，8 節で，AVSIMを用いることにより，精度が向上すると考えられると述べた．これと同様に，文対応においても，新聞記事に適した文対応付けアルゴリズムを用いることにより，抽出できる，正解である文対応の数が増えるものと考える．そのようなアルゴリズムを考案し，より多くの正解対応を求めることが今後の課題である．

13　おわりに

ノイズの多い日英新聞記事集合から，内容が対応した記事対応と文対応を得るための信頼性の高い尺度を提案した．それら尺度を用いることにより，1989年から2001年までの読売新聞と

17 今後の配布については第1著者まで問合せのこと．
18 これら内訳は機関名などから推測したものである．

The Daily Yomiuriとから記事対応と文対応を得た．それらのなかで，比較的良質と推定されるものが，記事対応は約4万7千あり，文対応は，1対1対応が約15万あり，1対1対応以外が約3万8千ある．これらは，現時点で一般に利用できる日英2言語コーパスとしては最大のものである．

我々は，今後，この日英対応付けコーパスを，より良質にしていくとともに，このコーパスを実際の応用に利用することを考えている．

参考文献

Brill, E. (1992). "A Simple Rule-Based Part of Speech Tagger." In *ANLP-92*, pp. 152–155.

Collier, N., Hirakawa, H., and Kumano, A. (1998). "Machine Translation vs. Dictionary Term Translation – a Comparison for English-Japanese News Article Alignment." In *COLING-ACL'98*, pp. 263–267.

Gale, W. A. and Church, K. W. (1993). "A Program for Aligning Sentences in Bilingual Corpora." *Computational Linguistics*, **19** (1), 75–102.

Matsumoto, K. and Tanaka, H. (2002). "Automatic Alignment of Japanese and English Newspaper Articles using an MT System and a Bilingual Company Name Dictionary." In *LREC-2002*, pp. 480–484.

Reynar, J. C. and Ratnaparkhi, A. (1997). "A Maximum Entropy Approach to Identifying Sentence Boundaries." In *ANLP-97*.

Robertson, S. E. and Walker, S. (1994). "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval." In *SIGIR'94*, pp. 232–241.

Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y., and Nagao, M. (1994). "Bilingual Text Matching using Bilingual Dictionary and Statistics." In *COLING'94*, pp. 1076–1082.

Yamamoto, E., Yamamoto, M., Umemura, K., and Church, K. W. (2000). "Dynamic Programming: A Method for Taking Advantage of Technical Terminology in Japanese Documents." In *IRAL-2000*, pp. 125–132.

高橋大和, 松尾義博, 古瀬蔵 (1999). "新聞記事における日英対応コーパスの自動構築." 言語処理学会第5回年次大会発表論文集, pp. 181–184.

日本電子工業振興協会 (2000). 自然言語処理システムに関する調査報告書.

付録

9節の実験より，比較的良質と推定される文対応は，1対1対応が約15万あり，1対1対応以外が約3万8千あることがわかった．更に，そこから，日英それぞれの文末が句点やピリオドな

どで終了しているものについて，1対1対応の上位15万対と1対1対応以外の上位3万対とを一般に公開していることを11節で述べた．

本付録では，この公開されている部分の文対応のサンプルと，公開されてはいないが，公開されている部分と公開されていない部分との境界付近にあるサンプルとを示す．公開されている部分のサンプルを示すことにより，どのような文対応が比較的良質とされているかの目安がつき，境界付近にあるサンプルを示すことにより，公開されている部分の文対応の最低品質の目安がつく．なお，2節で例示している日英対応を含む記事対応から得られる文対応は，1対1対応についても1対1対応以外についても，公開されている部分には入っていない．このことは，SntScoreを利用して文対応をソートすることにより，2節で示したような直訳とはいえないような文対応は，下位に位置付けられることを例証している．

公開されている部分の文対応のサンプル

1対1対応

1対1対応の上位15万対から無作為抽出された5対を以下に示す．

1. こうした現状を知った面川委員長らが、タイの気候、風土に合った伝統的な農業を取り戻す拠点となる農業学校の設置を計画。
To help these poorer farmers, the Kakuda agricultural association plans to set up an agricultural school in one of the northeast's villages, where it will serve as a center for the preservation of Thailand's traditional farming methods, assocation officials said.

2. 火薬庫と言われた住専の処理が動き出したことへの安心感も大きい。
The steps taken to deal with nonperforming loans left by jusen companies brought about a general sense of relief because these loans could have proved to be a powder keg ready to blow up in our faces.

3. 従って、不用意に競争を抑制すべきでないことも確かである。
Therefore, it is certain that they should not carelessly restrain competition.

4. WTOの農業協定は、食糧安全保障や環境保護などにも配慮しつつ、漸進的な自由化推進をうたっている。
Existing agricultural agreements reached by the WTO call for "gradual" promotion of liberalization, taking into consideration food security and environmental protection.

5. 管理部門の他の社員も同様のカードを支給されていたが、同容疑者のカード使用額は社内最高だったという。
The section chief spent more than any other Ajinomoto employee given a credit card for entertainment purposes, they added.

1対1対応以外

1対1対応以外の上位3万対から無作為抽出された5対を以下に示す．

1. だが、体当たり攻撃で旧秩序の拠点を打破せざるを得なかった政治・経済変革の第一段階は過ぎた。
今や創造の時代が始まった。
この時代が政治家に要求するのは、各種政治勢力と駆け引きを行い、妥協と合意を模索する手腕であり、相手の見解も考慮に入れることのできる能力である。
Now that the first phase of political and economic changes–the frontal attack against the bastions of the old order–is over, the time has come for constructive work, for mastering the skills of political maneuvering, compromise and agreement with various political forces, and the ability to take into account the viewpoint of one's

opponents.

2. 一因とされるのが、二十メートルにつき一メートル下がる砂浜の傾斜だ。
千葉市が手がけた「いなげの浜」が五十メートルにつき一メートル下がる構造なのに比べて、倍以上の急傾斜。
Experts say one of the main reasons for the erosion is the beach's relatively steep incline–one meter in 20, more than double the 1:50 incline at Inage-no-Hama beach, which the Chiba municipal government constructed.

3. 国際的行事などで不統一のままでいいのか。
現状を精査した上で、音楽上の議論を深めたい◆作・編曲家内藤孝敏さんら音楽家グループの提言だ。
A group of musicians, including composer and arranger Takatoshi Naito, thought it would be better for only one version of "Kimigayo" to be played at international events, so it proposed an in-depth study of the music.

4. 課税の減免などの特例は設けず一律適用し、ベンチャー助成は別途の政策などで手当てすべきではないか。
Instead, it should be applied on a uniform basis.
Measures for fostering venture businesses should be worked out as policy steps separate from the tax framework itself.

5. 国産米の供給比率は三月が五〇%、四月から六月までは三〇%だが、七、八月は六〇%程度まで上昇し、輸入米との比率が逆転しそうだ。
According to the agency's estimate, the ratio of domestic rice to total rice supplies is expected to rise to 60 percent in July and August.
This compares with a March ratio of 50 percent, and 30 percent for the months of April to June.

境界付近の文対応のサンプル

1対1対応

1対1対応の上位150001〜151000の間の1000対から無作為抽出された5対を以下に示す.

1. 落札総額は百三十六億千万円。
The joint venture's bids totaled 13.6 billion yen.

2. 燃えるゴミの中に金属、ガラス、プラスチックなど、焼却に適さないものが約一割混ざっている。
In fact, 10 percent of the combustible garbage collected contains metal, glass, plastics and other noncombustible articles.

3. 離れた工場の製品が神戸から輸出されていたのは、定期船の寄港が多く、納期が守りやすいほか、「できるだけ多くの荷物を積んでコンテナを有効利用する」（松下電器）などの理由があったためだ。
One of the reasons why products from such remote areas were exported via Kobe is that a vast number of shipping lines from around the world use the port.

4. しかし、積み替え作業などに一日約二百人の職員が必要で、時間がかかるほか、移動の際に市街地で渋滞に巻き込まれ、避難所への到着がしばしば遅れた。
But the delivery system was ineffective because of traffic congestion in quake-stricken areas, and about 200 ward employees spent their days trying to deliver food supplies.

5. ただ、政府米の一部をエサ米に充て、その分を買い入れる可能性はある。
"Under the rule, the government cannot buy rice harvested in 2000," an official of the Food Agency said.

1対1対応以外

1対1対応以外の上位30001〜31000の間の1000対から無作為抽出された5対を以下に示す.

1. ところが、九六米穀年度に入っての政府米売却量（輸入米も含む）は、六月までの八か月間で約三十五万トンと前年同期の半分以下に低迷している。

During the first eight months of the 1996 rice year, however, the agency sold only 350,000 tons of the regulated rice.
The sales were less than half of that attained during the same months in the previous year.

2. 村上氏が所属する江藤・亀井派の亀井政調会長は、国会近くのホテルで与党三党の政策責任者会議に出席していた。
LDP Policy Research Council Chairman Shizuka Kamei attended a meeting of chief policymakers of the three ruling parties in a hotel near the Diet building Thursday.
Murakami belonged to an LDP faction jointly led by Kamei and Takami Eto.

3. 経済状況の好転などを掲げ、選挙戦では三党体制による実績を強調した政府・与党だったが、有権者は、むしろ選挙目当てが明白な場当たり的な政策に対し、厳しい判断を下したと言ってよい。
On the hustings, the government and coalition parties trumpeted their achievements, such as a burgeoning economic recovery, in a bid to obtain public support.
Voters, however, harshly dismissed their campaign pledges as cosmetic–aimed only at winning the election.

4. 日数が短い都市はアジア地域とアメリカに集中し、最短のニューデリー（インド）が一・四日、次いで香港一・五日、台北（台湾）七・三日、ロサンゼルス（アメリカ）七・四日などとなっている。
A tendency toward shorter holidays was found in Asia and North America.
Workers in New Delhi planned the shortest holiday at an average of 1.4 days, followed by Hong Kong at 1.5 days, Taipei at 7.3 days and Los Angeles at 7.4 days.

5. イデオロギーの対立の終わりは、国家利益や民族感情の対立を表面化させるかもしれない。
What we may see is a world in which antagonism and brush-fire battles between smaller nations intensify and international conflict is fueled not by ideology but by national interest.
Our new world will be filled with promise and danger.

略歴

内山 将夫:　筑波大学第三学群情報学類卒業 (1992). 筑波大学大学院工学研究科博士課程修了 (1997). 博士 (工学). 信州大学工学部電気電子工学科助手 (1997). 郵政省通信総合研究所非常勤職員 (1999). 独立行政法人通信総合研究所任期付き研究員 (2001). 言語処理学会，情報処理学会，ACL，人工知能学会，日本音響学会，各会員.

井佐原 均:　1978 年京都大学工学部電気工学第二学科卒業．1980 年同大学院修士課程修了．博士（工学）．1980 年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所．現在，独立行政法人通信総合研究所けいはんな情報通信融合研究センター自然言語グループリーダー．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，各会員.