＜研究ノート＞

体育･スポーツ史研究関連史料からテキストデータへ
―コーパス言語学的アプローチによる オリンピズムの受容史研究を見据えて―

Formatting Text Data of the Material concerning the History of Physical Education and Sport
―The Corpus Linguistics' Approach to the Historical Study on the Reception of Olympism in Japan―

和田　浩一
Koichi WADA

Ⅰ．はじめに

A．研究の概要

本研究ノートは、コーパス言語学の研究手法を体育・スポーツ史研究に援用する際に必要となる、史料のテキストデータ化に焦点を当てたものである。史料のテキストデータ化とは、コンピュータが処理できるように、史料の文字情報を文字コードだけで構成された文字列のデータに変換することを指す。具体的には、１文（センテンス）に含まれる単語数「文長」と語彙の豊富さを示す指標「K 特性値」とを数値化するのに適したテキストデータを、体育・スポーツ史の研究者[1)]が〈納得できる形で〉作成する一つの試みである。本研究はコーパス言語学を専門とする研究者との共同研究の一部を構成しているが、史料のテキストデータ化の責任は、その史料の性格を熟知する当該研究者（筆者）が担うべきであろう。なお、本研究ノートでは、コーパスを「言語分析に利用できる電子化された言語資料の集積」（齊藤 2005、p. 21.）と定義する。

B．研究の背景

まず最初に、コーパス言語学の知見を体育・スポーツ史研究に援用するに至った経緯を簡単に説明しておきたい。

筆者は、近代オリンピックの創始者 Pierre de Coubertin（1863-1937）が名付けたオリンピズムという教育思想が日本でどのように受容されたのかを研究する中で（和田 2010a）、Mary Girard "Pierre de Coubertin: An Appreciation" (*Fortnightly Review* 74: 336-346, August 1903.）という英文雑誌記事を発見した。Coubertin の半生を描くこの記事は、1）国内外の主要な Coubertin 研究で取り上げられておらず（Lucas 1962、Boulongne 1975、MacAloon 1981、Callebat 1988、清水 1999、Bermond 2008）、オリンピック草創期における Coubertin 評価に新たな解釈をもたらす可能性を秘めている、2）明治期の人物批評の第一人者である鳥谷部春汀（1865-1908）による抄訳が『中学世界』（1903）に掲載されており、オリンピック参加以前の日本におけるクーベルタン理解に直結する文献である[2)]、という二つの意味で重要な史料である。

近代オリンピックが単なる国際スポーツイベントではなく、平和で秩序のある世界の構築を目指したスポーツによる教育改革運動であること（Coubertin 1896）を理解し、Coubertin を教育改革者として描く Girard の記事は驚きだった。なぜなら、この記事が書かれた 1903 年当時、国際オリンピック委員会（Comité International Olympique）は発足して 10 年目を迎えていたにもかかわらず、「世界にオリンピックの思想を広める」（Coubertin 1907）という委員会の目的が、ほとんどの委員に理解されていなかったからである（Coubertin 1931）。さらには、オリンピック復興者ではなく教育改革者という Coubertin への評価は、根底的な史料収集に基づいて彼の全生涯と業績とを俯瞰した Boulongne による博士論文（前出、1975 年）の提出を待たねばならなかった。そして、この英文記事を一読した直後、直感ではあるが「これは

Coubertin 自身が書いた記事ではないのか」との考えが頭をよぎった。

たとえ、史料の内容（テキスト）は同じでも「実は作者が別人だった」ということになれば、この史料が記された意図には別の解釈が成り立つ可能性が生まれ、別の解釈によって史料のもつ意味が変われば歴史認識にも変化が生じる。もし「Girard = Coubertin」という仮説が真であれば、オリンピック史における Coubertin とオリンピズムへの評価にも変化が生じよう。

「Girard = Coubertin」という仮説を検証するには、「誰が」これを記したのかという基本的な史料批判が必要となる。具体的には、1）この記事のオリンピック史上の意味を歴史的事実に基づきながら検討するとともに、2）Coubertin の著作等と比較しながら記事の内容を吟味するという作業である（和田 2010b）。本研究ノートは、このような伝統的な歴史学的手法に、近年めざましい進化を遂げつつあるコーパス言語学の研究手法を専門家の助けを借りて融合させる新しい史料批判の試みを見据えている。

C．コーパス言語学と著者推定

コーパス言語学とは、パソコンが発達・普及した 1990 年前後から急速に台頭した言語学で、コンピュータ処理が可能なコーパスを検索して言語分析・記述を行う歴史の浅い学問である。例えば英語学の分野においては語彙研究や文法研究、英語史研究、文体論研究に新しい風を吹き込むだけではなく、辞書の編纂や英語教育などにも活用されている（齊藤 2005、pp.3-4, 21-22, 121-265.）。

Girard の筆によるとされる記事が Coubertin の作であったかどうかを確かめる手法は、計量文献学の世界では著者推定と呼ばれており、19 世紀から現代に至るまでさまざまなジャンルのテキストに対して、その真贋を含めた分析が行われてきた（前川 1995；村上 2004、pp. 11-173.）。最近ではコーパス言語学の隆盛

に伴い、大量のテキストの中から一定の法則性や有用な情報を見出すテキストマイニング[3]の文脈で取り上げられることも多い（金 2008）。

D．先行研究の検討

体育・スポーツの歴史研究で扱う史料を、言葉を数量的に分析するコーパス言語学的なアプローチによって分析した先行研究は4点ある。19世紀および20世紀のフランス体育書における運動記述を分析した清水の研究（1988-1990年度;1991-1992年度）と、19世紀初頭のアメリカ新体育理論における術語体系を議論した小田切の研究（1990-1991年度；1994-1996年度）である。

科学研究費補助金の助成を受けたこれらの先駆的研究は、主に体育・スポーツの領域に特徴的な語彙の解明に焦点を当てたもので、本研究ノートが見据える史料批判を目指したものではない。また、体育・スポーツ史関係の史料をコーパス化するという作業自体は本研究での取り組みと共通しているが、コーパスから特徴的な用語を取り出し、これらの使用頻度を他のものと比較検討することが分析技術の中心となっている。

Ⅱ．分析方法と準備すべきデータ形式

A．分析方法

「Girard = Coubertin」という仮説を検証するため、Coubertin自身の手によるものであることが明らかなテキストデータを分析し、そこで得られる統計的諸特徴とGirardによる記事の諸特徴とを比較するという手順を取ることにした。もしも「Girard = Coubertin」が真であれば、両者のテキストは諸々の統計的数値において似通った分布を示すはずである。

Coubertinのデータは、翻訳者の名が併記されていない彼の英文雑誌記事から構成した。すなわち、*Century Magazine* から4

本(1896-1901年)、*Fortnightly Review* から16本(1897-1904年)、*American Monthly Review of Reviews* から12本（1896-1901）の計32本である[4]。

算出しようとする統計値は、「文長（一文に含まれる単語数)」の平均値、標準偏差、中央値、四分位範囲、さらにテキスト分析でよく用いられる「K特性値」の5つとした。

文章の計量的な分析による著者推定の試みを、国内外・時代・ジャンルを問わず幅広く報告している村上によれば、テキストデータの分析は単純な方法から複雑な方法まで多様にあり、これらの選択には、1）文化現象の内容や研究目的に応じた手法の検討、2）精度を含むデータの性格への配慮、3）できるだけ単純な方法での解明を試みる姿勢、が重要になるという。そして、著者の文体を理解するための具体的な情報として文長、単語長、単語の出現率、品詞の出現率、品詞の接続関係、語彙量、読点の付け方などを示している（村上 2002、p. 11-17.)。今回は、体育・スポーツ史研究においてコーパス言語学の知見を活用することが初めてということもあり、著者推定の方法として歴史がありかつこれまで頻用されてきた「文長」と「K特性値」を、上記仮説の検証で用いることにした[5]。複雑な手法を用いて導き出された分析結果が当該研究者（ここでは体育・スポーツ史研究者をはじめとする人文科学の研究者）に理解できないものであれば、本研究のような学際的研究は成立しないと言えるだろう。

B．準備すべきデータ形式

データ分析を担うコーパス言語学の専門家（共同研究者）から求められたデータ形式は、「1センテンス＋改行」であった。

Ⅲ．コーパス作成のアウトライン

Girard およびCoubertinの手による英文雑誌記事のコーパス

を作成し、これを分析に適したデータ形式に整形するための手順は、次のとおりである。1）テキスト（文字）入力の方針を決める、2）テキスト以外の情報の付与の規則を定める、3）作成したコーパスからテキスト整形を実施する。

本研究で対象とする文献史料は、英語だけではなくアクセント付き文字を含むフランス語も含んでいるので、コーパスのファイル形式はテキストファイル（文字コード：Windows 1252, iso-8859-1）とした。

「CiNii Articles - 日本の論文をさがす（http://ci.nii.ac.jp/）」でコーパスの作成や構築に関する論文を探すと、そのほとんどはテキスト入力と入力ミス訂正の能率化を検討するものとなっている。本研究ノートで特に焦点を当てようとする「作成したコーパスからのテキスト整形」、つまりコーパス言語学を専門としない研究者によるテキスト整形の実際を詳細に報告するものは皆無に等しい。

A．テキスト入力の方針と方法

Girard と Coubertin の筆による英文雑誌記事のコーパスは、誰もが容易に参照できるように、原文にできるだけ忠実な形式で保存することを原則とした。その上で、テキストファイルでは保持できない書誌や文字修飾、文章構造といったテキスト以外の情報を、次の方針・方法で処理することにした。

1. 英文雑誌記事のコピーを ScanSnap S1500（富士通製）で読み取る。
2. 読み取った結果を「ABBYY FineReader for ScanSnap(TM) 4.1」(OCR ソフト) で Microsoft Word 形式に変換する（ファイル A）。
3. ファイル A を Microsoft Word 2002 で開く。

4. 3の状態で表示されているWord 2002上で適当なテキストのまとまりを選択･コピーし、テキストエディタ「EmEditor Professional（Version 10.0.1、Emurasoft）に貼り付ける[6]。その際、ページ情報を「[改行] p. 204. [改行]」の形式で付与する。これを手作業で繰り返す（ファイルB）。
5. ファイルBをテキストエディタ「テキストエディタQX」（後述）で開き、印刷した結果と原文とを照らし合わせ、同テキストエディタ上で間違い箇所を修正する。

ここでの作業のポイントは4である。Microsoft Word 2002で開いたファイルAをテキストファイル（エンコード方法：西ヨーロッパ言語（Windows））で保存する方法は、アクセント付文字が正確に保存されず、使えないことが分かった。また、OCRソフトで読み取った直後のファイルAは、モニターでの見かけ上は史料のレイアウトに近くても、テキストデータに変換すると段落の順番が入れ替わるケースが頻発した。したがって、今回作成しようとする分量程度のコーパス作成では、一見時間がかかるように見える4の手作業が、後の入力ミスの訂正の手間を減らすことになり、一番能率がよいことが判明した。最終データには必要のないページ情報をテキストデータに付与したのは、原文とのチェック作業をしやすくするためである。

なお、すでに記したように、本研究でのテキストデータ処理にはワープロではなく、充実した検索や置換、grep検索、マクロ機能を備えたテキストエディタを用いる。具体的には「テキストエディタQX（v6.91、欧文モード[7]、新井健二araken氏作）」を用いた。テキストエディタQXはタブやコントロールコード、改行マーク、検索文字列、引用符、括弧、見出しなどの色を視認性がよいものに独自に設定することができ、データ処理作業の能率化を図れる[8]。

B．文字以外の情報の付与の規則化

本研究で求めようとする「文長」は1センテンス中の単語数から、「K特性値」は異なる単語数から算出される統計値である。構文解析を可能とするような品詞情報（品詞タグ）等を付与する必要はないので、テキスト以外の情報の付与は簡便な方法で行うことにした。テキスト以外に付与した情報とその方法は、次のとおりである[9]。

注）以下、半角スペースは「_」（アンダーバー）で表記する。

1. イタリック：
   Coubertin → {i}Coubertin{/}
2. 記事名、著者名、雑誌名、号数、出版年、見出し：
   Coubertin and Today, 2013. →［改行］t._Coubertin and Today, 2013.［改行］
3. 脚注：
   ［改行］f. ---------［改行］
   f._Coubertin, Pierre de. *Mémoires olympiques*, 1931.［改行］
   f. ---------［改行］
4. ダッシュ：
   —— → _---_
5. 注番号：
   Coubertin is Frenchman.[3] → Coubertin is Frenchman.*3
6. ハイフネーション：
   a. 複合修飾語、複合語、接頭辞、接尾辞などで用いられるハイフネーションはそのまま残す。
   b. ページ末のハイフネーションはそのまま残し、原文チェックの際の便宜を図る。
   c. aとb以外の理由で単語を分けているハイフネーション

は削除する。

7. 段落の頭には「半角スペース5個」を挿入する。

Ⅳ. 作成したコーパスからのテキスト整形

ここでの作業の意味は、Ⅲで記した方針に従って作成したコーパスを、「文長」と「K特性値」の分析に必要な「1センテンス＋改行」で構成されるデータ形式に整形することである。以下、コーパス言語学やコンピュータのプログラミング言語などに関する十分な知識をもたない体育・スポーツ史研究者が、〈納得できる形で〉データを整形する方法を示す。研究者自身が十分に理解できない複雑なプログラムにデータの整形を託すのではなく、一つ一つの作業が何をしているのかを確実に理解して作業を進めることによって、データ（史料）の正確性を担保したい。

作業の大まかな内容は、1）テキスト以外の情報を削除する、2）各センテンスの終わりに改行を入れる。3）約物（句読点や疑問符、括弧など）を削除する、ことである。

注1. テキスト整形ではテキストエディタの「検索」と「置換」の二つの機能を多用する。その際、特別に指示した場合を除き、正規表現[10]が使えるように設定しておく。なお、正規表現で用いるメタキャラクタ（. ^ * + ¥n ¥t ほか）の意味については、齊藤（2005、pp. 56-60.）などの説明を参照のこと。

注2. テキスト整形には文末処理が必要となるので、［改行(マーク)］を視認できるようにテキストエディタを設定しておく。

注3. ［改行］など、［　］で括られた文字はテキストエディタ上に表示される文字ではなく、筆者による説明である。

注4. 置換文字列の［なし］は、空欄（文字の入力なし）を意

味する。

（ここから）----------

A. アポストロフィー「'」と引用符「"」「'」の入力状態を確認し、必要があれば修正する。

「'」と「"」「'」は最終的には削除するが、正確なデータを作成するためにこの作業を最初に行う。なお、本研究ではアポストロフィー「'」を含む単語を1語として数え、例えば「Coubertin」と「Coubertin's」、「do not」と「don't」は異なる語と見なした。

1. 「's」が直前の単語とくっついているか。例えば「Coubertin_'s」となっていないか。
   検索文字列：_'s_
2. ダブルクォーテーション「"」の代わりに、二つのシングルクォーテーション「''」が使われていないか。
   検索文字列：''
3. ダブルクォーテーション「"」が、ダブルクォーテーション＋シングルクォーテーション「'"」、あるいはその逆「'"」になっていないか。
   検索文字列1：'"
   検索文字列2：'"
   ※. 本来、シングルクォーテーションとアポストロフィーは別物であるが、本研究では、テキストエディタ上で同じものとして（同じ文字コードで）扱った。
   ※. 1-3の検索でヒットしたケースがそれぞれ複数回あった。印刷結果上で判別しにくい文字列の組み合わせであり、OCRで誤変換したものが、2回の原文チェックを経ても修正されなかったのだと思われる。

B. イタリックの付与情報「{i}」「{/}」を削除する。

テキストエディタQXの「メモを削除して保存」機能を利用する。イタリック情報の付与は、「{i}」と「{/}」で挟まれた部分をイタリック体で印刷するテキストエディタQX（欧文モード）の機能を利用したものである。

C. 記事名、著者名、雑誌名、号数、出版年、見出しを一括置換で削除する。

検索文字列：¥nt¥._.*¥n → 置換文字列：¥n

意味：「[改行]t._[記事名など][改行]」を「改行」に置換

※. 検索文字列がヒットしなくなるまでこの作業を繰り返す。

D. 脚注を一括置換で削除する。

検索文字列：¥nf¥._.*¥n → 置換文字列：¥n

意味：「[改行]f._[脚注][改行]」を「改行」に置換

※. 検索文字列がヒットしなくなるまでこの作業を繰り返す。

E. 脚注番号を一括置換で削除する。

検索文字列：¥*[0-9]* → 置換文字列：[なし]

意味：「*[任意の数字]」を削除

F. 半角2個分以上のスペースを、半角スペース1個分「_」に置換する。

検索文字列：__ → 置換文字列：_

意味：[半角スペース2個]を[半角スペース1個]に置換

※. 検索文字列がヒットしなくなるまでこの作業を繰り返す。

G. 文末以外にピリオド「.」が使われている用例を抽出し、以下の手順で処理する。

文末の「.」はセンテンスの区切り記号として用いるので、これ以外の「.」を削除しておく。

1. ページ情報「[改行] p._ [ページ数]」の「.」を「@」に置換する。

   検索文字列：¥np¥._ → 置換文字列：¥np@_

   意味：「[改行] p._」を「[改行] p@_」に置換

   ※. データの整形中に原文と照らし合わせたいケースが生じたときの便宜を図るため、ページ情報はできるだけ長く保持しておきたい。

2. 「...」「._._.」(いずれも [三点リーダー]) が文末にあれば「.」に、文中にあれば「[なし]」に置換する。

   検索文字列 1：¥.¥.¥.

   検索文字列 2：¥._¥._¥.

   意味：「...」と「._._.」を検索

   ※. 一例ずつ確認しながら処理する。

3. 「.」の直前に「_」があれば、一括置換で削除する。

   検索文字列：_¥. → 置換文字列：¥.

   意味：「_.」を「.」に置換

4. 数字の区切り記号を統一する。

英語 (123,456,789.00) とフランス語 (123.456.789,00) とでは、数字の区切り記号が異なる。表記に「ゆれ」があれば、英語方式に修正・統一する。

検索文字列：[0-9][¥.,][0-9]

意味：「[0 を含む 1 桁の任意の数字]. (または「,」) [0 を含む 1 桁の任意の数字]」を検索

5. 「Dr.」など、大文字アルファベットを 1 文字以上含む文字列との組み合わせで各種略語に使用される「.」を、以下の要領で処理する。

「_ [大文字アルファベット] [1 文字以上の小文字アルファベット] .」で構成される語 A、および「[任意の文字] [大文字アルファベット] .」で構成される語 B が文末で使用されている箇所を探し出し、「.」をもう一つ加える（「.」→「..」）。

語 A　検索文字列：¥(_[A-Z][a-z]+¥)¥._

→置換文字列：¥1¥.¥._

意味：「_ [大文字アルファベット] [1 文字以上の小文字アルファベット] ._」を検索し、ヒットした語が文末にあれば「.」の直後に「.」を加える。

ヒットした用例：「Dr.」「St.」「Mr.」「Miss.」「Hon.」「Stockholm.」他

語 B　検索文字列：[A-Z]¥._

意味：「[大文字アルファベット] ._」を検索

ヒットした用例：「M.」「P.」「I.」「X.」「V.」「I.O.C.」「MM.」「B.C.」「A.D.」他

※. 文末の「.」かそれ以外の「.」かは、一例ずつ確認・判断して処理する。

※.「P.M.」「I.O.C.」など、語尾以外の位置にも「.」を含む語が文末以外にあった場合は、「.」を含まない「PM」「IOC」の形式に修正する。

※.「W. M. Sloane」の「W.」「M.」のような省略された人名のイニシャルについては、「Wname」「Mname」と表記し、「M.」(Monsieur の省略形)

等と区別することにした。

上記の処理を終えた後、次のように一括置換し、語Aと語Bの「.」を削除する。

語A　検索文字列：¥(_[A-Z][a-z]+¥)¥. → 置換文字列：¥1

意味：「_［大文字アルファベット］［1文字以上の小文字アルファベット］.」を「_［大文字アルファベット］［1文字以上の小文字アルファベット］」に置換

語B　検索文字列：¥([A-Z]¥)¥. → 置換文字列：¥1

意味：「［大文字アルファベット］._」を「［大文字アルファベット］」に置換

※. 文末の語Aと語Bについては「IOC..」の形になっているので、これらの置換で前の「.」が削除され、後ろの「.」が文末を示す記号として残る。

6. 「etc.」など、大文字アルファベットを含まない各種略語で使用される「.」を一例ずつ検索・確認し、文末以外で使用されていれば「.」を削除する。

検索文字列1：¥._[^A-Z]

意味：「._［大文字アルファベット以外の文字］」

検索文字列2：¥.[^_]

意味：「.［空白以外の文字］」

ヒットした用例：「etc.」「e.g.」「i.e.」「vol.」

※. 文中で「etc.」などが使われる場合、その多くは直後の単語が小文字で始まるという文法の性質を利用する[11]。

H. 括弧を次の要領で削除する。

テキストエディタQXの表示に関する書式設定で「対応する括

弧」欄にチェックを入れ[12]、同じく「色」の「強調括弧」を視認性のよい色に設定すると作業の能率が上がる。

1. 「" "」：一括置換で削除する。
   検索文字列：" → 置換文字列：［なし］
2. 「( )」：一括置換で削除する。
   検索文字列１：(→ 置換文字列：［なし］
   検索文字列２：) → 置換文字列：［なし］
3. 「[ ]」：一括置換で削除する。
   検索文字列１：[→ 置換文字列：［なし］
   検索文字列２：] → 置換文字列：［なし］
   ※.「[ ]」はメタキャラクターなので、検索機能で正規表現を使わない設定にしておく。
4. 「｛｝」：一括置換で削除する。
   検索文字列１：｛ → 置換文字列：［なし］
   検索文字列２：｛ → 置換文字列：［なし］
5. 「<< >>」［フランス語の引用符 ＝ フレンチダブルクォート］：一例ずつ検索・確認して削除する。
   検索文字列１：<<
   検索文字列２：>>
6. 「< >」：［フランス語の引用符 ＝ フレンチシングルクォート］：一例ずつ検索・確認して削除する。
   検索文字列１：<
   検索文字列２：>

I. 次の手順で文末の「?」と「!」の直後を改行する。

1. 「?」と「!」を一例ずつ検索・確認し、文中で使用されているものについてはそれぞれ「?#」と「!#」に置換する。
   検索文字列：[¥?!]
   意味：「?」または「!」を検索

2. 文末の「?」「!」の直後を一括置換で「改行」する。
   検索文字列：¥([¥?!]¥)_ → 置換文字列：¥1¥n
   意味：「?_」と「!_」を「?［改行］」と「!［改行］」に置換
3. I-1で置換した文中の「?#」「!#」を一括置換し、それぞれ「?」と「!」に戻す。
   検索文字列：¥([¥?!]¥)# → 置換文字列：¥1
   意味：「?#」と「!#」を「?」と「!」に置換

J. ページ情報とその前後を、次の手順で処理する。
1. ページ情報（例えば「［改行］p@_123.［改行］」）の直前にハイフネーション「-」がある場合、ページ情報「［改行］p@_123.［改行］」の削除後、1）複合修飾語、複合語、接頭辞、接尾辞などで必要なハイフネーションは、これを残して前後の単語をつなげる、2）行末調整のために単語を音節の区切りで分けているハイフネーションは削除し、前後の文字列を1語としてつなげる。
   検索文字列：-¥n
   意味：「-［改行］」を検索
2. ページ情報の直前が文末の「.」の場合、一括置換でページ情報を削除する。
   検索文字列：¥.¥np@_.*¥n → 置換文字列：¥.¥n
   意味：「.［改行］p@_123.［改行］」を「.［改行］」に置換
3. ページ情報で分けられたページ末とページ頭の単語を、一括置換により「_」で挟んでつなげる。
   検索文字列：¥np@_.*¥n → 置換文字列：_
   意味：「［改行］p@_123.［改行］」を「_」に置換
4. ページ情報とその前後の処理が正しく行われたかどうかを

確認し、必要があれば修正する。

検索文字列：@

意味：「@」を検索

K. セミコロン「;」とコロン「:」およびそれらの前後を、以下の要領に従って処理する。

これら2種類の約物をそれぞれ一例ずつ検索・確認し、約物の後に1センテンスあると判断した場合、約物の直後で「改行」する。本研究では、これらの約物の直後の単語が大文字アルファベットで始まりかつ、主語と述語とをそなえた文が続く場合、当該の約物の前後にそれぞれセンテンスが一つずつ（計2センテンス）あるものと解釈した[13]。

検索文字列1：;_[^a-z]

検索文字列2：:_[^a-z]

L. 文末に改行を入れるとともに、以下の約物を削除する。

1. 文末の「._」を一括置換で「.［改行］」にする。

検索文字列：¥._ → 置換文字列：¥.¥n

意味：「._」を「.［改行］」に置換

※.「.」と「_」を組み合わせているのは、数字の区切り文字で使用されている「.」を削除しないようにするため。

2. 各センテンスの、1）文頭が大文字になっているか、2）文末が「.」「?」「!」「;」「:」「_---_」のいずれかになっているかを確認する。

これまでの処理結果の〈おおまかな〉確認である。以下の「検索文字列1-3」の検索によって、1）と2）の規則から外れたケースがヒットする。ヒットした場合は原文を参照し、場合によっては1文の定義をさらに細かく定めて処理する。

検索文字列 1：¥n[^A-Z]

意味：「[改行]［大文字アルファベット以外の文字]」を検索

検索文字列 2：[^¥.¥?!:;]¥n

意味：「[「.」「?」「!」「;」「:」以外の文字]［改行]」を検索

検索文字列 3：[¥?!]¥n[^A-Z]

意味：「?（または「!」）［改行］［大文字アルファベット以外の文字]」を検索

3. 次の約物をそれぞれ一括置換する。

検索文字列：_--- → 置換文字列：［なし］

検索文字列：；→ 置換文字列：［なし］

検索文字列：：→ 置換文字列：［なし］

検索文字列：¥. → 置換文字列：［なし］

M. 「,」を一括置換で削除する。

検索文字列：,_ → 置換文字列：_

意味：「,_」を「_」に置換

※.「,」と「_」を組み合わせているのは、数字の区切り文字で使用されている「,」を削除しないようにするため。

N. 2 個以上の連続した半角スペースを「_」に置換する。

検索文字列：__ → 置換文字列：_

意味：「__」を「_」に置換

※. 検索でヒットしなくなるまで繰り返す。

O. 行頭の「_」を一括置換で削除する。

検索文字列：¥n_ → 置換文字列：¥n

意味：「[改行] _」を「[改行]」に置換
※．検索でヒットしなくなるまで繰り返す。

P．行末の「_」を一括置換で削除する。
検索文字列：_¥n → 置換文字列：¥n
意味：「_ [改行]」を「[改行]」に置換
※．検索でヒットしなくなるまで繰り返す。

Q．空行（改行だけの行）を一括置換で削除する。
検索文字列：¥n¥n → 置換文字列：¥n
意味：「[改行] [改行]」を「[改行]」に置換
※．検索でヒットしなくなるまで繰り返す。

（ここまで）----------

V．おわりに

言語処理（言語研究）に先立つテキスト処理（テキスト整形）を扱った本研究ノートは、「コーパス言語学の技術」と「体育・スポーツ史研究」の橋渡しの部分に焦点を当てた方法論的試行の中間報告である。これまで、コーパス言語学を専門としない研究者がコンピュータによる分析手法に適したテキストファイルを〈納得する形で〉作成するノウハウは、ほとんど公表されてこなかった。

村上（2002、pp. 13-17.）は筆者のようなコーパス言語学の門外漢に対し、言葉を計量的に分析する際の姿勢や態度に関して、次のような重要な示唆を与えている。

一つ目の示唆は、構築したコーパスの中身を一部修正したりデータを加工したりすることは頻繁に起こるということである。コーパスを利用する研究者は、たとえ言語学や情報学の専門家で

なくても、間違いなく情報を追加したり内容を修正したりできなければならないということだろう。

二つ目は、データは万能ではなく、データ分析の結果に対しては計量的な分析とは異なる観点からの裏付けが必要であるということである。今後算出する「文長」と「K 特性値」の結果が数字で表現されるがゆえに、体育・スポーツ史の研究者自身が「コーパス作成」と「テキスト整形」に《史料》としての正確性を担保するとともに《史料》に語らせるという責任を持たなければならない。

なお、本研究は JSPS 科研費 22500597「コーパス言語学的アプローチによるクーベルタン・オリンピズムの受容史研究」の助成を受けたものである。

Ⅵ. 注

1）「コーパス言語学やコンピュータのプログラミング言語などに関する十分な知識をもたない人文科学研究者」と読み替えることも可能であろう。
2）日本の近代オリンピック大会への初参加は、1912 年のストックホルム大会である。
3）テキストマイニング（text mining）とは、大量のテキストの中からコンピュータを用いて有益な情報を探し出す技術を指す。
4）記事の書誌情報は、Müller（1991）を参照のこと。
5）「文長」「K 特性値」とも、著名な統計学者ユール（1871-1951）がそれぞれ 1939 年と 1944 年に、著者推定のために開発した分析手法である（村上 2004、pp. 76-80.）。
6）これ以降の作業で使用するテキストエディタ「テキストエディタ QX（欧文モード）」は、選択・コピーした Microsoft Word 2002 上のアクセント付文字を正確に貼り付けることができなかった。Windows のアプリケーションソフト間でデータ交換・連携を実現する OLE（object linking and embedding）の機能が働いていないようだ。
7）欧文モードは作者による README.TXT やヘルプファイルには出てこない「隠し機能」となっている。
8）他のテキストエディタでも、機能上はまったく問題ない。

9）テキスト以外の情報はデータ分析に直接必要ないが、原文とのチェック作業を円滑に進めるための有益な情報となる。
10）通常使用する文字と「メタキャラクタ」と呼ばれる特別な意味や機能をもつ記号を組み合わせて、文字列のパターンを指定する表記法（齊藤2005、pp. 56-60.）。
11）「etc.」が文末で使われる場合、その直後の単語は文頭に位置するので通常は大文字のアルファベットで始まる。
12）「対応する括弧」にチェックを入れると、カーソル上に括弧 / 引用符があるとき、対応する括弧 / 引用符を強調表示する。
13）1センテンスの定義によって、処理の方法は異なる。

Ⅶ. 文献

小田切毅一「アメリカの新体育理論における術語体系に関する研究」（研究課題番号：02680104、1990-1991年度）
小田切毅一「アメリカ新体育論における用語の成立とその系譜に関する研究」（研究課題番号：06680105、1994-1996年度）
金明哲「フリーソフトによるデータ解析・マイニング：統計的テキスト解析（5）――統計法則と指標――」『ESTRELA』172、2008年、pp. 60-65.
齊藤俊男、中村純作、赤野一郎『改訂新版 英語コーパス言語学:基礎と実践』研究社、2005年
清水重勇「19世紀フランス体育書における運動記述の理論的特性に関する学説史的研究」（研究課題番号：63580095、1988-1990年度）
清水重勇「19世紀と20世紀のフランス体育書の文字列データ化による体系的文献解題法研究」（研究課題番号：03680114、1991-1992年度）
清水重勇『スポーツと近代教育：フランス体育思想史（下）』紫峰図書、1999年
前川守『文章を科学する』岩波書店、1995年、pp. 1-56.
村上征勝『文化を計る：文化計量学序説』朝倉書店、2002年
村上征勝『シェークスピアは誰ですか：計量文献学の世界』文藝春秋、2004年
和田浩一「オリンピズムという思想 —— 新しいオリンピズムの構想への序章」『現代スポーツ評論』23、2010年a、pp. 62-71.
和田浩一「Pierre de Coubertin: An Appreciation（*Fortnightly Review*, August 1903）の作者は誰なのか：記事の歴史的位置づけと内容の分析とによる作者同定の試み」スポーツ史学会第24回大会発表資料、2010

年 b（http://homepage3.nifty.com/wadaco/contenu/documents/20101128_wada_sh24.pdf）
Bermond, Daniel. *Pierre de Coubertin,* [Paris] : Perrin, 2008.
Boulongne, Yves-Pierre. *La vie et l'œuvre pédagogique de Pierre de Coubertin (1863-1937),* Ottawa : Leméac, 1975.
Callebat, Louis. *Pierre de Coubertin,* [Paris] : Fayard, 1988.
Coubertin, Pierre de. "The Olympic Games of 1896." *Century Illustrated Monthly Magazine* 53(31).1, November 1896, pp. 39-53.
Coubertin, Pierre de. "Critiques et calomnies." *Revue Olympique,* 7e année, janvier 1907, p. 198.
Coubertin, Pierre de. *Mémoires olympiques,* 1931. (Paris : "Revue EPS", 1996, pp. 9-10, 44.) ドイツ語版からの重訳に、カール・ディーム編・大島鎌吉訳『オリンピックの回想』（ベースボール・マガジン社、1962［1976］年）がある。
Lucas, John Apostal. *Baron Pierre de Coubertin and the formative years of the modern international Olympic movement 1883-1896,* Ann Arbor, Michigan : UMI, 1962.
MacAloon, John J. *This great symbol: Pierre de Coubertin and the origins of the modern Olympic Games,* Chicago : The University of Chicago, 1981.
Müller, Norbert; Schatz, Otto. *Bibliographie des œuvres de Pierre de Coubertin,* Lausanne : Comité International Olympique, 1991.