Statistical Domain Identification of the Voynich Manuscript: Distributional Mapping to 16th-Century Latin Alchemical and Medical Vocabulary（日英併記版）**Version 1.1.0**

著者：大井 啓嗣 **(Keishi Oi)**
ORCID：*0009-0006-7040-8353*

概要 (Abstract)

約600年間未解読であったヴォイニッチ手稿に対し、構造解明を基盤とする独自手法『OI-2026プロトコル』を適用し、中世のラテン語医学・錬金術文献コーパス(Stream B)との分布的対応関係を体系的に抽出した。多次元空間における文脈的類似度と、手稿の幾何学的配置から自動分類された機能的役割(品詞相当)の二重制約を通過した9,733語(全9,783スロットの99.4%)について、ラテン語コーパス内の対応候補を特定した。本稿ではこの対応関係を『分布的翻訳候補(distributional translation candidate)』と呼ぶ。これは意味論的な翻訳の確定ではなく、後続する歴史学的・薬草学的検証のための一次資料を提供するものである。さらに挿絵の枝葉端点とテキスト内『材料型』語彙の出現率の間に強い正の相関(R=0.7080)が観測され、手稿が記述と図像の両面で構造化された記録媒体であることが示唆された。これらの観察は、手稿が16世紀錬金術・初期医学の伝統に近しい記録体系である可能性を統計的に支持する。

Abstract: By applying the proprietary "OI-2026 Protocol," a method based on structural elucidation, to the Voynich manuscript—which has remained undeciphered for approximately 600 years—we systematically extracted its distributional correspondences with a corpus of medieval Latin medical and alchemical literature (Stream B). For 9,733 words (99.4% of the total 9,783 slots) that passed the dual constraints of contextual similarity in a multi-dimensional space and functional roles (equivalent to parts of speech) automatically classified from the geometric arrangement of the manuscript, corresponding candidates within the Latin corpus were identified. In this paper, these correspondences are referred to as "distributional translation candidates." This does not constitute a definitive semantic translation, but rather provides primary source material for subsequent historical and herbalistic validation. Furthermore, a strong positive correlation (R=0.7080) was observed between the endpoints of branches and leaves in the illustrations and the occurrence rate of "material-type" vocabulary in the text, suggesting that the manuscript is a structured recording medium in terms of both text and iconography. These observations statistically support the possibility that the manuscript is a recording system closely related to the traditions of 16th-century alchemy and early medicine.

キーワード (Keywords): ヴォイニッチ手稿 (Voynich Manuscript)/情報理論 (Information Theory)/実用薬物局方 / 物理操作手順書 (Practical Pharmacopoeia)/グラフ理論 / 端点 (Graph Theory / Endpoints)/OI-2026プロトコル (OI-2026 Protocol)

第**1**章：序論

1.1 ヴォイニッチ手稿研究の背景と意味論的アプローチの限界

15世紀初頭に制作されたと推定されるヴォイニッチ手稿は、1912年の再発見以来、言語学および暗号学の観点から多数の解読が試みられてきた。これまでの研究の大部分は、手稿のテキストを「人間の意思疎通を目的とした自然言語」、あるいは「既知の言語を秘匿した暗号」という前提に立脚している。すなわち、未知の記号列に対して、ラテン語やその他の言語の「意味（セマンティクス）」や「音声」を1対1で対応させようとするアプローチである。

しかし、この意味論的アプローチは、人間の優れたパターン認識能力に起因するアポフェニア（無作為なデータに規則性を見出す錯覚）を誘発しやすい。過去の解読案の多くは、特定の図像や文脈から局所的な単語の推測には成功したように見えながらも、システム全体の構文規則を証明できず、全域に適用した段階で例外処理が氾濫し、最終的に論理的破綻を迎えている。

1.2 構造的アプローチの確立：「**OI-2026**プロトコル」と本研究の目的

『OI-2026プロトコル』とは、著者の前報において、本手稿が自然言語の動的平衡から逸脱した『データマトリクス』として記述可能であり、便宜的に『初期化(Boot)』『代入(Set)』『推移(Transition)』『終了(Termination)』と呼称される4つの位置依存的な機能カテゴリによって特徴づけられることを数理的に示したものである。なお、これらの呼称は情報工学のメタファーであり、手稿が文字通りのオペレーティングシステム構造を持つことを主張するものではない。この成果により、テキスト構造の骨格(中核となる約17のシェル群と、それに従属する開いたペイロード集合からなる『プライマリ・レジスタ・マップ』)を分離することに成功した。しかし、それら構造の内部に具体的にいかなる単語が格納されているかを特定する作業が、本研究における最大の課題として残されていた。したがって、本稿の目的は、OI-2026プロトコルをさらに拡張・適用し、手稿のデータ構造内に残存する未定義スロットに対し、多次元空間での引力(コサイン類似度)と品詞(型)拘束という数学的二重条件を用いて、外部の実在コーパスとの分布的対応関係を体系的に抽出することにある。これにより、人間の意訳を介在させずに、手稿全域における対応候補の一覧を一次資料として提供する。本稿が達成するのは『翻訳の意味的確定』ではなく、後続する歴史学的検証の出発点となる『最も可能性の高いラテン語対応候補の数理的特定』である。前提として、特定されたラテン語の単語は意味論の確定ではなく、あくまで数学的に可能性の高い仮説止まりである。

1.3 比較対象の構築：**Stream A**（手稿データ）と**Stream B**（ラテン語コーパス）の定義

客観的な構造解析を実行するためには、抽出された特徴量と対峙させるための厳密な基準データ

が不可欠である。本研究では、手稿内の未知の文字群（EVA記号）の形態論的な振る舞いを定量化したデータを「Stream A」、比較対象となる現実の歴史的テキストの振る舞いを「Stream B」として定義した。Stream Aは、各記号がどのような接頭辞（Prefix）や接尾辞（Suffix）と結合しているか、あるいは独立状態（NULL）であるかを集計した頻度データである。これにより、手稿のシステム基板である基礎的な振る舞いを抽出する。比較対象となるStream Bには、手稿の放射性炭素年代測定（15世紀初頭）および図像的特徴（植物学、薬学、占星術など）のドメインと歴史的に合致する、中世の初期近代ラテン語コーパスを構築・採用した。具体的には「EMLAP（Early Modern Latin Alchemy and Pharmacy corpus）」および「GreLa」と呼ばれる文献群である。

EMLAPは、当時の革新的な医師・錬金術師であるパラケルスス（Paracelsus）、その著作をラテン語訳したゲラルドゥス・ドルン（Gerardus Dorn）、および偽ルルス（Pseudo-Lullus）などによる錬金術・医学の専門書145件で構成されている。現代の英語による注釈やメタデータを排除する過程で、数件の文献が意図せず除外された可能性はあるが、本研究の優先目的は純度の高いラテン語データセットを構築することにある。これは古代ギリシャから続く体液病理学と、新しい化学的医療（鉱物や化学物質の使用）の過渡期における実用的な記録であり、総単語数約396万語、約19万種類の語彙バリエーションを有する。

1.4 NMFから**TruncatedSVD**への移行と次元同期による客観的照合基盤

素性が全く異なるStream A（手稿の未知変数群）とStream B（約19万種のラテン語彙）を直接比較するためには、双方が持つ形態論的な振る舞いを、共通の数学的ベクトル空間へ変換する「次元削減（Dimensionality Reduction）」が必須となる。初期の解析設計において、我々は非負値行列因子分解（NMF: Non-negative Matrix Factorization）を採用していた。しかし、NMFはその数学的性質上、行列の要素が非負（ゼロまたは正の値）に制約される。この非負空間の制約下でコサイン類似度（Cosine Similarity）を算出すると、ベクトル間の類似度が「0から1の範囲」に限定されてしまう。これは、異なる語彙間の明確な非類似性や反発を示す「マイナスの値（負の相関）」を計算上表現できないという致命的な欠陥を意味しており、多次元トポロジーの照合においてその信憑性を著しく低下させる結果となった。

この数理的限界を解消するため、本研究は次元削減アルゴリズムを「TruncatedSVD（特異値分解）」へと変更した。SVDの適用により、ベクトル空間にマイナスの値が許容され、コサイン類似度における完全な引力（+1）から明確な斥力（-1）までの連続的な分布が正確に計算可能となった。このアルゴリズムの変更により、極めてスパース（疎）であったStream AとStream Bの初期データは、ゼロによる計算の歪みが根絶された「56次元の密行列（Dense Matrix）」という共通の数学的ベクトル空間へと変換（次元同期）された。さらに本研究は、後の解析フェーズにおいて文脈的特徴（マルコフ遷移および空間座標）を加え、この解像度を「128次元」へと拡張する。

なお、本稿で特定した対応語彙群は、あくまで分布的特徴に基づく数理的仮説に留まり、意味論的な確定を意味するものではない点に留意されたい。意味論的確定には後述する学際的な交差検証が不可欠であり、本稿の目的は統計的に最も蓋然性の高い仮説の提示に限定される。

--- (English Translation) —

Chapter 1: Introduction

1.1 Background of Voynich Manuscript Research and Limitations of Semantic Approaches

The Voynich manuscrip**t**, estimated to have been produced in the early 15th century, has been the subject of numerous decipherment attempts from linguistic and cryptographic perspectives since its rediscovery in 1912. The vast majority of previous studies operate on the premise that the text of the manuscript represents either a "natural language intended for human communication" or a "cipher concealing a known language." Specifically, these approaches attempt a one-to-one mapping of the "meaning" (semantics) or "phonetics" of Latin or other languages onto the unknown sequences of symbols.

However, this semantic approach is highly susceptible to inducing apophenia—the cognitive illusion of finding meaningful patterns in random or unstructured data. While many past decipherment proposals appeared successful in conjecturing localized words from specific illustrations or contexts, they failed to demonstrate grammatical rules valid for the entire system. Consequently, when applied globally to the entire manuscript, these methods became inundated with exception handling and ultimately faced logical collapse.

1.2 Establishment of a Structural Approach: The "OI-2026 Protocol" and Objectives of This Study

The "OI-2026 Protocol" refers to the methodology established in the author’s previous report, which mathematically demonstrated that this manuscript can be described as a "data matrix" deviating from the dynamic equilibrium of natural languages. This matrix is characterized by four position-dependent functional categories provisionally designated as "Boot," "Set," "Transition," and "Termination." Note that these designations serve as information engineering metaphors and do not imply that the manuscript literally possesses an operating system architecture. Through this framework, the foundational skeleton of the text structure—the "primary register map," comprising a core set of approximately 17 shells acting as informational receptacles and a subordinate open set of payloads—was successfully isolated. However, the task of identifying the specific words stored within these structural slots remained the primary challenge for the present study.

Therefore, the objective of this paper is to further extend and apply the OI-2026 Protocol to systematically extract the distributional correspondences between the remaining undefined slots within the manuscript's data structure and an external real-world corpus, utilizing the dual mathematical constraints of contextual attraction (cosine similarity) and part-of-speech (type) restriction in a multi-dimensional space. This approach provides a comprehensive mapping of corresponding candidates across the entire manuscript as a primary source, completely

bypassing subjective human interpretation. What this paper achieves is not the definitive semantic establishment of a translation, but rather the mathematical identification of the most probable Latin corresponding candidates, serving as a departure point for subsequent historical validation. As a prerequisite, the identified Latin words do not represent finalized semantics but remain strictly as mathematically probable hypotheses.

1.3 Construction of Comparative Baselines: Definitions of Stream A (Manuscript Data) and Stream B (Latin Corpus)

To execute an objective structural analysis, rigorous reference data to confront the extracted features is indispensable. In this study, data quantifying the morphological behavior of the unknown character groups (EVA symbols) within the manuscript is defined as "Stream A," while the behavior of real, historical texts used as a baseline for comparison is defined as "Stream B." Stream A consists of frequency data compiling how each symbol combines with prefixes or suffixes, or whether it exists in an independent state (NULL). This process extracts the fundamental behaviors that constitute the system substrate of the manuscript. For the comparative baseline, Stream B, we constructed and adopted an early modern Latin corpus that historically aligns with the domain of the manuscript's radiocarbon dating (early 15th century) and its iconographic characteristics (botany, pharmacology, astrology, etc.). Specifically, this comprises the literature groups known as "EMLAP" (Early Modern Latin Alchemy and Pharmacy corpus) and "GreLa."

EMLAP consists of 145 specialized texts on alchemy and medicine by figures such as Paracelsus, an innovative physician and alchemist of the era; Gerardus Dorn, who translated his works into Latin; and Pseudo-Lullus. While a few texts may have been inadvertently excluded during the process of removing modern English commentary and metadata, the primary priority of this study remains the construction of a high-purity Latin dataset. This corpus represents a practical record from the transitional period between classical humoral pathology and emerging chemical medicine (the utilization of minerals and chemical substances), comprising approximately 3.96 million total words with a vocabulary variation of roughly 190,000 unique terms.

1.4 Transition from NMF to TruncatedSVD and the Objective Matching Foundation via Dimensional Synchronization

Directly comparing Stream A (the unknown variables of the manuscript) and Stream B (approximately 190,000 Latin vocabulary terms), which possess completely different underlying properties, necessitates "dimensionality reduction" to transform the morphological behaviors of both sides into a shared mathematical vector space. In the initial analytical design, we adopted Non-negative Matrix Factorization (NMF). However, NMF mathematically restricts matrix elements to non-negative values (zero or positive values). Calculating cosine similarity under this non-negative spatial constraint limits the similarity between vectors to a range of 0 to 1. This represents a critical deficiency: the inability to computationally represent "negative values" (negative correlations) indicating explicit dissimilarity or repulsion between different

vocabularies, which severely diminished the credibility of multi-dimensional topological matching.

To resolve this mathematical limitation, this study transitioned the dimensionality reduction algorithm to TruncatedSVD (Truncated Singular Value Decomposition). The application of SVD allows negative values within the vector space, enabling the precise computation of a continuous distribution ranging from complete attraction (+1) to explicit repulsion (-1) in cosine similarity. This algorithmic shift transformed the initial data of Stream A and Stream B, which was extremely sparse, into a shared mathematical vector space designated as a "56-dimensional dense matrix," thereby eliminating computational distortions caused by zero elements. Furthermore, in a subsequent analysis phase, this study extends the resolution to "128 dimensions" by integrating contextual features (Markov transitions and spatial coordinates).

It must be noted that the corresponding vocabulary groups identified in this paper remain strictly mathematical hypotheses based on distributional characteristics and do not signify definitive semantic decipherment. Semantic validation requires subsequent interdisciplinary cross-verification, and the objective of this paper is confined to presenting the statistically most probable hypotheses.

第**2**章：統計的指標と生成規則の構造的制約

2.1 統計的な非言語性と「数学的真空」の証明

本手稿に対する客観的な全域翻訳を実行するにあたり、まず前報において証明された本手稿の基礎的な統計指標と構造の制約を再確認する。これらの指標は、本手稿がいかなる情報法則の支配下にあるかを示す基準となる。手稿のテキスト構造は、自然言語が必然的に内包する統計的なゆらぎから完全に逸脱している。前報での計測の通り、テキスト全体のワード・エントロピーは11.05 bitsという極めて偏った異常値を示し、単語の出現頻度分布を示すZipf則のアルファ係数は0.6〜0.8を記録している。これは自然言語特有の「頻出語のピーク」が存在しない、人工的に平坦化された構造である可能性を示唆している。さらに、全語彙におけるハパックス・レゴメナ（一度しか出現しない語）の割合は72.93%にもなる。

さらに本研究において特筆すべきは、文字列の組み合わせ空間における極限の隙間（疎性）である。長さ5の記号列における理論上の最大組み合わせ数（約6.75億通り）に対し、実際に手稿内で観測されるユニークな単語数は2,716通りに過ぎず、空間の占有率はわずか「0.000402%」にとどまる。この数学的真空の存在は、本手稿が自然言語であることを否定するだけでなく、カルダノ・グリルなどの乱数表を用いた無作為な文字生成による「捏造（Hoax）説」を数学的に可能性を低くする。もし無作為な乱数生成であれば、組み合わせの空間はより広範かつ無秩序に埋め尽くされるはずである。

2.2 決定論的な記号の連鎖と**4**段階の論理構造

このような数理的制約を生み出している生成規則の正体は、記号間の移行確率の解析によって明らかとなっている。本手稿のテキスト内には、特定の記号から次の記号への移行において、極めて強固で機械的な「固定された連鎖（ロック）」が存在する。

例えば、テキスト内で最も高頻度に出現する記号「4」の直後に「o」が配置される確率は、初期のセクション（Folio 1-60）において97.88%、後期のセクション（Folio 65-116）においても97.47%という極端な数値を記録している。執筆時期やセクションを跨いで、特定の記号の連鎖がわずか0.41%の誤差で維持される現象は、人間の自由意志や筆写の癖（ゆらぎ）を伴う言語活動では発生する可能性は低い。

この決定論的な記号の固定化は、テキストが執筆者の自由な記述ではなく、あらかじめ定義された機械的な手順に従って出力されていることを示唆している。具体的には、テキストの各まとまりが、メタファーではあるが「初期化（Boot）」「代入（Set）」「推移（Transition）」「終了（Termination）」という4段階の属性に分類されるという論理構造によって制御されており、それぞれの段階で配置が許可される記号が構造的に制限されているのである。

2.3 構造の解明：**17**の基本枠組みと基礎変数

この4段階の機能カテゴリに基づきテキスト全体を解析した結果、本手稿の情報構造は、中核となる約17のシェル群(『情報の器』として機能)と、それに従属する開いたペイロード集合という非対称な構造を示すことが確認された。シェル側は約17付近で結合多様性に明確な落差を示すのに対し、ペイロード側はべき乗則的に連続して減衰し、明確な境界を持たない。本研究では、この非対称構造を便宜的に『プライマリ・レジスタ・マップ』と呼称する。

本研究における『分布的翻訳』とは、未知の言語から文脈を意訳することではない。この数学的に特定された『17の基本枠組みと基礎変数』からなる構造体の空白（未知の変数領域）に対し、実在する歴史的テキストデータとの厳密な数理的照合を行う。これにより、各変数スロットに対応する可能性が最も高いラテン語候補を客観的に特定すると同時に、前報で仮定した『錬金術的ドメイン』の妥当性を検証する。本稿の最終的な目的は、後続する意味論的解読のための堅牢な一次資料を提供することである。

--- (English Translation) —

Chapter 2: Statistical Indicators and Structural Constraints of Generative Rules

2.1 Statistical Non-Linguistic Nature and Proof of the "Mathematical Vacuum"

In executing an objective full-text translation of this manuscript, we first re-examine the fundamental statistical indicators and structural constraints of the manuscript proven in our previous report. These indicators serve as criteria demonstrating under what information laws

the manuscript is governed. The text structure of the manuscript completely deviates from the statistical fluctuations inevitably inherent in natural languages. As measured in the previous report, the word entropy of the entire text exhibits an extremely skewed anomalous value of 11.05 bits, and the alpha coefficient of Zipf's law, which indicates the word frequency distribution, records between 0.6 and 0.8. This suggests the possibility of an artificially flattened structure lacking the "peak of frequent words" characteristic of natural languages. Furthermore, the proportion of hapax legomena (words occurring only once) across the entire vocabulary reaches as high as 72.93%.

Furthermore, particularly noteworthy in this study is the extreme gap (sparsity) within the combinatorial space of character strings. In contrast to the theoretical maximum number of combinations for a symbol sequence of length 5 (approximately 675 million), the number of unique words actually observed in the manuscript is merely 2,716, resulting in a spatial occupancy rate of only "0.000402%". The existence of this mathematical vacuum not only negates that the manuscript is a natural language but also mathematically reduces the probability of the "hoax theory" based on random character generation using random number tables such as the Cardan grille. If it were random number generation, the combinatorial space should be filled more broadly and chaotically.

2.2 Deterministic Symbol Chains and Four-Stage Logical Architecture

The true nature of the generative rules producing such mathematical constraints has been revealed through the analysis of transition probabilities between symbols. Within the text of the manuscript, there exists an extremely rigid and mechanical "fixed chain (lock)" in the transition from a specific symbol to the next.

For example, the probability of "o" being placed immediately after "4", the most frequently occurring symbol in the text, records extreme values of 97.88% in the early section (Folio 1-60) and 97.47% in the late section (Folio 65-116). The phenomenon where a specific symbol chain is maintained with an error margin of merely 0.41% across writing periods and sections is highly unlikely to occur in linguistic activities involving human free will or writing habits (fluctuations).

This deterministic fixation of symbols suggests that the text is not the free writing of the author but is output according to pre-defined mechanical procedures. Specifically, each unit of text is controlled by a logical architecture classified into four stages of attributes—metaphorically termed "Boot", "Set", "Transition", and "Termination"—and the symbols permitted to be placed at each stage are structurally restricted.

2.3 Elucidation of Structure: The 17 Operational Frameworks and Basic Variables

As a result of analyzing the entire text based on these four stages of functional categories, it was confirmed that the information structure of the manuscript exhibits an asymmetrical structure comprising a core set of approximately 17 shells (functioning as "informational receptacles") and a subordinate open set of payloads. While the shell side demonstrates a

distinct cliff in combinatorial diversity around 17, the payload side continuously attenuates in a power-law manner and possesses no clear boundary. In this study, we provisionally refer to this asymmetrical structure as the "primary register map."

"Distributional translation" in this study does not mean freely translating the context from an unknown language. We perform rigorous mathematical matching with actual historical text data against the blanks (unknown variable domains) of the structure composed of these mathematically identified "17 operational frameworks and basic variables." Through this, we objectively identify the most probable Latin candidates corresponding to each variable slot, while simultaneously validating the validity of the "alchemical domain" hypothesized in the previous report. The ultimate objective of this paper is to provide robust primary source material for subsequent semantic decipherment.

第**3**章：**56D**空間における初期アンカーの抽出

3.1 Stream Aのシステム基板に合わせた**56**次元空間の同期

前章までに証明された生成規則を前提とし、本章ではテキストの構造内に残された未知の記号群に対し、具体的なラテン語彙を特定していくプロセスを記述する。初めに、手稿内部の記号の形態論的な振る舞いを数値化したデータ（Stream A：3,808の未知の記号）と、16世紀のラテン語文献群から抽出した振る舞いのデータ（Stream B）を比較するための処理を行った。

Stream Aの抽出過程において、各未知の記号が持つ接頭辞・接尾辞との結合パターンおよび独立状態（NULL）を集計した結果、手稿のシステム基板は「56種類」の振る舞いを持つことが特定されている。この手稿固有の「56」という基準次元数に合わせるため、ラテン語コーパス（Stream B）に対しても「TruncatedSVD（特異値分解）」を適用し、次元を同期させた。なお、Stream Bのコーパス全体には約19万の語彙バリエーションが存在するが、行列計算においてノイズとなる極端な低頻度語をフィルタリングした結果、照合対象となる有効なラテン語彙は39,354語へと絞り込まれている。この次元同期により、データ間の計算上の歪みとなる疎行列の隙間（ゼロ要素）を解消し、Stream AとStream Bを「56次元の密行列（Dense Matrix）」という共通空間で客観的に照合する基盤を確立した。

3.2 相互最近傍と特異ギャップによる分布的対応候補の特定基準の抽出

構築された56次元空間において、Stream Aの全記号とStream Bの全ラテン語彙との間でコサイン類似度を用いた総当たり計算を行った。ここでは、人間の推論による恣意的な対応付けを排除するため、厳格な二段階の数理的条件を適用した。

第一の条件は「相互最近傍（MNN: Mutually Nearest Neighbors）」の抽出である。これは、手稿の記号とラテン語が互いに最も高い類似度を示す「1対1の双方向なペア」のみを採用し、片方向のみの類似はすべて破棄する処理である。第二の条件は、類似度における「特異ギャップ」の証明である。抽出されたペアについて、類似度の1位と2位の差（ギャップ）を計測し、空間全体の分布からZス

コア（標準スコア）を算出した。その結果、偶然では発生し得ない「Zスコア 2.0以上（上位約2%の異常値）」という特異な断崖を持つペアのみを最終合格とした。

これらの条件をクリアし、人間の介入を一切許さずに初期の足場として確定した7件のペアが『Voynich_Absolute_Anchors_Strict.csv』である。条件を満たさなかった記号は無理に解釈せず「未定義」として保留されており、これが手稿とラテン語を結びつける初期基準として機能する。

高次元空間における照合の性質上、ハブ化（Hubness）現象等により、意味的関連性が低くとも特定の高頻度語が対応候補として割り当てられるリスクが存在する。現段階の数理的アプローチのみでは個々の単語の真偽判定は困難であるため、本稿ではこれらを『分布的対応候補（仮定単語）』として保持し、最終的な真偽判定は次フェーズの意味論的検証に委ねる。

3.3 確定語彙の物理的削除（パージ）と反復アルゴリズム

初期のアンカーを仮確定させた後、より多くの記号を連鎖的に特定するための反復処理を導入した。まず、すでに特定された7件の記号とラテン語を、Stream AおよびStream Bの空間から物理的に削除（Drop）した。この削除処理の目的は、確定した語彙が持つ強大な引力が他の未知の記号を誤って吸い寄せてしまう「空間の重複」をシステムレベルで未然に遮断することである。縮小した新たな空間同士で再びコサイン類似度の総当たり計算を行い、前述の「相互最近傍（MNN）」および「Zスコア2.0以上」の条件を適用して新規のアンカーを抽出する。新たなアンカーが確定するたびに即座に空間から削除し、また総当たり計算を行うという反復処理（Whileループ）を、新規のペアが1つも発見されなくなるまで実行した。この厳密な連鎖処理によって特定された結果は『Voynich_Absolute_Anchors_Phase4.csv』として記録された。この数理的連鎖により、手稿の構造解明は、次なる高解像度の段階へと移行する。

--- (English Translation) ---

Chapter 3: Extraction of Initial Anchors in 56D Space

3.1 Synchronization of the 56-Dimensional Space to the System Substrate of Stream A

Assuming the generative rules proven in the previous chapters, this chapter describes the process of identifying specific Latin vocabulary for the unknown symbol groups remaining within the text structure. First, processing was performed to compare data quantifying the morphological behavior of the symbols within the manuscript (Stream A: 3,808 unknown symbols) with behavior data extracted from the 16th-century Latin literature corpus (Stream B).

In the extraction process of Stream A, aggregating the combination patterns with prefixes and suffixes, as well as the independent state (NULL) for each unknown symbol, identified that the manuscript's system substrate possesses 56 types of behavior. To align with this baseline dimensionality of 56 inherent to the manuscript, Truncated Singular Value Decomposition (TruncatedSVD) was applied to the Latin corpus (Stream B) to synchronize the dimensions.

Note that while the entire Stream B corpus contains approximately 190,000 vocabulary variations, filtering out extreme low-frequency words that act as noise in matrix calculations narrowed down the valid Latin vocabulary for matching to 39,354 words. This dimensional synchronization eliminated the gaps in sparse matrices (zero elements) that cause computational distortions between data, establishing a foundation to objectively match Stream A and Stream B within a common space called a 56-dimensional dense matrix.

3.2 Extraction of Identification Criteria for Distributional Translation Candidates via Mutually Nearest Neighbors and Singular Gaps

Within the constructed 56-dimensional space, brute-force calculations using cosine similarity were performed between all symbols of Stream A and all Latin vocabulary of Stream B. Here, to eliminate arbitrary mapping based on human inference, strict dual mathematical conditions were applied.

The first condition is the extraction of Mutually Nearest Neighbors (MNN). This is a process that adopts only one-to-one bidirectional pairs where a manuscript symbol and a Latin word exhibit the highest similarity to each other, discarding all unidirectional similarities. The second condition is the proof of a singular gap in similarity. For the extracted pairs, the difference (gap) between the first and second highest similarities was measured, and the Z-score (standard score) was calculated from the distribution of the entire space. As a result, only pairs possessing a singular cliff of a Z-score of 2.0 or higher (anomalous values in the top approximately 2%), which could not occur by chance, were ultimately accepted.

The 7 pairs that cleared these conditions and were established as initial footholds without permitting any human intervention are recorded in "Voynich_Absolute_Anchors_Strict.csv". Symbols that did not meet the conditions were not forcibly interpreted but retained as undefined, functioning as the initial baseline connecting the manuscript and Latin.

Due to the nature of matching in high-dimensional space, phenomena such as hubness present a risk where specific high-frequency words might be assigned as corresponding candidates even if their semantic relevance is low. Since determining the true or false nature of individual words is difficult with the mathematical approach alone at this stage, this paper retains these as distributional corresponding candidates (hypothetical words), leaving the final validation to the semantic verification in the subsequent phase.

3.3 Physical Deletion (Purging) of Established Vocabulary and the Iterative Algorithm

After provisionally establishing the initial anchors, an iterative process was introduced to systematically identify more symbols in a chain reaction. First, the 7 already identified symbols and Latin words were physically deleted (dropped) from the spaces of Stream A and Stream B. The purpose of this deletion process is to systematically preempt spatial overlap, where the massive gravitational pull of established vocabulary mistakenly attracts other unknown symbols. Brute-force calculations of cosine similarity were performed again between the newly reduced

spaces, and new anchors were extracted by applying the aforementioned conditions of Mutually Nearest Neighbors (MNN) and a Z-score of 2.0 or higher. An iterative process (While loop)—immediately deleting newly established anchors from the space and performing brute-force calculations again—was executed until not a single new pair was discovered. The results identified through this rigorous chain process were recorded as "Voynich_Absolute_Anchors_Phase4.csv". Through this mathematical chain, the structural elucidation of the manuscript transitions to the next high-resolution phase.

第**4**章：**128D**高解像度空間への拡張と連鎖的抽出

4.1 文脈的特徴量の抽出と**128**次元空間への拡張

前章における56次元空間での初期照合により、手稿を統制する基礎的な振る舞い（接頭辞および接尾辞の結合頻度）に基づく翻訳基準が抽出された。しかし、より広範な未知の記号を正確に特定するためには、単語単体の振る舞いだけでなく、テキスト全体における「文脈（コンテキスト）」の数学的構造を比較空間に組み込む必要がある。ここで最も留意すべきは、「この単語は動詞であるはずだ」といった人間の意味論的な推論や分類を排除することである。

前報の結果を鑑みて、本研究では手稿のテキスト（Stream A）とラテン語文献（Stream B）の双方から、純粋な物理的特徴量のみを新たに抽出した。第一の特徴は「マルコフ遷移（N-gram引力）」である。これは、ある単語の直前および直後に出現した単語のネットワーク（共起頻度）を数値化したものである。第二の特徴は「空間的配置」である。これは、ある単語が行や文の中において「先頭」「中間」「末尾」のどこに配置されやすいかという相対的な位置分布を示している。

これらの特徴量を基に、新たなクロス集計行列を生成した。そして、この巨大なデータに対して再び「TruncatedSVD（特異値分解）」を適用し、空間の解像度を「128次元」へと拡張・固定した。この処理により、手稿側とラテン語側の双方は、文字の遷移確率と空間配置という高度な文脈情報を含んだ、同一次元（128次元）の密な行列へと同期された。

4.2 超重力の遮断と連鎖的な照合処理の起動

解像度が引き上げられた新たな128次元空間において、全域の照合を実行する。本フェーズ（Phase 5）における入力データは以下の通りである。内部の未知の記号群データである『Stream_A_ContextDense_128D_v5.csv』（9,855の未知の記号）、外部の実在するラテン語彙群データである『Stream_B_ContextDense_128D_v5.csv』（173,867のラテン語彙）、そして前章のPhase 4までに初期の足場として仮確定した19件の翻訳基準を記録した『Voynich_Absolute_Anchors_Phase4.csv』である。

128次元空間での総当たり計算を開始する前に、まずこれら19件の確定済みペアを、双方の行列空間から物理的に完全に切除した。この処理の目的は、前述と同じく、すでに特定が確定している語

彙が持つ強大な引力（超重力）が、他の未知の記号を誤って吸い寄せてしまう「空間の重複」を数学的に未然に防ぐことにある。

4.3 128D空間における特異ギャップの証明と新たな翻訳基準の抽出

確定済みの要素が切除され、純化された新しい128次元空間同士で、コサイン類似度による総当たり計算を再度実行した。ここでも人間の恣意的な介入を防ぐため、前章と同様の条件を適用した。

第一に、「相互最近傍（MNN: 1対1の両思い）」の条件である。手稿の記号とラテン語が、互いに最も高い類似度を示す双方向のペアのみを抽出し、片方向のみの類似はすべて破棄した。 第二に、類似度における特異な断崖の証明である。抽出されたペアの類似度第1位と第2位の差（ギャップ）を計算し、空間全体の分布からZスコアを算出した。「Zスコア 2.0以上」という、偶然では発生し得ない極端な数学的断崖を持つペアのみを、新規の翻訳基準として確定した。

新たな翻訳基準が確定するたびに、即座にそれを空間から物理的に削除し、再び総当たり計算を行うという反復処理を、新たなペアが1件も抽出されなくなるまで実行した。この純粋な数理的連鎖によって抽出された最終結果は、『Voynich_Absolute_Anchors_Phase5_128D.csv』として記録された。条件をクリアできなかった残存記号は、意味論によるこじつけを一切許容せず、「未定義」の記号として保留され、次のフェーズへと引き継がれる。

--- (English Translation) ---

Chapter 4: Extension to the 128D High-Resolution Space and Chained Extraction

4.1 Extraction of Contextual Features and Extension to the 128-Dimensional Space

Through the initial matching in the 56-dimensional space in the previous chapter, translation criteria based on the fundamental behaviors (combination frequencies of prefixes and suffixes) governing the manuscript were extracted. However, to accurately identify a broader range of unknown symbols, it is necessary to incorporate into the comparison space not only the behavior of individual words but also the mathematical structure of the "context" across the entire text. What must be noted most here is the strict exclusion of human semantic inferences or classifications, such as assuming "this word must be a verb."

Considering the results of the previous report, this study newly extracted purely physical features from both the manuscript text (Stream A) and the Latin literature (Stream B). The first feature is "Markov transitions (N-gram attraction)." This quantifies the network (co-occurrence frequency) of words appearing immediately before and after a given word. The second feature is "spatial positioning." This indicates the relative positional distribution of where a word is likely to be placed—such as at the "beginning," "middle," or "end"—within a line or sentence.

Based on these features, a new cross-tabulation matrix was generated. Then, Truncated

Singular Value Decomposition (TruncatedSVD) was applied again to this massive data, extending and fixing the spatial resolution to "128 dimensions." Through this process, both the manuscript side and the Latin side were synchronized into dense matrices of the same dimensionality (128 dimensions), incorporating advanced contextual information consisting of character transition probabilities and spatial positioning.

4.2 Blocking of Supergravity and Initiation of Chained Matching Processes

In the newly established 128-dimensional space with increased resolution, matching across the entire domain is executed. The input data for this phase (Phase 5) are as follows: "Stream_A_ContextDense_128D_v5.csv" (9,855 unknown symbols), representing the internal unknown symbol group data; "Stream_B_ContextDense_128D_v5.csv" (173,867 Latin vocabulary items), representing the external real Latin vocabulary group data; and "Voynich_Absolute_Anchors_Phase4.csv", which records the 19 translation criteria provisionally established as initial footholds up to Phase 4 in the previous chapter.

Before initiating the brute-force calculations in the 128-dimensional space, these 19 established pairs were first physically and completely excised from both matrix spaces. The purpose of this process, identical to the previous chapter, is to mathematically preempt "spatial overlap," where the massive gravitational pull (supergravity) of already identified vocabulary mistakenly attracts other unknown symbols.

4.3 Proof of Singular Gaps in the 128D Space and Extraction of New Translation Criteria

With the established elements excised, brute-force calculations utilizing cosine similarity were re-executed between the purified new 128-dimensional spaces. Here again, to prevent arbitrary human intervention, the same conditions as in the previous chapter were applied.

The first condition is the extraction of "Mutually Nearest Neighbors (MNN)." Only bidirectional pairs where a manuscript symbol and a Latin word exhibited the highest similarity to each other were extracted, and all unidirectional similarities were discarded. The second condition is the proof of a singular cliff in similarity. For the extracted pairs, the difference (gap) between the first and second highest similarities was calculated, and the Z-score was derived from the distribution of the entire space. Only pairs possessing an extreme mathematical cliff of a "Z-score of 2.0 or higher," which could not occur by chance, were established as new translation criteria.

An iterative process—immediately removing a newly established translation criterion physically from the space and re-executing the brute-force calculation—was performed until not a single new pair could be extracted. The final results extracted through this purely mathematical chain process were recorded as "Voynich_Absolute_Anchors_Phase5_128D.csv". The remaining symbols that failed to clear the conditions were retained as "undefined" symbols, strictly

precluding any forced semantic rationalization, and are carried over to the next phase.

第5章：全体の構造逆解析と「実用薬物局方」の現像

5.1 全体構造の逆解析と未知記号の保護

前章までの処理により高次元空間から抽出された確実な翻訳基準を用いて、手稿全体のテキストに対する機械的な置換を実行した。辞書に存在しない未知の記号群は無理に当てはめず「未定義」として保護したまま出力した。この中間生成された記録が『Voynich_Decompiled_Record.txt』である。本処理の目的は、手稿を統制する構造のどこに「未知の空白」が残存しているかを客観的に定着させることにある。

5.2 特徴量の抽出と「型」の自動決定

残存する未定義を特定するためには、その空白が「名詞（材料）」や「動詞（操作）」といったいかなる品詞的役割を要求しているかを数理的に導き出すことが重要であり鍵となる。未定義の記号群に対し、4段階の論理構造（初期化/代入/推移/終了）における行内位置や、隣接する確定済みラテン語の引力といった「純粋な幾何学的特徴量」を抽出した。

抽出された特徴量に対して分類アルゴリズムを適用した。この際、分類の数を恣意的に決めるのではなく、「シルエットスコア」と呼ばれる空間の分離度を測る数理的評価関数を用いて、最も自然に分離される最適な「型（品詞などの機能的役割）」の数をシステムに決定させた。この客観的な動的分類によって各空白の型が特定された結果が『Voynich_Undefined_Variable_Types.csv』であり、それをテキスト構造に反映させたものが『Voynich_Schema_Mapped_Translation.txt』である。

5.3 直交する二重防壁と全域での分布的対応候補の特定

型の特定が完了したテキスト構造に対し、実在のラテン語彙を特定・結合させる最終翻訳処理を実行した。本研究はここで「マクロな位相（128次元SVD空間での文脈引力）」と「ミクロな機能（自動決定された型との整合性）」という、直交する二つの厳格な数理的制約（制約条件）を導入した。すなわち、全体の文脈的引力がどれほど高くとも、手稿側が要求する「型（品詞）」が一致しなければ破棄されるという厳格な照合である。

この二重条件を用いてラテン語コーパス(Stream B)と照合を行った結果、手稿内に残存していた全9,783件の未定義スロットに対し、99.4%(9,733件)の割合で、二重条件を通過する分布的対応候補を特定することができた。これは『意味論的な翻訳の確定』ではなく、『手稿の各スロットに対し、文脈類似度と型整合性の双方を満たすラテン語が、コーパス内に高い割合で存在する』という事実を示すものである。手稿全域の最終的な対応候補マッピングは『Voynich_Absolute_Translation_Final.txt』として記録された。これらの対応候補が意味論的に正し

い翻訳であるかは、本研究の数理的処理だけでは保証されず、歴史学的・薬草学的な多角的な検証を必要とする。

5.4 全域での分布的対応候補の特定：実用薬物局方としての記録の可能性

この99.4%という特定率が、文脈の推測によるこじつけではないことは、特定された語彙の振る舞いによって証明される。例えば、手稿内の「1oe」という未知の記号は、ラテン語の接続詞「quod（～ということ、～なので）」へと数理的に収束した。この「1oe」は前段の自動分類において「論理的な接続・前置（Type_6）」という型に固定されており、特定されたラテン語の実際の品詞と一致していることが確認できる。

しかし、複数の異なる未知記号が同一のラテン語に収束する「同音異義（Homophones）」とも取れる存在も数理的に発見された。これは同音異義語を確定させるものではない。<CUSCUTA>（ネナシカズラ / 材料名）に収束するEVA群は、soy9（位置：1r.2）/s9aly（位置：1r.3）/h1s9（位置：1v.1）/h98an9（位置：2r.1）/8oy1oy9（位置：2r.2）/hoom（位置：2v.1）/h1A（位置：2v.2）/k2cos（位置：3r.1）となる。これは、分布的対応候補の特定に失敗したのではなく、これらのEVA群に<CUSCUTA> と同じ文脈（レシピの材料スロット）で使われる、それに近い別の材料（近似の物質）」である可能性が極めて高いということを示している。

現在の「類似度ベースの照合」と「型枠への押し込み（二重防壁）」による抽出では、「型の役割（材料が入るスロット）」までは客観的に特定できても、そのスロットに代入される「個別の物質の違い」までは分離できないのである。しかし、ヴォイニッチ手稿が未知の暗号体系によって記述されている点や、文脈への強引な同化（過学習）が容易に成立し得るというテキストの特性を鑑みると 、同音異義語の可能性や高度に暗号化された可能性も排除しきれない。現時点では、そのどれをも判別可能な客観的な証拠はない。

その半面、他の抽出された対応候補群には『INCENDERE(点火せよ)』や『ALCALISATUS(アルカリ化された)』といった、中世錬金術・薬物処方に典型的な操作動詞・状態動詞が頻出することが確認された。この事実は、Stream Bとして錬金術・初期医学のラテン語コーパスを選定した妥当性を支持するものであり、手稿が物語や詩歌ではなく操作・材料の記録に類する性格を持つという仮説と整合する。ただし、個別の単語対応(例えば fachys → MOUENTE)が意味論的に正しいかについては、本研究の数理的処理だけでは確定できず、ラテン語学・科学史の専門家による精査を必要とする。これは現時点での『最も可能性の高い対応候補』であり、最終的な確定は次フェーズの学際的検証に委ねられる。

5.5 翻訳出力記録の提示と「レシピ的構造」の実証

前節で提示した、本手稿が中世の「実用薬物局方（作業手順書）」であるという仮説は、最終的な翻訳出力ファイル『Voynich_Multilingual_Pipeline.txt』の記録によって、より具体的に確認することができる。これまで多くの優れた言語学者や歴史学者が、手稿の文字列を自然言語による連続した文章として解読しようと試みてきた。その長年にわたる多大な努力と蓄積には深く敬意を表する。本

研究の数理的な制約を経て抽出されたラテン語彙の配列を観察すると、古典ラテン語のような複雑な文法規則（格変化や性数一致などの形態論的屈折）を伴う散文ではなく、特定の単語が規則的に並ぶ構造が浮かび上がってくる。

以下は、手稿の最も最初のページ（Folio 1r）の冒頭部分からの翻訳出力記録の抜粋である。未知の記号列（EVA）に対し、抽出された実在のラテン語（LAT）がどのように結びついているかを行単位で示している。

①EVA : fa19s 9 hae ay Akam 2oe !oy9 scs 9 hoy 2oe89
②LAT : <MOUENTE> <PERFECTIONEM> <EUADERET> <ADGLUTINATIS> <MARKASITA> <VARIARETUR> <LOCANDO> <PHOR> <PERFECTIONEM> <TACITUR> <INUENISTIS>

①EVA : soy9 Hay oy 9 hacy 1kam 2ay Ais Kay Kay 8aN
②LAT : <CUSCUTA> <UERTENTES> <DETINERE> <PERFECTIONEM> <LUXATIS> <EUANESCENT> <PRAEUIDUARE> <EUPHRAGIAE> <PASTORIS> <PASTORIS> <SURREPTUM>

この出力結果を見ると、主語と動詞が照応し、複雑な意味や文脈を形成するような文学的な文章構造は見受けられない。手稿の冒頭を飾る <1r.1> にも、物語の始まりを示すような言葉はなく、「動かすこと（MOUENTE）」「完璧さを（PERFECTIONEM）」「接着された（ADGLUTINATIS）」「白鉄鉱（MARKASITA）」といった、物理的な操作と材料を示す単語が連続して現れている。続く <1r.2> においても、「ネナシカズラ（CUSCUTA）」という明確な植物名から始まり、「回転させる（UERTENTES）」「保持する（DETINERE）」といった作業指示が続く。後半には「コゴメグサ（EUPHRAGIAE）」という別の薬草の名称も指定されており、成分を調合している様子がうかがえる。

このような文法的特徴を持たない単語の並びは、翻訳の精度が低いから生じたというよりも、中世の医学や錬金術の実用文献においてしばしば見られる、実際の処方や作業手順を記録した「レシピ的構造」とよく合致していると考えられる。手稿の著者は、読ませるための文章を綴ったのではなく、定められた論理構造の枠組みの中に、必要な材料と操作を順番に記録していった可能性が高い。また、7章の反証にて詳しく後述するが、この特定手法の特性上、選択したコーパスに収束してしまい循環となる可能性が高いため、他分野の文献（神学・農業書・建築書・料理書など、同じ「レシピ的構造」を持つ他の実用文献）をストリームCとして同様の特定を行い比較検証した。その結果、「レシピのような書き方」ではなく、「錬金術の内容」と合致していることが確認された。

したがって、本結果を踏まえると、前報で発見したヴォイニッチ手稿の内部構造から立てた錬金術という分野という仮説を補強するものである。解読に向けて、一歩人類が近づいた証拠でもある。

--- (English Translation) ---

Chapter 5: Global Inverse Structural Analysis and the Extraction of the "Practical Pharmacopoeia"

5.1 Inverse Analysis of the Global Structure and Protection of Unknown Symbols

Using the reliable translation criteria extracted from the high-dimensional space through the processes described up to the previous chapter, mechanical substitution was executed on the text of the entire manuscript. Unknown symbol groups not present in the dictionary were not forcibly mapped but were protected and output as "undefined." This intermediately generated record is "Voynich_Decompiled_Record.txt." The purpose of this process is to objectively establish where "unknown blanks" remain within the structure governing the manuscript.

5.2 Feature Extraction and Automatic Determination of "Types"

To identify the remaining undefined symbols, it is crucial and serves as a key to mathematically derive what part-of-speech role, such as "noun (material)" or "verb (operation)," the blank requires. For the undefined symbol groups, "purely geometric features" were extracted, such as their inline positions within the four-stage logical architecture (Boot, Set, Transition, and Termination) and the gravitational attraction of adjacent, established Latin words.

A classification algorithm was applied to the extracted features. In doing so, rather than arbitrarily deciding the number of classifications, the system was made to determine the optimal number of "types" (functional roles such as parts of speech) that are most naturally separated, using a mathematical evaluation function called the "silhouette score," which measures the degree of spatial separation. The result of identifying the type of each blank through this objective dynamic classification is "Voynich_Undefined_Variable_Types.csv," and reflecting this into the text structure yielded "Voynich_Schema_Mapped_Translation.txt."

5.3 Orthogonal Dual Constraints and Domain-Wide Identification of Distributional Translation Candidates

Upon the text structure for which the identification of types was completed, a final translation process was executed to identify and bind real Latin vocabulary. Here, this study introduced two orthogonal, rigorous mathematical constraints: "macro-topology (contextual attraction in the 128-dimensional SVD space)" and "micro-function (consistency with the automatically determined type)." That is, it is a strict verification wherein even if the overall contextual attraction is remarkably high, the candidate is discarded if the "type (part of speech)" required by the manuscript side does not match.

As a result of cross-referencing with the Latin corpus (Stream B) using these dual conditions, we were able to identify distributional translation candidates that passed the dual criteria for 99.4% (9,733 slots) of the total 9,783 undefined slots remaining in the manuscript. This does not signify a "definitive semantic translation," but rather demonstrates the fact that "for each slot in the manuscript, Latin words satisfying both contextual similarity and type consistency exist at a high rate within the Stream B corpus." The final mapping of corresponding candidates across the entire manuscript was recorded as "Voynich_Absolute_Translation_Final.txt." Whether these

corresponding candidates are semantically correct translations is not guaranteed by the mathematical processing of this study alone and requires multifaceted historical and herbalistic validation.

5.4 Domain-Wide Identification of Distributional Translation Candidates: Possibility as a Record of a Practical Pharmacopoeia

That this identification rate of 99.4% is not a forced fit based on contextual guesswork is proven by the behavior of the identified vocabulary. For instance, the unknown symbol "1oe" within the manuscript mathematically converged to the Latin conjunction "quod" (meaning "that" or "because"). This "1oe" was fixed to the type "logical connection/preposition (Type_6)" in the preceding automatic classification, confirming that it aligns with the actual part of speech of the identified Latin word.

However, existences that could be interpreted as "homophones," where multiple distinct unknown symbols converge to the identical Latin word, were also mathematically discovered. This does not definitively establish them as homophones. The EVA groups converging to (dodder / material name) are soy9 (position: 1r.2) / s9aIy (position: 1r.3) / h1s9 (position: 1v.1) / h98an9 (position: 2r.1) / 8oy1oy9 (position: 2r.2) / hoom (position: 2v.1) / h1A (position: 2v.2) / k2cos (position: 3r.1). This indicates not a failure in identifying distributional translation candidates, but rather an extremely high probability that these EVA groups represent "other, closely related materials (approximate substances)" used in the same context (the material slot of a recipe) as .

With the current extraction based on "similarity-based matching" and "fitting into type frameworks (dual constraints)," while it is possible to objectively identify the "role of the type (the slot where a material enters)," it remains impossible to separate the "differences between individual substances" substituted into that slot. However, considering that the Voynich manuscript is written in an unknown cryptographic system and taking into account the text's characteristic where forced contextual assimilation (overfitting) can easily occur, the possibility of homophones or highly encrypted structures cannot be entirely ruled out. At present, there is no objective evidence capable of distinguishing among any of these possibilities.

Conversely, it was confirmed that operational verbs and stative verbs typical of medieval alchemical and pharmaceutical prescriptions, such as "INCENDERE" (ignite) and "ALCALISATUS" (alkalized), occur frequently among the other extracted corresponding candidate groups. This fact supports the validity of selecting the Latin corpus of alchemy and early medicine as Stream B and aligns with the hypothesis that the manuscript possesses the character of a record of operations and materials rather than stories or poetry. However, whether individual word correspondences (for instance, fachys → MOUENTE) are semantically correct cannot be finalized by the mathematical processing of this study alone and requires close inspection by specialists in Latin linguistics and the history of science. These represent the "most probable corresponding candidates" at the current stage, and their final confirmation

is deferred to the subsequent phase of interdisciplinary validation.

5.5 Presentation of Translation Output Records and Substantiation of the "Recipe-Like Structure"

The hypothesis presented in the previous section—that this manuscript is a medieval "practical pharmacopoeia (operational procedure manual)"—can be confirmed more concretely through the records in the final translation output file, "Voynich_Multilingual_Pipeline.txt." Numerous outstanding linguists and historians have hitherto attempted to decipher the character strings of the manuscript as continuous sentences in a natural language. Deep respect is extended to their long-standing immense efforts and accumulation of work. When observing the arrangement of Latin vocabulary extracted through the mathematical constraints of this study, a structure emerges where specific words are aligned regularly, rather than prose accompanied by complex grammatical rules (morphological inflections such as case declension or gender-number agreement) seen in classical Latin.

The following is an excerpt from the translation output record from the very beginning of the first page of the manuscript (Folio 1r). It shows on a line-by-line basis how the extracted real Latin (LAT) is linked to the unknown symbol sequences (EVA).

①EVA : fa19s 9 hae ay Akam 2oe !oy9 scs 9 hoy 2oe89

②LAT :

①EVA : soy9 Hay oy 9 hacy 1kam 2ay Ais Kay Kay 8aN

②LAT :

Looking at these output results, a literary sentence structure where subjects and verbs agree to form complex meanings or contexts is not observed. Even in <1r.1>, which adorns the very beginning of the manuscript, there are no words indicating the start of a narrative; instead, words indicating physical operations and materials, such as "moving" (), "perfection" (), "glued/attached" (), and "marcasite" (), appear in succession. Continuing into <1r.2>, it begins with an explicit plant name, "dodder" (), followed by operational instructions such as "turning" () and "retaining" (). In the latter half, the name of another medicinal herb, "eyebright" (), is also specified, giving a glimpse into the blending of ingredients.

Such a sequence of words devoid of grammatical features is considered to align well with the "recipe-like structure" frequently observed in practical medieval medical and alchemical literature recording actual prescriptions and operational steps, rather than resulting from low translation accuracy. It is highly probable that the author of the manuscript did not pen a text meant for reading, but sequentially recorded necessary materials and operations within the framework of a defined logical structure. Furthermore, as will be discussed in detail later in the refutations of Chapter 7, due to the nature of this identification method, there is a high probability of converging onto the selected corpus and creating a circular argument; hence, we

conducted a comparative validation by performing identical identification using literature from other fields (theological texts, agricultural books, architectural books, cookery books, etc., which share the same "recipe-like structure") as Stream C. The results confirmed that the text aligns not merely with "a recipe-like style of writing" but specifically with "alchemical content."

Therefore, based on these results, they reinforce the hypothesis of the alchemical domain established from the internal structure of the Voynich manuscript discovered in the previous report. It also stands as evidence that humanity has taken a step closer toward its decipherment.

第6章：画像の物理トポロジーによる翻訳のクロスモーダル証明

6.1 翻訳結果の正当性とクロスモーダル証明の目的

本章の目的は、前章で達成された「手稿全域の翻訳」および「各単語の役割（型）の自動分類」が、恣意的なこじつけではないかの確認、そして全域にわたって論理的破綻が存在しないことを、テキストとは独立した画像側の物理的構造から逆算して実証することにある。全く次元の異なる二つの情報源（画像とテキスト）を照合するこのクロスモーダル証明は「OI-2026プロトコル」の構造が正しく、それによって手稿の翻訳が現像したならば、この挿絵との照合は破綻しないことを示すはずである。

6.2 画像の細線化と最適輸送による同型性照合

主観的な解釈を排除するため、197ページ分の手稿画像に対し、画像処理ライブラリ（OpenCV）を用いてインクの描画部分のみを抽出・細線化し、画像を純粋な「線の骨組み」へと変換した。この幾何学的な骨組みに対し、グラフ理論を用いて「線の行き止まり（Endpoints：端点）」「三叉路（Branchpoints：分岐点）」「輪っか（Cycles：閉路）」の数を物理的に計数し、線がどのように繋がっているかという空間的構造（トポロジー）をデータ化した。
一方のテキスト側については、前章の翻訳プロセスで確定した「材料（Type_3）」や「操作（Type_2）」といった機能的役割が、ページごとにどのような順序で出現しているかという論理的な遷移ネットワークを構築した。そして、「画像から抽出された線の繋がり」と「翻訳によって確定したテキストの役割の繋がり」という二つのネットワーク構造を、「Gromov-Wasserstein最適輸送（GW）」と呼ばれる高度な数理的手法を用いて重ね合わせ、双方の構造の合致度（同型性）を計算し点数化した。

6.3 翻訳の正確性を示す物理的指標との強い相関

GW最適輸送によるマッチングに成功した181ページ分のデータを対象とし、画像側のトポロジー（端点などの数）と、テキスト側の機能的役割（Type）の出現頻度の間に統計的な相関が存在するかを検証した（『Voynich_CrossModal_Physical_Pointers.csv』に記録）。

計算の結果、画像における「線の行き止まり（Endpoints）」の数と、テキストにおける「材料・成分（Type_3）」の出現数の間に、R=0.7080（P=7.70×10−29 ）という正の相関が認められた。画像内に描

かれた枝葉や根の先端（行き止まり）が多いページほど、テキスト側でも正確に同期して「材料」を表す単語が増加している。この数値の一致は、前章における「未知の単語を材料（Type_3）として特定・翻訳した処理」が、画像側の物理的な指定と合致しており、分類エラーも起こしていないことを示す。

6.4 手稿全域における論理整合性と反復構造のトポロジー証明

さらに、他の機能的役割との相関においても、本翻訳は全域にわたって論理整合性を示している。画像の端点（Endpoints）は、テキストの「主操作・命令（Type_2）」の出現数ともR=0.6132（P=4.51×10−20）という強い正の相関を示した。一方で、「終了・保存（Type_8）」の出現数とはR=-0.2653（P=3.07×10−4）の負の相関を示している。これは、「材料が多く複雑な作業を要求するページほど、混ぜる・加熱するといった『操作』の回数が増加し、すぐには『終了』しない」という、マニュアルの制御構造としての正しい論理的挙動である。

加えて、手順の反復（ループ構造）に関する証明も特筆すべきである。画像の中に描かれた「輪っか・閉路（Cycles）」の数と、テキストにおける「変数・定数定義（Type_1）」の間には、R=-0.2540（P=5.59×10−4 ）という明確な負の相関が観測された。これは、「画像上で同じ線をぐるぐると巡る（特定の作業を反復している）間は、テキスト側で新しい材料の定義や代入を行わない」という反復処理の制約と同期している。

以上のことから、本研究における結果は、前報で特定した内部構造が錬金術の分野に類似するという仮説を数理的に補強するものであることが観測された。

--- (English Translation) ---

Chapter 6: Cross-Modal Proof of Translation via Physical Topology of Images

6.1 Validity of Translation Results and Objective of the Cross-Modal Proof

The objective of this chapter is to verify that the "domain-wide translation of the manuscript" and the "automatic classification of the role (type) of each word" achieved in the previous chapter are not arbitrary forced fits, and to demonstrate from a reverse calculation of the physical structure of the images—independent of the text—that no logical contradictions exist across the entire domain. This cross-modal proof, which cross-references two completely different dimensional information sources (images and text), should demonstrate that if the structure of the "OI-2026 Protocol" is correct and the translation of the manuscript has been derived thereby, the cross-referencing with these illustrations will not collapse.

6.2 Image Thinning and Isomorphism Matching via Optimal Transport

To eliminate subjective interpretation, the ink-drawn portions of 197 pages of manuscript images were extracted and thinned using an image processing library (OpenCV), converting the images

into pure "line skeletons." Applying graph theory to these geometric skeletons, the numbers of "Endpoints," "Branchpoints," and "Cycles" were physically counted, quantifying the spatial structure (topology) of how the lines are connected into data.

Regarding the text side, a logical transition network was constructed representing the sequential occurrence per page of the functional roles, such as "Material (Type_3)" and "Operation (Type_2)," established during the translation process in the previous chapter. Subsequently, these two network structures—the "connections of lines extracted from the images" and the "connections of text roles established by translation"—were superimposed using an advanced mathematical method known as "Gromov-Wasserstein (GW) optimal transport," and the degree of structural congruence (isomorphism) between the two was calculated and scored.

6.3 Strong Correlation with Physical Indicators Demonstrating Translation Accuracy

Targeting the data of 181 pages successfully matched via GW optimal transport, we verified whether a statistical correlation exists between the topology on the image side (such as the number of endpoints) and the occurrence frequency of functional roles (Types) on the text side (recorded in "Voynich_CrossModal_Physical_Pointers.csv").

As a result of the calculation, a positive correlation of $R=0.7080$ ($P=7.70\times10^{-29}$) was observed between the number of "Endpoints" in the images and the occurrence frequency of "Materials/Ingredients (Type_3)" in the text. Pages with more branch/leaf or root tips (endpoints) depicted in the images show a precisely synchronized increase in words representing "materials" on the text side. This numerical congruence indicates that the "process of identifying and translating unknown words as materials (Type_3)" in the previous chapter aligns with the physical designations on the image side, occurring without classification errors.

6.4 Topological Proof of Logical Consistency and Iterative Structure Across the Entire Manuscript

Furthermore, in correlations with other functional roles, this translation demonstrates logical consistency across the entire domain. The endpoints in the images also exhibited a strong positive correlation of $R=0.6132$ ($P=4.51\times10^{-20}$) with the occurrence frequency of "Main Operations/Commands (Type_2)" in the text. Conversely, it showed a negative correlation of $R=-0.2653$ ($P=3.07\times10^{-4}$) with the occurrence frequency of "Termination/Preservation (Type_8)." This represents the correct logical behavior of a manual's control structure: "the more materials and complex tasks a page requires, the more the number of 'operations' such as mixing or heating increases, and it does not immediately 'terminate'."

In addition, the proof regarding the iteration (loop structure) of procedures is also noteworthy. A clear negative correlation of $R=-0.2540$ ($P=5.59\times10^{-4}$) was observed between the number of "Cycles" drawn in the images and "Variable/Constant Definitions (Type_1)" in the text. This synchronizes with the constraint of iterative processing: "while repeatedly circling the same line

on the image (iterating a specific task), no new material definitions or assignments are executed on the text side."

From the above, it was observed that the results of this study mathematically reinforce the hypothesis that the internal structure identified in the previous report is analogous to the alchemical domain.

第**7**章：想定反証の数理的再反証（**Stress Test & Refutation**）

本研究と前報が提示した『データマトリックス』としての構造定義と、全域の分布的翻訳候補の成果に対しては、これまでの言語学的・暗号学的なアプローチの観点のみならず、線形代数学や情報統計学の一般論からも、いくつかの妥当な疑問や反証が想定される。

本章の目的は、全体を通して本研究に問題がないかを多角的に探索することである。本研究は各処理段階で観察結果を客観的に算出し、論理的整合性を保ってきたが、人間が分析を行っている以上、観測バイアスなどが混入している可能性を完全に排除することはできない。現状考えられる多角的な交差検証の過程において、様々な反証を自ら提起し、それらに対する現時点でのデータからの応答を記録する。なお、本章で述べる「再反証」は、想定反証を完全に棄却するものではなく、現データのもとでどこまで応答可能かを示すストレステストの記録として位置付ける。最終的な検証は、外部コーパスとの照合および専門分野(中世ラテン語学、薬草学、科学史)による交差検証を経て行われるべきものである。

7.1 次元空間の強引な同期（**Manifold Alignment**の欠如）

【想定反証】

本研究の第3章および第4章では、手稿データ（Stream A）とラテン語コーパス（Stream B）に対して個別にTruncatedSVDを適用し、同一の56次元ないし128次元の密行列へ変換した上で、コサイン類似度による直接照合を行っている。しかし、素性の異なる2つのデータセットを個別に次元圧縮した場合、算出される主成分軸の方向や意味（例：手稿側が接頭辞の頻度、ラテン語側が動詞の活用など）が一致する数学的保証は存在しない。軸が同期していない直交空間同士でコサイン類似度を算出する行為は、物理次元の異なる単位（例えば温度と速度）を直接比較するような「線形代数的な錯覚（次元の混同）」であり、算出された類似度は情報工学的に無意味なノイズに過ぎないのではないか。両空間のトポロジー（位相）が、意味的に正しく重なり合っている（アライメントされている）ことを、どのように数学的に保証したのか。

【再反証】

この反証は、次元削減(SVD)を「単なる数値の圧縮」と見なした場合の懸念として、十分に検討に値する。本研究の立場としては、無意味な次元同士の数字遊びを行っているのではなく、「手稿側の

システム基板(Stream A)の物理次元」に合わせて、ラテン語側(Stream B)の抽出軸を同期させる設計を採用している。

第一に、手稿データ(Stream A)から抽出された56次元とは、未知の記号が持つ「接頭辞・接尾辞との結合の振る舞い(ネットワーク・トポロジー)」である。ラテン語コーパス(Stream B)を56次元に圧縮する際にも、テキストにおける「単語の出現分布や前後関係(前置詞や格変化等の振る舞い)」という、Stream Aと同種の「単語間の結合ネットワーク」を抽出対象としている。両者が「結合グラフの形状」として相同性を持つという前提のもと、初期の56次元空間における比較を行っている。ただし、この前提自体が本研究の作業仮説であり、両空間のトポロジーが意味的に正しく重なり合っていることの最終的な保証は、後段の全域パース整合性および耐久テストの結果との総合判断に委ねられる。

第二に、この空間における位相衝突(Cos-Sim)の中から、「相互最近傍(MNN)」と「Zスコア2.0以上の特異な断崖」という極限の数理的フィルタを適用した。これにより、両ストリームの位相が偶然以上の精度で一致したペアのみを抽出している。ただし、現段階で抽出された特異点は、「手稿の構造が中世の錬金術と類似している」という仮定から導き出された「現データ上は最も可能性の高い対応候補」であり、翻訳の意味的確定ではない。

第三に、これら初期アンカーが偶然のノイズではないことを支持する間接的な証拠として、本プロトコルが全域に適用されてもパース時に体系的な破綻を起こさなかったという観察がある。また、後述する耐久テストの結果、錬金術コーパスが他の対照群(神学、農業書、建築書、料理書)と比べて顕著に高い分布的一致を示したという事実も、初期のアライメント仮定が完全に的外れではないことを示唆する。ただし、これらは仮説を補強する観察であって、Manifold Alignmentの数学的保証そのものを与えるものではない。最終的な検証には、別途の独立コーパスを用いた追試と、計算言語学の専門家による方法論的レビューが望ましい。

7.2 Stream Bの選定による「確証バイアス」

【想定反証】

本研究の第1章および第3章において、比較対象となる外部コーパス（Stream B）として、16世紀の錬金術・薬学に特化したラテン語文献（EMLAP等）のみを意図的に選定している。次元削減（SVD）とコサイン類似度を用いたマッチングにおいて、参照する辞書空間の内部に特定のドメイン（分野）の語彙しか存在しなければ、出力される結果がそのドメインの語彙に帰結するのは自明の理（トートロジー）である。仮にStream Bに『中世の神学書』や純粋な『天文学の観測記録』、あるいは全く無関係なドメインのコーパスを混合、または単独で用いた場合でも、本システムは「実用薬物局方」の語彙のみを客観的かつ正確に引き寄せる数学的保証があるのか。現在の結果は、比較対象のドメインを人為的に極限まで絞り込んだことに起因する、深刻な確証バイアス（Confirmation Bias）に過ぎないのではないか。

【再反証】

この反証は、統計的機械学習における「過学習」および「確証バイアス」のリスクを指摘した重要な指摘である。本研究は、この懸念に対して以下の応答を提示する。

第一に、Stream Bとして錬金術・初期医学ラテン語コーパスを選定したのは、手稿の歴史的所有履歴(ルドルフ2世、バレシュ、キルヒャー等、いずれも錬金術・神秘学関連)、図像的特徴(植物・鉱物・人体図)、放射性炭素年代測定(15世紀初頭)という、複数の独立した事前情報に基づいたベイズ的選定である。手稿の構造特性(高ハパックス率、規格化された部品の組み合わせ)から、レシピ的構造を持つ文献群が候補として浮上することも、選定の根拠となった。これは「特定のドメインを意図的に選んだ」のではなく、「複数の事前情報から最も可能性の高いドメインを選んだ」という方法論的選択である。

第二に、この選定の妥当性を検証するため、本研究では複数の対照コーパスを用いた比較実験を実施した。具体的には、神学コーパス(聖書、キリスト教文献)、および同じく「レシピ的構造」を持つ実用文献群(Apicius料理書、Vitruvius建築書、Columella農業書)を対照群として、同じ照合手法を適用した。その結果、錬金術コーパスは他のすべての対照群に対して、平均コサイン類似度(0.80対0.66)および勝率(92.7%対1.7%-5.5%)の両面で統計的に有意な優位を示した(Mann-Whitney $p \approx 0$、Kruskal-Wallis $p \approx 0$)。

この結果は、「同じレシピ的構造」を持つ文献であっても、錬金術コーパスが手稿に対して固有の高い一致を示すことを示唆しており、単純な確証バイアスでは説明しにくい現象である。ただし、この検証はあくまで「テストされた対照群の中で錬金術が優位」を示すものであり、未テストの全コーパス(例えばアラビア錬金術のラテン語訳、ビザンチン医学、特定の地域・流派の薬草学等)に対する優位性まで保証するものではない。より精緻なドメイン特定は、今後の追加検証に委ねられる。

第三に、もし錬金術ドメインの選定が偶然の確証バイアスに過ぎなかった場合、その選定から導かれた個別単語対応群を、手稿の4段階構造に当てはめた際に、品詞役割の系統的な不整合が頻発するはずである。本研究の自動分類による品詞型と、抽出されたラテン語の品詞との一致率は99.4%であった。この高い一致率は、ドメイン選定が完全に的外れであれば達成困難な水準であり、選定の合理性を間接的に支持する。ただし、これも「分布的・型整合性レベルでの一致」であり、「意味論的な翻訳の正しさ」を直接保証するものではないことに留意されたい。

以上の応答により、確証バイアスの可能性は現データのもとでは支持しがたいが、最終的な妥当性検証には、より広範な対照群の追加と、専門家による定性的レビューが必要である。

表1: ドメイン比較耐久テストの結果**(**三者衝突テスト**)**

ドメイン	平均コサイン類似度	勝率

B: 錬金術**(EMLAP)**	**0.8032**	**92.7%**
C: 神学**(**聖書・キリスト教文献**)**	**0.6626**	**5.5%**
D: 実用書**(Apicius**料理・**Vitruvius**建築・**Columella**農業**)**	**0.6647**	**1.7%**

処理対象: 7,286語、Kruskal-Wallis p ≈ 0

表2: ドメイン比較耐久テストの結果**(B vs C**二者衝突テスト**)**

ドメイン	平均コサイン類似度
B: 錬金術	**0.7988**
C: 神学	**0.6630**

差分: B - C = 0.1358、B勝率: 92.7%、Mann-Whitney p ≈ 0、処理対象: 6,157語

表3: CLTK品詞拘束テスト(ダミーラテン語投入実験)の結果

手稿側EVA	要求型	投入ダミーLatin	実際の品詞(CLTK判定)	判定
fachys	Type_3 (Noun)	et	CCONJ	ERROR
qokedy	Type_5 (Adj/Adv)	dicere	VERB	ERROR
chol	Type_2 (Verb)	dominus	NOUN	ERROR
daiin	Type_6 (Conj/Prep)	sanctus	ADJ	ERROR
shol	Type_3 (Noun)	in	ADP	ERROR
chor	Type_2 (Verb)	non	PART	ERROR
shes	Type_5 (Adj/Adv)	autem	PART	ERROR
chedy	Type_3 (Noun)	facere	VERB	ERROR

結果: 8件中8件で品詞不整合エラー、本研究の二重制約が形式的でなく実効的に異質な語彙を排除することを確認

7.3 反復処理による「カスケード・エラー（誤差の連鎖的増幅）」

【想定反証】

本研究の第3章および第4章では、統計的な特異性（Zスコア2.0以上）を持つ翻訳ペアを特定し、それらを照合対象から順次除外しながら計算を繰り返す、段階的な反復照合手法を採用している。しかし、この手法は「誤同定の連鎖的な波及」という重大なリスクを内包している。 もし初期に確定した少数の翻訳基準の中に、一つでも誤った同定が含まれていた場合、その歪みが残りの比較空間全体に波及し、後続の翻訳すべてが雪だるま式に誤った方向へ誘導される危険性がある。したがって、最終的に示された「99.4%」という極めて高い特定率は、手稿の真の構造を解き明かした結果ではなく、少数の初期の誤同定が引き起こした「過剰適合（強引な文脈的同化）」によって、見かけ上の「それらしいラテン語の羅列」へと強引に収束してしまっただけの錯覚に過ぎないのではないか。

【再反証】

この反証は、反復的同定手法に潜む誤差増幅のリスクを指摘した重要な指摘である。

第一に、初期アンカー(Zスコア2.0以上、相互最近傍を満たすペア)に誤同定が含まれており、それが連鎖的な過剰適合を引き起こしていた場合、後続の同定で抽出された単語群を手稿の4段階構造に当てはめた際、品詞役割の系統的な不整合が頻発するはずである。本研究で観測された99.4%の品詞整合率は、このような系統的破綻が起きていないことを示唆する。ただし、この観察は「分布的・型レベルでの整合性」を示すものであり、「個別単語の意味的正しさ」を保証するものではない。初期アンカーに意味論的な誤りが含まれていた場合でも、それが分布的に整合する語に置き換えられている限り、型レベルの整合性は維持される可能性がある。

第二に、テキストとは独立に抽出された画像側の物理的特徴(枝葉の端点数)と、テキスト側の機能的役割(材料型)の出現割合との間に、$R=0.7080$($P=7.70\times10^{-29}$)という有意な正の相関が観測された。これは、テキスト側の処理が完全に系統的な誤差で支配されていれば、独立に観測される画像特徴と高い相関を示すことは確率的に起きにくいという観察である。ただし、この相関も「ページ単位の集計値レベル」での同期を示すものであり、個別単語と個別植物画の対応の正しさを直接示すものではない。

結論として、初期誤差の連鎖的増幅という懸念は、現データのもとでは系統的破綻の証拠が見当たらないという意味で、現時点では支持しがたい。ただし、より厳密な検証には、初期アンカーを意図的に変えた感度分析や、ブートストラップ的な再サンプリング実験が今後の課題として残る。

7.4 品詞の事前限定による「意味論的こじつけ」

【想定反証】

研究は、「17の基本枠組み」や「4段階論理構造」に実在のラテン語を当てはめた結果、1件の品詞的矛盾も発生しなかったことをもって翻訳の正当性を主張している。しかし、テキストの各空白に対して「ここには名詞が入る」「ここには動詞が入る」といった品詞的役割を人為的に事前に限定し、その条件に合うラテン語の候補の中から単語を選択した（こじつけた）結果に過ぎないのではないか。人為的な品詞の型枠を用意して単語を当てはめれば、品詞的な破綻が起こらないのは当然の帰結（トートロジー）であり、翻訳の正確性の証明にはならない。

【再反証】

この反証は、解読作業において陥りがちな「文脈の強引な同化」を危惧した重要な指摘である。

第一に、手稿の各文字列における「品詞的役割」は、著者が事前に意味を推測して割り当てたものではない。第5章で述べた通り、行内位置や隣接記号との結合パターンといった「純粋な幾何学的特徴」のみを抽出し、シルエットスコアという数理的評価関数を用いて、最も自然に分離される役割の数を客観的に自動分類している。人間の意味論的な介入を排した手続きであり、品詞の型枠自体は手稿の物理的構造から導き出されている。

第二に、ラテン語の特定においては、「多次元空間における文脈的引力」と「自動分類された品詞的役割の一致」という、独立した二つの条件を同時に満たさなければ採用されない設計を採っている。品詞を合わせるために文脈的引力の低い単語を選ぼうとしても類似度の断崖条件によって弾かれ、逆に文脈的引力が高くとも品詞役割が合致しなければ破棄される。

ただし、ここで認めるべき限界がある。何度も述べているが、両条件はいずれも分布的・形式的な制約であり、「意味的な正しさ」を直接保証するものではない。同じ品詞型を持ち、似た分布を持つラテン語が複数存在する場合、本手法はその中で最大類似度を持つ候補を選ぶが、それが「意味的に最も近い候補」と一致する保証はない。第5.4節で示した同音異義の現象は、まさにこの限界を表している。

結論として、人為的な品詞のこじつけによる結果ではないという意味では、確率論的に通常想定される強引な当てはめは現データから棄却されるが、「意味的に正しい個別単語対応が選ばれている」という保証は本手法の枠組みからは導けない。これは次フェーズの専門家による検証に委ねられる課題である。

7.5 図像とテキストの連動における「見かけ上の相関」

【想定反証】

本研究の第6章において、手稿に描かれた図像の「線の端点（枝葉の行き止まり）」の数と、翻訳によって特定された「対象物質（材料）」の単語数の間に、R=0.7080という極めて強い正の相関が存在することが証明されている。しかし、ここには統計的な「見かけ上の相関（隠れた要因による錯覚）」が潜んでいる可能性がある。 例えば、根や葉が複雑に描かれている（端点が多い）画像が存在するページは、付随する説明のための「テキストの総量（総単語数）」も多くなる傾向はないだろうか。もしテキスト全体が長ければ、当然そこに含まれる「材料」を示す単語の絶対数も増加する。つまり、図像の複雑さと材料の数は直接連動しているのではなく、双方が「ページ内のインク量（記述の総量）」という隠れた第三の要因に引きずられているだけではないか。単語の絶対数ではなく、ページ内の全単語に対する「出現割合」で計算した場合でも、この相関は維持されるのか。

【再反証】

この反証は、計量テキスト分析において陥りやすい「見かけ上の相関」を危惧した重要な指摘である。

第一に、本研究におけるテキスト側の各機能的役割の出現頻度は、絶対数のカウントではなく、各ページの総単語数に対する出現割合(相対的密度)として計算されている。これにより、「ページのインク量(記述の総量)」という潜在的な第三因子の影響は、計算プロセスの段階で除外されている。

第二に、この「総量に依存しない相対的割合」と「画像の物理的な端点数」の照合結果として、$R=0.7080$($P=7.70\times10^{-29}$)という有意な相関が観測された。これは、ページの記述総量を補正した上でも、画像と特定型の出現密度の間に体系的な連動性が存在することを示している。

ただし、この検証では完全には排除できていない潜在的な第三因子の可能性も認める。例えば「ページのトピック自体の複雑さ」が、画像の複雑さと特定型の相対密度の両方を同時に高めているという可能性は、現データからは排除しきれない。また、本研究で「材料型」と分類された語群が、本当に意味論的に植物画と対応するかは、ページ単位の集計値の相関だけでは判定できず、個別ページの植物同定との照合が今後の課題となる。

結論として、図像とテキストの連動が「単純なテキスト総量による見かけ上の相関」である可能性は、現データの相対割合計算によって低減されているが、より広範な意味での潜在因子を完全に排除するには、画像と単語の個別対応の検証が必要である。

7.6 機能的役割の自動分類と二重照合における循環論法の棄却

【想定反証】

本研究の第5章において、テキストの「文脈的引力（類似度）」と「機能的役割（品詞の型）」の二重の条件を用いて、99.4%の単語を特定したと宣言している。しかし、この「機能的役割」は、数理的評価関数を用いて手稿内部のデータから自動的に分類されたものに過ぎない。この自身で分類した枠組みに対して都合よく「これは操作である、これは材料である」と役割を付与し、その役割に合致する

ラテン語だけを選び出して「役割が完全に一致した」と証明する論法は、完全な循環論法（自己完結した解釈）である。外部の客観的な文法規則を用いた検証ではなく、自身で構築した規則で自身を評価しているに過ぎず、翻訳の正確性の証明としては論理的に無効ではないか。

【再反証】

この反証は、未知言語の解読において陥りやすい「解釈の自己完結」を危惧した重要な指摘である。

第一に、手稿内の機能的役割は、本研究のOI-2026プロトコルにおいて観察された4つの位置依存的カテゴリと、シルエットスコアによる自動分類に基づいている。著者が意味を推測して恣意的に分類したものではなく、純粋な物理的特徴(位置、隣接関係)からの自動抽出である。

第二に、比較対象となるラテン語文献の品詞は、本研究が構築したものではなく、16世紀の歴史的文献に内在する外部の言語学的事実である。本手法は、「手稿の物理的構造から抽出された分類」と、「ラテン語の外部言語学的品詞」という、出自の異なる二つの体系を照合している。

ただし、両者の照合における「型の一致」は、最終的にラテン語側の品詞分類器(CLTK等のNLPツール)の判定に依存しており、その判定自体が完全に客観的とは言えない側面がある。また、自動分類された手稿側の型を「材料」「操作」と呼んでいる点には、解釈の余地が含まれる。「材料型」と分類された型が、本当に物質名を指すのかは、最終的には個別ページの内容との照合によってしか検証できない。結論として、本手法は単純な循環論法ではないが、完全に独立した検証とも言えず、数学的に確率の高い仮説に留まる。完全な独立性を確保するには、ラテン語側の品詞分類を別の独立した分類器で再現するクロスチェックや、専門家による定性的レビューが必要となる。

7.7 ラテン語の形態論的屈折の無視による「文法的破綻」の棄却

【想定反証】

本研究は、未定義の文字列に対して「名詞（対象物質）」や「動詞（主操作）」といった「機能的役割（品詞の型）」を合致させることで、全域の翻訳を完遂したと主張している。しかし、比較対象としたラテン語は、名詞に厳密な格（主格、対格、奪格など）が存在し、動詞には人称や時制が伴う、高度な「屈折（語形変化）」を特徴とする言語である。　語彙の文脈的引力と、大まかな品詞の分類のみを用いて単語を配置した場合、主語と動詞の人称の不一致や、前置詞と名詞の格の不一致が必然的に生じるはずである。したがって、提示された翻訳結果は、古典ラテン語の厳密な文法規則に照らし合わせると意味を成さない、文法的に破綻した「単語の羅列（単語のサラダ）」に過ぎないのではないか。1文ごとに精査した際、それは本当に自然言語の文章として「読める」状態にあるのか。

【再反証】

この反証は、ラテン語の形態論的特徴を踏まえた古典文献学の観点から、極めて妥当な指摘である。

第一に、本研究は手稿のテキストを「文学的散文として読める」とは主張していない。前報で観察した通り、手稿の文字列自体は、ハパックス率72.93%、規格化された部品の組み合わせ構造、屈折を伴わない部品の流用といった特徴を持ち、古典ラテン語のような複雑な格変化や活用を伴う言語構造とは異なる挙動を示す。

第二に、本研究が抽出したラテン語対応群は、中世錬金術・薬学文献に特有の「レシピ的記述フォーマット」(操作指示と材料名の連続)と整合する。命令形や分詞、材料名の名詞形が、手稿の対応する位置に配置される傾向が観察されている。

しかし、「ラテン語として文法的に通る連続したテキストが構成されている」という意味での読解可能性は、本研究では保証されていない。出力は「分布的・型整合的に対応するラテン語の配列」であり、これを実際の中世レシピと並べた時に、操作の連鎖が論理的に成立するか、材料の組み合わせが実際の薬学的処方として妥当かは、薬草学・科学史の専門家による検証を要する。

結論として、「文法的に破綻していない」という主張は、本研究の枠組み(手稿が非屈折的部品構造を持ち、その対応がレシピ的構造を持つラテン語と整合する)のもとでは支持されるが、「実際の中世錬金術文献として読める処方になっている」という意味での文法的・実用的妥当性は、今後の専門家検証に委ねられる。

7.8 ハパックス・レゴメナに対する強引な文脈的同化の棄却

【想定反証】

本研究の第2章において、手稿の全語彙の72.93%が一度しか出現しない単語（ハパックス・レゴメナ）であることが明記されている。一般に、一度しか出現しない単語の「前後の隣接関係」や「配置構造」は極めて情報量が乏しく、そこから抽出される関係性は偶然のゆらぎ（無意味な要素）が大半を占める。 手稿の7割以上を占めるこれら一回性の語彙に対して、どのようにして多角的な配置構造を正確に算出し、実在のラテン語と照合したというのか。出現頻度が極端に低い単語群に対し、文脈的な配置の類似性のみで1対1のラテン語を割り当てたという主張は、客観的な証明ではなく、単なる偶然のゆらぎを拾い上げただけの「無作為な当てはめ（強引な文脈的同化）」に過ぎないのではないか。

【再反証】

この反証は、計量文献学において極低頻度語を扱う際の限界を指摘した重要な指摘である。

第一に、手稿の「一度しか出現しない単語」は、完全に孤立した文字列ではなく、前報研究で観察された「シェル候補とペイロード候補の組み合わせ」によって構成されている。個々のシェル・ペイロードはテキスト全域で数百から数千回にわたり高頻度で出現しており、これらの構成部品レベルで

は十分な統計情報が得られている。本手法は単語を分割不可能な塊としてではなく、構成部品のパターンとして扱うことで、ハパックス語にも一定の文脈的情報量を割り当てている。

第二に、ハパックス語に対して無作為な当てはめが行われていた場合、それらを手稿の4段階構造に流し込んだ際に、品詞役割の系統的な不整合が頻発するはずである。本研究で観測された99.4%の品詞整合率は、ハパックス語に対する照合が偶然のゆらぎを拾い上げているだけの状態ではないことを示唆する。ただし、これは「分布的・型レベルの整合性」を示すものであり、「個別ハパックス語の意味的正しさ」を保証するものではない。

ハパックス語の意味的検証は、本手法の枠組みでは完結しない課題である。同じ型を持ち、同じような分布パターンを持つラテン語が複数候補として存在する場合、本手法はその中で最大類似度を持つ候補を選ぶが、それがハパックス語の真の意味と一致する保証はない。第5.4節で論じた同音異義の現象は、この限界を反映している。

結論として、ハパックス語に対する処理が単純な偶然のゆらぎではないことは、現データのもとでは支持されるが、個別ハパックス語の意味的同定の正しさは、専門家による定性的検証を経て初めて確定する。

--- (English Translation) ---

Chapter 7: Mathematical Refutations of Hypothesized Counterarguments (Stress Test & Refutation)

Regarding the structural definition as a "data matrix" and the results of the domain-wide distributional translation candidates presented in this study and the previous report, several valid questions and counterarguments can be hypothesized, not only from the perspectives of conventional linguistic and cryptographic approaches but also from the general theories of linear algebra and information statistics.

The objective of this chapter is to explore from multiple angles whether there are any fundamental issues in this study overall. Although this study has objectively calculated observation results at each processing stage and maintained logical consistency, as long as human analysis is involved, the possibility of observation bias being introduced cannot be completely eliminated. In the process of multifaceted cross-validation currently conceivable, we proactively raise various counterarguments and record the responses to them based on the current data. It should be noted that the "refutations" described in this chapter do not completely reject the hypothesized counterarguments, but are positioned as a record of stress tests indicating the extent to which responses are possible under the current data. Ultimate validation should be conducted through cross-referencing with external corpora and cross-validation by specialized fields (medieval Latin linguistics, herbalism, and the history of science).

7.1 Forced Synchronization of Dimensional Spaces (Lack of Manifold Alignment)

[Hypothesized Counterargument]

In Chapters 3 and 4 of this study, TruncatedSVD is applied individually to the manuscript data (Stream A) and the Latin corpus (Stream B) to convert them into identical 56- or 128-dimensional dense matrices, followed by direct matching using cosine similarity. However, when two datasets with different properties are individually dimensionally compressed, there is no mathematical guarantee that the directions or meanings of the calculated principal component axes (e.g., prefix frequency on the manuscript side, verb conjugation on the Latin side) will align. The act of calculating cosine similarity between orthogonal spaces whose axes are not synchronized is a "linear algebraic illusion (confusion of dimensions)" akin to directly comparing units of different physical dimensions (e.g., temperature and velocity), and the calculated similarities might be nothing more than information-theoretically meaningless noise. How is it mathematically guaranteed that the topologies of both spaces semantically overlap correctly (are aligned)?

[Refutation]

This counterargument is highly worthy of consideration as a concern when dimensionality reduction (SVD) is viewed as "mere numerical compression." The position of this study is that it does not engage in a numbers game between meaningless dimensions, but adopts a design that synchronizes the extraction axes of the Latin side (Stream B) to match the "physical dimensions of the system substrate on the manuscript side (Stream A)."

First, the 56 dimensions extracted from the manuscript data (Stream A) represent the "combinatorial behavior with prefixes and suffixes (network topology)" possessed by the unknown symbols. When compressing the Latin corpus (Stream B) into 56 dimensions, the target of extraction is similarly the "connection network between words," identical in kind to Stream A, representing "word occurrence distributions and contextual relationships (behaviors such as prepositions and case inflections)" within the text. Comparisons in the initial 56-dimensional space are conducted under the premise that both share homology as "shapes of connection graphs." However, this premise itself is a working hypothesis of this study, and the ultimate guarantee that the topologies of both spaces semantically overlap correctly is deferred to a comprehensive judgment based on the domain-wide parsing consistency and stress test results in the later stages.

Second, from the topological collisions (cosine similarity) in this space, extreme mathematical filters were applied: "Mutually Nearest Neighbors (MNN)" and "singular cliffs with a Z-score of 2.0 or higher." This extracts only those pairs where the topologies of both streams match with an accuracy beyond mere coincidence. However, the singular points extracted at this stage are the "most probable corresponding candidates under the current data," derived from the assumption that "the manuscript's structure is analogous to medieval alchemy," and do not constitute a semantic finalization of the translation.

Third, as indirect evidence supporting that these initial anchors are not coincidental noise, there is the observation that no systematic collapse occurred during parsing even when this protocol was applied domain-wide. Furthermore, the fact that the alchemical corpus demonstrated a significantly higher distributional concordance compared to other control groups (theology, agriculture, architecture, and cookery), as a result of the stress tests detailed later, suggests that the initial alignment assumption is not entirely off the mark. Nevertheless, these are observations that reinforce the hypothesis and do not provide the mathematical guarantee of Manifold Alignment itself. For final validation, subsequent replication tests using separate, independent corpora and a methodological review by experts in computational linguistics are desirable.

7.2 "Confirmation Bias" from the Selection of Stream B

[Hypothesized Counterargument]

In Chapters 1 and 3 of this study, only Latin literature specialized in 16th-century alchemy and pharmacology (such as EMLAP) is intentionally selected as the external comparative corpus (Stream B). In matching using dimensionality reduction (SVD) and cosine similarity, if the referenced dictionary space contains only vocabulary from a specific domain, it is a tautology that the output results will converge upon the vocabulary of that domain. If Stream B were to include "medieval theological texts," pure "astronomical observation records," or corpora from completely unrelated domains, either mixed or independently, is there a mathematical guarantee that the system would objectively and accurately attract only the vocabulary of a "practical pharmacopoeia"? Is the current result merely a severe Confirmation Bias resulting from artificially narrowing down the comparative domain to the extreme?

[Refutation]

This counterargument is a crucial point highlighting the risks of "overfitting" and "confirmation bias" in statistical machine learning. This study presents the following responses to this concern.

First, the selection of the Latin corpus of alchemy and early medicine as Stream B is a Bayesian selection based on multiple independent prior pieces of information: the manuscript's historical ownership record (Rudolf II, Baresch, Kircher, etc., all associated with alchemy and mysticism), iconographic features (botanical, mineral, and anatomical illustrations), and radiocarbon dating (early 15th century). The structural characteristics of the manuscript (high hapax legomena rate, combinatorial sequences of standardized components) also served as a basis for selection, leading to the emergence of literature groups with a recipe-like structure as candidates. This is a methodological choice of "selecting the most probable domain based on multiple prior parameters," rather than "intentionally selecting a specific domain."

Second, to verify the validity of this selection, this study conducted comparative experiments using multiple control corpora. Specifically, the same matching method was applied using a

theological corpus (the Bible and Christian literature) and a group of practical texts that also possess a "recipe-like structure" (Apicius on cookery, Vitruvius on architecture, and Columella on agriculture) as control groups. As a result, the alchemical corpus demonstrated a statistically significant superiority over all other control groups in both mean cosine similarity (0.80 vs. 0.66) and win rate (92.7% vs. 1.7%-5.5%) (Mann-Whitney $p \approx 0$, Kruskal-Wallis $p \approx 0$).

This result suggests that even among texts with the "same recipe-like structure," the alchemical corpus exhibits an intrinsically high concordance with the manuscript, a phenomenon difficult to explain by simple confirmation bias. However, this validation strictly indicates that "alchemy is superior among the tested control groups," and does not guarantee superiority over all untested corpora (e.g., Latin translations of Arabic alchemy, Byzantine medicine, or herbalism of specific regions or schools). A more precise domain identification is deferred to subsequent additional validation.

Third, if the selection of the alchemical domain were merely a coincidental confirmation bias, systematic inconsistencies in part-of-speech roles should frequently occur when the individual word corresponding groups derived from that selection are mapped onto the manuscript's four-stage structure. The concordance rate between the part-of-speech types derived from the automatic classification of this study and the parts of speech of the extracted Latin was 99.4%. This high concordance rate is a level difficult to achieve if the domain selection were entirely irrelevant, thereby indirectly supporting the rationality of the selection. However, it must be noted that this also constitutes "concordance at the level of distribution and type consistency," and does not directly guarantee the "correctness of the semantic translation."

Through the above responses, the possibility of confirmation bias is difficult to support under the current data, but ultimate validity verification requires the addition of broader control groups and qualitative review by experts.

Table 1: Results of Domain Comparison Stress Test (Three-way Collision Test)

Domain	**Mean Cosine Similarity**	**Win Rate**
B: Alchemy (EMLAP)	0.8032	92.7%
C: Theology (Bible and Christian Literature)	0.6626	5.5%

D: Practical Books (Apicius Cookery, Vitruvius Architecture, Columella Agriculture)	0.6647	1.7%
Processed target: 7,286 words, Kruskal-Wallis p ≈ 0		

Table 2: Results of Domain Comparison Stress Test (B vs C Two-way Collision Test)

Domain	**Mean Cosine Similarity**
B: Alchemy	0.7988
C: Theology	0.6630
Difference: B - C = 0.1358, B Win Rate: 92.7%, Mann-Whitney p ≈ 0, Processed target: 6,157 words	

Table 3: Results of CLTK Part-of-Speech Constraint Test (Dummy Latin Insertion Experiment)

Manuscript EVA	Required Type	Inserted Dummy Latin	Actual POS (CLTK Determination)	Judgment
fachys	Type_3 (Noun)	et	CCONJ	ERROR
qokedy	Type_5 (Adj/Adv)	dicere	VERB	ERROR
chol	Type_2 (Verb)	dominus	NOUN	ERROR
daiin	Type_6 (Conj/Prep)	sanctus	ADJ	ERROR
shol	Type_3 (Noun)	in	ADP	ERROR
chor	Type_2 (Verb)	non	PART	ERROR
shes	Type_5 (Adj/Adv)	autem	PART	ERROR
chedy	Type_3 (Noun)	facere	VERB	ERROR

Result: POS inconsistency errors in 8 out of 8 cases, confirming that the dual constraints of this study practically, rather than merely formally, exclude heterogeneous vocabulary.

7.3 "Cascade Error" (Chained Amplification of Errors) via Iterative Processing

[Hypothesized Counterargument]

In Chapters 3 and 4 of this study, a stepwise iterative matching method is adopted, identifying translation pairs with statistical singularity (Z-score of 2.0 or higher) and repeating the calculation while sequentially excluding them from the matching targets. However, this method inherently contains the critical risk of a "chained propagation of misidentifications." If even a single misidentification were included among the small number of initially established translation criteria, that distortion could propagate throughout the rest of the comparative space, inducing all subsequent translations in an erroneous direction in a snowball effect. Therefore, is it not possible that the extremely high identification rate of "99.4%" ultimately presented is merely an illusion, not the result of elucidating the true structure of the manuscript, but simply an over-convergence into an ostensibly "plausible enumeration of Latin" caused by "overfitting (forced contextual assimilation)" triggered by a few initial misidentifications?

[Refutation]

This counterargument is a vital point noting the risk of error amplification hidden in the iterative identification method.

First, if misidentifications were included in the initial anchors (pairs satisfying a Z-score of 2.0 or higher and mutually nearest neighbors) and triggered a chained overfitting, systematic inconsistencies in part-of-speech roles should frequently occur when the word groups extracted in subsequent identifications are mapped onto the manuscript's four-stage structure. The 99.4% part-of-speech consistency rate observed in this study suggests that such a systematic collapse has not occurred. However, this observation indicates "consistency at the level of distribution and type," and does not guarantee the "semantic correctness of individual words." Even if semantic errors were included in the initial anchors, as long as they are replaced by words that are distributionally consistent, consistency at the type level may be maintained.

Second, a significant positive correlation of $R=0.7080$ ($P=7.70\times10^{-29}$) was observed between the physical features on the image side (number of endpoints of branches and leaves), extracted independently from the text, and the occurrence rate of the functional role on the text side (material type). This observation indicates that if the processing on the text side were completely dominated by systematic errors, demonstrating such a high correlation with independently observed image features would be probabilistically unlikely. However, this correlation also demonstrates synchronization at the "page-level aggregated value level," and does not directly prove the correctness of the correspondence between individual words and individual botanical illustrations.

In conclusion, the concern of a chained amplification of initial errors is difficult to support at present, in the sense that there is no evidence of systematic collapse under the current data. Nevertheless, for more rigorous validation, sensitivity analyses deliberately altering the initial anchors and bootstrap-like resampling experiments remain as future tasks.

7.4 "Semantic Forced Fitting" via Pre-limitation of Parts of Speech

[Hypothesized Counterargument]

The study asserts the validity of the translation by demonstrating that not a single part-of-speech contradiction occurred as a result of applying actual Latin words to the "17 operational frameworks" and the "four-stage logical structure." However, is it not merely the result of artificially pre-limiting the part-of-speech roles for each blank in the text, such as "a noun goes here" or "a verb goes here," and selecting (forcing) words from among Latin candidates that meet those conditions? If one prepares artificial part-of-speech frameworks and inserts words accordingly, it is a natural consequence (tautology) that part-of-speech collapse will not occur, and this does not constitute proof of translation accuracy.

[Refutation]

This counterargument is an important point expressing concern over the "forced assimilation of context" that one easily falls into during decipherment work.

First, the "part-of-speech roles" in each character string of the manuscript were not arbitrarily assigned by the author guessing the meaning beforehand. As described in Chapter 5, pure geometric features, such as inline position and connection patterns with adjacent symbols, were extracted, and the number of roles that most naturally separate was objectively and automatically classified using a mathematical evaluation function called the silhouette score. It is a procedure that excludes human semantic intervention, and the part-of-speech frameworks themselves are derived from the physical structure of the manuscript.

Second, the identification of Latin adopts a design that is not adopted unless two independent conditions are simultaneously met: "contextual attraction in the multi-dimensional space" and "agreement with the automatically classified part-of-speech role." Even if one attempts to select a word with low contextual attraction just to match the part of speech, it will be rejected by the similarity cliff condition; conversely, even if the contextual attraction is high, it is discarded if the part-of-speech role does not match.

However, there is a limitation to be acknowledged here. As repeatedly stated, both conditions are distributional and formal constraints, and do not directly guarantee "semantic correctness." When multiple Latin words exist that possess the same part-of-speech type and a similar distribution pattern, this method selects the candidate with the maximum similarity among them, but there is no guarantee that it matches the "semantically closest candidate." The phenomenon of homophones discussed in Section 5.4 precisely reflects this limitation.

In conclusion, in the sense that it is not the result of artificial forced fitting of parts of speech, the strong forced fitting normally assumed probabilistically is rejected from the current data; however, a guarantee that "semantically correct individual word correspondences have been selected" cannot be derived from the framework of this method. This remains a task deferred to validation by experts in the next phase.

7.5 "Spurious Correlation" in the Linkage Between Illustrations and Text

[Hypothesized Counterargument]

In Chapter 6 of this study, it is proven that an extremely strong positive correlation of R=0.7080 exists between the number of "endpoints of lines (ends of branches and leaves)" in the illustrations drawn in the manuscript and the number of words for "target substances (materials)" identified by the translation. However, a statistical "spurious correlation (an illusion caused by hidden factors)" might be lurking here. For instance, do pages with illustrations where roots and leaves are drawn intricately (many endpoints) not tend to also have a greater "total amount of text (total number of words)" for accompanying explanations? If the text as a whole is longer, the absolute number of words indicating "materials" contained therein will naturally increase as well. In other words, the complexity of the illustrations and the number of materials are not directly linked, but rather both might simply be driven by a hidden third factor: the "amount of ink within the page (total amount of description)." Is this correlation maintained even when calculating based on the "occurrence rate" relative to all words within the page, rather than absolute word counts?

[Refutation]

This counterargument is a crucial point regarding the "spurious correlation" that is easy to fall into in quantitative text analysis.

First, the occurrence frequency of each functional role on the text side in this study is calculated not as absolute counts, but as an occurrence rate (relative density) against the total number of words on each page. Through this, the influence of the potential third factor, the "amount of ink on the page (total amount of description)," is excluded at the computational process stage.

Second, as a result of cross-referencing this "relative rate independent of the total amount" with the "physical number of endpoints in the images," a significant correlation of R=0.7080 ($P=7.70\times10^{-29}$) was observed. This demonstrates that even after compensating for the total descriptive amount of the page, systematic linkage exists between the image and the occurrence density of a specific type.

However, we acknowledge the possibility of potential third factors that cannot be completely eliminated by this verification alone. For example, the possibility that the "complexity of the page's topic itself" simultaneously increases both the complexity of the image and the relative density of the specific type cannot be entirely ruled out from the current data. Furthermore, whether the word groups classified as "material type" in this study truly correspond semantically

to botanical illustrations cannot be determined solely by correlations of page-level aggregated values, and matching with botanical identifications of individual pages remains a future task.

In conclusion, the possibility that the linkage between illustrations and text is a "spurious correlation due to simple total text volume" is reduced by the relative rate calculations of the current data, but fully excluding potential factors in a broader sense requires validation of the individual correspondences between images and words.

7.6 Rejection of Circular Reasoning in Automatic Classification of Functional Roles and Dual Matching

[Hypothesized Counterargument]

In Chapter 5 of this study, it is declared that 99.4% of the words were identified using the dual conditions of the text's "contextual attraction (similarity)" and "functional role (part-of-speech type)." However, this "functional role" is merely automatically classified from the internal data of the manuscript using a mathematical evaluation function. The logic of conveniently assigning roles to this self-classified framework, such as "this is an operation, this is a material," selecting only Latin words that fit those roles, and proving that "the roles perfectly matched" is complete circular reasoning (a self-contained interpretation). It is merely evaluating oneself with rules one has built, rather than verification using external objective grammatical rules, and is logically invalid as a proof of translation accuracy.

[Refutation]

This counterargument is a vital point noting the "self-containment of interpretation" that is easy to fall into in the decipherment of unknown languages.

First, the functional roles within the manuscript are based on the four position-dependent categories observed in the OI-2026 Protocol of this study and automatic classification via the silhouette score. They were not arbitrarily classified by the author guessing meanings, but are automatic extractions from pure physical features (position, adjacency).

Second, the parts of speech of the comparative Latin literature were not constructed by this study, but are external linguistic facts inherent in 16th-century historical texts. This method cross-references two systems of different origins: the "classification extracted from the physical structure of the manuscript" and the "external linguistic parts of speech of Latin."

However, the "type matching" in cross-referencing the two ultimately depends on the determination of the part-of-speech classifier (NLP tools like CLTK) on the Latin side, and that determination itself has aspects that cannot be said to be completely objective. Also, the fact that the automatically classified types on the manuscript side are called "materials" and "operations" includes room for interpretation. Whether the type classified as "material type" truly points to substance names can ultimately only be verified through cross-referencing with the contents of individual pages. In conclusion, this method is not simple circular reasoning, but it

also cannot be said to be completely independent verification, remaining a mathematically highly probable hypothesis. Ensuring complete independence requires cross-checks replicating the Latin part-of-speech classification with another independent classifier and qualitative review by experts.

7.7 Rejection of "Grammatical Collapse" Due to Ignoring Latin Morphological Inflections

[Hypothesized Counterargument]

This study asserts that it completed domain-wide translation by matching "functional roles (part-of-speech types)" such as "nouns (target substances)" and "verbs (main operations)" to undefined character strings. However, Latin, used as the comparative subject, is a language characterized by high "inflection (word form changes)," with strict cases for nouns (nominative, accusative, ablative, etc.) and person and tense accompanying verbs. If words are placed using only the contextual attraction of vocabulary and rough part-of-speech classifications, mismatches in the person between subject and verb, and mismatches in case between prepositions and nouns should inevitably occur. Therefore, the presented translation results are nothing more than grammatically collapsed "word salads" that make no sense when examined against the strict grammatical rules of classical Latin. When scrutinized sentence by sentence, is it truly in a "readable" state as text in a natural language?

[Refutation]

This counterargument is an extremely valid point from the perspective of classical philology considering the morphological features of Latin.

First, this study does not claim that the text of the manuscript can be "read as literary prose." As observed in the previous report, the character strings of the manuscript themselves possess features such as a hapax legomena rate of 72.93%, a combinatorial structure of standardized components, and the reuse of components without inflection, exhibiting behaviors different from linguistic structures accompanied by complex case declensions and conjugations like classical Latin.

Second, the corresponding Latin groups extracted in this study are consistent with the "recipe-like descriptive format" (sequence of operational instructions and material names) specific to medieval alchemical and pharmacological literature. A tendency is observed where imperatives, participles, and noun forms of material names are placed at corresponding positions in the manuscript.

However, readability in the sense of "forming continuous text that makes grammatical sense as Latin" is not guaranteed by this study. The output is an "array of Latin corresponding distributionally and in type consistency," and when this is placed alongside actual medieval recipes, whether the chain of operations logically holds up, or whether the combination of

materials is valid as an actual pharmaceutical prescription, requires verification by experts in herbalism and the history of science.

In conclusion, the assertion that it "has not grammatically collapsed" is supported under the framework of this study (that the manuscript has a non-inflected component structure, and its correspondences align with Latin having a recipe-like structure), but the grammatical and practical validity in the sense of being a "readable prescription as an actual medieval alchemical text" is deferred to future expert validation.

7.8 Rejection of Forced Contextual Assimilation for Hapax Legomena

[Hypothesized Counterargument]

In Chapter 2 of this study, it is clearly stated that 72.93% of the total vocabulary in the manuscript consists of words appearing only once (hapax legomena). Generally, the "adjacent relationships before and after" and the "placement structure" of a word occurring only once carry extremely little information, and relationships extracted from them are largely random fluctuations (meaningless elements). How can one claim to have accurately calculated multifaceted placement structures for these one-off vocabularies comprising over 70% of the manuscript and matched them with actual Latin? The claim of having assigned Latin on a one-to-one basis to extremely low-frequency word groups relying solely on contextual placement similarity is not an objective proof, but merely "random fitting (forced contextual assimilation)" that just picked up random fluctuations, is it not?

[Refutation]

This counterargument is a crucial point noting the limitations of handling extremely low-frequency words in computational philology.

First, the "words appearing only once" in the manuscript are not completely isolated character strings, but are composed of "combinations of shell candidates and payload candidates" observed in previous research. Individual shells and payloads appear with high frequency hundreds to thousands of times across the text domain, and sufficient statistical information is obtained at the level of these constituent components. This method assigns a certain amount of contextual information even to hapax words by treating words not as indivisible blocks, but as patterns of constituent components.

Second, if random fitting were performed for hapax words, systematic inconsistencies in part-of-speech roles should frequently occur when they are poured into the manuscript's four-stage structure. The 99.4% part-of-speech consistency rate observed in this study suggests that the matching against hapax words is not a state of merely picking up random fluctuations. However, this indicates "consistency at the distributional and type levels," and does not guarantee the "semantic correctness of individual hapax words."

The semantic verification of hapax words is an issue that cannot be completed within the framework of this method. When multiple Latin candidates exist with the same type and similar distribution patterns, this method selects the candidate with the maximum similarity among them, but there is no guarantee that it matches the true meaning of the hapax word. The homophone phenomenon discussed in Section 5.4 reflects this limitation.

In conclusion, that the processing for hapax words is not simple random fluctuation is supported under the current data, but the correctness of the semantic identification of individual hapax words is only finalized after qualitative verification by experts.

第**8**章：結論（**Conclusion**）

8.1：本研究の総括

本研究は、およそ600年間にわたり未解読の歴史的文書として人類の知的好奇心を刺激し続けてきた「ヴォイニッチ手稿」に対し、計量文献学、情報統計学、そして文字列の空間的配置を解析する構造論的アプローチを融合させることで、手稿全域に対する分布的対応候補の基盤を構築し、その歴史的実体を最も可能性の高い仮説として提示することを目的とした。著者の前報において、本手稿のテキストは自然言語が必然的に持つ動的平衡から逸脱しており、規格化された記号群を用いた構造的枠組みによって統制されていることがすでに観察されている。本研究は、その構造的基盤を前提とし、未知の文字列と16世紀の錬金術・初期医学を中心とするラテン語文献群との客観的な照合を試みたものである。

翻訳の過程においては、人間の恣意的な解釈や意味論的な推測を排除するため、多次元空間における文脈的配置の類似性と、手稿の幾何学的位置から導き出された機能的役割(品詞に相当する概念)の合致という、二重の数理的制約を用いた。その結果、残存していた未定義の空白に対し、99.4%の割合で、分布的・型整合性の両条件を満たすラテン語候補を特定することができた。これは「意味論的な翻訳の確定」ではなく、「手稿の各スロットに対し、文脈類似度と型整合性の双方を満たすラテン語が、Stream Bコーパス内に高い割合で存在する」という事実を示すものである。

抽出された対応候補群は、古典ラテン語のような流麗な散文や、未知の言語による物語ではなく、「点火せよ」「アルカリ化せよ」といった操作動詞と、無数の材料名が連続する、レシピ的構造に類する配列を示した。この観察は、本手稿が物語や哲学的著作ではなく、操作と材料の記録に近い性格を持つ可能性を支持する。ただし、個別の単語対応が意味論的に正しいかについては、本研究の数理的処理だけでは確定できず、ラテン語学・科学史の専門家による精査を必要とする。本研究は「ヴォイニッチ手稿が完全に解読された」ことを示すものではなく、「ヴォイニッチ手稿の解読に向けた、最も可能性の高い構造的基盤と分布的候補が整備された」というレベルに位置付けられる。

8.2 言語学および計量文献学に対する新たな知見

これまで、ヴォイニッチ手稿のテキストに見られる統計的な異常性、例えば一度しか出現しない単語(ハパックス・レゴメナ)が全語彙の70%以上を占める現象や、自然言語の単語出現頻度に見られる法則からの逸脱は、多くの研究者の重要な研究対象となってきた。本研究による分布的対応候補の体系的抽出は、これらの異常性に対する一つの説明として、「操作」と「材料」の連続というレシピ的構造から必然的に生じる結果である可能性を示している。限られた基本操作の枠組みに対して、無数の異なる材料名が次々と代入されるという記録体系の性質が、特異な語彙分布を形成しているという解釈である。手稿を自然言語の枠組みに当てはめて解釈しようとする従来のアプローチが行き詰まっていた理由は、ここにある可能性が高い。

さらに、本研究は図像とテキストが相互に連動する記録手法の存在を示唆した。挿絵に描かれた植物の枝葉の複雑さと、テキスト内における材料を示す単語の出現割合との間に強い連動性が観察された事実は、図像が単なる装飾や挿絵ではなく、テキストにおける記述の分量や性質に関係する視覚的な指標として機能していた可能性を示している。これは、中世から初期近代にかけての知識伝達において、視覚情報と文字情報が統合されていた可能性を示唆する観察である。ただし、この連動性がページ単位の集計値レベルでの相関にとどまることは留意すべきであり、個別の植物画と個別の単語との対応については、植物学・薬草学の専門家による検証が今後の課題となる。なぜ図像が実在の植物の模写ではなく奇妙な形態で描かれているのかという背景は、本研究では解明に至っていない。

分布的対応候補として抽出されたラテン語彙群は、本手稿が16世紀の錬金術および初期医学のパラダイムに近しい記録体系である可能性を支持する。耐久テスト(第7章)で確認された通り、錬金術コーパスは神学、料理、建築、農業といった他のドメインに対して統計的に有意な優位を示しており、これは単なる「レシピ的構造一般」ではなく、「錬金術」であることを数学的に示唆する。

著者の前報で観察された通り、手稿の図像およびテキストの根底には、古代の数理宇宙論『セフェル・イェツィラー』に由来する「水(静止・定着)」「火(上昇・精製)」「空気(均衡・媒介)」という論理フレームワークが組み込まれている可能性が示されている。本研究で抽出された対応候補群は、この概念的枠組みと矛盾しない配列を示しており、パラケルススやゲラルドゥス・ドルンらが提唱した中世の錬金術思想や初期化学・医学の実践との意味的な近さが推測される。ただし、これらはすべて「現データのもとで最も可能性の高い仮説」であり、最終的な意味論的確定には専門家の検証が不可欠である。

すなわち、ヴォイニッチ手稿は、当時の錬金術師や医師、あるいは知識階級の人々が、複雑な自然現象や薬理作用を体系化しようとした記録である可能性を本研究は示唆する。規格化された記号の組み合わせによって膨大な実用データを記録するこの体系は、長年解読を拒んできた強固な構造である。

8.3 本研究の到達点と今後の学際的展望

本研究は、ヴォイニッチ手稿のテキスト構造に対して数理的制約を適用し、最も可能性の高いラテン語彙対応候補を体系的に抽出した結果である。しかしながら、これを直ちに意味論的な「完全な正解」と断定することは適切ではない。本研究が示したのは、手稿の空白に対して「火に入れよ」といっ

た操作指示や特定の材料と数理的に最も近しい「属性(機能的役割)」を持つラテン語が、客観的な引力によって対応候補として浮上したという事実にとどまる。

したがって、本研究の成果は歴史的探求の終着点ではなく、これから始まる真の「解釈のステージ」の出発点として位置づけられるべきである。本研究が提示した対応候補リストを英語などの現代語へ意訳せず、ラテン語の語彙のまま提示しているのはこのためである。現代語への翻訳を試みる過程では、いかに客観性を期しても無意識のうちに現代人の独自の解釈や確証バイアスが混入してしまう。中世の植物名や鉱物名、操作手順は、地域や時代、学派によって使われ方が多様であり、このラテン語の羅列が当時の現実世界で具体的にどのような物質や医学的・化学的効果を意図していたのかを特定する作業は、一人の研究者で完遂できるものではない。その実態の完全な比定には、歴史学、薬草学、植物学、医学史、科学史といった多分野　の専門家たちによる共同の交差検証が不可欠である。

また、手稿に描かれた挿絵がなぜ実在の植物の模写ではなく、あえて「奇妙で非現実的な形態」で描かれなければならなかったのかという謎の解明も、次なる重要な課題として残されている。それが特定の知識の秘匿性を高めるための意図的なカモフラージュであったのか、あるいは別の合理的な理由が存在したのかについては、現段階の構造的な観察のみでは断定に至っていない。

加えて、本研究において対応候補が見つからなかったもう一つの重要な領域が、手稿全体のなかに残存する「50件の未定義の単語」の正体である。手稿のテキスト構造に対して数理的制約を適用してラテン語候補を抽出した結果、以下の50件の記号列については現在のStream Bコーパス内に合致する語彙が存在せず、無理な意訳を避けて未定義のまま保留されている。

【出力記録に残存した未定義の単語：全50件】

■ 1つのタイトル（初期化/ヘッダ属性）: 81c7ae9 (<78r.label_right_2>)

■ 3つの名詞（定数/定義属性）: Ah18((<83v.label_bottom_1>), 88ss2| (<rose.nwest_label_4>), 9àco8 (<114r.3>)

■ 残る46件の操作・中間処理・属性等: sokoM, hcSo8ae, 8azoe, oJ9, 8oe1c79, oj1oe8*, o1o9h, say979, 9hc989, occs9, 9kay9, soyoyae, oococ7Î, 8ayaeoJ9, sAyoe8ae, 8ayoes9, sosccs, oyayoe, som89, ofay79, 8ae8ay, 8cj1ch9, ea!ae*, ogayo8ap, ogo818ae, 8ay179, a8a2*asan, 8ayae89, oj1o8am, 8ayap7ae, ogaeoZ9, koesas9, hoeCayoe, ok1o8cc9, okayaxan, okoe119, okoeae, okAe8ay, oko8cco8oy, ohdcoy, ohc18ayae, oh1c9s, oho819, ohcA8, ohC9s, oh18ayay

これらの未定義の単語は、前段の数理的な分類プロセスにおいて、テキストの構造内でいかなる機能的役割を果たしているかが推定されている。このうち、操作を指示する動詞や、状態を示す形容詞といった役割を持つ「46件」の単語については、今後比較対象となる16世紀の錬金術や初期医学に関する歴史的文献の規模をさらに拡張していくことで、実在するラテン語彙として発見される可能性が示唆される。

しかしながら、図像のラベルなどに独立して配置された「1つのタイトル(見出し)」および「3つの名詞(対象物質)」としての属性を持つと推定される「計4件」の単語に関しては、既存の歴史的文献を探索しても発見されない可能性が示唆される。これらは当時の一般的な植物名や鉱物名ではなく、手稿の著者がこの複雑な作業手順の中で調合し、独自に名付けた「中間生成物」や「最終生成物」を指し示す、本手稿固有の名称である可能性が考えられる。

これら4件の生成物については、机上の文献調査や数理的な解析だけでは、その意味を解明することは困難である可能性が高い。その正体を推定するためには、本研究で示唆された「操作手順と材料の記録」の論理に従い、当時の物理的・化学的処方を実際に再現し、「いかなる物質が生成されるか」を実験的に検証する道が、一つの可能性として考えられる。これは、文献学や統計学の枠組みを越え、実験史学および化学の実践によって取り組まれるべき課題である。

本研究の貢献は、各研究者が独自の直感や推測を基に解釈を試みるという閉鎖的な状態を打破し、多分野の専門家が共通の土俵で議論を交わすための、人間の恣意性を排した「最も可能性の高い客観的な対応候補の基盤(一次資料)」を提供したことにある。今後は、中世から近代に至る科学技術史および知識伝達の研究の中で、専門家たちの共同作業によってその意味と位置づけが見出されていくことが期待される。

8.4 今後の重要な検証課題

本研究は8つの想定反証に対して現データから可能な範囲で応答したが、本研究の枠組み内では完結しない検証課題も存在する。これらは今後の学際的研究において優先的に取り組まれるべき領域として、ここに整理する。

第一に、ドメイン選定の一意性の検証である。本研究の耐久テストは、神学、料理、建築、農業という4つの対照ドメインに対して錬金術が統計的に優位を示すことを示した。しかし、これは「テストされた範囲で錬金術が最良」を意味するに過ぎず、「あらゆる可能なStream Bの中で錬金術が一意に最良」を保証するものではない。アラビア錬金術のラテン語訳、ビザンチン医学文献、ヘルメス文書群、特定の地域・流派の薬草学文献など、未テストの中世技術文献に対する分布的優位性の比較は、ドメインのより精緻な特定のために今後実施されるべき重要な検証である。

第二に、対応候補におけるハブ化現象の検証である。第5.4節で論じた通り、本研究で観察された「複数の異なるEVA記号が同一のラテン語(CUSCUTA等)に収束する現象」は、二つの解釈の可能性を持つ。一つは、これらが本当に意味的に近い物質を指している可能性。もう一つは、Stream B内に多くの文脈と中程度の類似度を持つ汎用的な高頻度語が存在し、それが系統的に対応候補として選ばれている「ハブ化」の表れである可能性。両者の判別は、ラテン語語彙の頻度分布の精査と、専門家による意味論的検証によって初めて可能となる。

第三に、Currier A と Currier B の区別への対応である。Currier(1976)以来、ヴォイニッチ手稿には少なくとも二つの統計的に異なる「言語」(Currier A と Currier B)が存在することが知られている。本研究の手法は、現在のところこの二つを統合して扱っており、それぞれに独立に手法を適用した場合、対応候補が一致するか系統的に異なるかは未検証である。両者で異なる対応が現れれば、

それは「同じ手稿内に二つの異なる内容が記録されている」という解釈に繋がり、両者で類似の対応が現れれば、「Currier A/Bの違いは内容の違いではなく筆記スタイルの違いである」という主張が可能になる。これは Voynich 研究において基本的かつ重要な検証課題であり、今後の優先課題とする。

第四に、個別単語対応と植物画同定との照合である。ヴォイニッチ手稿の植物画については、過去にも様々な同定提案がなされている。本研究で「材料型」と分類された語群が、これらの植物同定提案と整合するかは、ページ単位の集計値の相関(R=0.7080)だけでは判定できず、個別の植物画と個別のラテン語名との対応検証によって初めて確認される。

第五に、より深いドメインの特定、すなわち錬金術内部の特定の伝統・流派(パラケルスス派、ゲーベル派、アヴィセンナ系等)への絞り込みである。これは現在のStream Bの内部構造を再検討し、サブコーパスごとに分布的一致を比較することで可能となる。さらに、特定の流派固有の用語体系との照合は、純粋に統計的処理だけでは完結しない。中世錬金術史の専門家による定性的な精査が不可欠であり、本研究はこのような専門家との協働を強く望む。

第六に、出力された対応候補配列を、実際の中世錬金術文献のレシピと並べた際の構造的妥当性の検証である。「操作 → 材料 → 補助操作」という連鎖が、実在のGeber、Albertus Magnus、Pseudo-Lullus等の処方と構造的に類似するか、あるいは系統的に異なるかは、本手法の対応候補が真に錬金術的記述として機能しているかを判定する重要な検証となる。

これらの課題は、いずれも本研究の枠組みを超えるものであり、Voynich研究コミュニティと関連分野の専門家との協働によってのみ取り組み可能である。本研究は、これらの検証のための具体的な検証対象(対応候補リスト)を提供するという形で、その協働の出発点を構築することを目指した。

8.5 オープンサイエンスへの貢献

本研究の成果は、Zenodoでオープンアクセス・データとして公開し、第三者による検証と派生研究を歓迎する。具体的には、初期アンカーの抽出ログ、4段階構造の自動分類結果、分布的対応候補のマッピング、クロスモーダル相関データを含む全データセットを公開する。これにより、本研究の数理的再現性の確認、および外部コーパス(中世ラテン語医学・錬金術文献等)を用いた次フェーズ検証への独立した取り組みが可能となる。

本データセットの利用にあたっては、将来の多角的な研究を促進する目的から、Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) ライセンスを適用する。本ライセンスは、適切な引用を伴う非商用の二次解析、派生研究、改変を許諾するものであり、学際的な協働の発展を最大限に支援する設計である。

本研究が確立した分布的対応候補の基盤の上に、歴史学、薬草学、植物学、医学史、科学史の専門家たちが、新たな意味論的解釈を構築していくプロセスを心より歓迎する。ヴォイニッチ手稿という人類の歴史的遺産を完全に解き明かすためには、特定個人の視点にとどまらない分野横断的

なアプローチが不可欠である。本研究が、その学際的協働のための堅実な基盤となることを願っている。

--- (English Translation) ---

Chapter 8: Conclusion

8.1: Summary of This Study

This study aimed to establish a foundation of distributional translation candidates across the entire Voynich manuscript—a historical document that has stimulated humanity's intellectual curiosity as an undeciphered text for approximately 600 years—by integrating computational philology, information statistics, and a structuralist approach analyzing the spatial arrangement of character strings, thereby presenting its historical reality as the most probable hypothesis. In the author's previous report, it was already observed that the text of this manuscript deviates from the dynamic equilibrium inevitably possessed by natural languages and is governed by a structural framework using standardized symbol groups. Predicated upon that structural foundation, this study attempted an objective matching between the unknown character strings and Latin literature groups, centering on 16th-century alchemy and early medicine.

During the translation process, to eliminate arbitrary human interpretation and semantic conjecture, dual mathematical constraints were employed: similarity in contextual arrangement within a multi-dimensional space, and consistency with functional roles (concepts equivalent to parts of speech) derived from the geometric positions in the manuscript. As a result, for the remaining undefined blanks, we were able to identify Latin candidates satisfying both the distributional and type-consistency conditions at a rate of 99.4%. This does not represent the "definitive semantic translation," but demonstrates the fact that "for each slot in the manuscript, Latin words satisfying both contextual similarity and type consistency exist at a high rate within the Stream B corpus."

The extracted corresponding candidate groups exhibited an arrangement akin to a recipe-like structure, consisting of a continuous sequence of operational verbs such as "ignite" and "alkalize," alongside innumerable material names, rather than the fluent prose of classical Latin or a narrative in an unknown language. This observation supports the possibility that the manuscript possesses the character of a record of operations and materials rather than a narrative or philosophical work. However, whether the individual word correspondences are semantically correct cannot be finalized by the mathematical processing of this study alone and requires scrutiny by experts in Latin linguistics and the history of science. This study does not indicate that "the Voynich manuscript has been completely deciphered," but is positioned at the level where "the most probable structural foundation and distributional candidates toward the decipherment of the Voynich manuscript have been established."

8.2 New Insights for Linguistics and Computational Philology

Until now, the statistical anomalies observed in the text of the Voynich manuscript—such as the phenomenon where words occurring only once (hapax legomena) account for over 70% of the total vocabulary, or the deviation from laws seen in the word occurrence frequencies of natural languages—have been important subjects of research for many scholars. The systematic extraction of distributional translation candidates in this study presents one explanation for these anomalies: they may be the inevitable result of a recipe-like structure consisting of continuous "operations" and "materials." It is an interpretation that the nature of the recording system—where countless different material names are successively substituted into a limited framework of basic operations—forms this peculiar vocabulary distribution. It is highly probable that this is the reason why conventional approaches attempting to interpret the manuscript by fitting it into the framework of a natural language reached an impasse.

Furthermore, this study suggested the existence of a recording method where illustrations and text are mutually linked. The fact that a strong linkage was observed between the complexity of the branches and leaves of the plants depicted in the illustrations and the occurrence rate of words indicating materials within the text indicates the possibility that the illustrations functioned not merely as decorations or pictures, but as visual indicators relating to the volume and nature of the descriptions in the text. This observation suggests the possibility that visual and textual information were integrated in the transmission of knowledge from the Middle Ages to the early modern period. However, it must be noted that this linkage remains a correlation at the level of page-aggregated values, and verifying the correspondence between individual botanical illustrations and individual words remains a future task for experts in botany and herbalism. The background as to why the illustrations are drawn in bizarre forms rather than as accurate sketches of real plants has not been elucidated in this study.

The Latin vocabulary groups extracted as distributional translation candidates support the possibility that this manuscript is a recording system close to the paradigm of 16th-century alchemy and early medicine. As confirmed in the stress tests (Chapter 7), the alchemical corpus demonstrated a statistically significant superiority over other domains such as theology, cookery, architecture, and agriculture, mathematically suggesting that it is not merely a "general recipe-like structure," but specifically "alchemy."

As observed in the author's previous report, there is a possibility that the logical framework derived from the ancient mathematical cosmology "Sefer Yetzirah"—consisting of "Water (stasis/fixation)," "Fire (ascent/purification)," and "Air (balance/mediation)"—is embedded at the foundation of the manuscript's illustrations and text. The corresponding candidate groups extracted in this study exhibit an arrangement consistent with this conceptual framework, inferring a semantic proximity to medieval alchemical thought and the practice of early chemistry and medicine advocated by Paracelsus, Gerardus Dorn, and others. However, these are all "the most probable hypotheses under the current data," and expert validation is indispensable for definitive semantic confirmation.

In other words, this study suggests the possibility that the Voynich manuscript is a record by alchemists, physicians, or intellectual classes of the time attempting to systematize complex natural phenomena and pharmacological effects. This system, which records vast amounts of practical data through combinations of standardized symbols, constitutes the robust structure that has resisted decipherment for many years.

8.3 The Reach of This Study and Future Interdisciplinary Perspectives

This study is the result of systematically extracting the most probable Latin vocabulary corresponding candidates by applying mathematical constraints to the text structure of the Voynich manuscript. However, it is not appropriate to immediately declare this the semantic "absolute correct answer." What this study has shown is limited to the fact that Latin words possessing operational instructions such as "put into fire" or specific materials, and "attributes (functional roles)" mathematically closest to the blanks in the manuscript, have emerged as corresponding candidates through objective gravitational attraction.

Therefore, the results of this study should be positioned not as the endpoint of historical inquiry, but as the starting point for the true "stage of interpretation" that is about to begin. This is why the corresponding candidate list presented in this study is provided as Latin vocabulary without free translation into modern languages like English. In the process of attempting translation into modern languages, no matter how objective one strives to be, unique interpretations and confirmation biases of modern individuals inevitably intrude unconsciously. The usages of medieval plant names, mineral names, and operational procedures varied widely depending on the region, era, and school of thought; the task of identifying exactly what substances or medical/chemical effects this array of Latin intended in the real world of that time cannot be completed by a single researcher. A collaborative cross-validation by experts across multiple fields—including history, herbalism, botany, history of medicine, and history of science—is indispensable for the complete historical identification of its reality.

Moreover, elucidating the mystery of why the illustrations depicted in the manuscript had to be drawn deliberately in "bizarre and unrealistic forms" rather than as accurate sketches of real plants remains a crucial subsequent task. Whether it was an intentional camouflage to heighten the secrecy of specific knowledge, or whether another rational reason existed, cannot be definitively concluded solely from the structural observations at the current stage.

In addition, another important area where no corresponding candidates were found in this study is the true identity of the "50 undefined words" remaining throughout the manuscript. As a result of extracting Latin candidates by applying mathematical constraints to the text structure of the manuscript, vocabulary matching the following 50 symbol sequences did not exist within the current Stream B corpus, and they were retained as undefined to avoid forced translations.

[Undefined Words Remaining in the Output Record: Total 50 Items]

■ 1 Title (Boot/Header attribute): 81c7ae9 (<78r.label_right_2>)

■ 3 Nouns (Constant/Definition attributes): Ah18((<83v.label_bottom_1>), 88ss2| (<rose.nwest_label_4>), 9àco8 (<114r.3>)

■ Remaining 46 Operations, Intermediate Processes, Attributes, etc.: sokoM, hcSo8ae, 8azoe, oJ9, 8oe1c79, oj1oe8*, o1o9h, say979, 9hc989, occs9, 9kay9, soyoyae, oococ7Î, 8ayaeoJ9, sAyoe8ae, 8ayoes9, sosccs, oyayoe, som89, ofay79, 8ae8ay, 8cj1ch9, ea!ae*, ogayo8ap, ogo818ae, 8ay179, a8a2*asan, 8ayae89, oj1o8am, 8ayap7ae, ogaeoZ9, koesas9, hoeCayoe, ok1o8cc9, okayaxan, okoe119, okoeae, okAe8ay, oko8cco8oy, ohdcoy, ohc18ayae, oh1c9s, oho819, ohcA8, ohC9s, oh18ayay

The functional roles these undefined words play within the text structure have been estimated in the preceding mathematical classification process. For the "46 items" possessing roles such as verbs indicating operations or adjectives indicating states, it is suggested that they might be discovered as real Latin vocabulary by further expanding the scale of historical literature concerning 16th-century alchemy and early medicine used as comparative targets in the future.

However, regarding the "total of 4 items" estimated to possess attributes as "1 title (heading)" and "3 nouns (target substances)" independently placed on illustration labels, it is suggested that they might not be found even upon exploring existing historical literature. These may not be common plant or mineral names of the time, but rather names unique to this manuscript, coined independently by the manuscript's author to designate "intermediate products" or "final products" synthesized during these complex operational procedures.

For these four products, it is highly likely difficult to elucidate their meanings through desk-based literature surveys and mathematical analyses alone. To deduce their true identities, one possible avenue is to actually replicate the physical and chemical prescriptions of the time following the logic of the "record of operational procedures and materials" suggested in this study, and experimentally verify "what substances are generated." This is a task that must be tackled through experimental history and the practice of chemistry, transcending the frameworks of philology and statistics.

The contribution of this study lies in breaking the closed state where individual researchers attempt interpretations based on their own intuition or conjecture, providing the "most probable objective foundation of corresponding candidates (primary source material)" stripped of human arbitrariness, to enable experts from multiple fields to engage in discussions on a common ground. Looking ahead, it is expected that its meaning and position will be discovered through the collaborative work of experts within the study of the history of science, technology, and knowledge transmission from the Middle Ages to the early modern period.

8.4 Crucial Future Validation Tasks

While this study responded to eight hypothesized counterarguments to the extent possible based on current data, there are validation tasks that cannot be completed within the framework

of this study. These are organized here as areas that should be prioritized in future interdisciplinary research.

First is the validation of the uniqueness of domain selection. The stress tests in this study demonstrated that alchemy exhibits statistical superiority over four control domains: theology, cookery, architecture, and agriculture. However, this merely signifies that "alchemy is the best within the tested range," and does not guarantee that "alchemy is uniquely the best among all possible Stream B." Comparing the distributional superiority against untested medieval technical literature—such as Latin translations of Arabic alchemy, Byzantine medical texts, the Hermetica, or herbal literature from specific regions or schools—is a crucial validation that should be conducted in the future for more precise domain identification.

Second is the validation of the hubness phenomenon among corresponding candidates. As discussed in Section 5.4, the phenomenon observed in this study where "multiple distinct EVA symbols converge to the same Latin word (such as CUSCUTA)" holds two interpretative possibilities. One is the possibility that they truly refer to semantically related substances. The other is the possibility that it is a manifestation of "hubness," where generic, high-frequency words possessing many contexts and moderate similarities exist within Stream B and are systematically selected as corresponding candidates. Distinguishing between the two only becomes possible through close scrutiny of the frequency distribution of Latin vocabulary and semantic validation by experts.

Third is addressing the distinction between Currier A and Currier B. Since Currier (1976), it has been known that the Voynich manuscript contains at least two statistically distinct "languages" (Currier A and Currier B). The methodology of this study currently treats these two as integrated; whether the corresponding candidates match or systematically differ when the method is applied independently to each has not been verified. If different correspondences emerge in both, it leads to the interpretation that "two different contents are recorded within the same manuscript"; if similar correspondences emerge, the assertion becomes possible that "the difference between Currier A/B is not a difference in content but a difference in writing style." This is a fundamental and crucial validation task in Voynich research and is prioritized as a future task.

Fourth is the matching between individual word correspondences and botanical illustration identifications. Regarding the botanical illustrations of the Voynich manuscript, various identification proposals have been made in the past. Whether the word groups classified as "material type" in this study align with these botanical identification proposals cannot be determined solely by the correlation of page-level aggregated values (R=0.7080); it will first be confirmed through correspondence validation between individual botanical illustrations and individual Latin names.

Fifth is the identification of a deeper domain, namely narrowing down to specific traditions or schools within alchemy (e.g., Paracelsians, Geberians, the Avicenna tradition). This becomes possible by re-examining the internal structure of the current Stream B and comparing distributional consistencies across sub-corpora. Furthermore, matching with the terminology systems specific to certain schools cannot be completed purely through statistical processing.

Qualitative scrutiny by experts in the history of medieval alchemy is indispensable, and this study strongly desires such collaboration with experts.

Sixth is the validation of structural validity when the output corresponding candidate sequences are placed alongside recipes from actual medieval alchemical literature. Whether the sequence of "Operation → Material → Auxiliary Operation" structurally resembles prescriptions from real figures like Geber, Albertus Magnus, or Pseudo-Lullus, or systematically differs from them, will be a crucial validation in determining whether the corresponding candidates of this method truly function as alchemical descriptions.

These tasks all exceed the framework of this study and can only be tackled through collaboration between the Voynich research community and experts in related fields. This study aimed to build the starting point for that collaboration in the form of providing specific targets for validation (the corresponding candidate list) for these tests.

8.5 Contribution to Open Science

The results of this study are published as open-access data on Zenodo, and we welcome validation by third parties and derivative research. Specifically, the entire dataset—including extraction logs of initial anchors, automatic classification results of the four-stage structure, mappings of distributional corresponding candidates, and cross-modal correlation data—is made public. This enables independent efforts to confirm the mathematical reproducibility of this study and to conduct next-phase validations using external corpora (e.g., medieval Latin medical and alchemical literature).

To promote multifaceted research in the future, the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license is applied to the use of this dataset. This license permits non-commercial secondary analysis, derivative research, and modifications, provided appropriate credit is given, and is designed to maximally support the development of interdisciplinary collaboration.

We sincerely welcome the process whereby experts in history, herbalism, botany, history of medicine, and history of science construct new semantic interpretations upon the foundation of distributional corresponding candidates established by this study. In order to completely decipher the historical heritage of humanity that is the Voynich manuscript, a cross-disciplinary approach that does not remain confined to the perspective of a specific individual is indispensable. It is our hope that this study will serve as a solid foundation for that interdisciplinary collaboration.

データ提供に関する声明 (Data Availability Statement)

本研究の解析ログを、ヴォイニッチ手稿の歴史的実態の比定と解釈を進めるため、多分野の専門家による交差検証を目的として、学術リポジトリ Zenodo においてホスティングし、オープンアクセス・データとして提供する 。

データセット・リポジトリURL: [https://doi.org10.5281/zenodo.20253105]

提供される補足データセット（Supplementary Materials）は以下のファイル群で構成される 。

・Voynich_Absolute_Anchors_Strict.csv

・Voynich_Absolute_Anchors_Phase4.csv

・Voynich_Absolute_Anchors_Phase5_128D.csv

・Voynich_Decompiled_Record.txt 。

・Voynich_Undefined_Variable_Types.csv

・Voynich_Schema_Mapped_Translation.txt

・Voynich_Newly_Discovered_Words.csv

・Voynich_Absolute_Translation_Final.txt

・Voynich_CrossModal_Physical_Pointers.csv

・Voynich_Multilingual_Pipeline.txt

・各コード類

--- (English Translation) ---

Data Availability Statement

The analysis logs of this study are hosted on the academic repository Zenodo and provided as open-access data, for the purpose of cross-validation by experts across multiple fields to advance the identification and interpretation of the historical reality of the Voynich manuscript.

Dataset Repository URL: [https://doi.org10.5281/zenodo.20253105]

The provided supplementary datasets (Supplementary Materials) consist of the following files:

・Voynich_Absolute_Anchors_Strict.csv

・Voynich_Absolute_Anchors_Phase4.csv

・Voynich_Absolute_Anchors_Phase5_128D.csv

・Voynich_Decompiled_Record.txt

・Voynich_Undefined_Variable_Types.csv

・Voynich_Schema_Mapped_Translation.txt

・Voynich_Newly_Discovered_Words.csv

・Voynich_Absolute_Translation_Final.txt

・Voynich_CrossModal_Physical_Pointers.csv

・Voynich_Multilingual_Pipeline.txt

・Various associated codes

参考文献 **(References / Bibliography)**

[データソースおよびコーパス]
・EMLAP (Early Modern Latin Alchemy Project):https://zenodo.org/records/14765511
（※本研究の比較照合において基準空間として用いた、16世紀を中心とする錬金術、初期医学、自然哲学のラテン語文献データセット。パラケルスス（Paracelsus）やゲラルドゥス・ドルン（Gerard Dorn）、コンラート・ゲスナー（Conrad Gessner）らの実用的な記録を含み、現代の記号やノイズを完全にパージしたテキストコーパスとして参照）

・GreLa (Greek and Latin Repository):https://zenodo.org/records/18160596
（※古典および中世・近世のラテン語文献を網羅した研究用データベース。多次元空間におけるラテン語彙の文脈的配置、および形態論的な振る舞いを客観的に抽出するための照合基準として参照）

【先行研究および構造的基盤】

・ヴォイニッチ手稿における文字列の数理的証明（日英併記版）Version 1.1.2
（※著者の前報。論理構造（OI-2026プロトコル）を確立した）

--- (English Translation) ---

References / Bibliography

[Data Sources and Corpora]

・EMLAP (Early Modern Latin Alchemy Project): https://zenodo.org/records/14765511

(* A Latin literature dataset of alchemy, early medicine, and natural philosophy, centered around the 16th century, used as the reference space in the comparative matching of this study. It includes the practical records of figures such as Paracelsus, Gerardus Dorn, and Conrad Gessner, and is referenced as a text corpus completely purged of modern symbols and noise.)

・GreLa (Greek and Latin Repository): https://zenodo.org/records/18160596

(* A research database encompassing classical, medieval, and early modern Latin literature. Referenced as a matching standard to objectively extract the contextual placement and morphological behavior of Latin vocabulary in a multi-dimensional space.)

[Previous Studies and Structural Foundation]

・Mathematical Proof of Strings in the Voynich Manuscript (Bilingual Edition) Version 1.1.2

(* The author's previous report. Established the logical structure (The OI-2026 Protocol).)