成蹊大学理工学研究報告
J. Fac. Sci. Tech., Seikei Univ.
Vol.47 No.2 (2010) pp.15-22

各種直交変換を用いた文章間の類似性判定

松澤　当[*1]，樋口　政和[*2]，村上　仁己[*3]

The Judgment of Document Similarities Using Various Orthogonal Transformations

Ataru MATSUZAWA[*1], Masakazu HIGUCHI[*2], Hitomi MURAKAMI[*3],

ABSTRACT：The objective of this work is to propose a new method to judge similarities between two documents by simple calculation using orthogonal transformations and show the performance for actual plagiarisms of sentences. Specifically, Japanese characters in document are represented by Shift-JIS code in computer and then transformed by an orthogonal transformation algorithms, such as Fourier, Cosine, Haar wavelet and Hadamard system. As a criterion of similarity judgments of documents, the correlation of these transformed coefficients are applied. In order to examine the performance of the method, such correlation calculations are performed for actual plagiarisms. As a result of the experiments it is confirmed that Fourier transformation is effective for documents that include various types of character modifications especially for plagiarisms that have shifted the order of the words.

Keywords：document similarity, orthogonal transformation, character code, plagiarism

（Received September 22, 2010）

１．序　論

近年，インターネット上では急激に情報量が増大し，「情報大爆発」が懸念されている。そのため重複した無駄な情報を見つけ除去することは重要である。また最近では，学生など多くの人によるレポートなどのコピー&ペーストによる不正も増えている。このような背景から，似通った文章を発見することの重要性はさらに増大している。本研究では掲示板などコンピュータ上の文章の類似性を計算により求める方法を提案し，各種類似した文章に対してシミュレーション実験を行い，提案方式の特性と有効性を調べた結果を報告する。

音声認識等，文字の研究の分野では形態素解析を用いる手法が多く行われている。形態素解析では，言語を品詞に分解し辞書にあるデータベースと比較することで解析が行われる。そのため形態素解析を用いる方法では，その特性は対象となる言語知識，辞書に依存する。我々は文献[1]と同様に，コンピュータ上の文字・文章を文字コードとして扱い，これら入力信号をいくつかの方式で直交変換し，文章の類似性を各種直交変換後領域での特徴抽出によって求める方法を考案した。「情報大爆発」を考慮し辞書や対象言語に依存せずに，シンプルに文字・文章の類似性を判定する提案手法の効果を実験的に測定した。また，類似性を判定するため，類似性の高い文章として盗作文をモデルとして用いた。

[*1]：理工学研究科理工学専攻修士学生
[*2]：情報科学博士研究員
[*3]：情報科学科教授 (hi-murakami@st.seikei.ne.jp)

２．文字コードについて

文字コードには複数種類があり，日本語を割り当てるものには Shift-JIS，EUC-JP，JIS，UTF-8 がある。その他アルファベットを割り当てるものには ASCII コード，多言語を扱うものとして Unicode がある。文字コードはコンピュータ上に整数で割り当てられ，半角文字では 1 バイト，全角文字では 2 バイト表現となる。文字コードにより異なるが，アルファベット，数字，カタカナは半角，全角両方の割り当てがあり，平仮名，漢字は全角文字に割り当てられる。それら整数値が割り当てられたデータ列を入力信号とし各種直交変換を行い，文字列の周波数特性を調べることで類似性を判定することを試みた。文字コードは，全角文字と半角文字の両方に対応し，日本文および英文に対し実験が行うことが可能なため，本研究に適していると考えられる Shift-JIS を用いた。

３．各種直交変換

本研究で文字コードのデータ列に対して行った変換は，各種ディジタル信号処理に用いられる代表的な，以下に示す離散フーリエ変換（DFT），離散コサイン変換（DCT），順序高速ハールウェーブレット変換（HWT），アダマール変換（HT）の 4 種類であり[2][3][4]，それらの変換式を以下に示す。

$$F_j = \sum_{k=0}^{n-1} f_k e^{-\frac{2\pi i}{n}jk} \quad (1)$$

$$Fc_j = \sum_{k=0}^{n-1} f_k \cos\{ \frac{\pi}{n}(k + \frac{1}{2}) j\} \quad (2)$$

式(1)，式(2)はそれぞれ離散フーリエ変換，離散コサイン変換式である。ここで，f_k は入力信号のデータ列，F_j，Fc_j はそれぞれ入力信号に対する離散フーリエ変換係数，離散コサイン変換係数を表す。n は入力データ数，i は虚数単位である。式(1)により変換された信号は複素数となるので，類似性比較ではその振幅スペクトルを計算した。

$$f'_j = \begin{cases} \dfrac{f_{2j} + f_{2j+1}}{2} \ (0 \le j \le \dfrac{N}{2} - 1) \\ \dfrac{f_{2j} - f_{2j+1}}{2} \ (\dfrac{N}{2} \le j \le N - 1) \end{cases} \quad (3)$$

式(3)はハールウェーブレット変換を表す。N は 2 の累乗のデータ数である。

$$F_h = \frac{1}{\sqrt{N}} Hf \quad (4)$$

式(4)にアダマール変換を示す。F_h はアダマール変換係数であり，f は入力信号である。式(4)では F_h および F を N 次元ベクトルとして計算した。H は $N \times N$ のアダマール行列であり，N は 2 の累乗である。$N = 4$ の場合のアダマール行列を式(5)に示す。

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad (5)$$

４．盗作文をモデルにした実験

4．1　盗作文の特徴

本実験では類似性の高い文章のモデルとして盗作文を用いた。次に盗作文の特徴を示す[5]。

オリジナルの文に，①0~5 文字程度の単語が差し替えられる（以下‘差し替え’，と称する），②文中の語が入れ替えられる（‘入れ替え’），また③0~5 文字程度の単語が差し込まれる（‘差し込み’），という特徴が代表的である。その他にも漢字や平仮名カタカナの変換，言い回しの違いなどが挙げられるが，本実験では盗作文に見られる代表的な特徴である①~③について調べる[6]。

4．2　盗作文をモデルにした実験方法

本研究では 4 つの変換を用い 5 つの異なる特徴を持つ文章に対して実験を行った。アダマール変換およびハールウェーブレット変換の入力データ数が 2 の累乗という条件，また実際の盗作文に近い文字数という理由から，用いる入力文字数は 128 文字とした。以下にここで実際に‘差し替え’に用いた 128 文字の文章例の 1 つを示す。

具体例：「文章間の類似性を判定するにあたって，辞書に依存せず判断することを試みる。対象とする文字はコンピュータ上の文字で，各文字には文字コードが割り当てられている。文字コードは整数で割り当てられていて，半角は 1 つ全角は 2 つの数字が割り当てられている。例文 1 作成終了，」

実験 1 では類似文を相対的な尺度で評価するため，盗作文でない自然な 2 つの文章（以下，非類似文と称する）を用いた。言い回しや語尾の異なるジャンルの文章（音楽・テレビ・スポーツなど）を 15 種類準備し，そこから文字コード化された二つを選び，全 105 通りの相互相関を計算する。ここでは，全角文字のみの場合，半角文字のみの場合，全角文字と半角文字が混在の 3 通りについてそれぞれ相互相関を計算した。

実験 2~4 では 2 つの類似文[5]に対し評価を行った。この類似文の例を，以下に示す。

類似文の例：

A: (オリジナル)　山川惣治は CBA の説に沿って，空飛ぶ円盤を以下のように考えていました。

(盗作)　山川自身は CBA の説に沿って，空飛ぶ円盤を以下のように考えていたらしい。

B: (オリジナル)　彼が初めて円盤を目撃したのはその 2 日後のことです。
(盗作)　しかし，その 2 日後，彼もやっと，円盤の目撃を実現させる。

具体的には，実験 2 では‘差し替え’，実験 3 では‘入れ替え’，実験 4 では‘差し込み’に対する実験である。入力データはオリジナル文章とそれを改変した文章であり，それらに対し先に述べた 4 種類の直交変換を行う。そして変換係数同士に対し相関係数を計算し評価を行う。日本語による盗作文を調べた結果，全角文字のみの文章が比較的多くみられたため，実験 2~4 では全角文字のみの入力データに対して実験を行った。

実験 2 ではオリジナルデータに 1~5 文字の範囲で異なる語を差し替え，オリジナルとの相互相関を計算する。その際，改変する 1~5 文字それぞれ文の左から挿入し，右へシフトさせる。図 1 に文字数 5，改変文字数 2 の場合の例を示す。図の右側の文章が改変したものである。

図 1　差し替えの例

実験 3 では，②入れ替え型の文章に対する実験を行った。オリジナルデータの 1~5 文字に対して入れ替えを行い，相関係数を求める。入れ替える方法としてオリジナルの文を先頭から順に文末の方へ入れ替える。図 2 に文字数 5，改変文字数 2 の場合の例を示す。図の右側の文章が改変したものである。

図 2　入れ替えの例

実験 4 として③差込み型文章の実験を行った。オリジナル文章に異なる語を差し込む際に，差し込んだ文字数だけ文全体がシフトする。そこに着目し，重複しないオリジナルの 128 文字の文章に対し自己相関をとった。この実験ではシフト回数が増すごとに比較する文章が類似していない文へと変化する。シフトする際比較する文字数は減少し，実験では文字数が 1 になるまで繰り返す。

図 3 に文字数 5 の場合の例を示す。

図 3 差し込みの例

５．実験結果

5. 1　実験 1（非類似文）の結果

実験 1 の結果を表 1~5 に示す。ここで用いた，非類似文 A，B（全角文字のみ），C（全角文字・半角文字混在）を以下に示す。

非類似文（全角文字のみ）の例：

A:「この後も，東北の日本海側は曇りや雨でしょう。雷が鳴ったり，雨脚の強まる恐れがあります。そのほかは晴れる所が多いですが，大気の状態が不安定となり，所々でにわか雨や雷雨がありそうです。局地的に激しく降ることもありますので，急な強い雨にご注意ください。最高気温は」

B:「同じ C 言語を使ったプログラムでも，コンソールアプリケーションと WINDOWS(GUI) アプリケーションでは，作り方が全く異なります。本書は, WINDOWS(GUI) アプリケーションの作り方を初歩の初歩から解説したものです。ただ，前提として C 言語そのものは知」

非類似文（全角文字・半角文字混在）の例：

C:「2010 サッカーFIFA ワールドカップ・テレビ放送予定。6/11-7/11，一ヶ月にわたってグループリーグ 48 試合，決勝トーナメント 16 試合，計 64 試合の激闘が繰り広げられる第 19 回ワールドカップ。地上波テレビ放送，BS デジタル放送，CS 放送（スカパー！）で生中継される全 64 試合の日程・」

この結果は類似文と非類似文を比較する指標となる。また，代表的な自然な文章の例として，全角文字と半角文字が混在している文字データについての結果を図 4~8 に示す。これらは 105 通りの各組み合わせに対して相関係数を求めたものをプロットした図である。

表 1　非類似文入力信号に対する無変換の
相関係数平均・標準偏差

	全角	半角	全角＋半角
平均	0.471	0.030	0.030
標準偏差	0.086	0.092	0.186

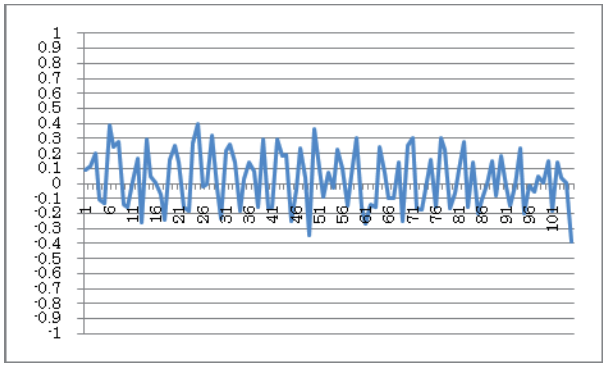

図 4　無変換における半角・全角文字混在文章の相関係数

一般的な文章の文字コードをそのまま入力信号とした場合の相関係数は図 4 のようになった。表 1 から全角および半角のみに比べ，全角・半角混在文章の標準偏差が高いことがわかる。

表 2　実験 1 の入力信号に対する DFT 係数の
相関係数平均・標準偏差

	全角	半角	全角＋半角
平均	0.890	0.995	0.751
標準偏差	0.052	0.004	0.076

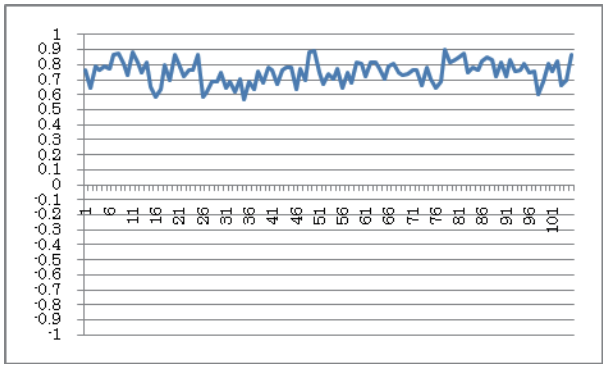

図 5　DFT における半角・全角文字混在文章の相関係数

表 1 と表 2 を比較すると DFT の周波数領域における相関係数は無変換に比べ高くなることがわかった。全角文字における非類似文章に対しては，約 0.9 程度の相関が示される。

表 1，2 から無変換に比べ DFT を施した相関係数の標準偏差がどの文章に対しても低いことがわかる。このことから文字コードに対し DFT を施すことで相関係数が無変換の場合より狭い範囲に集中することがわかった。

表 3　実験 1 の入力信号に対する DCT 係数の
相関係数平均・標準偏差

	全角	半角	全角＋半角
平均	0.788	0.988	0.548
標準偏差	0.055	0.009	0.118

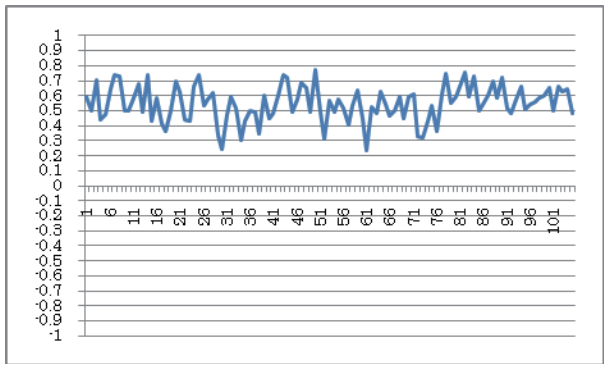

図 6　DCT における半角・全角文字混在文章の相関係数

表 2 と表 3 を比較すると DCT と DFT では比較的近い値をとる。更にその二つの変換の平均と標準偏差を比較すると，平均においては DCT の方が低く，一方標準偏差は DFT の方が低い。つまり DCT は DFT に比べ相関係数が低くなり，その変動範囲は広くなる。

表 4　実験 1 の入力信号に対する HWT 係数の
相関係数平均・標準偏差

	全角	半角	全角＋半角
平均	0.409	0.585	0.038
標準偏差	0.088	0.145	0.210

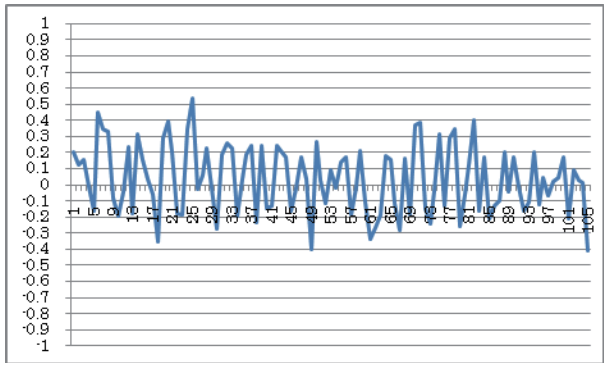

図 7　HWT における半角・全角文字混在文章の相関係数

表 1 と表 4 から HWT は無変換に比べ標準偏差が高くデータによりばらつきが大きく変動することがわかる。DFT および DCT と比べて平均値は低い値となり，無変換に近い値となった。

表 5　実験 1 の入力信号に対する HT 係数の
相関係数の平均・標準偏差

	全角	半角	全角＋半角
平均	0.701	0.977	0.391
標準偏差	0.067	0.017	0.139

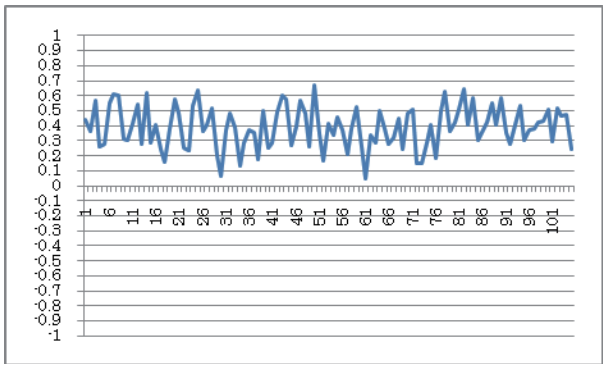

図 8　HT における半角・全角文字混在文章の相関係数

表 5 と表 3 および表 4 を比較すると HT は平均および標準偏差が DCT と HWT の間の値をとる。半角文字においては DFT，DCT と同様に 1 に近い高い値をとる。

5. 2　実験 2（①差し替え）の結果

実験 2 では各比較に対し $1 \leqq n \leqq 5$ の間で文字数を変化させて実験を行ったが，この程度の文字数の変化では結果に大きな差異は見られなかった。そこで代表的なものとして $n = 5$ の場合結果を表示する。グラフの横軸は改変した文字のシフト回数を示す。

図 9~13 より，128 文字に対し 5 文字程度の改変を施した文章に対する比較は，変換の有無にかかわらず 0.9 以上の高い相関を得ることがわかった。したがってこの程度の変化に対しては，直交変換ありと変換なしの間に大きな差はでないといえる。

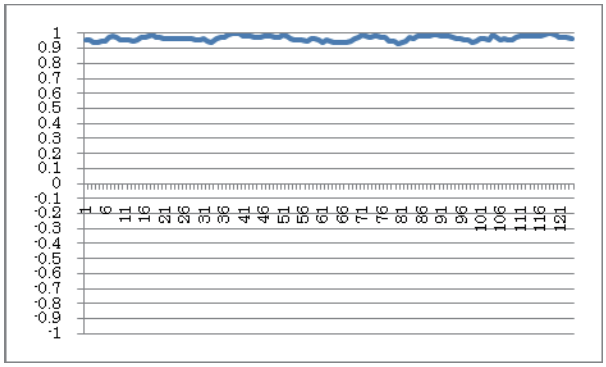

図 9　‘差し替え’文章の実験結果（無変換）

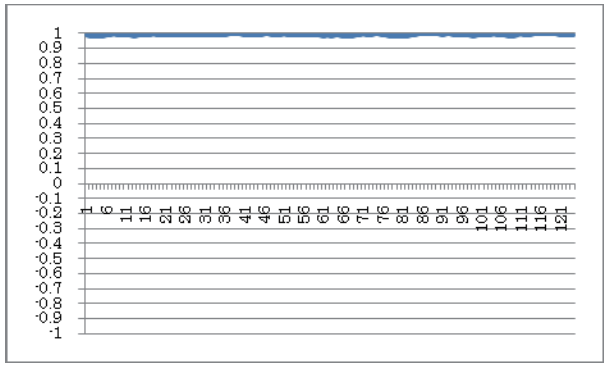

図 10　‘差し替え’文章の実験結果（DFT）

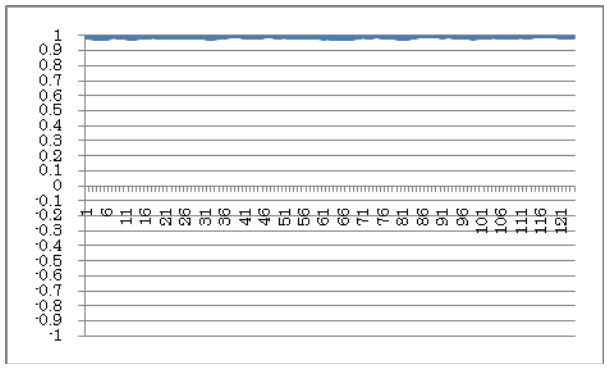

図 11　‘差し替え’文章の実験結果（DCT）

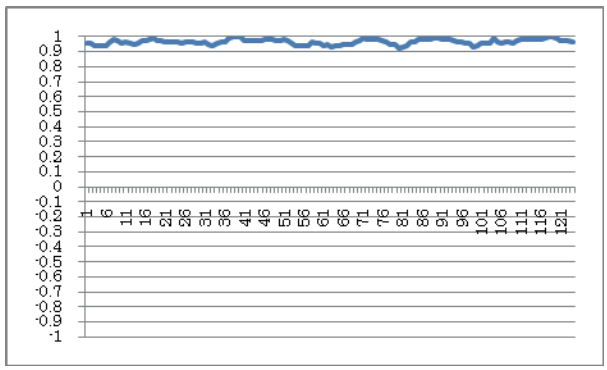

図 12　‘差し替え’文章の実験結果（HWT）

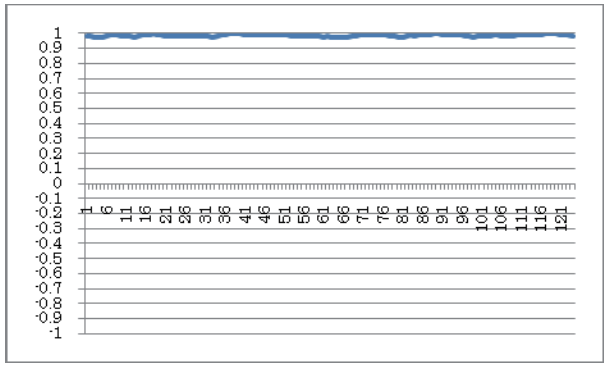

図 13　‘差し替え’文章の実験結果（HT）

5. 3　実験 3（②入れ替え）の結果

この実験結果も差し替え同様 n による差異は見られなかったため，n=5 の場合のみの結果を示す。図において，グラフの横軸は入れ替え文字のシフト回数を示す。

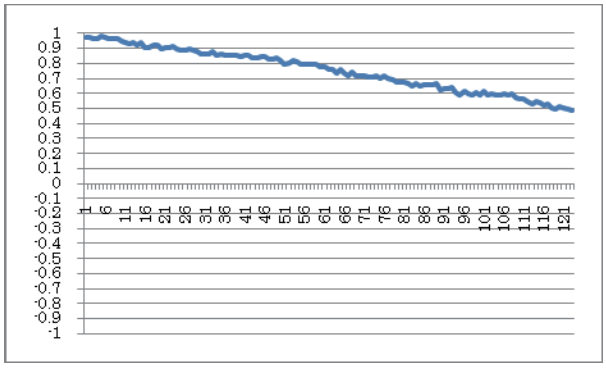

図 14　‘入れ替え’文章の実験結果（無変換）

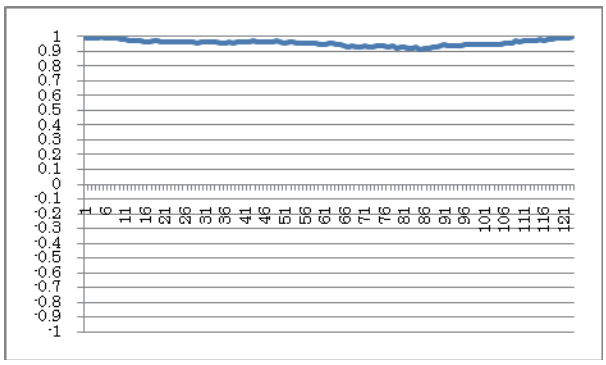

図 15　‘入れ替え’文章の実験結果（DFT）

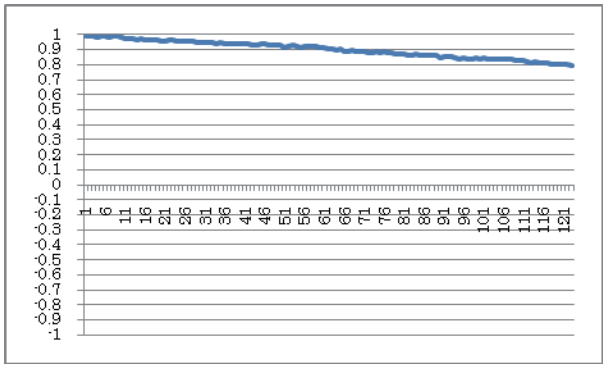

図 16　‘入れ替え’文章の実験結果（DCT）

図 14~18 をみると図 15 以外はシフト回数が増すごとに相関係数が減少する。無変換では 0.48, DCT では 0.79, HWT では 0.43，HT では 0.71 まで減少した。この値はそれぞれ表 1~5 の対応する変換における全角の平均にほぼ等しい。n = 5 程度なら文頭から区切った文字が文末へたどり着き，任意に二つに分割した文章が入れ替わった場合（例えば‘あいうえ’が‘うえあい’となる場合）

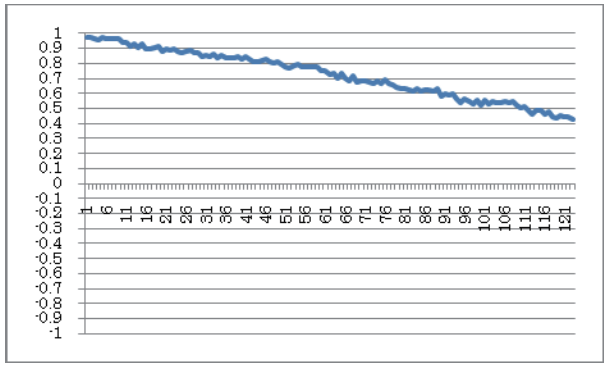

図 17　‘入れ替え’文章の実験結果（HWT）

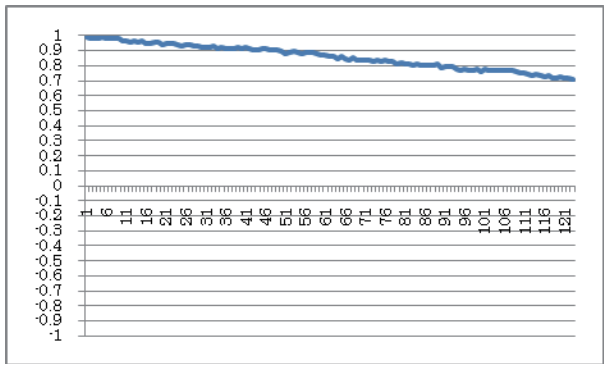

図 18　‘入れ替え’文章の実験結果（HT）

の相関係数は，非類似文章とほぼ一致する。

図 15（DFT）に着目すると，シフト回数が 80 回を超えたあたりからそれまで減少傾向にあった相関係数が増加し始めた。n = 1 ~ 5 どの場合においても最終的に相関係数は 1 となった。確認のため様々な文字数 n で試してみたところ，二つに分割した文章が入れ替わったとき相関係数は必ず 1 となることがわかった。

相関係数が最も低いときで 0.92 となり，表 2 の全角の平均値 0.89 を上回っている。

5. 4　実験 4（③差し込み）の結果

以下に差し込みの文章に対する結果を示す。差し込みの文章の類似性判定として自己相関を求めた。この実験ではデータをずらすたびデータ数が減少する。HWT と HT はデータ数が 2 のべき乗のときのみでしか計算できないため図 22, 23 のようなグラフとなった。図において，横軸は入力データのシフト回数を示す。

無変換，DFT，DCT についての結果である図 19~21 に着目すると，DFT 以外のデータは一回シフトし文章全体がずれたとき相関係数の減少の幅が大きいのに対し，DFT ではこの減少幅が少ない。今回の実験では DFT においてシフト回数が 52 のとき相関係数が 0.88 となり表 2 の全角の平均値を下まわった。

実験 4 においてシフト回数が増すたびグラフの振幅が増すことがわかるが，これはデータ数が減ったためわずかな文字コードの変化で計算結果の値が大きく変化するためと考えられる。

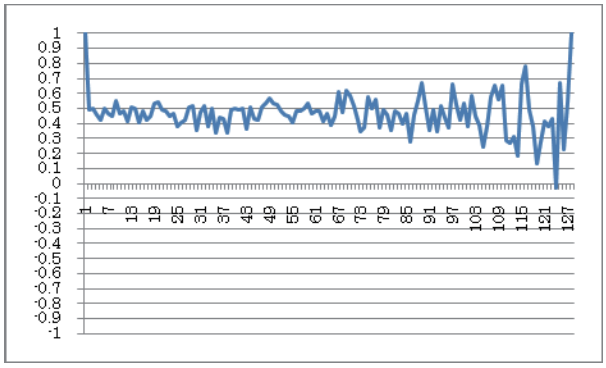

図 19　‘差し込み’文章の実験結果（無変換）

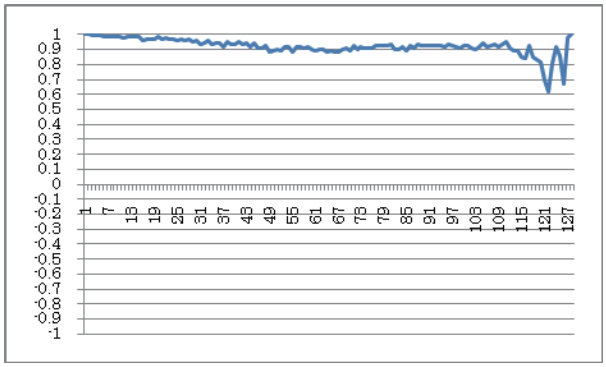

図 20　‘差し込み’文章の実験結果（DFT）

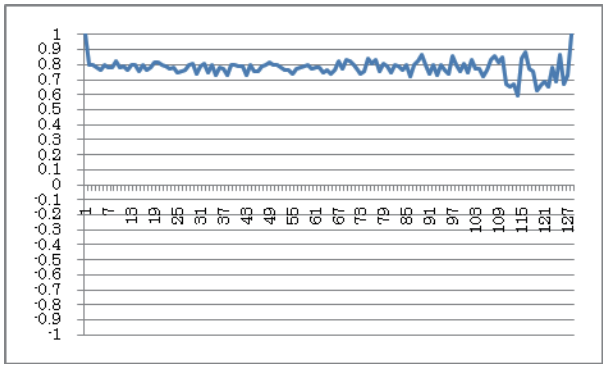

図 21　‘差し込み’文章の実験結果（DCT）

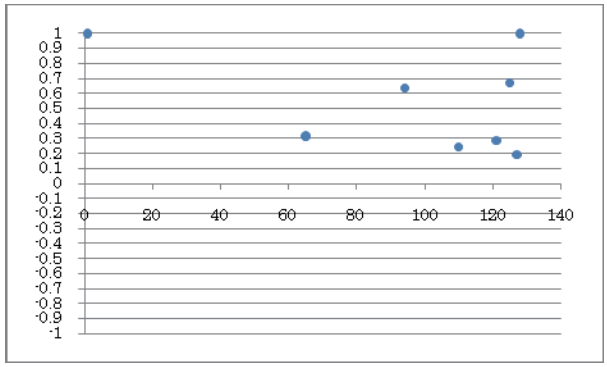

図 22　‘差し込み’文章の実験結果（HWT）

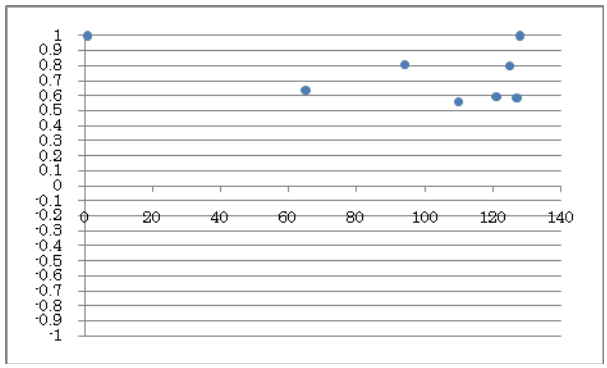

図 23　‘差し込み’文章の実験結果（HWT）

6．実際の盗作文に対する実験

この節では，提案した方式が実際の盗作文に対してどのような評価を得るか実験を行った。入力データは 4．1 節で紹介したオリジナル文章と盗作文の組であり，今回はこれらの例文を含む 3 組の入力データで実験を行った。

入力データ 1 は全角文字のみ（日本語のみ）の文章であり，盗作の特徴としてみられるのが①‘差し替え’の文章である。入力データ 2 は全角文字と半角文字（アルファベット）の混在の文章であり，盗作の特徴としては①‘差し替え’の文章である。さらにアルファベットの部分が全角から半角へ変換されていて変換された分ずれるため，③‘差し込み’の文章の特徴も持つ。入力データ 3 は全角文字と半角文字（アラビア数字）の混在の文章であり，盗作文の特徴としては②‘入れ替え’の文章である。実験結果を表 6 に示す。

表 6　盗作文に対する各種変換における相関係数

	無変換	DFT	DCT	HWT	HT
入力データ1	0.818	0.961	0.941	0.801	0.940
入力データ2	−0.224	0.929	0.336	−0.148	0.373
入力データ3	−0.273	0.859	0.510	0.087	0.588

表 6 における入力データ 1 は全角文字のみの比較のため，非類似文章と比較する際は実験 1 の結果である表 1〜5 の「全角」の部分に対応する。表 6 における入力データ 2 および入力データ 3 は全角文字と半角文字の両方を含む比較のため，非類似文章と比較する際は実験 1 の結果である表 1〜5 の「全角＋半角」の部分に対応する。

入力データ 1 における結果は，変換の有無に関わらず実験 1 の結果に比べ高い値となった。これは実験 2 の‘差し替’の実験と同様，改変部分が少なく，全て全角のため高い相関を得たと考えられる。入力データ 2 の結果において，フーリエ変換は実験 1 の結果に比べ，標準偏差のばらつきの範囲を上回る高い値を得た。入力データ 3 において，フーリエ変換とアダマール変換が実験 1 に比べ高い相関係数の値を得た。アダマール変換はばらつきの範囲を考えれば類似していない文章に当てはまる範囲であるが，フーリエ変換においてはばらつきの範囲を上回る高い値を得た。この実験においてもフーリエ変換の有効性を示すことができた。

7．考　察

ここで行った実験を纏めると，以下のように整理できる。

5.1 節において非類似文章に対する実験では，提案した手法の有効性を確かめる基準となる値を得た。

5.2 節において，改変する文字数が少ない場合は直交変換の種類に依存せず，高い値となる。ただしこれは文字数が 128 という条件の下での結果である。

5.3 節の‘入れ替え’の文章では，フーリエ変換が類似性判定に有効であることが示された。二つに分割した文字を入れ替えた場合 DFT が必ず 1 になったのは，フーリエ変換が信号を周期関数として扱うためであると考えられる。入力信号を任意の点で区切り入れ替えた場合，位相は変化するが振幅は変化しない。そのためオリジナルの文章と改変した文章の周波数が等しくなると考えられる。

5.4 節ではシフトした文章に対しフーリエ変換が有効であることが示された。この理由も 5.3 節同様，フーリエ変換が入力信号を周期関数として扱うためと考えられる。

8．まとめ

日本語文章に対する類似性判定の 1 手法として，形態素解析を用いない，文字データ列の直交変換係数での相関特性を用いる簡単な手法を提案，その特性を吟味した。これらの結果は，以下のように纏められる。

文字を入れ替えた文章，異なる文が差し込まれ全体的に文字がシフトした文章には，フーリエ変換が有効である。さらに実際の盗作文においても，その有効性を示すことができた。本研究では，主に全角文字に対する各種直交変換の比較を行った。今後は半角文字含む文章に対する特性比較を深めることが課題である。また，さらに多くの盗作文に対して評価を行うことも課題である。

今回は文字コードとして Shift-JIS を用いた。手法を変えずに類似文検索精度をあげるための一つの手段として，文字コードを変えることが考えられる。文字コードが一つの値で，更に文字コードの割り当てが現在よりも大きなばらつきを持てば，周波数としての特徴が既存の文字コードより大きくなると考えられる。このことから，今回の実験では文字コードを直接入力データとして使用したが，文字コードに何らかの前処理を施し有効性を高めることも視野に入れる必要がある。

謝　辞

この研究の 1 部は，文科省戦略的研究基盤形成支援事業によって行われた。記して，感謝する

参考文献

[1] 近山　隆, 柳内啓司, 三輪　誠 : “自然言語テキストの言語知識を用いない類似性判定の試み”, 情報処理学会夏のプログラミング・シンポジウム報告集, 2003, pp.81-86.

[2] 和田成夫 :“よくわかる信号処理”, 森北出版株式会社, 東京 pp.79-84 (2009).

[3] 小暮陽三, 佐々木良輔(編) : “なっとくするフーリエ変換”, 講談社, 東京(1999).

[4] 貴家仁志 : “ディジタル信号処理”, 昭晃堂, 東京, pp.95-99(1997)

[5] 漫棚通信ブログ版 : http://mandanatsusin.cocolog-nifty.com/blog/2007/06/post_a49c.html

[6] 松澤　当, 樋口政和, 村上仁己 “各種直交変換を用いた文章間の類似性判定” 映像メディア学会年次大会 2010 年８月