Research Article

The Theoretical Background of (Dis) similarity Algorithms: Analyzing the Voynich Manuscript as an Application

Ramon Carbó-Dorca[1]

1. Institut de Química Computacional i Catàlisi (IQCC), University of Girona, Girona, Spain

This paper gathers, in self-contained form, the mathematical background of tagged sets, a generalization of fuzzy sets, and of the (dis) similarity measures built on them: inward products, scalar products and norms of arbitrary order, Minkowskian metric vectors, Gram matrices and hypermatrices, and collective distances and similarity indices. As an application, the text of the Voynich manuscript is compared with the Universal Declaration of Human Rights in 62 languages and orthographies, with medieval Catalan, Castilian, and Latin texts, and with ciphers and text generators of known nature. At the character level, the manuscript lies more than eight standard deviations from the nearest reference text, although distance alone does not single it out, since a natural language in an unusual orthography can lie as far. The anomaly is located inside the word, in the information carried by its fourth and fifth characters, and a slot-grammar generator fitted to the manuscript reproduces most of it, far better than the same construction reproduces any language tested. At the word level, the vocabulary statistics of the manuscript are those of a language, and its adjacent words carry as little mutual information as those of the sparsest Latin technical prose measured. The analysis neither deciphers the manuscript nor proves it meaningless; it characterizes the text as more compatible with a word-building procedure than with a simply enciphered language.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction

In 1980 [1] the author's laboratory used ideas from the AI of the time to develop, within quantum mechanics and using the first-order electronic density function, two numerical algorithms to compare

molecules, which the author called quantum (dis) similarity techniques. Over time, the same author has published more than 150 papers on quantum similarity and related problems such as QSAR.

Among these works, several articles [2][3] describe extensions of the initial theoretical background, applicable to comparing objects of any kind, provided they could be mathematically described by a numerical entity in the form of a vector or a function, which the author called an object tag.

Moreover, after using the usual two-object evaluation framework along the way to classify and relate objects, several new mathematical extensions of the initial algorithms have also been studied; see, for example, references [4][5][6][7]. These extensions aimed to expand the usual two-object comparison framework to a many-object perspective.

Because the mathematical structure of these (dis) similarity comparison techniques applies to any set of tagged objects, this review of the theoretical background of (dis) similarity mathematical structure could be used as the basic knowledge to compare the Voynich manuscript with texts in several languages, attempting to elucidate, if possible, the nature of this enigmatic written book. Thus, the present article can be viewed from two aspects. One provides a self-contained description of the mathematics of (dis) similarity theory; the other discusses the Voynich manuscript.

Sections 2 to 5 describe the mathematical background of the (dis) similarity framework, and Section 6 applies it to the Voynich manuscript. Every element of the theory is used there except the collective distances of Section 5.1, which are kept for the sake of completeness.

A word of caution is due here. Some definitions in the following theoretical sections are well known and have different names in other mathematical contexts. Here, we link them to support the study of relationships among object sets, based on vector descriptions of varied types of tagged object sets. A good deal of the definitions that structure the present paper can be found in extended form in the article [8].

The result is a framework that benefits research in many scientific branches. Particularly in those involving algorithmic relationships between objects of any kind, whenever they can be attached to vectors of any dimension. This attachment includes infinite-dimensional vectors in the form of functions, as occurs when submicroscopic objects of quantum nature are studied, and then density functions are chosen as object tags.

2. Vector Semispaces

A vector semispace is the non-negative orthant of a vector space: a set of vectors closed under addition and under multiplication by non-negative scalars, with elements taken from the non-negative rational number set $\mathbf{Q}^+$. In a semispace, addition forms a semigroup, so negative vectors and vector subtraction are not defined. One can therefore use a symbol like: $V_N\left(\mathbf{Q}^+\right)$ to represent such an $N-$dimensional vector semispace.

Then, a vector as an element of $V_N\left(\mathbf{Q}^+\right)$and of other true vector spaces will be defined here with Dirac's bra notation as a row $N-$tuple of rational numbers, that is:

$$\forall \langle \mathbf{v}| \in V_N\left(\mathbf{Q}^+\right) : \langle \mathbf{v}| = (v_1, v_2, v_3, \ldots, v_N) \wedge \{v_I \,| I = 1, N\} \subset \mathbf{Q}^+. \tag{1}$$

It will be the same to have chosen as the main elements of $V_N\left(\mathbf{Q}^+\right)$ the column transpose vectors, that can be described by Dirac's ket symbols, or: $|\mathbf{v}\rangle = \left\langle \mathbf{v}^T\right|$. We generally chose the vector row form instead of the dual column representation for easier typographical handling. This choice does not prevent the use of the ket form when necessary.

Moreover, a semispace has also to be seen as a subset of a vector space, where the non-negative reference rational subset of semispaces is substituted by the whole rational field $\mathbf{Q}$, that is: $V_N\left(\mathbf{Q}^+\right) \to V_N\left(\mathbf{Q}\right)$.

2.1. Object Description in a Vector Semispace Structure: Tagged Sets

When an arbitrary set of objects: $\Omega = \{\omega_J \,| J = 1, M\}$ has to be studied within some algorithm involving some attached mathematical structure, then to every object in Ω one can assign a vector of some semispace $V_N\left(\mathbf{Q}^+\right)$, in such a way that a vector subset of $V_N\left(\mathbf{Q}^+\right)$ like:

$$T = \{\langle \mathbf{t}_J| \,| J = 1, M\} \subset V_N\left(\mathbf{Q}^+\right), \tag{2}$$

can be associated with the object set as follows:

$$\begin{aligned} \forall J = 1, M : \omega_J \in \Omega \Rightarrow \exists \langle \mathbf{t}_J| \in T \to \\ \exists \zeta_J = (\omega_J; \langle \mathbf{t}_J|) \to \zeta_J \in Z = \Omega \times T. \end{aligned} \tag{3}$$

The set T will be called a tag set, and the set of pairs Z can be named a tagged object set; strictly, it is the subset of the Cartesian product of both sets formed by the pairs that associate each object with its own tag.

Tagged sets can be considered a generalized version of the well-known fuzzy sets [9].

2.2. From semispaces to spaces: $N-$dimensional polyhedra

Also called polytopes, $N-$dimensional polyhedra can be defined by a set of M vectors belonging to an $N-$dimensional semispace. For example, the set of vectors earlier associated with the vector tag set T described in equation (3) can be considered a polytope, constructed with respect to an arbitrarily located vector origin.

However, the vector set Tcan also be used to compute a centroid vector using the arithmetic mean of its elements, that is:

$$\langle \mathbf{c}| = M^{-1} \sum_{J=1}^{M} \langle \mathbf{t}_J| \in V_N\left(\mathbf{Q}^+\right). \tag{4}$$

Once the centroid of the $N-$dimensional polyhedron is known, one can shift the origin of the set T with it, generating a new set no longer belonging to the semispace but to the extended space $V_N\left(\mathbf{Q}\right)$:

$$S = \{\langle \mathbf{s}_J| = \left(\langle \mathbf{t}_J| - \langle \mathbf{c}|\right) | J = 1, M\} \subset V_N\left(\mathbf{Q}\right), \tag{5}$$

assuming the property that the new polyhedron will possess a zero centroid:

$$\langle \mathbf{z}| = M^{-1} \sum_{J=1}^{M} \langle \mathbf{s}_J| = M^{-1} \sum_{J=1}^{M} \left(\langle \mathbf{t}_J| - \langle \mathbf{c}|\right) = \langle \mathbf{c}| - MM^{-1} \langle \mathbf{c}| = \langle \mathbf{0}|. \tag{6}$$

Therefore, this origin shift can be used to transform any tag set into an origin-centered polyhedron constructed within the vector space $V_N\left(\mathbf{Q}\right)$instead of the semispace $V_N\left(\mathbf{Q}^+\right)$.

So, even if one starts constructing object tags in a semispace, one can perform this general change of origin via the centroid and a subsequent origin shift. This transforms the original semispace vector set into a vector space tag set, while preserving the initial polyhedral structure, that is, all the distances between the tag vectors. Scalar products, cosines, and the Gram matrix of Section 4 are not preserved by the shift and must be computed again from the new origin. The advantage of shifting from an arbitrary origin to the true space origin consists in that any arbitrarily constructed tag set polyhedron can be brought to a general common structure centered at the zero vector. In this way, the universe of tag sets can be uniformized around a new common origin.

3. Inward Product of Vectors

Among the operations that can be built with vectors, one already described by Hadamard [10], Schur [11], and more recently by Horn [12], but named here the inward product, as in the initial reference [4], has

proven a helpful tool for developing generalized ways to compare tagged object sets.

The inward product of two vectors multiplies two $N-$dimensional vectors, giving a new vector of the same dimension by a simple algorithm:

$$\forall \{\langle \mathbf{a}|, \langle \mathbf{b}|\} \subset V_N(\mathbf{Q}) : \langle \mathbf{p}| = \langle \mathbf{a}| * \langle \mathbf{b}| \in V_N(\mathbf{Q}) \leftarrow \{p_I = a_I b_I \,|I = 1, N\} \tag{7}$$

Owing to the properties of the vector elements, this operation can be considered as an extended product of the elements of the set $\mathbf{Q}$, permitting us to consider that the vectors of any vector space can behave as scalars with respect to the inward product.

Such a trivial definition can be extended to an arbitrary number of vectors, and therefore one can write the inward product of M vectors in a semispace as:

$$\begin{aligned} &\forall \{\langle \mathbf{t}_J| \,|J = 1, M\} \subset V_N\left(\mathbf{Q}^+\right) \Rightarrow \\ &\quad \underset{J=1}{\overset{M}{*}} \langle \mathbf{t}_J| = \langle \mathbf{p}| \in V_N\left(\mathbf{Q}^+\right) \leftarrow \left\{p_I = \prod_{J=1}^{M} t_{JI} \,|I = 1, N\right\}. \end{aligned} \tag{8}$$

Such an extended inward product property also permits defining the *inward power P* of a vector easily, as:

$$\forall \langle \mathbf{t}| \in V_N\left(\mathbf{Q}^+\right) \Rightarrow \left\langle \mathbf{t}^P\right| = \underset{K=1}{\overset{P}{*}} \langle \mathbf{t}| \in V_N\left(\mathbf{Q}^+\right) \leftarrow \left\{t_I^P = \prod_{K=1}^{P} t_I \,|I = 1, N\right\}. \tag{9}$$

These definitions and the ones of the following sections hold the same when vectors of the space $V_N(\mathbf{Q})$ are used.

The neutral element of the inward product is the unity vector, whose elements are all equal to one, since its inward product with any vector returns the same vector. It will appear below as the metric vector of the Euclidean case.

3.1. Generalized Scalar Products, Norms, and Metric Vector

3.1.1. Sum of the elements of a vector

Inward products can easily be seen as an intermediate operation leading to a scalar product of vectors. To see this possibility, one must first define another trivial operation on a vector: the sum of its elements, which can be written simply as:

$$\forall \langle \mathbf{t}| \in V_N\left(\mathbf{Q}^+\right) : \langle \langle \mathbf{t}| \rangle = \sum_{I=1}^{N} t_I \in \mathbf{Q}^+. \tag{10}$$

The sum of the elements of any vector acts like a linear operator that transforms any vector into a scalar of the definition field, which in our case is $\mathbf{Q}^+$. Again, the same sum of vector elements operator can be used in vector spaces $V_N(\mathbf{Q})$.

3.1.2. Scalar product of two vectors

To obtain the scalar product of two vectors, one can start from the definition of their inward product as in equation (7) and apply to the resultant vector the complete sum of its elements; the result coincides with the scalar product of two vectors of $V_N\left(\mathbf{Q}^+\right)$. That is, using the bra-ket product:$\langle\mathbf{a}|\mathbf{b}\rangle$ in Dirac's symbolism for the scalar product:

$$\forall\left\{\langle\mathbf{a}|,\langle\mathbf{b}|\right\}\subset V_N\left(\mathbf{Q}^+\right):\sigma=\langle\mathbf{a}|\mathbf{b}\rangle=\langle\langle\mathbf{a}|*\langle\mathbf{b}|\rangle=\sum_{I=1}^{N}a_Ib_I\rightarrow\sigma\in\mathbf{Q}^+. \tag{11}$$

3.1.3. Scalar product of many vectors

Following the way a classical two-vector scalar product is defined, one can use the definition of the multiple vector inward product in equation (8) and then take the complete sum of the resulting vector, that is:

$$\begin{aligned}&\forall\left\{\langle\mathbf{t}_J|\,|J=1,M\right\}\subset V_N\left(\mathbf{Q}^+\right)\Rightarrow\\&\qquad\pi=\langle\mathbf{t}_1\mathbf{t}_2\mathbf{t}_3\ldots\mathbf{t}_M\rangle=\left\langle\overset{M}{\underset{J=1}{*}}\langle\mathbf{t}_J|\right\rangle=\sum_{I=1}^{N}\left(\prod_{J=1}^{M}t_{JI}\right)\rightarrow\pi\in\mathbf{Q}^+.\end{aligned} \tag{12}$$

So, the classical scalar product can be easily generalized to any set of vectors of arbitrary cardinality M and can be obviously extended in vector spaces $V_N(\mathbf{Q})$.

Two remarks will be useful later. First, the scalar product of many vectors is a two-vector scalar product in disguise: it equals the scalar product of any one of the vectors with the inward product of all the others. Comparing an object with a pair, a triple, or a larger set of other objects thus amounts to comparing it with a single composite vector. Second, it is a symmetric multilinear form rather than a scalar product in the axiomatic sense: in a vector space, and for an odd number of vectors, it can be negative even when all the vectors are equal.

3.1.4. The vector norm of arbitrary order

In the same way as the scalar product could be generalized to an arbitrary vector subset cardinality, the norm of any order can be constructed from the definition of the arbitrary power of a vector as in equation

(9). It is a matter of computing the P-th power of any vector and applying the complete sum of the elements of the resultant vector. One can write thus:

$$\forall \langle \mathbf{t}| \in V_N\left(\mathbf{Q}^+\right) \Rightarrow E_P\left[\langle \mathbf{t}|\right] = \left\langle \left\langle \mathbf{t}^P\right|\right\rangle = \sum_{I=1}^{N} t_I^P \in \mathbf{Q}^+. \tag{13}$$

In the case that a vector belonging to a vector space $V_N\left(\mathbf{Q}\right)$ is to be normalized via an algorithm like the equation (13), and the norm is to be kept non-negative, the powers of the elements shall be kept positive and the transformation of the vector elements: $\forall I = 1, N : \left\{t_I^P\right\} \rightarrow \left\{\left|t_I\right|^P\right\}$ has to be taken into account. However, there is another alternative but bold way to define the norms of the equation (13) non-negative, simply using the absolute value of the sum of the elements of the P-th power vector. A norm so defined is non-negative, but it can vanish for a non-zero vector when the order is odd. Note also that no root is taken: the norm of order P defined here is the P-th power of the P-norm of the mathematical literature.

3.1.5. Scalar products and norms in Minkowskian spaces

The scalar products and norms defined so far must be considered performed within a Euclidean metric.

All of them, though, can be transformed into a Minkowskian metric by constructing a metric vector $\langle \mathbf{m}| \in V_N\left(\mathbf{Q}\right)$, which can be included in the expressions associated with equations (11),(12) and (13).

This metric vector can no longer be considered as in the semispace $V_N\left(\mathbf{Q}^+\right)$ but rather in the vector space $V_N\left(\mathbf{Q}\right)$. It can be constructed as a unity vector with some of its elements being negative, for instance:

$$\begin{aligned} \langle \mathbf{m}| &= \langle \langle \mathbf{1}_{N_E}| ; - \langle \mathbf{1}_{N_M}| | \\ &= \left((1, 1, \ldots, 1)_{N_E} ; (-1, -1, \ldots, -1)_{N_M} \right) \leftarrow N_E + N_M = N. \end{aligned} \tag{14}$$

Once the metric vector, as in the equation (14) is defined, then the Minkowskian scalar product can also be constructed, including the metric vector as a factor within the vector products. That is, using instead of the equation (12) the expression:

$$\begin{aligned} \forall \left\{ \langle \mathbf{t}_J| \, | J = 1, M \right\} \subset V_N\left(\mathbf{Q}^+\right) \Rightarrow \mu = \langle \mathbf{t}_1 \mathbf{t}_2 \mathbf{t}_3 \ldots \mathbf{t}_M \mathbf{m} \rangle = \\ \left\langle \left(\overset{M}{\underset{J=1}{*}} \langle \mathbf{t}_J| \right) * \langle \mathbf{m}| \right\rangle = \sum_{I=1}^{N} \left(\prod_{J=1}^{M} t_{JI} \right) m_I \rightarrow \mu \in \mathbf{Q}. \end{aligned} \tag{15}$$

Equation (15), when the metric vector equals the unity vector, that is, if: $\langle \mathbf{m}| = \langle \mathbf{1}_N|$ corresponds to the Euclidean metric expression of equation (12).

In the same way, the corresponding Euclidean norms, as are defined in equation (13), can be generalized into a Minkowskian framework using:

$$\forall \langle \mathbf{t}| \in V_N\left(\mathbf{Q}^{+}\right) \Rightarrow M_P\left[\langle \mathbf{t}|\right] = \left\langle \left\langle \mathbf{t}^P \right| * \langle \mathbf{m}| \right\rangle = \sum_{I=1}^{N} t_I^P m_I \in \mathbf{Q}. \tag{16}$$

Note that now the Minkowskian norms can be negative or zero. Taking the absolute value removes the sign but does not return the Euclidean norm, which is recovered only when the metric vector is the unity vector in equation (16).

Three remarks on terminology and scope are in order. In the mathematical literature, the name Minkowski distance usually denotes the distance derived from the P-norm; here, as in relativity, Minkowskian refers to a metric vector with elements of both signs. With such a metric, the Gram matrix of Section 4.1 is no longer guaranteed to be non-negative definite, and a cosine-like index built from Minkowskian scalar products and norms is not bounded by one and loses its meaning when a norm is zero or negative. Finally, the metric vector need not be restricted to elements of unit modulus: any vector of positive weights defines a weighted Euclidean metric, and the familiar standardization of descriptors by their standard deviations is an origin shift to the centroid, as in Section 2.2, followed by a metric vector of this kind.

4. Comparison of Two Objects

One can use the previous sections' definitions to compare two or more tagged objects in the framework using what we call classical (dis) similarity measures. However, before introducing the mathematical structures that demand further specific application algorithms (these depend on the problem under study and will not be discussed here), which one can later employ to manipulate and extract the appropriate information, we must first define a matrix common to any $N-$dimensional polyhedron tag set.

4.1. Gram matrix of an $N-$dimensional polyhedron tag set

The Gram matrix [13] of a vector tag set can be readily constructed once a vector space is such that it holds some scalar product definition, as it has been done here in the section 3.1.2.

Once the scalar product is defined, it can be computed for every pair of vectors in a given subset of cardinality M of the corresponding vector space, as the polyhedron tag sets already defined. The result of

such a collective calculation is (in our case) a $(M \times M)$ symmetric matrix, which can be easily constructed by the algorithm:

$$\forall \{\langle \mathbf{t}_J | \, | J = 1, M\} \subset V_N \left(\mathbf{Q}^+\right) : \mathbf{\Gamma} = \{\Gamma_{JK} = \langle \mathbf{t}_J | \mathbf{t}_K \rangle \, | J, K = 1, M\}, \tag{17}$$

With the possibility that the vector tag set can be defined in $V_N(\mathbf{Q})$, as it has been indicated in previous definitions.

The Gram matrix in equation (17) because of the symmetry of the scalar products as defined here in equation (11), corresponds to a symmetric matrix, that is:

$$\forall J, K = 1, M : \langle \mathbf{t}_J | \mathbf{t}_K \rangle = \langle \mathbf{t}_K | \mathbf{t}_J \rangle \rightarrow \mathbf{\Gamma} = \mathbf{\Gamma}^T. \tag{18}$$

Gram matrices can also be expressed via the product of the $(N \times M)$ matrix, whose columns are the tag set ket vectors, by its $(M \times N)$transpose, whose rows are the bra vectors:

$$\mathbf{T} = (|\mathbf{t}_1\rangle, |\mathbf{t}_2\rangle, |\mathbf{t}_3\rangle, \ldots, |\mathbf{t}_M\rangle) \Rightarrow \mathbf{\Gamma} = \mathbf{T}^T \mathbf{T}, \tag{19}$$

as a consequence, the Gram matrix must be non-negative definite.

Only when the vector tag set is linearly independent is the Gram matrix positive definite with a non-zero determinant.

4.2. The Gram matrix as a new representation of any tag set

Observing the Gram matrix as defined in equations (17), (18) and (19), one can focus on the nature of every row (or, being this matrix symmetric, on every column). Then every column of the matrix (17) can be written as a new vector with the simple algorithm, involving the equivalent nature (considering the applications one is describing) of rows and columns of the Gram matrix:

$$\mathbf{\Gamma} = (|\boldsymbol{\gamma}_1\rangle, |\boldsymbol{\gamma}_2\rangle, |\boldsymbol{\gamma}_3\rangle, \ldots, |\boldsymbol{\gamma}_M\rangle) \rightarrow \mathbf{\Gamma}^T = \{\langle \boldsymbol{\gamma}_J | \, | J = 1, M\}. \tag{20}$$

However, every column, as written in equation (20) can be constructed in turn by:

$$\forall J = 1, M : \langle \boldsymbol{\gamma}_J | = \{\langle \mathbf{t}_J | \mathbf{t}_K \rangle \, | K = 1, M\}, \tag{21}$$

That is: every row $\langle \boldsymbol{\gamma}_J |$ of the Gram matrix corresponds to the scalar products of the associated tag vector $\langle \mathbf{t}_J |$ with all elements of the tag set. Thus, one can use the rows of the Gram matrix as representations of the vectors of the original tag set, constructing a new tag set Γ which can also be attached to the original object set:

$$\Gamma = \{\langle \boldsymbol{\gamma}_J | \, | J = 1, M\}, \tag{22}$$

where the new tag elements are the original vector tags, but now connected with all the original tag set vectors, as shown in the equation (21).

One can call the new tag set Γas the Gram tag set. The Gram tag set can soundly substitute for any tag set polyhedron attached to any set of objects, acquiring for every original tag a relationship with every vector in the original tag set T.

In fact, in the present approach one not only constructs a new Gram tag set, but also provides a general way to make any tag set homogeneous for any arbitrarily attached object set in a broader sense than the origin shifted set Sdefined in equation (5).

Of course, the Gram matrix can be computed using the original tag set or the origin-shifted set by the procedure discussed in the equation (5) of section 2.2.

The two Gram matrices are different. If H is the centering matrix, that is, the unit matrix minus the matrix with all its elements equal to 1/M, the Gram matrix of the origin-shifted set is the product HΓH, with Γ the Gram matrix of the original set. It is always singular, since its rows add up to zero.

Moreover, the centroid of the Gram tag set is related to the centroid of the original tag set, because:

$$\begin{gathered} \langle \boldsymbol{\gamma}_C | = M^{-1} \sum_{J=1}^{M} \langle \boldsymbol{\gamma}_J | \rightarrow \langle \boldsymbol{\gamma}_C | = \left\{ \gamma_{CI} = M^{-1} \sum_{J=1}^{M} \langle \mathbf{t}_J | \mathbf{t}_I \rangle \, | I = 1, M \right\} \\ \left\{ \gamma_{CI} = \left\langle \left(M^{-1} \sum_{J=1}^{M} \langle \mathbf{t}_J | \right) * \langle \mathbf{t}_I | \right\rangle | I = 1, M \right\} \\ \{ \gamma_{CI} = \langle \mathbf{t}_C | \mathbf{t}_I \rangle \, | I = 1, M \} \end{gathered} \tag{23}$$

That is, the centroid elements of the Gram tag set correspond to the scalar products of the original tag set centroid with all vector elements of the original tag set.

Two further properties of the Gram tag set may be noted. Since each Gram tag is the original tag multiplied by the matrix of the whole tag set, the squared distance between two Gram tags is the squared distance between the original tags, measured with a metric that the tag set itself defines; and the Gram matrix of the Gram tag set is the square of the original Gram matrix. The Gram tag set has, therefore, the same rank as the original one, and it does not escape the dimensionality paradox discussed in Section 4.3.

4.3. The dimensionality paradox

After discussing the linking of the information within objects described by vector tags, it will be of interest to delve a bit further into the connection $object \Leftrightarrow vector\ tag$, an element present in all the

scientific applications using the present framework.

In chemistry, for example, molecules are usually represented as objects described by tag vectors, constructed from parameters obtained from many sources. See, for example, the exhaustive compilation in reference [14] of available parameters. This procedure allows one to find relations using well-known methods, generically called QSPR (Quantitative Structure-Property Relationships).

The basic idea is the one associated with the present formalism: every different molecular structure of a set made by M of them is associated with an $N-$dimensional vector. This simple assembly connection gives rise to the so-called dimensionality paradox [15].

If every molecule in the object set is completely different from the rest of the set's objects under study, then to represent such a uniqueness vectorially, the vectors associated with the tag set must be linearly independent.

Such a tag set independence is possible only when the tag set space or semispace has a dimension equal to or greater than the number of molecules in the object set, $N \geq M$, according to the properties of the attached Gram matrix of the tag set vectors studied in the preceding section.

Unfortunately, in usual QSPR procedures, the tag set vectors' dimension is smaller than the number of objects in the study. Since the 19^{th} century, many QSPR relationships have been developed using simple monodimensional tag vectors (scalars).

Thus, current QSPR procedures appear associated with a dimensionality paradox: the method tries to relate a large number of independent objects using tag set vectors with dimensions much smaller than the number of objects. The paradox is overcome only when considering quantum mechanical density functions, which can act as infinite-dimensional tag vectors attached to the molecules of the object set. In this case, QSPR becomes Quantum QSPR (QQSPR); see earlier developments [16].

Therefore, the dimensionality paradox may also appear in other procedures and applications generally based on the tagged set formalism discussed so far. Removing it may be difficult unless the studied objects, such as molecular structures, can be tagged with adequate functions.

One must be aware of this geometrical limitation related to the dimensionality paradox when handling objects' relationships using finite-dimensional tag vectors.

4.4. Dissimilarity measures

One can first associate a dissimilarity measure with a simple squared distance, eliminating the triangle inequality from the definition axioms, and write for this purpose:

$$\forall \{\langle \mathbf{a}|, \langle \mathbf{b}|\} \subset V_N(\mathbf{Q}) : D^{[2]}(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}|\mathbf{a}\rangle + \langle \mathbf{b}|\mathbf{b}\rangle - 2\langle \mathbf{a}|\mathbf{b}\rangle, \tag{24}$$

and one can see how the elements of the Gram matrix (18) coincide with the elements intervening within the distance construction in equation (24).

Then, if the set $\{\langle \mathbf{a}|, \langle \mathbf{b}|\}$ corresponds to a pair of tags attached to two objects $\{\alpha, \beta\}$, the equation (24) permits computing the amount of dissimilarity between them, as when: $D^{[2]}(\mathbf{a}, \mathbf{b}) = 0$, both tags are the same, and one can admit that their associated objects are (within the chosen tag structure and information content) equivalent or equal: $\{\alpha \equiv \beta\}$.

Because the distance is unbounded, it can grow to infinity, so it measures the intensity of the difference between the two objects within the chosen tag structure. One needs to be aware of what one obtains when a distance like the equation (24) is computed. The result reflects the objects only through the pair of vector tags chosen to describe them.

As equation (24) can be written as the second-order norm of the difference of the involved vectors, one can also obtain:

$$\begin{aligned} \forall \{\langle \mathbf{a}|, \langle \mathbf{b}|\} \subset V_N(\mathbf{Q}) : \langle \mathbf{\Delta}| = \langle \mathbf{a}| - \langle \mathbf{b}| \in V_N(\mathbf{Q}) \\ \Rightarrow D^{[2]}(\mathbf{a}, \mathbf{b}) = \langle \langle \mathbf{\Delta}| * \langle \mathbf{\Delta}|\rangle = \left\langle \left\langle \mathbf{\Delta}^2 \right|\right\rangle, \end{aligned} \tag{25}$$

therefore, it is easy to extend the order of the distance between two vectors arbitrarily, using equations (13) and (25):

$$\forall \{\langle \mathbf{a}|, \langle \mathbf{b}|\} \subset V_N(\mathbf{Q}) : D^{[P]}(\mathbf{a}, \mathbf{b}) = \left\langle \left\langle \mathbf{\Delta}^P \right|\right\rangle. \tag{26}$$

One must also use the absolute value of the vector's elements when the power is an odd natural number, so the norm remains non-negative. In Minkowskian distances, where the metric vector shall be used as in the equation (16), even if the Euclidean distance is non-zero, a chosen Minkowskian one might be null.

For an even order, the distance can be expanded binomially into scalar products of many vectors. At order four, for instance, the distance between two tag vectors a and b is (aaaa) - 4(aaab) + 6(aabb) - 4(abbb) + (bbbb), where each parenthesis is a four-vector scalar product. The distances of even order are thus contained in the Gram hypermatrices of Section 5.2, in the same way as the squared distance is contained in the Gram matrix.

4.5. Dissimilarity measures and the Gram matrix

Suppose one knows the Gram matrix $\mathbf{\Gamma} = \{\gamma_{JI} = \langle \mathbf{t}_J | \mathbf{t}_I \rangle \,|I, J = 1, M\}$of some tag set $T = \{\langle \mathbf{t}_J | \,|J = 1, M\}$. A vector with the second-order Euclidean norms can also be constructed: $\langle \mathbf{N}| = \{N_J = \langle \mathbf{t}_J | \mathbf{t}_J \rangle \,|J = 1, M\}$.

With this norm vector, one can construct a symmetric $(M \times M)$matrix $\mathbf{N}$, using the tensor sum of a vector [17][18], in this case:

$$\mathbf{N} = \langle \mathbf{N}| \,[\![+]\!]\, \langle \mathbf{N}| = \{N_{JI} = N_J + N_I = \langle J|J\rangle + \langle I|I\rangle \,|J, I = 1, M\} \tag{27}$$

then the $(M \times M)$ matrix $\mathbf{D}^{[2]}$of the second-order distances of a given tag set, defined in general according to the equation (24) can be constructed by the matrix composition:

$$\mathbf{D}^{[2]} = \mathbf{N} - 2\mathbf{\Gamma} = \left\{ D_{JI}^{[2]} = N_{JI} - 2\Gamma_{JI} \,|J, I = 1, M \right\}. \tag{28}$$

This relation is unaffected by an origin shift, because the distances are. Applied to the origin-shifted set, it gives the shifted Gram matrix as −½ HDH, where D is the distance matrix just defined and H is the centering matrix of Section 4.2. The eigenvectors of that matrix, scaled by the square roots of its eigenvalues, are the coordinates of classical multidimensional scaling, which is thus part of the present formalism and will be used in Section 6.

4.6. Similarity measures

Complementary to the dissimilarity measures in the previous section, one can use another classical scalar expression to compare the tag vectors of a pair of objects: a similarity measure.

One possible way to do this is to use the squared cosine of the angle between two vectors. One can write:

$$\forall \{\langle \mathbf{a}| \,, \langle \mathbf{b}| \} \subset V_N \left(\mathbf{Q} \right) : C^2 \left(\mathbf{a}, \mathbf{b} \right) = \frac{\left| \langle \mathbf{a} | \mathbf{b} \rangle \right|^2}{\langle \mathbf{a} | \mathbf{a} \rangle \, \langle \mathbf{b} | \mathbf{b} \rangle}. \tag{29}$$

In this case, the similarity measure of the equation (29) is constrained within the unit interval: $C^2 \left(\mathbf{a}, \mathbf{b} \right) \in [0, 1]$. One can interpret the extreme values as completely dissimilar objects when the measure is 0, and completely similar when the value reaches 1. Squared distances and cosines behave inversely: for two tag vectors of unit norm, the squared distance equals twice the difference between one and the cosine, so that one measure grows as the other decreases. Unlike the distance, the cosine depends on the origin, and in a semispace it cannot be negative.

5. Comparison of Many Objects

The two kinds of (dis) similarity measures discussed in the previous section involve only two tagged objects. This constitutes the classical, algorithmic, usual way to compare objects in any tagged set.

However, other methods might compare several objects at once and, as in the two-object case, one can consider the two ways of obtaining (dis) similarity information from the tag vectors in this less familiar way. We now discuss several aspects of this possibility.

5.1. Collective distances

When computing the squared distance matrix using equations (25) or (26) with the whole set of pairs of vector tags, a symmetric $(M \times M)$ matrix with non-negative elements can be computed. One can then write:

$$\forall \{\langle \mathbf{t}_J| \, | J = 1, M\} : \mathbf{D}^{[P]} = \left\{ D^{[P]}_{JK} = D^{[P]}\left(\mathbf{t}_J; \mathbf{t}_K\right) | J, K = 1, M \right\}. \tag{30}$$

These distance matrices, involving all possible two-object dissimilarities between the vector tags, can be used to construct collective dissimilarity measures.

First, define the axioms a collective distance measure must satisfy. One can write them in this way:

Suppose a tag set as the one used throughout this article, composed of Mvectors: $T_M = \{\langle \mathbf{t}_J| \, | J = 1, M\}$. One can use a subset of these elements and form: $T_{M;m} = \{\langle \mathbf{t}_j| \, | j = J_1, J_m\}$. For example, when $m = 2$, the subsets $T_{M;2}$will be all the vector pairs forming the distances of the matrix (30).

A collective $m-$distance of order P: $D^{[P]}_m\left(T_{M,m}\right)$can be associated with these constraints, which were described in [19]:

1. $D^{[P]}_m\left(T_{M,m}\right) = 0$ when $T_{M,m}$ contains all the m elements equal to a unique vector $\langle \mathbf{t}|$.
2. $D^{[P]}_m\left(T_{M,m}\right) > 0$ in the general case where $T_{M;m} = \{\langle \mathbf{t}_j| \, | j = J_1, J_m\}$ contains at least two different vectors.
3. If by $\Xi\left[T_{M;m}\right]$ is understood as any permutation of the order of the vector elements within $T_{M,m}$, then: $\forall \Xi : D^{[P]}_m\left(\Xi\left[T_{M,m}\right]\right) = D^{[P]}_m\left(T_{M,m}\right)$.

These three axioms are sufficient to generalize the two-object squared distances discussed and used in classical Euclidean environments. Now, from the two-object distances, one can always construct a

symmetric distance matrix; for example, when considering a subset of cardinality m one can construct an $(m \times m)$ distance matrix, using the elements of the tag subset $T_{M;m} = \{\langle \mathbf{t}_j | \, | j = J_1, J_m\}$, with a generic two-object element:

$$D^{[P]}_{m;KL} = \left\langle \left| \langle \mathbf{t}_K | - \langle \mathbf{t}_L | \right|^P \right\rangle \Leftarrow \forall \{K, L \, | J_1, J_m\} \wedge \{\langle \mathbf{t}_K | \, ; \langle \mathbf{t}_L | \} \subset T_{M;m}, \tag{31}$$

which can be used to construct the two-object matrix:

$$\begin{aligned} &\mathbf{D}^{[P]}_m = \left\{ D^{[P]}_{m;KL} \, | KL = J_1, J_m \right\} \\ &\rightarrow D^{[P]}_{m;KL} = D^{[P]}_{m;LK} \wedge \forall L = J_1, J_m : D^{[P]}_{m;LL} = 0 \\ &\rightarrow \left(\mathbf{D}^{[P]}_m \right)^T = \mathbf{D}^{[P]}_m, \end{aligned} \tag{32}$$

and finally, the complete sum of the matrix $\mathbf{D}^{[P]}_m$ corresponds to a collective distance involving the tag subset $T_{M;m}$:

$$d^{[P]}_m \left[T_{M;m} \right] = \left\langle \mathbf{D}^{[P]}_m \right\rangle = \sum_{K=J_1}^{J_m} \sum_{L=J_1}^{J_m} D^{[P]}_{m;KL}, \tag{33}$$

or, using only the upper triangle of the distance matrix, half this value:

$$D^{[P]}_m \left[T_{M;m} \right] = \sum_{K=J_1}^{J_m - 1} \sum_{L=K+1}^{J_m} D^{[P]}_{m;KL}. \tag{34}$$

Everywhere, the formulation can be developed in Minkowskian metrics, but enumerating this possibility is just a simple exercise in rewriting the expressions above, taking into account the inclusion of the Minkowski metric vector in the norm description, leading to the distance's matrix in the expression (31).

Collective distances are included here for the sake of completeness; they are not used in the application of Section 6. It may be noted that, at order two, the collective distance of a subset of m vectors, in its upper-triangle form, is m times the sum of the squared distances of the vectors to their own centroid, that is, m times the trace of the origin-shifted Gram matrix of the subset. This connects the present section with Sections 2.2 and 4.2.

5.2. Generalized Gram hypermatrices

Generalized cosine similarity measures must first consider the generalized scalar products and norms of a subset of tag vectors in a tagged-set environment, as in the previous section.

Again, one can choose a tag subset like $T_{M;m} = \{\langle \mathbf{t}_j | \, | j = J_1, J_m\}$, where the cardinality m, if equal to 2, permits developing the two-object (dis) similarity measures of sections 4.4 and 4.6. In the section 5.1,

collective distances in the case that the choice is $m > 2$were discussed.

We can now introduce collective similarity measures. However, first we can define what can be taken as collective Gram matrices, using the already defined collective or generalized scalar products of several vectors. One can describe a collective scalar product of m vectors of a known tag set with a number of elements $M > m$, as the complete sum of the inward product of the involved $m$vectors:

$$\begin{array}{l}(T_{M;m}) \equiv (\mathbf{t}_{J_1}\mathbf{t}_{J_2}\mathbf{t}_{J_3}\ldots\mathbf{t}_{J_m}) = \\ \langle\langle\mathbf{t}_{J_1}| * \langle\mathbf{t}_{J_2}| * \langle\mathbf{t}_{J_3}| * \ldots \langle\mathbf{t}_{J_m}|\rangle \Leftarrow \forall\left\{\langle\mathbf{t}_J|\right\} \in T_M\,.\end{array} \tag{35}$$

The tensor whose elements are the scalar products (35) corresponds to the collection of all the scalar products one can extract from the whole tag set of vectors, but in sets of $m$vectors, one can describe it as:

$$\mathbf{\Gamma}_m = \left\{\gamma_{L_1L_2L_3\ldots L_m}\,|\forall L_J\,|J = J_1, J_m\right\}. \tag{36}$$

The tensor $\mathbf{\Gamma}_m$ permits imagining the possibility of computing similarity measures involving an arbitrary number of $m$objects in a known tag set. However, it can also be considered as a generalized Gram matrix, while the usual Gram matrix becomes $\mathbf{\Gamma}_2$.

The tensor $\mathbf{\Gamma}_3$, for example, corresponds to a set of MGram $(M \times M)$matrices associated with each tag vector element:

$$\begin{array}{l}\mathbf{\Gamma}_{M;3} = \left\{\gamma_{JKL} = \langle\langle\mathbf{t}_J| * \langle\mathbf{t}_K| * \langle\mathbf{t}_L|\rangle\,|J, K, L = 1, M\right\} \\ \equiv \left\{\mathbf{\Gamma}_{M;3,J} = \left\{\gamma_{JKL}\,|K, L = 1, M\right\}|J = 1, M\right\},\end{array} \tag{37}$$

and the set of Gram matrices $\left\{\mathbf{\Gamma}_{M;3,J}\,|J = 1, M\right\}$ represents the J-th object in a larger dimension expansion than the vectors $\left\{\langle\boldsymbol{\gamma}_J|\,|J = 1, M\right\}$of the initial Gram matrix, as discussed in Section 4.2, when the rows or columns of the Gram matrix were taken as alternative tag representations of every object. That is, with this representation, every object has a matrix as an attached tag.

Each of these matrices is the ordinary Gram matrix of the tag set, computed with the tag vector of the J-th object in the role of the metric vector of Section 3.1.5: in the third-order hypermatrix, every object acts as the metric for all the others. In a semispace, the weights are non-negative, and every such matrix is non-negative definite.

Beyond this three-object tag representation, one can also use tensor representations including four, five... object tags, if needed.

5.3. Many-object similarities

With this previous description of the generalized Gram matrices, one can easily describe a similarity measure involving m objects associated with the corresponding tag vectors:

$$C^m = \frac{\left(\mathbf{t}_{J_1}\mathbf{t}_{J_2}\mathbf{t}_{J_3}\dots\mathbf{t}_{J_m}\right)^m}{\prod_{K=J_1}^{J_m}\left(\langle\langle\mathbf{t}_K^m|\rangle\right)} = \frac{\left(\mathbf{t}_{J_1}\mathbf{t}_{J_2}\mathbf{t}_{J_3}\dots\mathbf{t}_{J_m}\right)^m}{\prod_{K=J_1}^{J_m} E_m\left[\langle\mathbf{t}_K|\right]}. \tag{38}$$

For example, the similarity measure for three objects can be written easily. Suppose that the three objects are described by the vectors: $\{\langle\mathbf{a}|, \langle\mathbf{b}|, \langle\mathbf{c}|\}$, then a third-order similarity measure can be defined with the expression:

$$C^3\left(\mathbf{abc}\right) = \frac{\left(\mathbf{abc}\right)^3}{\left(\mathbf{aaa}\right)\left(\mathbf{bbb}\right)\left(\mathbf{ccc}\right)}. \tag{39}$$

In a semispace, the generalized Hölder inequality bounds the scalar product of m vectors by the product of the m-th roots of their norms of order m, so the similarity measure lies in the unit interval for every m and reaches one only when all the vectors are proportional. In a vector space with m odd, the scalar product and the norms can be negative or zero, and the bound holds only if absolute values are taken; this is the formal reason for working in a semispace.

6. Application: analysis of the Voynich manuscript

6.1. The manuscript description

The Voynich manuscript (Yale University, Beinecke Library, MS 408; [a high-quality facsimile has been published by Editorial Siloé (2017) Burgos (Spain)]) is a codex of about 240 vellum pages, radiocarbon-dated to 1404–1438 [20], whose text is written in an unknown script and has resisted every attempt to read it since its modern rediscovery in 1912. The text is accompanied by drawings of plants, astronomical diagrams, bathing figures and what appear to be recipes, which have led to the conventional division of the book into herbal, astronomical, biological (or balneological), cosmological, pharmaceutical and recipe sections. Two statistically distinct varieties of the text, called languages A and B by Currier [21], are present, and several scribal hands have been identified [22].

6.2. Previous studies of the Voynich manuscript

The statistical literature on the text is extensive [23][24][25][26][27][28][29][30][31][32][33][34][35]. Its principal results are that the conditional entropy of its characters is lower than that of any European language, that the characters are strongly tied to fixed positions inside the word, that word frequencies follow Zipf's law, and that models of three families have been proposed to account for these facts: a natural language under some encoding, an artificial language or syllabary, and a meaningless text produced by a mechanical procedure, whether a table-and-grille device [28] or the copying with modification of words already on the page [29]. These results were obtained with different tools, on different transliterations, and against different and usually small sets of reference languages, so comparisons of the manuscript with real languages have mostly been made by inspecting scatter plots of one or two statistics.

6.3. Purpose of this analysis using (dis) similarity procedures

The purpose of the present section is to show that the (dis) similarity framework of Sections 2-5 can treat this problem in the same way quantum similarity treats a set of molecules: each text becomes a tagged object, the set of texts becomes a polyhedron in a vector semispace, and the questions "is the manuscript a language?" and "which objects does it resemble?" become questions about distances, cosines and collective measures within that polyhedron, with known-nature controls included in the object set so that every measure is calibrated. The historical constraint is also kept in mind: if a language is hidden in the manuscript, it must be one of those of its time and place (Latin, Catalan, Aragonese, Occitan, French, Galician-Portuguese, the Italian dialects, Middle English), and medieval Catalan and Castilian texts in original spelling were therefore included as specific tests. Modern German, French, Italian, Portuguese, and Occitan are among the reference languages; period texts in these languages, and in German in particular, were not measured, which is a limitation of the present study. All computations described below were performed by the AI system Claude Fable (Anthropic), versions 5.1 and 5.2, under the author's direction, with Python code that implements the formalism of Sections 3-5 in about one hundred lines and is independent of the Voynich problem; the code and the numerical results are available (see the statement on data availability at the end of the article).

6.4. Objects and tag vectors

6.4.1. The object set

The object set comprises four kinds of texts.

- The manuscript in the Takahashi EVA transliteration (32,840 running-text words, about 166,000 characters), as a whole, split by Currier language (A, B), split by section (herbal, biological, stars/recipes, pharmaceutical), with the compound glyphs (benches and gallows ligatures) merged into single symbols, and in two scrambled versions: with the order of the words destroyed, and with the order of the letters inside every word destroyed. Two other transliterations, Zandbergen–Landini ZL 3b (EVA) and Claston v101 (a different alphabet), were used in Section 6.12 to check that the results do not depend on the transliteration.
- Reference languages: the Universal Declaration of Human Rights in 50 languages [36], extended to 62 with tonal and diacritic-rich Latin orthographies (Hmong, Hani, Ewe, Fon, Twi, Bambara, Lingala, Wolof, Shona, Xhosa, Mandarin in pinyin with and without tone marks) and to 65 with three genre-matched Latin and Catalan medical texts (Macer Floridus, De viribus herbarum; Pliny, Naturalis Historia XXV–XXVII; Lluís Alcanyís, Regiment preservatiu e curatiu de la pestilència, Valencia 1490); plus four medieval Catalan texts in original spelling (Muntaner, c. 1325; Bernat Metge, Lo somni, 1399; Curial e Güelfa, c. 1445; Tirant lo Blanch, 1490) and four medieval Castilian texts in original spelling spanning the three centuries before the manuscript (the Siete Partidas of Alfonso X, c. 1265; Juan Ruiz, Libro de buen amor, c. 1330-43; Juan de Mena, Laberinto de Fortuna, 1444; Fernando de Rojas, La Celestina, 1499) [37]. The 62 Declarations and the three medical texts form the 65 reference texts that define the language cloud; for brevity, they are also called the reference languages below. The eight medieval Catalan and Castilian texts are objects to be tested against that cloud, not part of it.
- Transformations of known languages, serving as controls of known nature: a simple substitution cipher of Catalan (which must leave every descriptor unchanged, and does), a verbose cipher (each letter replaced by a fixed pair of symbols), a mixed cipher (vowels one-to-one, consonants one-to-two), Catalan with the vowels removed, Catalan with the standard late-medieval scribal abbreviations simulated, Latin with the letters shuffled inside words, Vietnamese with its diacritics stripped.
- Generated texts: Markov “monkey” texts of orders 1-3 trained on Latin, a table-and-grille generator in the manner of Rugg [28], a crude self-citation generator in the manner of Timm and Schinner [29], a

table of random digits, and — the most important control — a generator fitted to the manuscript's own word grammar, described in Section 6.7.

All texts were measured on non-overlapping chunks of 1,900 words, and the means over the chunks (the first twelve of a text at the character level, the first eight at the word level) were used as the object's tag. A text shorter than one chunk was used whole; this is the case of 51 of the 62 Declarations (the 62 have between 1,006 and 2,355 words, with a median of 1,607), so that sample-size effects are similar, but not identical, for every object. Section 6.6 reports the effect of measuring every text on chunks of 1,000 words and of using all the chunks of the manuscript. For the medieval and the medical texts, the seven spectral descriptors were computed on the first chunk only.

6.4.2. The tag vectors

Each text is attached to a vector of $N = 25$ non-negative descriptors at the character level, all invariant under any permutation of the alphabet, so that a script that cannot be read can be compared with scripts that can:

- character entropies: the first-order entropy h1, the conditional entropies h2 and h3 (with the space as a symbol), the conditional entropy within words, and the word entropy;
- word shape: mean and standard deviation of the word length, share of words of one or two letters, type–token ratio;
- positional structure: the mean position concentration of the characters (how strongly each is tied to the initial, medial or final slot of the word), the neighbor density of the vocabulary (share of word types with another type at edit distance one), and the mean probability of the most likely successor of a character;
- the information profile: the conditional entropy of the n-th character of a word given the preceding n − 1, for n = 1, …, 6;
- spectral descriptors of the within-word transition matrix: its six largest singular values and the modulus of its second eigenvalue.

A second block of eight descriptors above the character level was added later (Section 6.11): the mutual information between adjacent words, the share of words recurring within the previous thirty positions, the lag-one autocorrelation of word length, and the edit-distance deficit between a word and its nearest neighbor among the previous twenty (each measured as the excess over the same chunk shuffled, so that

any process that draws words independently scores zero by construction), together with the hapax share, the share of the ten most frequent words, the Zipf slope over the first two hundred ranks, and the burstiness of the fifty most frequent words.

For the similarity indices and the pseudo-norms, each descriptor is divided by its mean over the 65 reference texts. This scaling keeps all components non-negative, as the odd-order products of Section 3 require; it makes the components commensurable; and it makes "a language" the unit of every coordinate, so that the tag set is a polyhedron in the semispace of Section 2 in which the centroid of the reference texts is the unity vector. The tagged set so defined is exactly the structure of Section 2.1, and the Gram matrix of Section 4.1 is only non-negative definite, of rank at most N: with $N = 25$ (or 33) and $M = 98$ objects, of which the 65 reference texts are the ones that define the cloud, the dimensionality paradox of Section 4.3 is not avoided; Section 6.6 examines how the conclusions change when the descriptors are reduced to subsets, which is the practical test of whether the tag dimension is sufficient.

For the distances, the descriptors are instead standardized: the mean over the reference texts is subtracted, and the result is divided by their standard deviation. In the terms of Sections 2.2 and 3.1.5, this is an origin shift to the centroid of the reference set followed by a metric vector of positive weights; it leaves the semispace, which the distances do not require.

6.5. The measures as implemented

With the scaled tag vectors, the inward product of Section 3 and its complete sum, optionally weighted by a Minkowskian metric vector, give the generalized scalar product of Section 3.1.3 (equation 12). From it follows the norm of order P, (with absolute values when P is odd, as noted in Section 3.1.4), the multiple similarity index of order P, which for $P = 2$ is the squared-cosine similarity of Section 4.6 (here used without squaring, since all components are non-negative) and for $P = 3$ the three-object similarity of Section 5.3 (both indices are used here without the final power: the scalar product of the vectors is divided by the product of the corresponding roots of their norms, so that the three-object index, denoted C_3 below, is the cube root of the measure of Section 5.3), and the distance of order P of Section 4.4. No roots were taken in the norms and the distances, except where stated in Section 6.10. We computed pairwise indices of orders 2, 3, and 4; the full three-object tensor over the whole object set (the generalized Gram hypermatrix of Section 5.2); and distances of orders 1 to 12, Euclidean and Minkowskian.

The statistic used to answer the question is not an object's position in the cloud but its distance to the nearest reference text, measured from the mean of the nearest-neighbor distances among the reference texts themselves and expressed in units of their standard deviation; we denote it z. The reference texts have z near zero on average by construction, and z cannot fall below about −1.2, the value of an object at zero distance from a reference text; an object of a different kind should have a large z; and a working method must give a large z to known non-languages and a small z to known languages and to transformations of languages that preserve the statistics. A single convention is used in all the tables and figures of this section: the reference set consists of the 65 reference texts, and when one of them is itself evaluated, it is left out of the set, both in the search for its nearest neighbor and in the means and standard deviations used to standardize. At order 2, the mean nearest-neighbor distance among the 65 texts is 10.5 and its standard deviation 9.1, and 62 of the 65 texts have $z \leq 2$ when evaluated in this way. The distribution of nearest-neighbor distances is skewed, so z is a ranking device and not a probability. We obtained confidence intervals by drawing 1,000 random subsets of 80 % of the reference texts (reference uncertainty) and, for objects with three or more chunks, by resampling their chunks (object uncertainty).

6.6. Distance to the language cloud

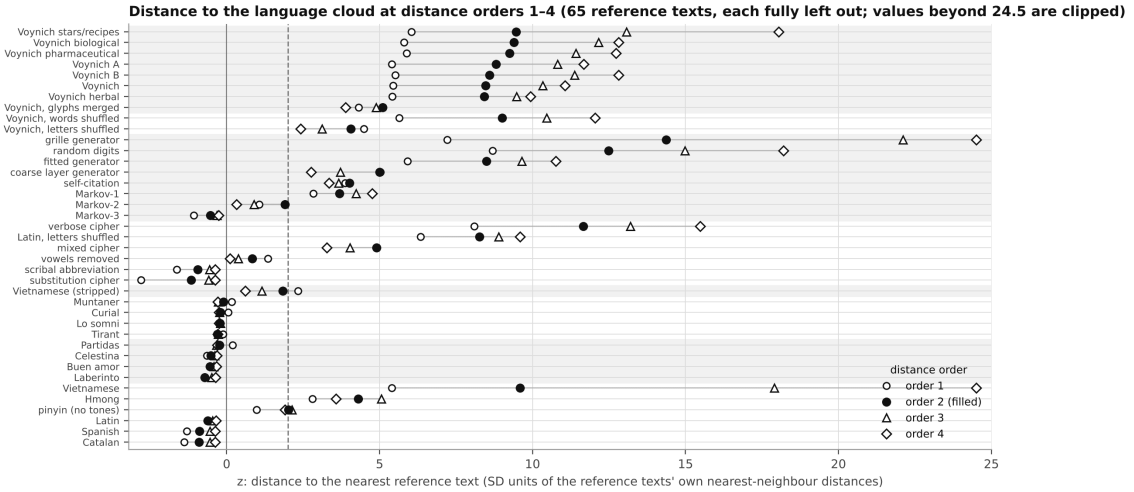


Figure 1. *Distance of each object to its nearest reference text, in standard deviations of the reference texts' own nearest-neighbor distances, for distance orders 1 to 4 (filled symbol: order 2), in the character-level descriptor space with the 65 reference texts (each of them fully left out when it is evaluated, as in Table 1); values beyond 24.5 are clipped. The dashed line at z = 2 is a conventional threshold.*

doi.org/10.32388/DLIILJ

Table 1 gives the order-2 distance z with the 65 reference texts, its 95 % intervals, and the pairwise index with the manuscript, for a selection of objects; Figure 1 shows all the objects outside the reference set, and six reference texts, at four orders.

object	z (order 2)	95 % interval, reference	95 % interval, chunks	C_2 with the manuscript
Voynich (all)	8.5	[4.9,11.8]	[7.9,9.1]	1.000
Voynich A	8.8	[5.0,12.5]	[7.5,10.1]	1.000
Voynich B	8.6	[5.0,13.4]	[8.2,9.1]	0.998
Voynich, words shuffled	9.0	[5.2,12.0]	[8.7,9.3]	0.999
Voynich, letters shuffled in words	4.1 (2.4 at order 4)	[2.2,10.2]	[3.8,4.4]	0.704
Fitted slot-grammar generator	8.5	[4.9,11.3]	[8.4,8.7]	0.998
Grille/table generator	14.4	[8.3,19.0]	—	0.985
Catalan, verbose cipher (1→2)	11.7	[6.8,16.4]	—	0.700
Catalan, substitution cipher	-1.2	[-1.5, -0.7]	—	0.774
Catalan, scribal abbreviation	-0.9	[-1.2, -0.5]	—	0.761
Catalan, vowels removed	0.8	[0.3,1.4]	—	0.741
Markov order-2 monkey (Latin)	1.9	[1.0,2.6]	—	0.949
Random number table	12.5	[7.0,21.8]	—	0.630
Catalan (UDHR)	-0.9	[-1.1, -0.6]	—	0.774
Latin (UDHR)	-0.6	[-0.9, -0.4]	—	0.880
Tirant lo Blanch (1490)	-0.3	[-0.5, 0.0]	[-0.3, -0.3]	0.856
Siete Partidas (c. 1265)	-0.2	[-0.4, 0.0]	[-0.4, 0.0]	0.820
Libro de buen amor (c. 1330-43)	-0.5	[-0.7, 0.3]	[-0.6, -0.5]	0.849
Laberinto de Fortuna (1444)	-0.7	[-0.9, 0.3]	[-0.7, -0.7]	0.840
La Celestina (1499)	-0.5	[-0.7, 0.2]	[-0.5, -0.5]	0.829
Pliny XXV–XXVII	-0.6	[-0.8, 1.4]	[-0.7, -0.5]	0.865

object	z (order 2)	95 % interval, reference	95 % interval, chunks	C_2 with the manuscript
Hmong (RPA, tone letters)	4.3	[2.2,6.4]	—	0.911
Mandarin, pinyin without tones	2.0	[1.5,6.0]	—	0.640
Vietnamese	9.6	[7.3,18.7]	—	0.629
Vietnamese, diacritics stripped	1.8	[0.9,2.4]	—	0.782

Table 1. *Order-2 distance z to the nearest reference text (65 reference texts, each fully left out when it is evaluated), with 95 % intervals from the resampling of the reference set and of the object's chunks, and the order-2 similarity index C_2 with the manuscript (descriptors divided by their means over the 65 reference texts). Among the 65 reference texts, C_2 between two texts has mean 0.92 and 5th percentile 0.79.*

The table reads as a validation followed by a verdict. The substitution cipher lies at zero distance from Catalan, as it must; the medieval texts, the abbreviated and the devowelled Catalan sit inside the cloud at every order; the Markov texts trained on Latin approach the cloud as their order grows (z = 3.7, 1.9, and −0.5 for Markov orders 1, 2, and 3), which is a fair statement of what the descriptors measure; the random digits, the grille generator, the verbose cipher and, less far, the mixed cipher (z = 4.9) lie outside. The manuscript lies outside in every split and section (z between 8.4 and 9.5 at order 2 for the whole text, the two Currier languages, and the four sections), with a 95 % interval under resampling of the reference set that does not fall below $z \approx 5$, and the distance grows with the order (5.4, 8.5, 10.3, and 11.1 at orders 1–4). Scrambling the order of the words changes little (z = 9.0): only the entropy h3 looks across the word boundary, and the chunks of the shuffled text mix the vocabulary of the whole manuscript. Scrambling the letters inside the words halves the distance (z = 4.1, and 2.4 at order 4), whereas the same scrambling takes Latin out of the cloud (z = 8.3). The manuscript's anomaly, therefore, lies largely in the order of the letters inside the word. This last result depends on the reference set: against the 62 Declarations alone, the letter-shuffled manuscript lies at z = 12.3, because the Declarations are narrowly spread on the entropy h3, on which every letter-shuffled text is far out, and the three medical texts widen that spread by a fifth.

doi.org/10.32388/DLIILJ

The four medieval Castilian texts behave exactly as the Catalan ones. All lie inside the cloud at every order of the distance (z between −0.2 and −0.7 at order 2, with intervals not above 0.3), have pairwise indices of 0.98 with modern Catalan and modern Spanish and of 0.82–0.85 with the manuscript, three-object indices with the pair (Catalan, Latin) of 0.86–0.89 (Section 6.9, Table 2; the manuscript 0.64), and flat order trajectories (Section 6.10, Table 3). The Siete Partidas, a law code whose formulae ("e por ende", "como quier que", "segund que") recur at every paragraph, have the lowest conditional entropy of any Romance text measured here, h2 = 2.90 bits, and still lie 0.9 bits above the manuscript and comfortably inside the cloud: this bounds what formulaic repetition alone can do to the character statistics of a real language.

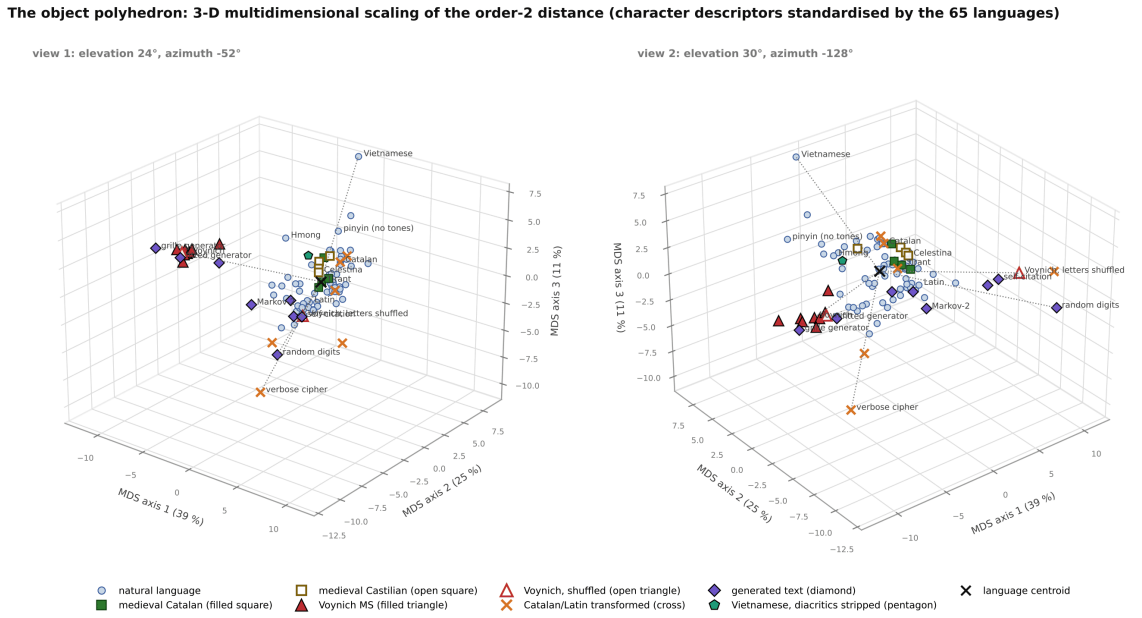


Figure 2. *The object polyhedron in three dimensions: classical multidimensional scaling of the order-2 distance in the standardized character-descriptor space (the three axes carry 39, 25 and 11 % of the variance). Dotted rays join the centroid of the 65 reference texts (×) to the main outliers; the medieval Catalan and Castilian texts (green, gold) lie inside the language cloud.*

Figure 2 shows the polyhedron from Section 2.2, embedded in three dimensions by the classical multidimensional scaling of Section 4.5. The language cloud is compact, and the medieval texts are inside it; the manuscript, its splits, and its fitted generator (Section 6.7) form one tight group far along the first axis, with the grille generator beside them, while the verbose cipher, the tonal orthographies, and the random texts leave the cloud along other axes. Measured from the centroid of the reference texts in the

doi.org/10.32388/DLIILJ

full standardized space of 25 descriptors, that is, in the origin-shifted space of Section 2.2, the manuscript and Vietnamese lie 89° apart, essentially orthogonal, whereas the manuscript, its fitted generator, and the grille generator lie within 20° of one another (11° between the manuscript and its generator), and the verbose cipher lies 76° from the manuscript. The outliers are not merely at different distances from the languages; they are on different sides of them, which is the geometric content of the results of Sections 6.9 and 6.10.

The robustness of this result was examined by repeating the computation under four alternative scalings (division by the mean or by the maximum instead of standardization; log(1+x); absolute standardized scores) and five descriptor subsets (omission of the last two or of all six information-profile descriptors; entropies only; spectral descriptors only; everything except entropies and spectra), and by leaving out each descriptor in turn. The manuscript stays outside the cloud under every scaling (z between 6.9 and 11.1), when any single descriptor is left out (z between 5.9 and 10.0), and in every subset but two (z between 5.7 and 14.5). On the seven spectral descriptors alone, it is indistinguishable from a language (z = 0.0), and so are its fitted generator and the grille generator (z = −0.1 and −0.2), though not the verbose cipher (z = 3.2) or the random digits (z = 5.8). The spectrum of the row-stochastic transition matrix, the descriptor most attractive on paper, says little about the kind of text that produced it; the information lies in the entropies and, above all, in the positional profile. And with all six information-profile descriptors removed, the harshest reasonable subset, the manuscript sits at z = 2.3, barely above the conventional threshold: the result leans heavily on the information profile of the word. The outliers among the reference texts lean on other coordinates: on the information profile alone, the manuscript lies at z = 31, and Hmong and Vietnamese at 2.8 and 2.1; on the spectral descriptors alone, Vietnamese lies at z = 92.

Three checks on the sampling were made. (i) When every text is measured again on chunks of 1,000 words, the length of the shortest Declaration, the manuscript lies at z = 7.7 against the 62 Declarations (8.5 with chunks of 1,900 words), Currier A at 8.4, and Currier B at 8.6. (ii) With all seventeen chunks of the manuscript instead of the first twelve, z = 8.6, and the seventeen chunks taken one by one lie between z = 6.5 and z = 10.5. (iii) Long texts of other genres, which are not part of the reference set, fall inside the cloud [38]: modern Catalan and Spanish newspaper prose and an English novel (z between −0.6 and −0.4), and Latin technical prose by Pliny, Cato, Columella, Isidore, and Apicius (z between −1.0 and −0.3). On h3, these texts lie 0.9 to 3.3 standard deviations above the mean of the reference texts: the Declaration is a single formulaic text in many translations, and on that descriptor it is narrower than language in general.

With the compound glyphs merged into single symbols, the manuscript lies at z = 5.1, nearer the cloud but still outside it; this is the transliteration choice to which the result is most sensitive (Section 6.12).

Two first conclusions follow. The manuscript is not a natural language written in a plain alphabetic script, nor a simple substitution of one, which would leave every descriptor unchanged, nor Catalan abbreviated or devowelled as tested here; since every text of its time and place measured here has h2 in the band 2.9–3.35 bits, and the manuscript has 2.0 (1.8 to 2.1 in its sections), this exclusion plausibly extends to the whole historical list, although period texts were measured only in Catalan, Castilian, and Latin. And the distance alone cannot say what the text is: a table-and-grille generator and a verbose cipher of Catalan also lie outside the cloud, and so does Vietnamese, a natural language in an orthography rich in diacritics (z = 9.6), so that at this level the character descriptors give an unequivocal “not plain language” but not an unequivocal “not a language”. The remaining sections address this second point.

6.7. Where the dissimilarity lives: the word as a slot grammar

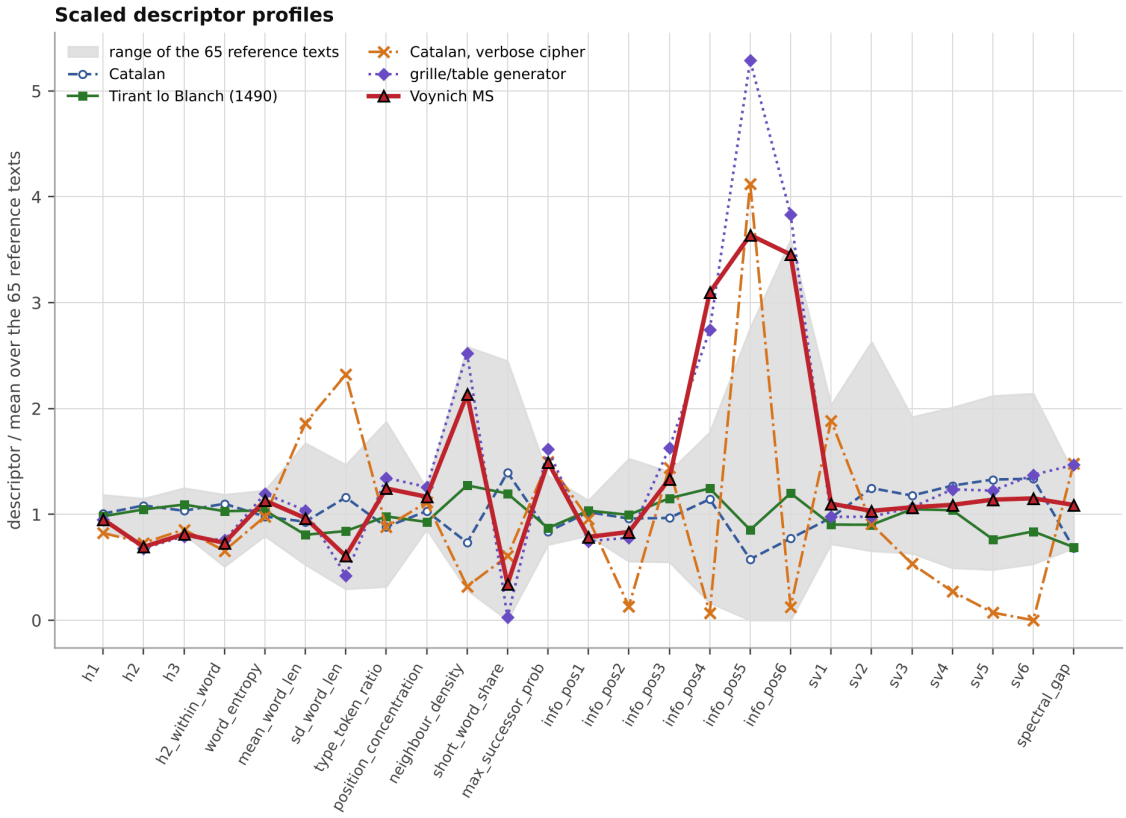


Figure 3. *Scaled descriptor profiles (each descriptor divided by its mean over the 65 reference texts). The manuscript and the grille generator follow the same pattern; the verbose cipher reaches the same h2 by a different route; the shaded band is the range of the 65 reference texts.*

The coordinates that carry the distance are those of the information profile. Between the manuscript and its nearest reference text, Hmong, the information carried by characters 5, 4, and 1 of the word accounts for two thirds of the standardized squared distance (28, 25, and 14 %); between the manuscript and Catalan, the largest terms are the information carried by characters 5 and 4 (19 and 18 %), h2 (10 %), and the information carried by character 6 and the probability of the most likely successor (8 % each). The entropies h2 and h3 are low (0.69 and 0.81 of the mean of the reference texts, h2 lying below all of them), but the information carried by characters 4, 5, and 6 is 3.1, 3.6, and 3.5 times that mean, and for characters 4 and 5 it lies outside the range of the reference texts. In a chunk of a language, the word is almost determined by its first three letters; in the manuscript it is not. The table-and-grille generator reproduces this shape because a three-slot word resets the uncertainty at every slot boundary.

doi.org/10.32388/DLIILJ

Since the grille generator's syllable lists were written by hand from Voynich word shapes, this agreement is partly circular. We obtained the non-circular result with a generator fitted to the manuscript's own word grammar. Every EVA character was assigned to one of the three layers of Stolfi's crust–mantle–core description [27] (core: the gallows k, t, p, f; mantle: c, e, h, which form the benches and the e-groups; crust: the rest), and every word was cut at each change of layer. The words of the manuscript fall into 134 layer patterns, the commonest being crust alone (24 %), crust–core–mantle–crust (16 %), crust–core–crust (14 %), crust–mantle–crust (13 %) and mantle–crust (11 %). For each pattern and slot, we counted the strings observed in that slot. The generator draws a pattern from the empirical pattern distribution and then fills each slot independently from the strings observed in that slot of that pattern. Nothing else is fitted.

This generator reproduces the 25 character descriptors of the manuscript closely: its index with the manuscript is C_2 = 0.998 (C_3 with the manuscript taken twice, 0.997), its z and its intervals in Table 1 nearly coincide with those of the manuscript, and its standardized squared distance to the manuscript is 5.1 ± 0.3 over ten draws, in a space where the mean nearest-neighbor distance between two reference texts is 10.5 and the two Currier languages lie 4.7 apart. The generator is thus about as far from the manuscript as Currier A is from Currier B, and half as far as a reference text is from its nearest neighbor. A coarser version that pools the slot strings per layer instead of per pattern and slot fails (distance 53), so the slot-specific inventories carry the structure. The generator contains only the pattern frequencies and the per-slot inventories, so most of the character-level structure of the manuscript follows from those two things.

The agreement of the similarity index is not by itself informative, since the manuscript with its words merely shuffled also gives C_2 = 0.999 (Table 1). The generator was therefore submitted to five controls, in the standardized space of the 65 reference texts and with texts of the length of the manuscript (32,840 words). (i) Sampling floor: redrawing whole words of the manuscript at random gives a distance of 2.2 to the manuscript, and its two halves, taken as alternate chunks, lie 0.9 apart; the generator gives 5.1 ± 0.3, and when it is fitted on one half of the text it lies 5.7 from that half and 2.9 ± 0.1 from the other, so that the agreement does not come from overfitting. (ii) Calibration: a text produced by the generator, which is a slot grammar by construction, is reproduced at a distance of 0.2 ± 0.1 by a generator fitted to it; the distance of 5.1 is therefore a real residual, and the manuscript is not exactly a slot grammar with independent slots. (iii) Random classes: with the letters of the manuscript assigned at random to three classes, under the condition that the words are cut as often as Stolfi's layers cut them (1.76 cuts per word), the same construction gives distances between 8.1 and 33.7 in twenty trials (median 17.5): Stolfi's layers

are a real feature of the text. (iv) Languages: the same construction was applied to modern Catalan, Spanish, and English prose and to Latin technical prose by Isidore and by Pliny [38]. With vowels and consonants as the classes, or with vowels, sonorants, and the other consonants, it misses its own source by 26 to 48; with twenty random partitions of each text, matched to the manuscript in cutting rate, by 7.0 to 49 (medians 13 to 30). None of these 120 random partitions, of the manuscript or of a language, reproduces its text as well as Stolfi's layers reproduce the manuscript. (v) Dependence between slots: the mutual information between the contents of adjacent slots, in excess of what random pairing gives, is 0.33 bits per word in the manuscript, that is, 0.19 bits per junction between slots; it is 0.52 to 1.14 bits per junction for the five language texts with their phonological classes, and 0.30 to 0.98 for the random classes of the manuscript.

The same construction was also applied to texts of the length of a Declaration. Seventeen pieces of 1,900 words of the manuscript are reproduced at a distance of 1.2 on average (0.7 to 2.6), with 0.14 bits of dependence per junction (0.10 to 0.18). For thirty of the Declarations, with vowels and consonants as the classes, the generator misses its source by 35 to 135, and the dependence is 0.30 to 1.62 bits per junction. The thirty include the tonal orthographies, whose monosyllabic words have a rigid structure of onset, nucleus, and tone that might be expected to behave as a slot grammar: Hmong gives 135 and 0.82, pinyin without tones 36 and 1.03, and Vietnamese 80 and 1.62. At this length, however, the control with random classes loses its power: for five of the thirty Declarations, all among the shortest, one of four random partitions reproduces the text about as well as the resampling of its words does.

A slot grammar with Stolfi's three layers therefore accounts for most of the manuscript's character-level statistics, far better than the same construction accounts for any language or orthography tested, because the junctions between Stolfi's layers are places where the parts of the manuscript's words combine almost freely. The combination is not entirely free: the dependence of 0.19 bits per junction remains to be explained, and because the generator ignores it, 56 % of the word types it produces (20 % of its tokens) do not occur in the manuscript, against 46 % and 13 % for a text that is a slot grammar by construction, and its type–token ratio rises from 0.44 to 0.52.

6.8. The Minkowskian signature

The metric vector of Section 3.1.5 was used to write a hypothesis into the metric. The six descriptors expected to be low in the manuscript (h2, h3, the within-word h2, the information carried by characters 1 and 2, the share of short words) were placed in the negative block, and the remaining nineteen in the

positive block. With this signature, the order-2 pseudo-norm of a reference text lies between 5 and 35 (mean 14.5), that of the eight medieval Catalan and Castilian texts between 10 and 15, that of the manuscript at 55, the grille generator at 76, and the verbose cipher at 38, just above the highest reference text (Vietnamese, 35). The sign of the pseudo-norm does not change for the manuscript, because its positive block (above all the information carried by characters 4–6, more than three times the mean of the reference texts) grows more than its negative block shrinks; the pseudo-norm therefore works as a one-number separator of the manuscript and the grille generator rather than as the sign test that had been hoped for. The Minkowskian cosine (which, as noted in Section 3.1.5, is not bounded by one) between the manuscript and Tirant lo Blanch is 0.987 against 0.856 Euclidean: the signature cancels the part of the difference that lies along the hypothesized coordinates, which is the intended effect, and what remains still places the manuscript at $z = 11.0$ in the signed Minkowskian distance (which is not a distance in the strict sense, since it is negative between some pairs of reference texts). A signature chosen to reverse the sign would have to be tuned to the manuscript, and would then be a description, not a test.

6.9. Many-object measures: the three-object index and the Gram hypermatrix

The three-object index $C_3(x, L_\alpha, L_\beta)$ of Section 5.3 asks a question that pairwise comparison with single languages cannot ask: given a pair of languages, does a third object relate to the pair the way a language does? Table 2 gives it for the pair (Catalan, Latin) and, as a check that the effect is not specific to Romance, for the pair (Hawaiian, Tagalog). By the first remark of Section 3.1.3, the index compares the object with the single composite vector formed by the inward product of the two languages.

third object	C_3 with (Catalan, Latin)	C_3 with (Hawaiian, Tagalog)
Voynich (all)	0.639	0.429
Voynich A	0.631	0.420
Voynich B	0.621	0.412
Fitted slot-grammar generator	0.626	0.407
Tirant lo Blanch (1490)	0.876	0.701
Siete Partidas (c. 1265)	0.858	0.705
Libro de buen amor (c. 1330–43)	0.885	0.694
Laberinto de Fortuna (1444)	0.888	0.689
La Celestina (1499)	0.885	0.707
English	0.878	0.695
Hawaiian	0.693	0.726
Hmong (RPA, tone letters)	0.657	0.462
Mandarin, pinyin without tones	0.651	0.647
Vietnamese	0.676	0.592
Voynich, letters shuffled in words	0.727	0.517
Grille/table generator	0.565	0.361
Catalan, verbose cipher (1→2)	0.521	0.401
Random number table	0.639	0.451

Table 2. *Three-object similarity index C_3 of each object with two fixed language pairs (descriptors divided by their means over the 65 reference texts). Among the 43,680 triples of reference texts, the index has a mean of 0.79, a 5th percentile of 0.59, and a minimum of 0.36.*

The manuscript's index with the two Romance languages (0.64) lies near the 11th percentile of all triples of reference texts, and just below the index of every reference text with that pair: the lowest are those of

Maori (0.65), the two pinyin orthographies (0.65 and 0.67), Hmong (0.66), Vietnamese (0.68), and Hawaiian (0.69), while the Romance and Germanic languages lie between 0.84 and 0.89. A text that was a Romance language in disguise would be expected to keep a high triple index with Romance pairs even as its pairwise indices fell; it does not, and the drop is not specific to Romance, since the index is lower still with the Polynesian pair (0.43), again just below every reference text (the lowest being Hmong, 0.46, and Finnish, 0.49). Nor does the manuscript recover with any other pair: over all 2,080 pairs of reference texts, its highest index is 0.82 (with Zulu and Shona) and its median 0.63, whereas every reference text reaches at least 0.85 with some pair of the others and has a median of at least 0.65, the lowest in both respects being Hmong and Maori. The margins are small, but the manuscript in every split and section, its fitted generator, the grille generator, the verbose cipher, and the random digits lie on one side of them, and all the reference texts on the other; the manuscript with merged glyphs (0.87 and 0.72) falls on the side of the reference texts.

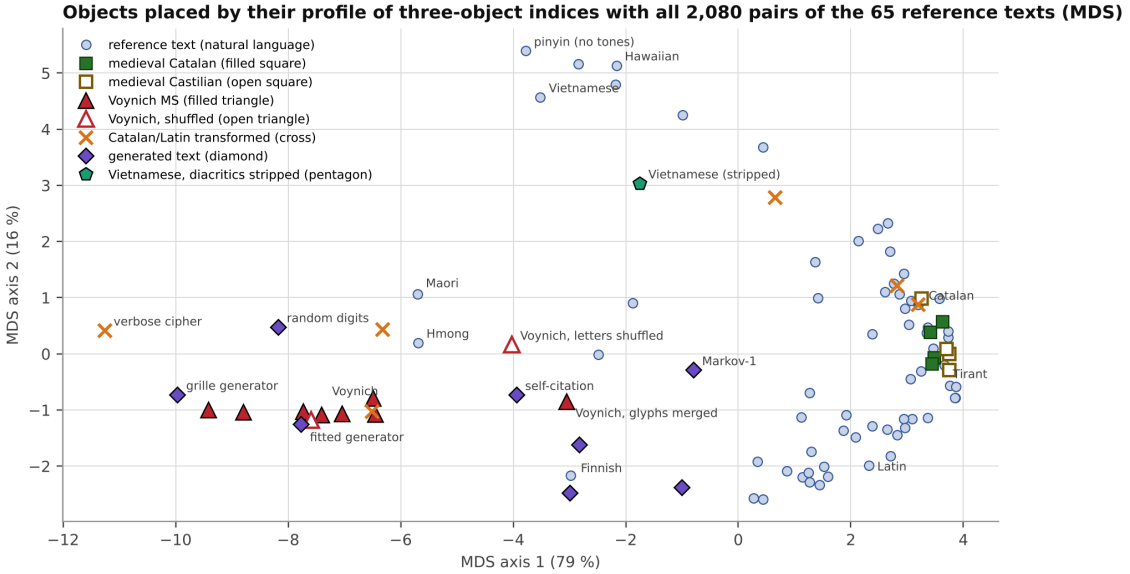


Figure 4. *Objects placed by their profile of three-object indices with all 2,080 pairs of the 65 reference texts (multidimensional scaling of the distance between profiles; the two axes carry 79 and 16 % of the variance).*

The full Gram hypermatrix of order three (Section 5.2) was then used systematically. Each object was represented by the vector of its triple indices with all 2,080 pairs of the 65 reference texts, the third-order analog of the Gram tag set of Section 4.2, and distances were computed between these profiles (Figure 4); a reference text is evaluated without the pairs that contain it. This representation asks whether two

objects deviate from the reference population in the same way, and its verdict differs from that of the pairwise distance. The two tonal orthographies that the pairwise distance had placed outside the cloud, Vietnamese and Hmong, come back inside (z = 0.9 and 0.6), because they deviate as other reference texts do; the manuscript stays outside (z = 3.7; A 4.7, B 5.1), together with its fitted generator (6.0); and the verbose cipher and the hand-written grille are placed very far away (z = 35 and 17). The profile is not, however, a clean separator of “a strange writing system for a language” from “the manuscript”: the manuscript with merged glyphs falls inside (z = 1.7), Vietnamese with its diacritics stripped falls outside (z = 2.5), and one ordinary reference text, Finnish, lies far outside (z = 11), because no other reference text deviates as it does. In the terms of Section 4.2, the profile is a distance under a metric defined by the reference population: objects that deviate as some reference text does are drawn in, and objects that deviate alone are pushed out, whatever their nature.

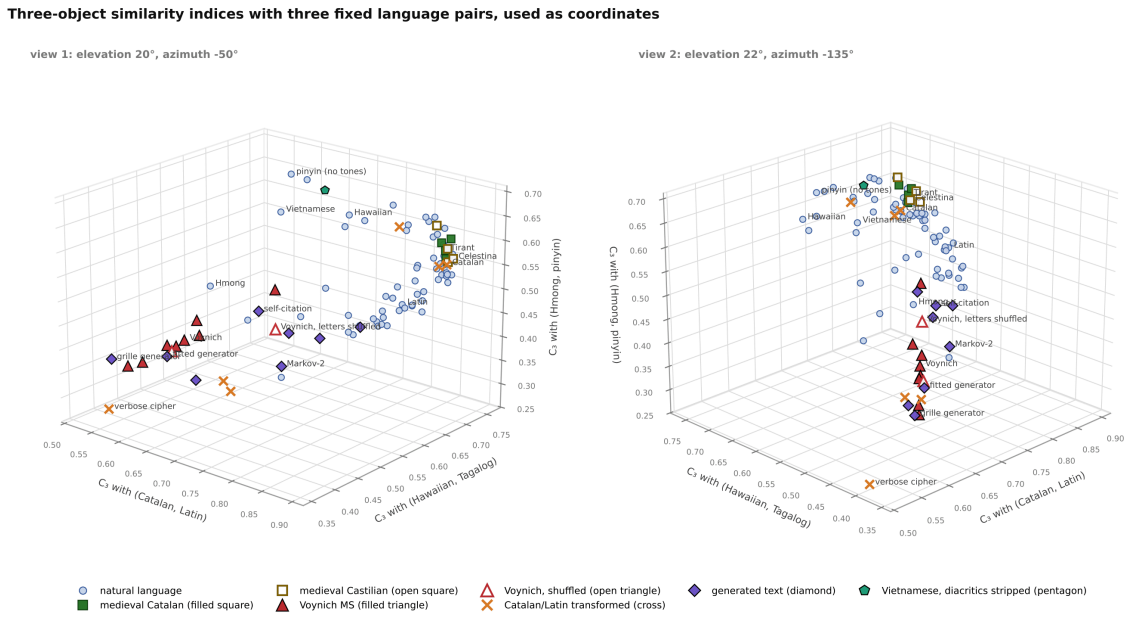


Figure 5. *Three entries of each object’s three-object matrix are used directly as coordinates: index C_3 with the fixed language pairs (Catalan, Latin), (Hawaiian, Tagalog), and (Hmong, Pinyin without tones). No embedding is involved. Most reference texts and the medieval texts occupy the upper corner (the medians of the reference texts on the three axes are 0.85, 0.66, and 0.55); the manuscript lies at (0.64, 0.43, 0.43), between the grille generator (0.57, 0.36, 0.40) and the reference texts, below every reference text on the first two axes and below all but two of them (Finnish and Maori) on the third.*

doi.org/10.32388/DLIILJ

Multiple indices of orders 4 to 8 were also computed, between an object and n − 1 distinct reference texts, averaged over 400 random tuples. All indices decay with n, as a product of n vectors must, and the spread among the reference texts' own tuples grows faster than the gap between the manuscript and the baseline: the manuscript's separation falls from 1.85 standard deviations at n = 2 to 0.75 at n = 8. Higher-order multiple indices therefore dilute the pairwise information rather than sharpen it; the three-object index is useful because it answers a different question, not because it is of higher order.

6.10. Distances of arbitrary order: the order trajectory

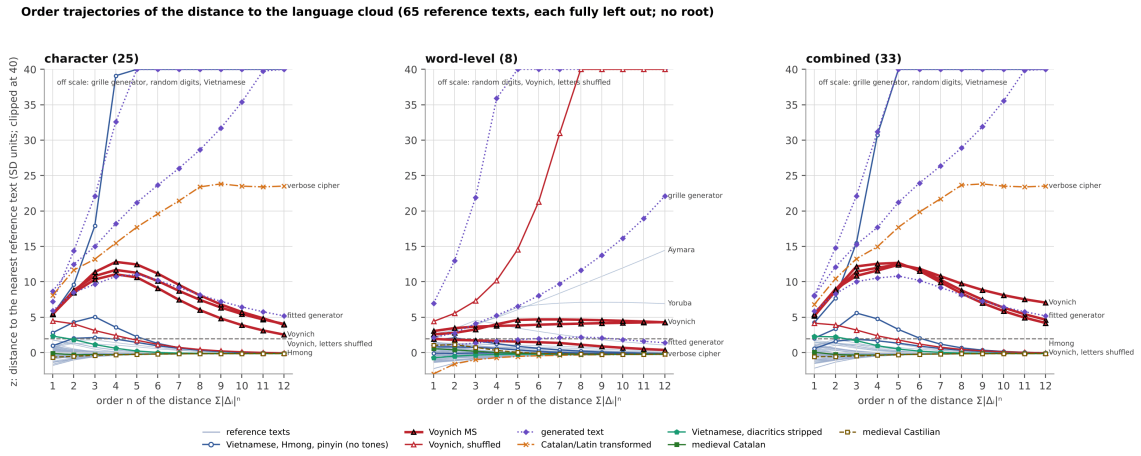


Figure 6. *The distance z to the nearest reference text against the order n of the distance $\Sigma|\Delta_i|^n$ (no root), in the character (25), word-level (8) and combined (33) descriptor spaces, with the 65 reference texts, each fully left out when it is evaluated. Thin blue lines: the reference texts; values beyond 40 are clipped.*

Because $\Sigma|\Delta_i|^n$ is dominated by the largest coordinate as *n* grows (the limit of its n-th root is the Chebyshev norm), how z changes with the order shows how an object's deviation from the language cloud is distributed across the descriptors. Four kinds of trajectory appear in the character space (Table 3). Growing trajectories (z in the tens, the hundreds, or the thousands by order 12) belong to Vietnamese, the hand-written grille, the random digits and, alone among the sections of the manuscript, the stars/recipes section: one coordinate is grossly off. The trajectory of the verbose cipher grows to a plateau. Fading trajectories that fall to zero by order 8-12 belong to Hmong, pinyin, stripped Vietnamese, the mixed cipher, the self-citation generator, and the manuscript with merged glyphs or with shuffled letters; the ordinary languages, including the eight medieval Catalan and Castilian texts, stay flat near zero: many small deviations. The manuscript as a whole, its two Currier languages, and three of its four

doi.org/10.32388/DLIILJ

sections show a peaked trajectory, rising to a maximum at order 4–5 (11.1 for the whole text) and declining afterward (2.5 at order 12), and its fitted slot-grammar generator traces a similar curve (maximum 11.0 at order 5, and 5.2 at order 12): a handful of comparable deviations and no single dominant one. The peaked shape is not exclusive to the manuscript, however: Latin with its letters shuffled (maximum 9.8 at order 5, and 6.8 at order 12) and the first-order Markov text (4.9 and 2.3) show it too.

object	z at order 2	z maximum	order of peak	z at order 12	z(12)/z(2)	shape
Voynich (all)	8.5	11.1	4	2.5	0.30	peaks
Voynich A	8.8	11.7	4	4.1	0.46	peaks
Voynich B	8.6	12.8	4	3.9	0.46	peaks
Voynich, stars/recipes section	9.5	29.2	12	29.2	3.08	grows
Voynich, glyphs merged	5.1	5.1	2	0.0	≈ 0	fades
Voynich, letters shuffled in words	4.1	4.5	1	0.0	≈ 0	fades
Fitted slot-grammar generator	8.5	11.0	5	5.2	0.61	peaks
Grille/table generator	14.4	482.6	12	482.6	33.6	grows
Random number table	12.5	44.9	12	44.9	3.60	grows
Catalan, verbose cipher	11.7	23.8	9	23.5	2.02	plateau
Latin, letters shuffled in words	8.3	9.8	5	6.8	0.82	peaks
Vietnamese	9.6	20,428	12	20,428	2,130	grows
Hmong (RPA)	4.3	5.1	3	−0.1	≈ 0	fades
Mandarin, pinyin without tones	2.0	2.1	3	0.0	≈ 0	fades
Tirant lo Blanch (1490)	−0.3	−0.1	—	−0.1	—	flat
Siete Partidas (c. 1265)	−0.2	0.2	—	−0.1	—	flat
La Celestina (1499)	−0.5	−0.1	—	−0.1	—	flat
Catalan	−0.9	−0.1	—	−0.1	—	flat

Table 3. *Order trajectories in the character descriptor space (65 reference texts, each fully left out when it is evaluated; no root). The values at order 2 are those of Table 1.*

The shape adds something to the value. Hmong and the manuscript were already apart in value at order 2 (4.3 against 8.5), and the trajectories separate them further: that of Hmong fades to zero by order 10, as those of the other reference texts do, while that of the manuscript is still above 2 at order 12, as is that of

doi.org/10.32388/DLIILJ

its fitted generator. The one reference text whose trajectory keeps growing, Vietnamese, does so for a single known cause, its orthography rich in diacritic letters, which the spectral descriptors register (Section 6.6), and it returns inside the cloud when the diacritics are stripped (z = 1.8); the manuscript's anomaly is reduced, but not removed, by merging its compound glyphs or by shuffling its letters. Two numbers summarize a trajectory, the order of the maximum and the ratio z(12)/z(2): 4 and 0.30 for the manuscript, 5 and 0.61 for its generator, 12 and more than 2,000 for Vietnamese, 3 and about 0 for Hmong. The trajectory is a useful description of how a deviation is distributed, to be read together with the three-object index of Section 6.9 and after the plain distance of Section 6.6.

Three cautions limit this reading. First, the shapes are produced by the absence of the root: a standardized deviation larger than one grows with n, and a smaller one vanishes, so that the standard deviation of the reference set is the unit that separates growth from decay. When the n-th root is taken, the peak, the explosion, and the fading all disappear, and every object keeps roughly the z it has at order 1. At n = 1, 2, 4, 8, and 12, the manuscript gives 5.4, 5.4, 4.7, 4.1, and 3.9; its fitted generator 5.9, 5.4, 4.7, 4.3, and 4.3; the grille generator 7.2, 7.6, 6.9, 7.0, and 7.2; the verbose cipher 8.1, 6.6, 5.3, 5.2, and 5.1; Vietnamese 5.4, 5.7, 7.0, 8.3, and 8.8; and Hmong 2.8, 3.4, 3.2, 2.4, and 2.2. A weaker pattern survives: Vietnamese, with one dominant coordinate, rises, while the manuscript, its generator, and the verbose cipher fall.

Second, the unit itself depends on the reference set. Against the 62 Declarations alone, the trajectory of the letter-shuffled manuscript grows without limit (z = 12.3 at order 2 and 821 at order 12), whereas with the 65 reference texts it fades (4.1 and zero), because the three medical texts widen the spread of the reference set on h3. Third, the shapes do not sort the objects by their nature: the manuscript with merged glyphs has its maximum at order 2 and fades, as Hmong does; one section of the manuscript grows; and a language with its letters shuffled shows the peak of the manuscript. The unrooted trajectory is thus an amplifier of the standardized deviations rather than an invariant fingerprint.

6.11. Above the character: word-level descriptors and genre

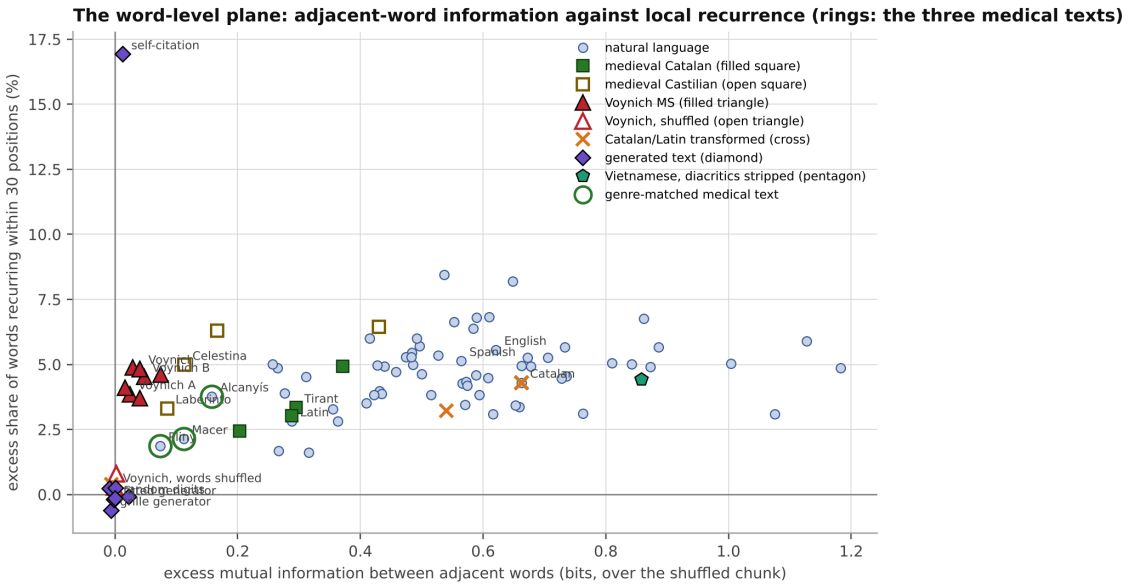


Figure 7. *The two most telling word-level descriptors: the excess mutual information between adjacent words and the excess share of words recurring within the previous thirty positions. Languages occupy the right; generators that draw words independently are at the bottom left; the manuscript has the local recurrence of a language and almost as little adjacent-word information as a generator; the three Latin and Catalan medical texts (rings) lie between it and the other languages.*

Of the eight word-level descriptors, the manuscript splits into two. Its vocabulary statistics match those of a language: hapax share 0.31, share of the ten most frequent words 0.19, Zipf slope 0.79, burstiness 1.11, and an excess of 4.8 % of words recurring within thirty positions, all within the range of the reference texts. Its adjacent-word mutual information, however, is 0.041 bits (chunks 0.020–0.102), below every one of the 62 Declarations (minimum 0.256), below the medieval Catalan narratives (0.20–0.37), and below the four medieval Castilian texts (0.085–0.43), although the verse of the Laberinto de Fortuna comes close to it (0.085, with chunks between 0.024 and 0.153). The manuscript's words recur locally and cluster like topics, yet hardly predict their neighbors. The fitted slot-grammar generator, which draws each word independently, has by construction neither recurrence nor adjacent-word information. In the word-level space, the manuscript lies at z = 3.5, and its generator at z = 1.1 with the 65 reference texts but at 4.5 against the 62 Declarations alone: the Latin medical texts, with their low adjacent-word information and low recurrence, are the nearest neighbors of an independent process. The self-citation generator is the

only process in the population with the right sign on both axes, though the crude version used here overshoots the recurrence (17 % against 5 %) and has almost no adjacent-word information (0.013 bits).

text	words	h2	z char	MI excess (bits)	chunk range	recurrence ≤ 30 (%)	hapax share
Macer Floridus, De viribus herbarum	14,123	3.28	−0.6	0.112	0.072–0.178	2.1	0.39
Pliny, Nat. Hist. XXV–XXVII	18,241	3.34	−0.6	0.073	0.029–0.146	1.9	0.47
Alcanyís, Regiment (1490)	5,199	3.18	0.4	0.158	0.147–0.169	3.8	0.32
Pliny, Nat. Hist. II	17,216	3.34	−0.6	0.037	0.018–0.064	2.5	0.47
Cato, De agri cultura	14,930	3.23	−0.9	0.144	0.053–0.200	9.0	0.34
Tirant lo Blanch (1490)	26,540	3.07	−0.3	0.295	0.193–0.413	3.3	0.22
Siete Partidas (c. 1265)	12,242	2.90	−0.2	0.430	0.222–0.702	6.4	0.15
Libro de buen amor (c. 1330–43)	43,718	3.08	−0.5	0.166	0.105–0.313	6.3	0.28
Laberinto de Fortuna (1444)	15,239	3.08	−0.7	0.085	0.024–0.153	3.3	0.34
La Celestina (1499)	64,468	3.09	−0.5	0.113	0.073–0.177	5.0	0.31
Catalan (UDHR)	1,685	3.18	−0.9	0.663	—	4.3	0.20
Voynich (all)	32,840	2.03	8.5	0.041	0.020–0.102	4.8	0.31

Table 4. *Genre-matched reference texts, two further Latin texts [38], and medieval Catalan and Castilian texts, compared with the manuscript at the character and the word level.*

This descriptor is sensitive to genre, and the genre-matched texts correct an overstatement that a

reference set of legal declarations and chivalric narratives would have produced. Latin technical prose carries far less adjacent-word information than Catalan or English: Pliny's medicinal books give 0.073 bits, with chunks as low as 0.029, and Macer's verse herbal 0.112 (Table 4). Two further Latin texts were measured as a check [38]: Cato's De agri cultura gives 0.144 bits, and the second book of Pliny's Naturalis Historia, which is continuous prose, 0.037 (chunks 0.018–0.064), the value of the manuscript. The descriptor falls as the vocabulary of a chunk grows richer (the hapax share of Pliny's chunks is 0.47, against 0.31 for the manuscript), but the result does not depend on that: when only the thirty most frequent words of each chunk are kept and all the others are pooled into a single symbol, the manuscript gives 0.029 bits, Pliny's second book 0.011, Cato 0.046, and modern Catalan, Spanish, and English prose between 0.19 and 0.31. The descriptor measures the density of function-word syntax, which is low in an inflected language with free word order. What can be stated is therefore that the manuscript's adjacent-word information is at the lowest end of Latin technical prose and not below it, together with a local recurrence (4.8 %) about twice that of Pliny and Macer (1.9–2.5 %) and a hapax share (0.31) below theirs (0.39–0.47): its vocabulary is more repetitive and more locally clustered than theirs, with as little dependence between neighbors. Nor is a position outside the word-level cloud diagnostic by itself: Cato's manual, whose formulaic recipes give a recurrence of 9 %, lies at $z = 6.0$ in the word-level space, farther out than the manuscript. At the character level, the three medical texts and the two further Latin texts are ordinary languages (h2 3.18–3.34, $z \leq 0.5$) and change nothing: Latin technical prose is the kind of plaintext whose word-level statistics come closest to the manuscript's, which is at least consistent with its pictures, but it would still fail at the character level, which is where the exclusion of a plain or simply enciphered language rests.

6.12. Independence from the transliteration

The first objection a reader will raise is that every descriptor was computed on one EVA transliteration and that EVA's way of cutting glyphs into letters could manufacture the slot grammar. The two other complete transliteration files [30] were therefore measured with the same code: ZL 3b, Zandbergen and Landini's curated EVA transliteration, and v101, Claston's, which uses a different alphabet, groups strokes into glyphs differently, and keeps apart variants that EVA merges.

object	z char (order 2)	z, orders 1 / 3 / 4	z word	cosine with Takahashi	h2	MI excess
Voynich, Takahashi (all)	14.3	8.1 / 20.7 / 28.2	3.5	1.0000	2.030	0.041
Voynich, ZL 3b (all)	14.2	8.2 / 19.0 / 25.6	2.6	0.9998	2.034	0.042
Voynich, v101 (all)	9.1	6.7 / 10.1 / 11.5	3.5	0.9638	2.350	0.036
Fitted generator on ZL	14.0	8.1 / 19.3 / 27.7	2.5	0.9988	2.125	−0.001
Hmong (RPA)	9.9	4.9 / 11.1 / 10.6	0.8	0.9000	2.634	0.862
Mandarin, pinyin without tones	2.6	1.5 / 2.8 / 2.9	−0.1	0.5650	2.170	0.706
Catalan	−1.0	−1.6 / −0.6 / −0.4	−0.2	0.7454	3.176	0.663

Table 5. *Three transliterations on the 18 non-spectral character descriptors (65 reference texts, each fully left out when it is evaluated; the absolute z values differ from those of Table 1 because the seven spectral descriptors are omitted, and comparisons within the table are what matter).*

ZL and Takahashi are nearly indistinguishable (cosine 0.9998; the scaled descriptors differ by less than 0.1, the largest differences being in the standard deviation of the word length, 0.70 against 0.61, and in the neighbor density, 2.04 against 2.13), and on the full set of 25 descriptors ZL lies at z = 8.5, as Takahashi does. The v101 alphabet gives shorter words and a higher conditional entropy, as a coarser grouping must, yet stays far outside the language cloud with cosine 0.964 to the Takahashi profile, and the descriptors that carry the result survive the change of alphabet: neighbor density 2.2 times the mean of the reference texts, information in characters 4 and 5 2.7 and 3.2 times, most-likely-successor probability 1.4 times. A slot-grammar generator fitted on the ZL words (173 patterns) reproduces ZL as the first one reproduced Takahashi (cosine 0.999, similar order trajectory). The word-level descriptors, which rest on word identity, do not depend on the alphabet, though they do depend on where a transliteration places the word breaks. The v101 result and the merged-glyph result of Section 6.6 (z = 5.1 against 8.5) point the same way: the more the strokes are grouped into single symbols, the nearer the text comes to the

languages, without reaching them. The slot-grammar result is therefore a property of the manuscript's text, although its size depends on the alphabet used to write it down.

6.13. A codicological check: the sheet as the unit of production

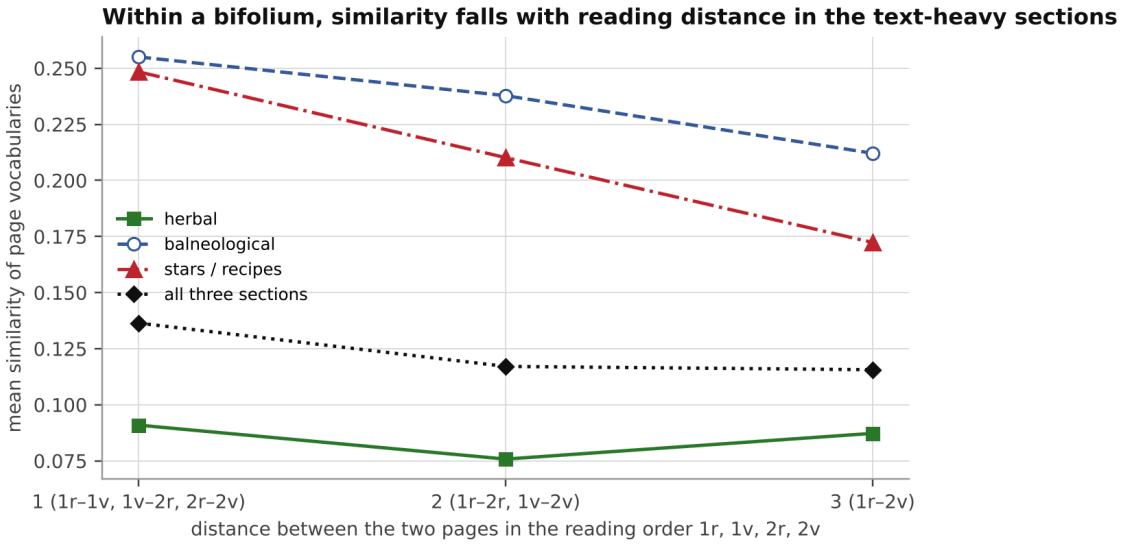


Figure 8. *Similarity (TF-IDF cosine of page vocabularies) between pages of the same bifolium as a function of their distance in the order 1r, 1v, 2r, 2v, by section.*

A recent study [31] found, by latent semantic analysis, that the two halves of a Voynich bifolium are more alike in vocabulary than the pages that face each other in the present binding, the reverse of what two Latin manuscripts show, and concluded that the book is misbound and was meant to be read as loose four-page sheets. We reproduced the pattern here with a plain cosine between page vocabularies, and added controls. In the herbal section the low score of facing pages is largely the interleaving of scribes (scribes as attributed in [22]; facing pages by the same scribe 0.084, by different scribes 0.019, and pages of the same scribe on other sheets 0.060); in every section, pages of the same sheet are more alike than pages by the same scribe on other sheets (one-sided Mann–Whitney test; $p < 0.001$, $p = 0.012$, and $p = 0.018$ in the herbal, balneological, and stars sections); in the two text-heavy sections, facing pages by the same scribe are no more alike than pages of that scribe on other sheets (0.19 against 0.21, and 0.18 against 0.18); and the re-ordered sequences proposed in [31] improve the facing score by about what selection alone yields: in the two quires tested here (13 and 20), the best of all 120 or 720 possible orders of the sheets raises the mean similarity of facing pages from 0.19 to 0.25 and from 0.17 to 0.25, while the present

order ranks at the 53rd and the 41st percentile of all orders. Within a sheet, similarity falls with distance in the two text-heavy sections (Figure 8). The sheet thus appears to be a four-page unit, and the binding carries no continuity; with three distances per curve and no intervals, the order of production within the sheet is suggested rather than shown. Whether the sheet was meant to be read in that order, or merely written in it by a scribe drawing each word from what he had just written, a bag-of-words similarity cannot decide; both readings fit; the second does not require assuming a binder unable to read the book.

6.14. Discussion

Taken together, the (dis)similarity framework supports the following, in descending order of confidence.

- Established with a margin. The Voynich text is not a natural language written in an ordinary alphabet, nor a simple transformation of one. Against 65 reference texts, and unlike the eight medieval Catalan and Castilian texts tested, it lies more than eight standard deviations outside the language cloud at the character level (z = 8.5, with a 95 % interval of 4.9 to 11.8 under resampling of the reference set), and no substitution, abbreviation, or devowelling of Catalan moves a text anywhere near it. The languages of its time and place measured here are excluded in that form. The exclusion holds for every scaling, order, chunk length, and transliteration tried; it leans on the information profile of the word, without which the manuscript falls to z = 2.3.
- Established. The anomaly lives inside the word and is largely a slot grammar. A generator that knows nothing but the manuscript's slot patterns and the inventory of each slot, filled independently, reproduces the 25 character descriptors at half the distance that separates a reference text from its nearest neighbor, across two transliterations, while the same construction fails for random letter classes, for five long texts in four languages, and for the thirty Declarations tested, the tonal orthographies included. The combination of the slots is not entirely free: a dependence of about 0.2 bits per junction between slots remains to be explained.
- Indicated. Unusual writing systems do not behave like this. Tonal and diacritic-rich orthographies can lie as far from the cloud as the manuscript does, but on other coordinates: Vietnamese leaves the cloud along the spectral descriptors and returns to it when the diacritics are removed, and Hmong lies outside without sharing the manuscript's information profile (z = 2.8 on those six descriptors, against 31). With no pair of reference texts does the manuscript reach the three-object index that every reference text reaches with some pair, and an independent-slot generator, which reproduces the manuscript, fails for every orthography tested, the tonal ones included. Neither the three-object

profile nor the order trajectory is, however, a clean separator: an ordinary language, Finnish, lies outside the first, and the manuscript with merged glyphs inside it.

- Dependent on genre. At the word level, the manuscript has the vocabulary statistics of a language, a local recurrence about twice that of Latin technical prose, and an adjacent-word information far below that of narrative or legal prose but at the lowest end of Latin technical prose, where the second book of Pliny's Naturalis Historia matches it. Any one-to-one encoding of words preserves that information, so the word level does not exclude a plaintext as syntactically sparse as Latin technical prose or a list of remedies; such a plaintext would still fail at the character level. A hidden language with denser syntax would have to be hidden by something that scrambles word order, or by a many-to-one or one-to-many treatment of words (homophones, nulls, or word breaks that do not coincide with those of the plaintext).
- Consistent with everything, not proven. The one kind of process in the object set that produces this combination, words that recur locally and cluster like topics yet do not predict their neighbors, over a rigid slot grammar, with the sheet as the unit of production, is copying with modification [29]. The generator of that kind used here is crude, overshoots the recurrence, and does not reproduce the character-level structure ($z = 4.0$ in Figure 1); the quantitative fit remains to be shown with a calibrated model.

From a formal perspective, three results deserve to be singled out. The pairwise distance of Section 4.4 serves as a first screen but cannot by itself separate a strange orthography from the manuscript; the three-object index and the Gram hypermatrix of Section 5 ask a different question, how an object deviates from a population rather than how far it is from it, and they bring the tonal orthographies back among the languages while leaving the manuscript outside, although they also leave outside an ordinary language that deviates alone. The distances of arbitrary order of Section 4.4, taken without the root, amplify the standardized deviations, and their trajectory describes how a deviation is distributed, although its shape depends on the root and on the reference set; and multiple indices of order above three dilute rather than sharpen the information. Finally, the Minkowskian metric vector of Section 3.1.5 can carry a hypothesis into the metric, but the sign of the pseudo-norm is not a free test, since a signature tuned to reverse it would describe the object rather than test it. The controls of known nature, which the quantum similarity practice of comparing a molecule with a family of related ones suggested from the outset, are what turn these measures from descriptions into tests.

The analysis does not prove the manuscript meaningless, nor does it decipher it. Whether a meaning is nevertheless hidden under a scheme that destroys word order, or the correspondence between words, cannot be excluded by statistics of this sort. The manuscript's text has been characterized, with controls and margins, as more compatible with the output of a word-building procedure than with a simply enciphered language. The contribution of the present framework is that this statement is calibrated: every measure is accompanied by objects of known nature.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Data Availability

The texts analyzed are public: the Universal Declaration of Human Rights corpus [36], the editions of the medieval and medical texts listed in [37], the additional texts listed in [38], and the transliteration files of the manuscript [30]. The Python code that implements the formalism of Sections 3–5, the descriptor vectors of all the objects, the scripts that produce the tables and the figures of Section 6 from them, and the numerical results are deposited with the preprint of this article (file Voynich_Similarity_I.zip) and are available from the author upon request.

Author Contributions

R. C.-D. wrote the whole mathematical elements of the first part of this study, as a compilation of many ideas which have been published over the years since his first commitment to (dis)similarity problems. These endeavors have allowed the author to add new ideas and construct a coherent body of possible application nuances of the theoretical background in any scientific field. Under the author's direction, the AI system Claude Fable 5.1 (Anthropic) chose the languages entering the object set of the present paper, applied all elements of the (dis)similarity approach to the Voynich manuscript, defined the parameters for building the tag set of the included object set languages, established the basis for the present

computational elements, built the programming body, analyzed the results, and drafted the text of Section 6. A later version, Claude Fable 5.2, reviewed the whole manuscript, proposed the remarks added to Sections 2-5, recomputed all the numerical results of Section 6 with a single reference set, and ran the additional controls of Sections 6.6, 6.7, 6.9, and 6.11. The author has reviewed all the material and takes full responsibility for its content.

Use of Generative AI

During the preparation of this study, the author used the generative AI system Claude Fable (Anthropic), versions 5.1 and 5.2, to write and run the Python code of Section 6, to analyze its results, to draft the text of that section, and to review the whole manuscript, as detailed under Contributions. The figures of Section 6 are ordinary plots of the computed data, drawn by that code; no generative imagery is involved. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Acknowledgements

The author thanks Ms. Blanca Cercas for her daily efforts; without her commitment, this article would never have been written. The author also thanks the AI system Claude Fable, versions 5.1 and 5.2 (Anthropic), for its dedicated, tireless work.

References

1. ^*Carbó R, Leyda L, Arnau M (1980). "How Similar Is a Molecule to Another? An Electron Density Measure of Similarity Between Two Molecular Structures." Intl J Quant Chem. **17**:1185–1189.*
2. ^*Carbó-Dorca R (1997). "Fuzzy Sets and Boolean Tagged Sets." J Math Chem. **22**:143–147.*
3. ^*Carbó-Dorca R (1998). "Fuzzy Sets and Boolean Tagged Sets, Vector Semispaces and Convex Sets, QSM an d ASA Density Functions, Diagonal Vector Spaces and Quantum Chemistry." Adv Molec Simil. **2**:43–72. JAI P ress. ISBN 0-7623-0258-5.*
4. ^a, ^b*Carbó-Dorca R (2001). "Inward Matrix Products: Extensions and Applications to Quantum Mechanical F oundations of QSAR." J Mol Struct Teochem. **537**:41–54.*
5. ^*Carbó-Dorca R (2002). "Shell Partition and Metric Semispaces: Minkowski Norms, Root Scalar Products, D istances and Cosines of Arbitrary Order." J Math Chem. **32**:201–223.*

6. ^ Carbó-Dorca R, Besalú E (2012). "Shells, Point Cloud Huts, Generalized Scalar Products, Cosines and Similarity Tensor Representations in Vector Semispaces." J Math Chem. **50**:210–219.
7. ^ Carbó-Dorca R (2021). "Generalized Scalar Products in Minkowski Metric Spaces." J Math Chem. **59**:1029–1045.
8. ^ Carbó-Dorca R (2005). "Mathematical Elements of Quantum Electronic Density Functions." Adv Quant Chem. **49**:121–207.
9. ^ Zadeh LA (1965). "Fuzzy Sets." Information and Control. **8**:338–353.
10. ^ Hadamard J (1899). "Théorème sur les séries entières" [Theorem on Power Series]. Acta Math. **22**:55–63.
11. ^ Schur I (1911). "Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen" [Remarks on the Theory of Bounded Bilinear Forms with Infinitely Many Variables]. Journal für die reine und angewandte Mathematik [Journal for Pure and Applied Mathematics]. **140**:1–28.
12. ^ Horn RA, Johnson CR (1991). Topics in Matrix Analysis. Cambridge: Cambridge University Press. Chapter 5 ("The Hadamard product").
13. ^ Hogben L, editor (2007). Handbook of Linear Algebra. Boca Raton (FL): Chapman & Hall/CRC. pp. 66-5–66-7.
14. ^ Todeschini R, Consonni V (2009). Molecular Descriptors for Chemoinformatics. 2nd ed. Weinheim: Wiley-VCH.
15. ^ Carbó-Dorca R (2021). "Molecular Spaces and the Dimension Paradox." Pure and Applied Chemistry. **93**:1189–1196.
16. ^ Carbó-Dorca R, Amat L, Besalú E, Gironés X, Robert D (2000). "Quantum Mechanical Origin of QSAR: Theory and Applications." J Mol Struct. **504**:181–228.
17. ^ Mallows C (2015). "Which Young Tableaux Can Represent an Outer Sum?" Journal of Integer Sequences. **18**:Article 15.9.1, 1–8.
18. ^ Carbó-Dorca R (2016). "A Study on Goldbach Conjecture." Journal of Mathematical Chemistry. **54**:1798–1809.
19. ^ Carbó-Dorca R, Barragán D (2015). "Communications on Quantum Similarity (4): Collective Distances Computed by Means of Similarity Matrices, as Generators of Intrinsic Ordering Among Quantum Multimolecular Polyhedra." WIREs Comput Mol Sci. **5**:380–404.
20. ^ Hodgins G (2009). "Radiocarbon Dating of the Voynich Manuscript Parchment, 2009: Vellum Dated to 1404–1438 at 95 % Confidence." Reported in D. Stolte, "UA experts determine age of book 'nobody can read'", UA News, 10 February 2011.

21. ^Currier PH (1976). "Some Important New Statistical Findings." In: M. E. D'Imperio, editor. New Research on the Voynich Manuscript: Proceedings of a Seminar. Washington DC.

22. a, bDavis LF (2020). "How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript." Manuscript Studies. **5**:164–180. doi:10.1353/mns.2020.0011.

23. ^D'Imperio ME (1978). The Voynich Manuscript: An Elegant Enigma. Fort George G. Meade: National Security Agency / Central Security Service.

24. ^Bennett WR (1976). Scientific and Engineering Problem-Solving with the Computer. Englewood Cliffs: Prentice-Hall.

25. ^Reddy S, Knight K (2011). "What We Know About the Voynich Manuscript." In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011). Portland, pp. 78–86.

26. ^Lindemann L, Bowern C (2020). "Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript." arXiv:2010.14697; Bowern C, Lindemann L (2021). "The Linguistics of the Voynich Manuscript." Annual Review of Linguistics. **7**:285–308.

27. a, bStolfi J (2000). "A Prefix–Midfix–Suffix Decomposition of Voynichese Words" and "The Crust–Mantle–Core Word Grammar." University of Campinas. www.ic.unicamp.br/~stolfi/voynich/.

28. a, b, cRugg G (2004). "An Elegant Hoax? A Possible Solution to the Voynich Manuscript." Cryptologia. **28**:31–46; Schinner A (2007). "The Voynich Manuscript: Evidence of the Hoax Hypothesis." Cryptologia. **31**:95–107.

29. a, b, c, dTimm T, Schinner A (2020). "A Possible Generating Algorithm of the Voynich Manuscript." Cryptologia. **44**:1–19; Timm T (2014). "How the Voynich Manuscript Was Created." arXiv:1407.6639.

30. a, b, cZandbergen R (1998). "The Voynich Manuscript." www.voynich.nu; Landini G, Zandbergen R (1998). "The European Voynich Alphabet (EVA)."

31. a, b, cLayfield C, Davis LF (2026). "Singulion Structure and the Voynich Manuscript." Digital Medievalist. **19**. doi:10.4000/16k0a.

32. ^Montemurro MA, Zanette DH (2013). "Keywords and Co-occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis." PLoS ONE. **8**:e66344.

33. ^Amancio DR, Altmann EG, Rybski D, Oliveira Jr ON, da F Costa L (2013). "Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript." PLoS ONE. **8**:e67310.

34. ^Landini G (2001). "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis." Cryptologia. **25**:275–295.

35. ^Hauer B, Kondrak G (2016). "Decoding Anagrammed Texts Written in an Unknown Language and Script." *Transactions of the Association for Computational Linguistics.* **4**:75–86.

36. a, bUnited Nations. "Universal Declaration of Human Rights." Translations in the UDHR Corpus Distributed with the Natural Language Toolkit (NLTK).

37. a, bTexts used as medieval references: R. Muntaner, Crònica (c. 1325); B. Metge, Lo somni (1399); Curial e Güe lfa (c. 1445); J. Martorell, Tirant lo Blanch (Valencia 1490); L. Alcanyís, Regiment preservatiu e curatiu de la p estilència (Valencia 1490); Macer Floridus, De viribus herbarum (11th c.); Pliny the Elder, Naturalis Historia X XV–XXVII; Alfonso X, Las siete partidas (c. 1265), Partida I tít. 1–3 and Partida II tít. 1–6; Juan Ruiz, Libro de buen amor (c. 1330–43); Juan de Mena, Laberinto de Fortuna (1444); Fernando de Rojas, La Celestina (Burg os 1499; ed. J. Cejador, Madrid 1913). The Castilian texts in the editions of es.wikisource.org, original spelling.

38. a, b, c, d, eTexts used as additional controls: the CESS-CAT and CESS-ESP corpora of Catalan and Spanish ne wspaper prose and J. Austen, Emma (1815), as distributed with the Natural Language Toolkit (NLTK); Pliny t he Elder, Naturalis Historia II–V; Cato, De agri cultura; Columella, De re rustica X–XII; Isidore of Seville, Ety mologiae IV, XI, XII, XVI, and XVII; and Apicius, De re coquinaria I–V, in the texts of The Latin Library (www. thelatinlibrary.com).

Supplementary data: available at https://doi.org/10.32388/DLIILJ

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.