Title	XMLによる初期刊本の本文記述の方法論の確立と印刷史研究への応用
Sub Title	Towards method of transcription of early printed books in XML and its application to the study of the printing history
Author	安形, 麻理(Agata, Mari)
Publisher	
Publication year	2015
Jtitle	科学研究費補助金研究成果報告書 (2014.)
JaLC DOI	
Abstract	初期刊本の画像データを用いた活字の識別の正確かつ効率的な手法を開発した。この手法により, 一般のOCRソフトでは処理できない典型的な初期刊本についても, 大規模なテキストデータ化が可能になると期待される。 次に, 西洋最初の印刷本であるグーテンベルク聖書の画像を対象に本活字識別手法を応用した。識別結果に基づき, 活字を客観的な基準で分析するため, 活字画像のクラスタリングを行い, 活字の鋳造方法についての先行研究を検証した。また, 識別結果に基づきトランスクリプションデータを作成し, XMLによる本文記述を行った。 An efficient and precise method of identifying individual type of the early printed books was developed, which is indispensable in making transcription of early printed books, since ordinary OCR software cannot deal with them. The proposed method is expected to enable to make transcription data of the early printed books on large scale. The proposed method was applied to the digital images of the first printed book in Europe, the Gutenberg Bible. Cluster analysis of the type images were conducted in order to shed some light objectively on the early metods of making types. Furthremore, Based on the result of the type image recognition, transcription data was also made, and then described in XML format.
Notes	研究種目：若手研究(B) 研究期間：2008～2014 課題番号：20700225 研究分野：書誌学
Genre	Research Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KAKEN_20700225seika

様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２２ 日現在

機関番号：３２６１２
研究種目：若手研究(B)
研究期間：2008～2014
課題番号：２０７００２２５

研究課題名（和文）ＸＭＬによる初期刊本の本文記述の方法論の確立と印刷史研究への応用

研究課題名（英文）Towards method of transcription of early printed books in XML and its application to the study of the printing history

研究代表者

安形　麻理（agata, mari）

慶應義塾大学・文学部・准教授

研究者番号：７０４３３７２９

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：　初期刊本の画像データを用いた活字の識別の正確かつ効率的な手法を開発した。この手法により、一般のOCRソフトでは処理できない典型的な初期刊本についても、大規模なテキストデータ化が可能になると期待される。
　次に、西洋最初の印刷本であるグーテンベルク聖書の画像を対象に本活字識別手法を応用した。識別結果に基づき、活字を客観的な基準で分析するため、活字画像のクラスタリングを行い、活字の鋳造方法についての先行研究を検証した。また、識別結果に基づきトランスクリプションデータを作成し、XMLによる本文記述を行った。

研究成果の概要（英文）：An efficient and precise method of identifying individual type of the early printed books was developed, which is indispensable in making transcription of early printed books, since ordinary OCR software cannot deal with them. The proposed method is expected to enable to make transcription data of the early printed books on large scale.
The proposed method was applied to the digital images of the first printed book in Europe, the Gutenberg Bible. Cluster analysis of the type images were conducted in order to shed some light objectively on the early metods of making types. Furthremore, Based on the result of the type image recognition, transcription data was also made, and then described in XML format.

研究分野：書誌学

キーワード：書誌学　印刷史　トランスクリプション　グーテンベルク聖書　初期刊本　デジタル画像

１．研究開始当初の背景

15 世紀半ばの活版印刷術の発明から 1500 年末までに印刷された書物は初期刊本（incunabula）と呼ばれ、その後の印刷本とは区別される。初期刊本時代に書物の形態が大きく変化し、印刷本の発展を方向付けることになるが、写本からの移行期であることから、写本ともそれ以降の印刷本とも異なる固有の特徴を持っているためである。

こうした重要性の認識から、近年、初期刊本のデジタル画像化は急速に進展しており、申請者による画像を用いた刊本の校合手法の開発や 1)、画像による活字自動同定の試みなど 2)、研究への応用も成果を挙げつつある。画像によって書物史・印刷史の新たな研究が可能になるのは確実であるが、一方で画像のみによる研究には限界があり、本文データの整備が求められる。

ただし、必要とされるのは、テクスト（本文）、パラテクスト（序文やレイアウトや装飾など本文以外の形態的特徴）、コンテクスト（その書物を取り巻く社会的・文化的な文脈や読者の反応）という書物の三つの層を表現できる本文データである。単純なテキスト化や画像データでは、この三層を扱い、検索や分析が可能な形で表現することはできないため、構造化されたタグ付けが必要になる。本文だけをとっても、初期刊本には同一文字の異なる形が複数あり、その使用方法はコンテクストと密接に関係するため、厳密な区別が必要である。また、現存する諸本は同一版でも少しずつ異なる独自の本文を持ち、その違いが印刷工程を解明する手がかりとなるため、違いを提示しなければならない。

現在まで初期刊本のトランスクリプション・データはほとんど作られていない。そのため、例えばシェイクスピア研究においては綴りの特徴の違いから植字工の分担作業が明らかになるなどの成果が挙がっているが、初期刊本に関する同様の研究は進んでいない。今後、初期刊本のデジタル化の発展が予想されるなか、書誌学的な研究を行う上で有用な本文データの作成の方法論を確立しておくことは急務だと考えられる。

２．研究の目的

本研究では、三つの具体的な課題を設定した。

① 最初の課題は、書誌学的な研究を行うにあたって有益であるような初期刊本のトランスクリプションの要件を明確化し、実現に際しての課題を整理することで、構築の方法論を確立することである。

初期刊本では同一の文字に複数の異字体が使われ、多種の短縮語・省略語が使用されている。先行研究では、異字体や短縮語の出現率や使用方法、規則からの逸脱、植字ミスや印刷中の修正（stop-press variant）、現存本間での違いがその書物の植字・印刷の工程を解明する手段となることが示されている。そこで、印刷史を研究する上で重要なこうした着眼点を効果的に記述するための方策を検討する。

② 二つ目の課題は、その方法論に基づき、グーテンベルク聖書のトランスクリプションを作成することである。同一版の諸現存本の間の異同を、その原因や、印刷の順序、技術的要因などの書誌学的分析・研究成果と関連付けてタグ付けするための方法を検討し、実装する。

これによって、方法論を検証・評価することが可能になる。同時に、必要性は認識されながらも実現されてこなかった、グーテンベルク聖書の本文のトランスクリプションを作成することで、今後のグーテンベルク聖書研究の基盤の一つを作ることができると期待できる。

③ 三つ目の課題は、構築したトランスクリプションを用いた調査を行うことにより、グーテンベルク聖書の印刷工程の解明に寄与することである。

３．研究の方法

タグの検討と整理

二方向からアプローチした。一つは、文献調査および初期刊本や同時代の写本のデジタル画像や現資料を調査することによって、これまでの印刷史研究において着目されてきた点を整理し、表現すべきタグとみなした。さらに、初期刊本研究の先導的な研究者からの意見聴取を行った。

もう一つは、前述の TEI のガイドラインを基本的な枠組みとしたうえで、先行事例を参考にして、実際にどのようなタグをどのように付与すべきかを検討するという方向である。レイアウト情報や異字体の記述の方法に関しては、写本のトランスクリプションや松田隆美による 16 世紀の英語の挿絵入り本の XML デジタル・エディションの作成事例を参考にした。

（1）　効率的なトランスクリプション・データ作成方法の検討

グーテンベルク聖書を対象に、四種類の方法でトランスクリプション・データの作成を試みることで、効率的な作成方法を検討した。いずれの方法でも、異字体を識別するためにデータ入力者の訓練を行い、入力マニュアルを整備した。

① 現行のウルガタ聖書の電子テキストデータをもとに人手で修正
② デジタル画像を見ながらすべて人手で入力
③ 既存の光学的文字認識（OCR）ツールを利用
④ 高精細画像データを用いて活字画像を自動認識するための手法の考案、および、それによるデータ作成。なお、高精細画像は慶應義塾図書館から研究用

とでの提供を受けることができた。

（2）　XML による本文の記述

グーテンベルク聖書の本文のテキストデータには既存のものがないため、（2）で検討した方法に基づいてトランスクリプション・データを作成し、（1）で検討したタグ付けを行い、XML の形式で記述した。

（3）　グーテンベルク聖書の書誌学的分析

本研究の（1）～（3）までの成果および先行研究の成果を合わせ、グーテンベルク聖書の印刷工程の詳細についての書誌学的な分析を行った。具体的には、活字の鋳造方法や印刷中の修正作業の分布、修正作業への植字職人の関与などについて分析した。

４．研究成果

（1）タグ

先行研究からは活字の形の識別が重要であることが明らかであるため、当初は異形活字についてもすべてタグ付けすることが適切であると考えていた。たとえば、下の図は、n や m などの省略を示す横棒が上についた小文字 a の四つのバリエーションを示している。①②は文字の左端の角にダイヤモンド型のひげがついているが、③④には付いていない。先行研究により、ひげの有無は、隣にくる文字種との関係で詳細な植字の規則があったことがわかっている。

しかし、（4）で後述するように、本研究に

①	②	③	④

図 1　省略記号付の小文字 a

より、文字の種類は従来考えられてきたよりも多い可能性が明らかになった。また、Agüera y Arcas と Needham は、グーテンベルクの別の印刷物中の「i」の活字画像のクラスタリングを行い、数百のバリエーションがあるという結果を得たことから、金属製の母型と鋳造機を用いたという活字鋳造方法の定説に疑問を呈している。

文字種を詳細に区別してタグ付けを行うことは、作業効率が非常に悪いだけでなく、活字鋳造方法をめぐる新説を考慮すると、不適切である可能性がある。そこで、本研究では、文字種の細かい区別を示すタグは付与しないこととした。

その他の初期刊本についての先行研究で扱われてきたさまざまな点、つまり、省略語、短縮語、活字の向きの誤り、印刷中の修正、手書き文字、手書きの修正、改行は採用することにより、トランスクリプションの有用性を高めると考えられる。

（2）トランスクリプション・データの作成

入力方法として検討した四種類の方法について検討した結果、手法①（現行のウルガタ聖書のデータを利用）は本文内容は大きくは変わらないものの、グーテンベルク聖書における短縮語の多用のため、②（すべて人手で入力）よりも非効率であることがわかった。また、③（OCR）についても、フリーウェアや市販のソフトウェアを施行したが、いわゆる「ひげ文字」と言われる最初期の印刷本に一般的に使われていたゴシック・テクストゥーラ体の書体が使われていること、特殊記号や連字が多用されていること、などの李に対しては非常に認識率が低いことを確認した。

そこで、技術面では他の研究者からの協力を得、素材としては慶應義塾図書館所蔵本の高精細画像の提供を受け、④独自の活字画像自動認識の手法の考案に特に努力を傾注した。その結果、初期刊本の活字の識別に関して効率的な手法を開発することができた。その手順は以下の通りである。

a) 活字画像のコラム単位での分割
b) 装飾や手書き文字の除去などの前処理
c) 傾き補正や明るさの正規化等
d) 前処理後の画像をオープンソースの OCR ソフトの Tesseract-OCR 3.02 に投入し、活字境界識別と文字認識の実行と修正
　この段階では、新たな学習データを作成し、jTessBoxEditor 1.1 を用いて人手で修正し、修正データで学習を行い、それに基づき自動識別をするというサイクルで精度を高めることができた。その結果、弁別が非常に困難な一部の文字を除けば修正が不要な文字認識ができるようになった。
e）テンプレートマッチング

このように、活字画像を自動識別し人手で修正する半自動化によって、活字境界識別やトランスクリプションにかかる労力や時間を大幅に軽減し、正確なデータを作成することができた。同様の手法は、他の初期刊本にも適用できると期待できる。

（3）XML による本文の記述

前述の（2）の方法でテキストデータ化したデータに対してタグを付与し、XML 形式での記述を行った。

（4）グーテンベルク聖書の書誌学的分析

グーテンベルク聖書に使われている文字種は、研究者によって多少意見が異なるが、およそ 200 種類だとされている。Paul Schwenke が 1923 年に発表した一覧表は今日でもよく参照されている。

本研究では、最初期の活字鋳造方法に関する議論をふまえ、まず活字について検討することとした。(2) で識別した活字画像を使い、同じ文字の異なる形状の活字を弁別するた

めに、クラスタリングを行った。SIFT (Scale-Invariant Feature Transform)を用いて局所特徴検出、特徴量記述を行い、そこから活字画像同士の距離を算出した。その距離行列に基づき、オープンソースの統計パッケージ R 3.1.1 上でウォード法によって階層的クラスタリングを行なった。

下の図 2 は、10 ページ分に出現した活字「g」の活字画像を対象としたクラスタリングの結果をデンドログラムである。デンドログラムは高い位置で二つのクラスターに分割できる。上のクラスターには左端のひげがない字形、下にはひげがある字形が集まった。代表的な例として最も端の画像を図に添付した。

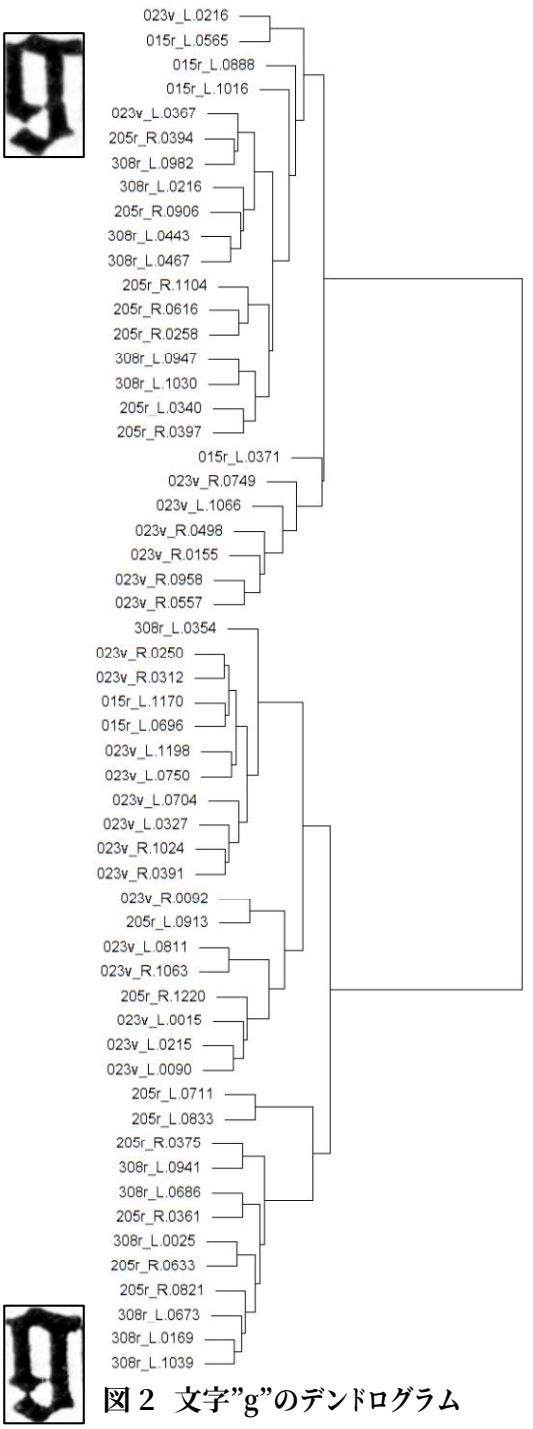

図 2　文字”g”のデンドログラム

また、従来考えられてきたよりも文字種が多いことが明らかになった。たとえば、図 1 の③と④は従来の一覧表では一種類とされてきたものである。さらに、省略を示す横棒の位置や長さにさまざまなバリエーションがあることが明らかになった。ただし、このことが、金属製の母型と鋳造機を用いた際にも生じうるのか、あるいは Agüera y Arcas と Needham が主張するように、一度しか使用できない母型から作られたことを意味するのかについては、さらなる分析が必要である。

本研究の手法で活字を分析することで、Agüera y Arcas と Needham の提示した新説を検討し、活字鋳造方法の解明に寄与する可能性が示された。

また、印刷中に行われた本文の修正作業に着目すると、従来考えられてきた分業のユニットによって、行われている修正の種類に違いがあることがわかった。植字の規則のみを理解していれば行うことができる修正が多いユニットと、ラテン語を理解していなければ行うことができない修正が行われているユニットがあること、特定の箇所で、植字の規則を誤解して本来は正しかったものを間違った活字に差し替える誤修正が行われていることが明らかになった。このことからは、印刷中に行われた修正作業は、親方や校正係など一人の人間の指示によるものというよりは、そのユニットを担当している職人の裁量に任された部分が多かったのではないかと推測できる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① 安形麻理、デジタルが切り開く書誌学の最前線（特集「特集：「書誌」万考 - 書物学・書誌学のいま」）、現代の図書館、第 53 巻 2 号、2014、（掲載決定）、査読無

② 安形輝、安形麻理、活字の識別とその応用：グーテンベルク聖書の活字のクラスタリング、日本図書館情報学会 2014 年度研究大会、2014 年 11 月 29 日、梅花女子大学、第 62 回日本図書館情報学会研究大会発表論文集、p. 117-120、査読無

③ Mari Agata, “Improvements, corrections, and changes in the Gutenberg Bible” in *Scribes, Printers, and the Accidentals of Their Texts*、Thaisen, Jacob; Rutkowska, Hanna eds. Frankfurt am Main, Peter Lang、2011、 p. 135-155 (Series: Studies in English Medieval Language and Literature - Volume 33)、査読有

〔学会発表〕（計３件）

① 安形麻理、デジタル技術を応用した初期印刷本の印刷工程の解明、国際アーサー王学会日本支部 2014 年度年次大会、2014 年 12 月 13 日、龍谷大学大宮学舎（京都府・京都市）

② 安形輝、安形麻理、活字の識別とその応用：グーテンベルク聖書の活字のクラスタリング、日本図書館情報学会 2014 年度研究大会、2014 年 11 月 29 日、梅花女子大学（大阪府・茨木市）

③ 安形麻理、ヨーロッパ初期印刷本研究とデジタル化の技法：グーテンベルク聖書の画像を用いた校合と XML によるコーディング、日本オリエント学会第 54 回大会、2012 年 11 月 25 日、東海大学湘南キャンパス（神奈川県・平塚市）

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
安形　麻理（AGATA, Mari）
慶應義塾大学・文学部・准教授
研究者番号：70433729

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：