

# The Voynich Manuscript

### Evelyn Keane

### C00239302

### Research report

Table of Contents

[The Voynich Manuscript 1](#_Toc89346286)

[Evelyn Keane 1](#_Toc89346287)

[C00239302 1](#_Toc89346288)

[Research report 1](#_Toc89346289)

[The Voynich MS - General Introduction 4](#_Toc89346290)

[What does the Voynich MS look like? 6](#_Toc89346291)

[What does the Voynich MS say? 7](#_Toc89346292)

[Description of the Voynich MS 8](#_Toc89346293)

[Introduction 8](#_Toc89346294)

[Terminology 8](#_Toc89346295)

[Illustrations in the Manuscript 9](#_Toc89346296)

[Text and script of the Voynich MS 17](#_Toc89346297)

[Text Analysis - the Writing System 17](#_Toc89346298)

[Introduction 18](#_Toc89346299)

[Text and Layout 18](#_Toc89346300)

[*Main text writing* 18](#_Toc89346301)

[The character set 22](#_Toc89346302)

[The handwriting 26](#_Toc89346303)

[*General* 26](#_Toc89346304)

[Extraneous writing 31](#_Toc89346305)

[Text Analysis 39](#_Toc89346306)

[Introduction 39](#_Toc89346307)

[1. Introductory information 39](#_Toc89346308)

[2. Character statistics 41](#_Toc89346309)

[3. + 4. Word statistics 41](#_Toc89346310)

[5. Sentences, paragraphs, sections 42](#_Toc89346311)

[Analysis Section ( 2/5 ) - Character statistics 42](#_Toc89346312)

[2.1 Introduction 42](#_Toc89346313)

[2.2 Observations in the printed literature 43](#_Toc89346314)

[2.3 Character frequencies 45](#_Toc89346315)

[2.4 Entropy 46](#_Toc89346316)

[2.5 Vowel/consonant detection 50](#_Toc89346317)

[2.6 Hidden Markov Modelling 51](#_Toc89346318)

[2.7 Other material 51](#_Toc89346319)

[Analysis Section ( 3/5 ) - Word structure 52](#_Toc89346320)

[3.1 Introduction 53](#_Toc89346321)

[3.2 Jorge Stolfi's ground-breaking work 56](#_Toc89346322)

[3.3 Hidden Markov Modelling 58](#_Toc89346323)

[3.4 Summary 58](#_Toc89346324)

[Analysis Section ( 4/5 ) - Word statistics 58](#_Toc89346325)

[4.1 Introduction 58](#_Toc89346326)

[4.2 Observations in printed literature at the word level 59](#_Toc89346327)

[4.3 Are word spaces significant? 59](#_Toc89346328)

[4.4 Are words really words? 60](#_Toc89346329)

[4.5 Word frequencies 60](#_Toc89346330)

[4.6 Word length distribution 60](#_Toc89346331)

[4.7 Word entropy 61](#_Toc89346332)

[4.8 Zipf's laws 61](#_Toc89346333)

[Analysis Section ( 5/5 ) - Sentences, paragraphs, sections 61](#_Toc89346334)

[5.1 Introduction 62](#_Toc89346335)

[5.2 Observations in printed literature at syntax level 62](#_Toc89346336)

[5.3 Repeating Phrases 63](#_Toc89346337)

[5.4 Cluster analysis of the MS text 63](#_Toc89346338)

[5.5 Long-range correlations 64](#_Toc89346339)

[5.6 Other analyses 65](#_Toc89346340)

The Voynich MS - General Introduction

*cumque in mea Bibliotheca Sphinx quaedam, Scripturae incognitorum characterum inutiliter occupasset locum, ...  
Ex pictura herbarum, quarum plurimus est in Codice numerus, imaginum diversarum, Astrorum, aliarumque rerum, faciem chymicorum arcanorum referentium, conjicio totum esse medicinalem;[[1]](#endnote-1)*

In 1639, Prague citizen Georgius Barschius wrote a letter to the Jesuit Athanasius Kircher in Rome, which included the above lines. He explained that he owned a mysterious book that was written in an unknown script and that was profusely illustrated with pictures of plants, stars, and chemical secrets.

Text

Description automatically generatedA picture containing text, whiteboard

Description automatically generated

Barschius said that he could not read the text, and he hoped that Kircher would be able to translate this book for him. As far as we can tell, Kircher did not succeed in this.

This book later passed through various hands, and it ended up in the Beinecke Rare Book and Manuscript Library of Yale University[[2]](#endnote-2). It is now known as the "Voynich Manuscript (MS)". It is a medieval handwritten book of almost 250 pages, and even today the text cannot be understood. It has become quite famous, and it is recognised as one of the main unsolved problems in the history of cryptography.

Text, letter

Description automatically generated

While nobody has been able to find an explanation for the text, the book just seems to be waiting for someone to achieve this. In fact, there are plenty of people who believe that they have done this. Many new translations of parts of the MS, individual pages, or even individual words are proposed each year. The problem is that none of these is sufficiently convincing to be accepted.

The book is preserved in the Beinecke Library since 1969, where it is officially known as MS 408, but it is better known by its nickname. It was named after Wilfrid Voynich, an antiquarian book dealer who brought it to light in 1912. We may use his own words to describe this event[[3]](#endnote-3), although we have to keep in mind that not all of his statements are necessarily reflecting the truth. This is because he bought the MS under the condition of absolute secrecy about the details of the sale.

*In 1912 [...] I came across a most remarkable collection of preciously illuminated manuscripts. For many decades these volumes had lain buried in the chests in which I found them in an ancient castle in Southern Europe [...]*

*While examining the manuscripts, with a view to the acquisition of at least a part of the collection, my attention was especially drawn by one volume. It was such an ugly duckling compared with the other manuscripts, with their rich decorations in gold and colours that my interest was aroused at once. I found that it was written entirely in cipher. Even a necessarily brief examination of the vellum upon which it was written, the calligraphy, the drawings and the pigments suggested to me as the date of its origin the latter part of the thirteenth century. The drawings indicated it to be an encyclopaedic work on natural philosophy. [...]*

*the fact that this was a thirteenth century manuscript in cipher convinced me that it must be a work of exceptional importance, and to my knowledge the existence of a manuscript of such an early date written entirely in cipher was unknown, so I included it among the manuscripts which I purchased from this collection. [...]*

*two problems presented themselves - the text must be unravelled and the history of the manuscript must be traced. [...]*

*It was not until some time after the manuscript came into my hands that I read the document bearing the date 1665 (or 1666), which was attached to the front cover. [...]*

*This document, which is a letter from Joannes Marcus Marci to Athanasius Kircher making a gift of the manuscript to him, is of great significance…*

The Prague physician and scientist Johannes Marcus Marci had been a faithful correspondent of Athanasius Kircher for 25 years. Sometime before his death he sent the manuscript to Kircher. The MS was accompanied by the letter mentioned by Voynich[[4]](#endnote-4), in which Marci explains how he had inherited the manuscript from a close friend, who had tried to decipher it until the end of his life, and had also asked for Kircher's help. This friend was Georgius Barschius, whose letter was cited above.

Voynich took the manuscript to London in 1912, and later (January 1915) to the United States. He always called it his 'cipher MS', and occasionally he provided photographic copies of pages of it to experts in various disciplines. The manuscript became famous when, in the 1920's, William Romaine Newbold proposed a spectacular partial translation of its text. This 'solution' was disproved by John M. Manly in 1931.

The manuscript has attracted the attention of the code-breaking experts ever since 1917, and in the 1940's and 1960's the eminent cryptanalysts William F. Friedman and Elizebeth Smith Friedman made several valiant attempts at deciphering its text. They were aided by groups of codebreaking experts, but also they did not find any solution.

In 1961 the book was bought by the famous antiquarian H. P. Kraus for the sum of $24,500. He tried to sell it for $160,000 but was unable to find a buyer. Finally, in 1969 he donated it to the Beinecke Rare Book and MS Library of Yale University.

In 2009 one of the many questions surrounding the manuscript could be resolved. The parchment of the manuscript was radio-carbon dated resulting in a date range of 1404-1438 with 95% confidence.

What does the Voynich MS look like?

The following is only a very brief description of the MS.

The Voynich MS is a parchment codex of 22.5 x 16 cm, with its leaves numbered up to 116, of which 14 are now missing. Its cover, also parchment, is blank: it does not indicate any title or author. The manuscript is written in an elegant, but otherwise unknown script.

Text, letter

Description automatically generated

The text, written from left to right, appears to be arranged in short paragraphs. The manuscript is profusely illustrated, and from these illustrations it appears to be a scientific or medical work from the Middle Ages. Illustrations of similar type are mostly grouped together in the manuscript, and one may tentatively identify the following sections, based on these illustrations:

• A herbal section, with drawings of herbs, some of which look realistic, while others appear imaginary;

• An astronomical section, with illustrations of Sun, Moon, stars and zodiac symbols;

• A cosmological section, with mostly circular drawings;

• A so-called biological section, which contains some possibly anatomical drawings with small human (mostly feminine) figures populating systems of tubes transporting liquids;

• A pharmaceutical section, so called because it has drawings of containers, next to which various small parts of herbs (leaves, roots) have been aligned;

• A recipes section, which contains over 300 short paragraphs, each accompanied by a star in the left margin.

What does the Voynich MS say?

The text of the manuscript is still a mystery. It is tempting to assume that the text relates to the illustrations, but this is not certain. There have been many suggestions about the historical importance of the Voynich MS, ranging between opposite extremes, including:

* early discoveries and inventions by the 13th century friar Roger Bacon, written in a very complicated code;
* nonsense written by a medieval quack, to impress his clientele;
* a rare prayer book from the Cathars, not destroyed by the Inquisition, written in a pidgin version of a Germanic/Romance creole;
* meaningless strings of characters cleverly composed by John Dee and/or his associate Edward Kelley, for monetary gain.

This list could be extended further. So far, we are not able to answer the most basic question, namely whether the text is plain language, encrypted language, constructed language or entirely meaningless. While there are numerous sources presenting a great variety of proposed solutions, occasionally accompanied by long discussions, none of these can be considered correct. A correct explanation of the text should be easily recognised as such, and not require a lot of explanation or discussion.

The manuscript continues to attract people from all over the world, primarily because of the mystery of its unreadable text, but there is more to it. For some reason, it allows just about everyone to recognise something in it. It has something that makes so many people believe that they can solve this mystery.

Voynich once stated that the book would become more valuable as soon as it has been deciphered, but this is hardly true. Its mystery and its resistance to translation is what makes it special[[5]](#endnote-5).

Description of the Voynich MS

Introduction

The Voynich MS is a parchment codex in octavo, measuring 225 x 160 mm. It is about 5 cm thick. It has a parchment cover without any indication of its origin (year, title, or author). This cover, which has its own story to tell, is described below.

The text block of the MS consists of 102 folios (originally probably 116, of which 14 are missing) organised in 18 gatherings or quires (originally probably 20 of which 2 are completely missing). The MS is written in an elegant, but otherwise unknown script and almost all of its pages contain illustrations of not easily identifiable herbs, constellations or systems of tubes transporting liquids and populated by small female figures. The illustrations are largely unique to this MS, though a number of similarities with illustrations in other manuscripts have been identified[[6]](#endnote-6).

In the following, the composition of the MS will be looked at in detail, followed by relatively brief descriptions of the illustrations and the script. Both of the latter topics will discussed in more detail on dedicated pages.

Terminology

The Voynich MS is a book or codex which is composed of parchment leaves or folios, combined into gatherings or quires. A 'standard' quire in the MS consists of a stack of four sheets or bifolios which is folded in the middle to form 8 folios. Each quire is sewn onto a set of three thongs, at the combined fold. A standard Voynich MS bifolio is roughly 32 cm wide and 23 cm high, meaning that a folio is roughly 23 by 16 cm.

All folios in the MS have writing and/or illustrations on both sides, and the individual sides of each folio will be referred to here as 'pages'. Thus, a standard quire has 16 'pages'. The notation used to identify a 'page' in the Voynich MS is the character *f* (for folio) followed by the folio number, followed by *r* (for *recto* - the front) or *v* (for *verso* - the reverse). Thus, the first quire starts with 'pages' f1r, f1v, f2r, f2v, f3r, etc, and ends with f7v, f8r, f8v. The 'pages' f1r, f1v, f8r and f8v together form one bifolio.

The quires in the Voynich MS have been numbered 1 to 20 (with 16 and 18 missing). Like many other medieval codices, several quires in the Voynich MS do not consist of the standard 8 folios. What is more unusual, though, is that several bifolios are wider than the standard size. These have additional folds and consequently more than the normal four 'pages'. They are referred to as foldouts. These foldouts have different dimensions, with widths of the corresponding bifolios ranging from three to five 'pages' (instead of two). In addition, there is one (approximately) 45 by 45 cm bifolio which has an additional horizontal fold.

*Quire marks*

Quire marks are written on one folio in each quire. It is not known when or by whom these were written, but they appear older than the folio numbers. They are indicated with an Arabic numeral followed by a 9 for Latin -us, and sometimes an 'm' in between[[7]](#endnote-7). Quire marks 16 and 18 are missing, and from the missing folio numbers between quires 15 and 17 and between 17 and 19, one may tentatively conclude that these quires consisted of one bifolio each (which may or may not have been foldout folios), and that they were lost. This also suggests that, at the time when the folio numbers were added, these quires were not missing.

The quire marks are written on the verso side of the last folio of each quire, with two exceptions. One exception is quire 9. In this case the quire number would have been in the 'correct' place if this single-sheet multiple foldout quire had been bound into the MS in a different manner. The other exception is quire 20 (the final quire), where the quire number is on the recto side of the first folio.

Illustrations in the Manuscript

Almost all pages in the MS are illustrated. Illustrations of a similar type are mostly grouped together in the MS, although there are a few text-only pages among them. On the basis of these groups of illustrations, one may tentatively identify the following sections in the MS:

* A ***herbal*** or botanical section, with drawings of herbs, some of which look realistic, while others appear imaginary;
* An ***astronomical*** section, with illustrations of Sun, Moon, stars and zodiac symbols;
* A ***cosmological*** section, with mostly circular drawings;
* A so-called ***biological*** section, which contains some possibly anatomical drawings with small human (mostly feminine) figures populating systems of tubes transporting liquids;
* A ***pharmaceutical*** section, so called because it has drawings of containers, next to which various small parts of herbs (leaves, roots) have been aligned;
* A ***recipes*** section, which contains over 300 short paragraphs, each accompanied by the drawing of a star in the margin.

These groups of illustrations are now summarised in some more detail.

*Herbal illustrations*

A picture containing text

Description automatically generatedA picture containing text, fabric

Description automatically generated Text

Description automatically generated

Herbal pages typically contain one, in a few cases also two, page-filling pictures of herbs with some paragraphs of text that carefully avoids the drawings.

About half the pages in the MS are of herbal nature, namely [Quire 1](http://www.voynich.nu/q01/index.html) (except [f1r](http://www.voynich.nu/q01/index.html#f1r) which is text-only), all of [Quire 2](http://www.voynich.nu/q02/index.html), [Quire 3](http://www.voynich.nu/q03/index.html), [Quire 4](http://www.voynich.nu/q04/index.html), [Quire 5](http://www.voynich.nu/q05/index.html), [Quire 6](http://www.voynich.nu/q06/index.html), [Quire 7](http://www.voynich.nu/q07/index.html), [f57r](http://www.voynich.nu/q08/index.html#f57r), [f65](http://www.voynich.nu/q08/index.html#f65), [f66v](http://www.voynich.nu/q08/index.html#f66v), [f87](http://www.voynich.nu/q15/index.html#f87), [f90](http://www.voynich.nu/q15/index.html#f90) and all of [Quire 17](http://www.voynich.nu/q17/index.html).

D'Imperio describes these drawings in Section 3.3.1[[8]](#endnote-8).

*Astronomical illustrations*

A picture containing text

Description automatically generated A picture containing text, whiteboard

Description automatically generated

Astronomical pages feature drawings of the Sun or the Moon, or arrangements of stars. It is sometimes hard to draw a clear line between astronomical and cosmological pages (see below). The twelve astronomical pages that show illustrations of the zodiac cycle are usually called 'astrological', and are discussed separately.

The following pages are classified as astronomical: [f67](http://www.voynich.nu/q09/index.html#f67) and [f68](http://www.voynich.nu/q09/index.html#f68), except [f67v2](http://www.voynich.nu/q09/index.html#f67v2) and [f68v3](http://www.voynich.nu/q09/index.html#f68v3) (both cosmological).

D'Imperio describes the astronomical and astrological drawings in Section 3.3.3.[[9]](#endnote-9)

*Cosmological illustrations*

A picture containing text, old

Description automatically generated A picture containing text

Description automatically generated

The use of the term 'cosmological' for these illustrations was first introduced in Newbold (1928)[[10]](#endnote-10). Cosmological pages feature geometric designs which cannot always be easily classified. Most of them are circular, with the notable exception of a large and complicated design of nine connecting circles (also called 'network of rosettes') on a multiple fold-out sheet.

The following pages are classified as cosmological: [f67v2](http://www.voynich.nu/q09/index.html#f67v2), [f68v3](http://www.voynich.nu/q09/index.html#f68v3), [f69](http://www.voynich.nu/q10/index.html#f69), [f70r1](http://www.voynich.nu/q10/index.html#f70r1), [f70r2](http://www.voynich.nu/q10/index.html#f70r2), [f85r2](http://www.voynich.nu/q14/index.html#f85r2), [network of rosettes](http://www.voynich.nu/q14/index.html#fRos), [f86v4](http://www.voynich.nu/q14/index.html#f86v4) and [f86v3](http://www.voynich.nu/q14/index.html#f86v3)

D'Imperio has an extensive description of these drawings in section 3.3.4, calling them cosmological or meteorological.

*Astrological illustrations (zodiac signs)*

Map

Description automatically generated A close-up of a book

Description automatically generated with medium confidence

The so-called astrological pages contain concentric circles with 30 small (mostly feminine) figures holding stars, and an emblem of a zodiac sign in the centre. The human figures are very similar to those drawn in the biological section (see below).

The following pages are classified as astrological: [f70v2](http://www.voynich.nu/q10/index.html#f70v2), [f70v1](http://www.voynich.nu/q10/index.html#f70v1), [f71](http://www.voynich.nu/q11/index.html#f71), [f72](http://www.voynich.nu/q11/index.html#f72) and [f73](http://www.voynich.nu/q12/index.html#f73).

*Biological illustrations*

Text

Description automatically generated Text

Description automatically generated Text, letter

Description automatically generated

Perhaps the most unusual section of the Voynich MS is the biological section which contains drawings of human figures (mostly unclothed and mostly female) in arrangements of pipes or vessels, and pools that seem like baths or clouds. The section is also called balneological by some authors. Many illustrations leave the impression of representing a natural or (al)chemical process, including representations of human organs.

The biological section comprises the entire [Quire 13](http://www.voynich.nu/q13/index.html).

D'Imperio describes these illustrations in Section 3.3.5..

*Pharmaceutical illustrations*

Text

Description automatically generated A picture containing text, indoor

Description automatically generated

This section combines collections of apparent containers or jars, and parts of herbs, such as individual leaves and roots.

The following pages are classified as pharmaceutical: [f88](http://www.voynich.nu/q15/index.html#f88), [f89](http://www.voynich.nu/q15/index.html#f89) and [all of Quire 19](http://www.voynich.nu/q19/index.html).

D'Imperio briefly describes these illustrations in Section 3.3.2.

*Text-only with marginal stars*

Text, letter

Description automatically generated Text, letter

Description automatically generated

Some pages contain only text, with stars drawn in the margin. About half the stars have a red dot in the centre, or are coloured red, and many of them have a tail. This section of the manuscript is at the end, and is usually referred to as the 'recipes' section.

The following pages are text-only with marginal stars: [f58](http://www.voynich.nu/q08/index.html#f58) and [all of Quire 20](http://www.voynich.nu/q20/index.html) except [f116v](http://www.voynich.nu/q20/index.html#f116v). D'Imperio mentions these pages in Section 3.3.7[[11]](#endnote-11).

*Text-only*

Some pages contain no illustration, but only text.

The following pages are text-only: [f1r](http://www.voynich.nu/q01/index.html#f1r), [f76r](http://www.voynich.nu/q13/index.html#f76r), [f85r1](http://www.voynich.nu/q14/index.html#f85r1), [f86v6](http://www.voynich.nu/q14/index.html#f86v6) and [f86v5](http://www.voynich.nu/q14/index.html#f86v5).

*Missing folios*

The following 14 folios are missing from the MS:

* [f12](http://www.voynich.nu/q02/index.html#f12), apparently cut out, stub still visible;
* [folios 59-64](http://www.voynich.nu/q08/index.html#f59), three bifolios that should have been in the centre of quire nr. 8;
* [fol. 74](http://www.voynich.nu/q12/index.html#f74), cut out, stub still visible. Cut marks are clearly visible on f75r;
* Quire 16, which appears to have consisted of one bifolio composed of f91 and f92;
* Quire 18, which appears to have consisted of one bifolio composed of f97 and f98;
* [folios 109 and 110](http://www.voynich.nu/q20/index.html#f109), that should have been in the centre of quire nr. 20.

Text and script of the Voynich MS

A picture containing text, whiteboard, receipt

Description automatically generated Text, letter

Description automatically generated

Almost the entire text of the Voynich MS is written in a script that is not found in any other surviving document. The very few examples of writing in the Latin script are either in the margins, or look as if they could have been added by a later owner. Most of the text has been written in a line-by-line manner, undoubtedly from top to bottom and from left to right. There is no sign at all of the application of the usual techniques for creating straight lines in medieval manuscripts, namely ruling or pricking. On many pages this results in wavy lines, although the left margins of the text is usually quite straight. Especially on the herbal pages it appears as if the illustration (or at least its outline) was on the page before the written text, and the text carefully avoids these illustrations.

The text appears to consist of words separated by spaces. Some of these words are very frequent and may be found throughout the entire MS. Others occur only once in the entire MS. In general, the word frequency distribution shows a continuum which is typical for normal language.

In some places single words are written near elements of the illustrations. These are usually called labels, and this convention will be followed throughout this site. Other more complicated pages containing circular diagrams have parts of the text written along radii or circumferences of these circles. In various places in the MS short words or even single characters in the Voynich script form what may be called 'sequences'.

There are a few fragments of text in the MS which are not written in the Voynich script. Most of this so-called 'extraneous writing' is barely intelligible.

Text Analysis - the Writing System

Introduction

The main mystery of the Voynich MS is clearly its unknown writing. This topic is addressed from three different aspects, on three (sets of) pages:

* A look at the writing system, describing its main properties, and similarities and differences with other known writing systems;
* Transliteration of the text;
* Statistical analysis of the text of the MS (further subdivided into five areas).

This page addresses the first part, the analysis of the writing system. Following are the main topics of this page:

• Text and Layout

• The character set

• The handwriting

• Extraneous writing

Text and Layout

*Main text writing*

Text, letter

Description automatically generated

Almost the entire Voynich MS is written in a script that is not found in any other surviving (old) document. The text of the MS has been written mostly in a line-by-line manner, obviously from top to bottom and from left to right. The majority of this text is written in short paragraphs, which are often separated from each other by a larger line spacing. Herbal pages typically have two or three such paragraphs, which tend to occupy mostly the upper half of the page, clearly avoiding the herb drawing. Pages in the so-called biological section tend to have much more text, filling the entire page, sometimes with just three paragraphs, but occasionally also more. The text tends to have a straight left margin, and is only roughly right-justified, except for the last line of each paragraph which tends to be shorter.

The text consists of groups of characters separated by spaces, and these groups seem to form words. The same words tend to appear throughout the MS, with a frequency distribution that is quite normal for a meaningful text[[12]](#endnote-12).

*Labels*

In some places, single 'words' are written near elements of drawings. These have come to be called 'labels'. There is a clear suggestion that these words provide the name of the object in question. Many of the label words also occur in the running text, though when they do, this is rarely in the immediate vicinity of where the label occurs.

Diagram

Description automatically generated with low confidence

The following types of labels may be found in the Voynich MS:

* A few whole plant labels in the herbal sections, on [f2r](http://www.voynich.nu/q01/index.html#f2r), [f41v](http://www.voynich.nu/q06/index.html#f41v) (uncertain) and [f65r](http://www.voynich.nu/q08/index.html#f65r);
* A label near the dead body on [f66r](http://www.voynich.nu/q08/index.html#f66r);
* Star labels on the astronomical and cosmological pages (see sample picture above);
* Star labels in the zodiac section, on [f70v2](http://www.voynich.nu/q10/index.html#f70v2), [f70v1](http://www.voynich.nu/q10/index.html#f70v1), [f71](http://www.voynich.nu/q11/index.html#f71) (all), [f72](http://www.voynich.nu/q11/index.html#f72) (all), [f73](http://www.voynich.nu/q12/index.html#f73) (all);
* Labels near items in the illustrations of the biological section;
* Labels near plant parts in the illustrations of the so-called pharmaceutical section. They are also written inside, or near the containers that have been drawn in these pages.

There are also places in the MS, for example in the cosmological section, where single words appear as elements in the overall design, but not necessarily near an identifiable object.

*Titles*

The term 'titles' was introduced by John Grove. This term is related to the layout of the last lines of some paragraphs. Normally, these last lines are left-justified and do not reach the right margin. Three alternative formats are used occasionally:

* The last line is short and centred  
  A picture containing text, alcohol, beverage

  Description automatically generated
* The last line is short and right-justified  
  A picture containing text, meter, device, alcohol

  Description automatically generated
* The last line is left-justified, but has additional words that are right-justified  
  A close-up of a document

  Description automatically generated with low confidence

The last example is strictly speaking what John Grove called titles, but all three cases are of interest. Counted are 17 pages that include centred end lines, 11 pages with right-justified end lines and 5 pages that use 'titles'.

*Text in circles*

Almost all astronomical, astrological and cosmological pages have circular drawings, some with text written in normal paragraphs, but all with text that has been integrated in the drawings. Frequently, text is written along the circumference of these circles, and occasionally also along radial lies. The following figure shows an example of text along the circumferences and along radii of such a circular drawing.

A picture containing text

Description automatically generated

*Character or word sequences*

In some circular designs and occasionally also in in the margins of some pages, sequences of single characters or short words may be found. These are usually referred to as key-like sequences. They have been included in the table at the bottom of this page, with the keyword 'SEQ'.

The circular diagram on [f57v](http://www.voynich.nu/q08/index.html#f57v) deserves special mention here, as there appear to be several sequences of words and characters integrated in this figure. Since there is no reason to call these 'additional' or 'extraneous', these alone are not included in the table.

*Non-sequential writing*

In a few places it appears as if the first characters of lines were written in a vertical column first, possibly in order to create a straight left margin. The remainder of the text was then added later. The following example is one paragraph on [f88v](http://www.voynich.nu/q15/index.html#f88v). Especially in the last two lines of this fragment, there is a strong suggestion that the initial characters were written first, and the remainder of the line was written later.

Text

Description automatically generated

In addition, there are several places in the MS where writing seems to have been added 'afterwards' or where the text was perhaps not written in a strictly top-down line-by-line manner. One clear example is [f105r](http://www.voynich.nu/q20/index.html#f105r), where one may observe a text break and a use of different inks before and after it, and also a partial line added above the last paragraph in the darker ink. There are more examples of this type of change in the MS.

Text, letter

Description automatically generated

The character set

While the earlier owners of the Voynich MS (Barschius, Kircher) may possibly have believed that the writing represented a language unknown to them, nowadays we know that there is no other old MS that uses the same writing. Ever since it was brought to light by Voynich, people have compared the script with examples of writing. A summary of this is provided in section 4 of D'Imperio (1978)[[13]](#endnote-13), and this may serve as the basis for the following.

* A number of characters are very similar to Latin characters, and look like 'a', 'c', 'i' (undotted), 'm' 'n' and 'o'. (a e i iin in o);
* Others look like numerals, such as 2, 4, 8 and 9. Of '4' both an early Arabic and a modern variety exist. ( r l q d y);
* Some characters look like abbreviations found in Latin medieval manuscripts. They are briefly discussed below;
* The remaining set has been compared with alchemical symbols or early renaissance ciphers. These are also briefly addressed below.

*Latin abbreviations*

In medieval manuscripts numerous more or less standard abbreviations and ligatures were in use. A collection of such abbreviations may be found in the dictionary of Cappelli (1912)[[14]](#endnote-14), which lists, in alphabetical order, a large number of abbreviations found in Latin and Italian manuscripts. While browsing through this dictionary, one will immediately recognise the similarity between some of its entries and the writing in the Voynich MS. D'Imperio shows a few cases in Fig.17. Some of these are also included below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **c** | cum, con |  | **ch** | ra, ci, cri |  | **Co** | co, quo |
| **Ca** | ca | **Cy** | cus | **s** | cun, con, ... |
| **m** | -nd- , -nt- | **g** | eius | **y** | con, cum, -us, ... |
| **n** | ter, in-, ... | **in** | -um | **Sh** | termi |

*The so-called gallows characters*

A particular set of characters in the Voynich MS is usually called 'gallows' characters. They ascend above the majority of other characters. These are not typically found in medieval manuscripts, also not as abbreviation signs. There are four of these:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

They also occur in combination with the character ch, in which case they have been called 'pedestalled gallows':

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

While these characters are not typical abbreviations, one illustration in Cappelli (1912) provides a striking comparison: A picture containing text, device

Description automatically generated

*Alchemical symbols*

Figure 42 in D'Imperio, taken from Gessmann (1922)[[15]](#endnote-15) shows some similarities between Voynich MS characters and a few alchemical symbols. There are only a few examples, and it is not certain whether this is coincidental or not, since these alchemical symbols were not known to have been in use in the early 15th century.

*14th and 15th Century cipher*

Potentially of great interest is the comparison between the characters in the Voynich MS and the symbols found in early renaissance cipher systems. D'Imperio (Fig.39) shows examples from a cipher of Parma (1379), A Venetian cipher (1411) and the Code of Urbino (1440). While these do not show exactly the same characters as the Voynich MS, there are some striking similarities, and the author of the MS may well have been inspired by these, or similar, examples.

Of particular interest is the codex of Tranchedino (MS Vindobonensis 2398 in Vienna), which has been issued in a facsimile edition[[16]](#endnote-16). This MS lists numerous, rather similar sets of ciphers to be used with different correspondents, and they are dated from 1450 to 1496. None of these ciphers uses the gallows characters mentioned above, but the common sequence qo (see also the various illustrations of Voynich MS text above) appears several times, typically representing a single character. Other typical features of the ciphers in Tranchedino are that they include nulls (characters that are introduced but have no meaning), and that double characters are usually represented by a single code character.

*Initials*

The initial characters of many paragraphs are larger than usual, and are sometimes embellished with additional curls or dots. Not all characters in the MS can appear in a paragraph-initial position, and indeed these characters form only a very small subset.

|  |  |  |
| --- | --- | --- |
| Text, letter  Description automatically generated | A picture containing text, alcohol  Description automatically generated | Text  Description automatically generated |

In addition to this, there are two initials highlighted in red on the first page of the MS ([f1r](http://www.voynich.nu/q01/index.html#f1r)), which are not standard characters in the Voynich MS alphabet. These have been the subject of some speculation, and René Zandbergen has pointed out how one of them appears almost identical to the old Aries symbol in Greek astronomical manuscripts[[17]](#endnote-17). More recently, they have been observed in Spanish manuscripts - see for example this blog page by J.K.Petersen[[18]](#endnote-18). These symbols have been used for centuries as emphasis or paragraph markers, and the strong suggestion is that the Voynich MS scribe was familiar with these.

*Rare characters*

The text of the MS includes numerous rare characters. Many of these appear only once in the MS, and could be considered aberrant forms of standard characters. Others appear several times, and are concentrated in some pages of the MS. One of these has been called the 'picnic table' (x).

The handwriting

*Apparent lack of corrections*

One unusual feature of the writing of the Voynich MS is that it appears to have no corrections. The first recorded mention of this is in a letter preserved in the Beinecke Library, from Anne Nill to Theodore Petersen, dated 19 Feb. 1953. She writes:

*I remember I talked too much, but did I really say "the ms. does not include a single erasure or correction"; whatever I said, this is my present opinion: in all my experience of manuscripts I have never come across one in which corrections and erasures are so unobtrusive as they are in this ms. if it contains any. I have looked through it again - of course not every word, and have nothing to add to the one or two probable corrections I recorded c. 1936 when I worked with photostats.   
   
In this connection I must add that we still have a few Voynich estate mss. (requiring additional research which I never found time for, before I attempt to sell them) which include one beautifully written text on fine vellum and some well and some poorly written manuscripts; and in none of them is there any difficulty in detecting corrections, erasures, deletions or transpositions.*

This apparent lack of corrections has occasionally been taken as evidence that the scribe could not understand what he wrote, or even that the text is meaningless.

Ever since high-resolution images of the MS have been publicly available, closer scrutiny by many people has revealed a few cases where it appears that the text has been amended. Such corrections are few and minor.

|  |  |  |  |
| --- | --- | --- | --- |
| **f16r** | **f20v** | **f24v** | **f39r** |
| A picture containing fabric  Description automatically generated | A picture containing text, coin, fabric  Description automatically generated | A picture containing fabric  Description automatically generated | A picture containing text, fabric  Description automatically generated |
| **f39r** | **f42r** | **f50v** | **f79r** |
| A picture containing text, fabric  Description automatically generated | A picture containing text  Description automatically generated | A picture containing text, fabric  Description automatically generated | A picture containing text, fabric  Description automatically generated |
| **f80r** | **f83r** | **f102v2** | **f112r** |
| A picture containing text  Description automatically generated | A picture containing text, ground, fabric  Description automatically generated | A picture containing ground, fabric, stone  Description automatically generated | A picture containing fabric  Description automatically generated |

Extraneous writing

This general term is used to indicate several different types of additional writing that may be found in the MS. Only very few barely legible phrases in the normal (non-Voynich) alphabet may be observed in the Voynich MS. These are listed below, including the numbering of folios and quires.

*Month names in the zodiac section*

Month names have been written in a later hand in the central drawings of each of the zodiac pages: [fol. 70v2,1](http://www.voynich.nu/q10/index.html#f70), [fol. 71](http://www.voynich.nu/q11/index.html#f71), [fol. 72](http://www.voynich.nu/q11/index.html#f72) and [fol. 73](http://www.voynich.nu/q12/index.html#f73).

The language, which is certainly a Romance language or dialect, has been much debated, including suggestions of Spanish, Occitan, and French. The most convincing argument is presented at this web page[[19]](#endnote-19) and thus the language would appear to be Northern French. This is further confirmed by the appearance of very similar month names on an astrolabe that originates from Northern France[[20]](#endnote-20). Following is a table of the readings of the month names in the Voynich MS zodiac pages:

|  |  |  |
| --- | --- | --- |
| **Folio** | **Sign** | **Month name** |
| f70v2 | Pisces | mars |
| f70v1 | Aries | aberil |
| f71r | Aries | aberil |
| f71v | Taurus | may |
| f72r1 | Taurus | may |
| f72r2 | Gemini | jong |
| f72r3 | Cancer | iollet |
| f72v3 | Leo | augst |
| f72v2 | Virgo | septe(m)b(r) |
| f72v1 | Libra | octe(m)bre |
| f73r | Scorpius | nove(m)bre |
| f73v | Sagittarius | decebre |

*Apparent colour annotations*

On some of the earlier herbal pages in the beginning of the MS we find individual letters, and in one case the word 'rot', in or near leaves, flowers and stem of plants. These appear like colour annotations. A detailed list of these annotations has been added to the table at the bottom of this section with the key word 'COL'.

*Other 'plain text' writing*

On [f1r](http://www.voynich.nu/q01/index.html#f1) we find the faded or erased *ex libris* of Jacobus de Tepenec. In addition, in the right margin we find faded or erased character tables, which are presumably a decryption attempt by a later owner.

*Mixed writings in or near margins*

On [f17r](http://www.voynich.nu/q03/index.html#f17) we find a largely unreadable comment near the top margin. Nick Pelling was the first to observe that this also includes one or two words in the Voynich script, barely visible in normal light, but clear under UV illumination. They seem to read: oteeeon oiil but in particular the second word is not clear. See also Pelling (2006) p. 164, where he reads oteeeol aim [(17)](http://www.voynich.nu/writing.html#n17).

In the left margins of [f49v](http://www.voynich.nu/q07/index.html#f49v) and [f76r](http://www.voynich.nu/q13/index.html#f76r) we find detached sequences of characters, in the former case accompanied by the Arabic numerals 1-5.

On [f66r](http://www.voynich.nu/q08/index.html#f66r) we find both characters and short words in the left margin. Near the bottom of the same page we find some apparently German words, which have been partially amended, near the dead body of a man or woman, and some other objects. This was first interpreted by R. Salomon as 'der Mussdel', and some now propose the reading 'Musmel'. Neither are generally accepted. There is also some writing in the Voynich script, this time clearly visible.

The text on [f116v](http://www.voynich.nu/q20/index.html#f116v) is a short paragraph of text including what appears to be German, Latin and two words in the Voynich script. An additional line is slightly offset above this, in the top margin. It is perhaps the most debated text in the entire MS.

*Other*

On [f66v](http://www.voynich.nu/q08/index.html#f66v) and [f86v3](http://www.voynich.nu/q14/index.html#f86v3) we find strange, unreadable yet similar scribbles of which it is hard to say whether they represent readable text or not. They are usually referred to as chicken scratches. These two have been added to the below table with the key word 'SCR'. More chicken scratches may be observed above the roots of [fol. 43r](http://www.voynich.nu/q06/index.html#f43), but these really look like scratches, so are not counted.

*Overview of additional or extraneous writing*

The following table combines all types of additional or extraneous writing into a single overview. This includes observations brought together by many different observers. As already indicated above, the sequence(s) of characters integrated in the circular design of [f57v](http://www.voynich.nu/q08/index.html#f57v) is not included here.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fol.** | **Type** | **Lang** | **What / where** | **Comment** |
| 1r | SEQ | V+L | Three vertical sequences of single characters in the right margin, faded or erased | Possibly a decryption attempt by a later owner |
| 1r | WRI | L | The ex libris of Jacobus de Tepenec, followed by the number 19 (TBC) faded or erased, in the bottom margin |  |
| 1v | COL | L | A single 'g' under the paint of the second leftmost green leaf | The suggestion is that this indicates that it should be painted green. |
| 1v | COL | L | A less clear single letter, possibly a 'j', inside the lowest yellow leaf right of the stem | It is not clear that this is ink under the paint. It is hard to identify this as a colour indication. |
| 2r | COL | V | Voynich text (ios an on ?) under green paint of bottom right leaf (middle petal) | The suggestion is that this could also be a painting instruction |
| 2v | COL | L | Cursive ‘fo’ or 'fa' in dark ink just above the first word of the second paragraph. | Similar to 'fo' (meaning folio) annotations on all pages of the alchemical herbal Firenze MS 106. |
| 4r | LET | L | A capital F in the rightmost flower |  |
| 4r | COL | L | The word 'rot' (German for red) written vertically in the stem of the plant (near the bottom) |  |
| 7r | COL | L | Characters probably forming the word 'rot' (German for red) under the paint in the left half of the root. | Could be an additional colour annotation similar to f4r |
| 9v | COL | L | Several characters in the top left flower under the blue paint: 'por' in the top petal, 'p' in the lower left petal and 'r' in the lower right petal | Readings are tentative. These may also be colour annotations. |
| 9v | COL | L | An unclear letter of scribble between the top two petals of the top left flower | Meaning not clear. |
| 9v | COL | L | A single 'g' to the right of the top right flower | Seems to be another colour annotation. |
| 11v | MAR | Nr | The nr. 88 (or Voynichese dd ?) close to the edge of the page, at the height of the second last line. | Seems to refer to the Voynich character sequence dd found on that line. |
| 17r | MAR | L+V | Small writing in the top margin of the page. Under UV illumination, one or two words in Voynich writing, largely faded, can be seen at the end of this fragment. |  |
| 20r | COL | L | A 'p' (or possibly an 'r') at the top of the root system | Could be another colour annotation. |
| 28v | SCR | ? | Some apparent symbols in the middle of the flower | If this is writing, the script has not yet been identfied. |
| 29r | COL | L | An 'r' in the higher part of the root system | Could be another colour annotation. |
| 32r | COL | L | A 'p' and what looks like a 'v' or an 'r' in the bottom right flower. There could be another character after the 'p'. | These could be more colour annotations. |
| 39v | LET | L | A capital B in the white space between the two green part of the bottom right leaf. |  |
| 49v | SEQ | V+L | A vertical sequence of single characters in Voynich script in the left margin, aligned with and just before each line of writing. In addition, the numbers 1 to 5 to the left of the top five characters. | Both may have been added after the main text was added. |
| 57v | OTH | ? | A symbol in the lower right corner of the page | Tentative reading: 17 (-mus) |
| 66r | SEQ | V | Sequences of words and characters in the left margin of the page |  |
| 66r | SEQ | L+V | Writing near the lower left corner of the page, with a reclining person and some other small objects | The well-known 'musdel' reference |
| 66v | SCR | ? | Unintelligible scribbles left of the root of the plant | Similar to those on f86v3 |
| 67r2 | FOL | L | A pencilled 'b' in the upper right corner of the page |  |
| 68r1 | FOL | L | A pencilled 'a' in the upper left corner of the page |  |
| 68r2 | FOL | L | A pencilled 'b' in the upper left corner of the page |  |
| 68r3 | FOL | L | A pencilled 'c' in the upper left corner of the page |  |
| 70r1 | FOL | L | A pencilled 'a' in the upper left corner of the page |  |
| 70r2 | FOL | L | A pencilled 'b' in the upper left corner of the page |  |
| 70v2 | MON | L | The word 'mars' near the Pisces emblem |  |
| 70v1 | MON | L | The word 'aberil' near the Aries emblem |  |
| 71r | MON | L | The word 'aberil' near the Aries emblem |  |
| 71v | MON | L | The word 'may' near the Taurus emblem |  |
| 72r1 | MON | L | The word 'may' near the Taurus emblem |  |
| 72r2 | MON | L | The word 'jong' near the Gemini emblem | This is the strongest evidence that the language is northern French |
| 72r3 | MON | L | The word 'iollet' near the Cancer emblem |  |
| 72v3 | MON | L | The word 'augst' near the Leo emblem |  |
| 72v2 | MON | L | The word 'septe(m)b(r) near the Virgo emblem |  |
| 72v1 | MON | L | The word 'octe(m)bre near the Libra emblem |  |
| 73r | MON | L | The word 'nove(m)bre near the Scorpius emblem |  |
| 73v | MON | L | The word 'decembre' near the Sagittarius emblem |  |
| 76r | SEQ | V+L | A vertical sequence of single characters in Voynich script in the left margin, aligned with some of the lines of writing. |  |
| 86v3 | SCR | ? | Unintelligible scribbles in the middle of the page | Similar to those on f66v |
| 99v | COL | V? | What looks like Voynichese qo in the fourth root of the third row | If this is writing, it is very small |
| 116v | MAR | L+V | Recipe or spell, in a mixture of pseudo-Latin and German, with two words in Voynich script. Additional marginal drawings. |  |

Text Analysis

Introduction

The main mystery of the Voynich MS is clearly its unknown writing.

This page addresses the statistical analysis of the text. A great number of such statistical analyses has been made over the last 100+ years. Different techniques have been employed, either in order to decipher the text, or just to better understand its properties.

Much of the material summarised here includes hypotheses about the MS text or tentative conclusions. Hypotheses and tentative conclusions will not be the main focus of this part of the site. The aim is to collect as much information as possible. Conclusions can only be drawn after taking into account all of the available statistics.

This topic is addressed in five different areas, and the present page is an entry point into these areas, as follows:

1. Introductory information

2. Character statistics

3. Word structure

4. Word statistics

5. Sentences, paragraphs, sections

Before this, two disclaimers:

1. It is not possible for any one person to read everything that has been written on this topic, and for this reason this section can always be found wanting.
2. It is difficult to present the multitude of analyses that have been performed in an orderly fashion. Beside the five general areas indicated above, there are a number of studies that cannot be classified easily.

In the following, the five areas are briefly summarised.

Introductory information

This part introduces the most common concepts used in the analysis section: Currier languages, entropy, Zipf law, etc.

*Some additional words about transliteration*

The transliteration alphabet used throughout the site is the EVA (English-Voynich Alphabet, previously Extended) alphabet. In some places, I will use small graphic files for the Voynich characters. In the present analysis section, the Voynich characters are rendered by the "Voynich EVA Hand 1" True Type font created by Gabriel Landini. This is demonstrated below, using the first paragraph of text on folio 1r of the manuscript:

The following figure was created using the EVA True Type font. The EVA text representing this section is given below it. It is then repeated, but using the EVA True Type font for the rendition.

Text, letter

Description automatically generated

|  |  |
| --- | --- |
| fachys ykal ar ataiin Shol Shory cThres y kor Sholdy  sory cThar or y kair chtaiin Shar are cThar cThar dan  syaiir Sheky or ykaiin Shod cThoary cThes daraiin sa  o'oiin oteey oteor roloty cTh\*ar daiin otaiin or okan  sair y chear cThaiin cPhar cFhaiin ydaraiShy | |
| fachys ykal ar ataiin Shol Shory cThres y kor Sholdy  sory cThar or y kair chtaiin Shar are cThar cThar dan  syaiir Sheky or ykaiin Shod cThoary cThes daraiin sa  o'oiin oteey oteor roloty cTh\*ar daiin otaiin or okan  sair y chear cThaiin cPhar cFhaiin ydaraiShy |

The choice of the transliteration alphabet will have an impact on numerical analysis done on the Voynich MS text. This is particularly important for the calculation of the word length distribution, since the number of characters to represent one 'glyph' of the Voynich MS text is different for each alphabet. It does play a role in other statistics as well. In general, the EVA alphabet is not the most suitable for performing statistics.

Character statistics

This includes, among others:

* qualitative observations of the properties of some characters
* frequency distribution of single characters
* properties of combinations of two or three characters (entropy)
* vowel / consonant analysis

Word statistics

When people talk about a 'word' in the Voynich MS, they refer to a string of characters separated from other such words by a space in the writing. Whether these strings of characters actually represent words as we understand it, is not certain.

The analysis of the apparent words in the Voynich MS is discussed in two separate sections. The first treats the word structure, a unique property of the Voynich MS text.

The second section includes:

* some qualitative observations of the properties of some words
* the question whether spaces in the MS are 'word separators'
* frequency distribution of words
* application of Zipf's laws

Sentences, paragraphs, sections

This includes topics like:

* The search for common word combinations. (The apparent lack of common phrases is one of the main anomalies of the Voynich MS text).
* Currier's 'line as a functional unit'
* Page cluster analysis based on word statistics
* Long-range character and word correlations
* Any other work that is mixed or difficult to classify

Analysis Section ( 2/5 ) - Character statistics

2.1 Introduction

This page will first list a number of observations about the Voynich MS character statistics that may be found in the printed literature, and then concentrate on more quantitative analysis results.

2.2 Observations in the printed literature

*Tiltman (1967)[[21]](#endnote-21)*

(Note: Tiltman treats f as a variant form of k and p as a variant form of t. In the following, characters or sequences in parentheses represent such variant forms).

* cKh (cFh) and cTh (cPh) appear to be infixes of k (f) and t (p) within ch. The variant symbol represented by m appears most commonly at the end of a line, rarely anywhere else.
* Paragraphs nearly always begin with k (f) or t (p), most commonly in the second variant forms, which also occur frequently in words in the top lines of paragraphs where there is some extra space.
* y occurs quite frequently as the initial symbol of a line followed immediately by a combination of symbols which seem to be happy without it in any part of a line away from the beginning. Otherwise it occurs chiefly before spaces very frequently preceded immediately by d. Hence my belief that these two have some separative or conjunctive function. (I have to admit, however, that y also seems sometimes to take the place of o before k or t (though rarely, if ever, after q); this is particularly noticeable in some of the captions to illustrations in the astronomical section of the manuscript - these most commonly begin ok (of) or ot (op) and it is here that we occasionally see yk (yf) or yt (yp).
* o, which has a very common and very definite function in "roots", seems to occur frequently in "suffixes" in rather similar usage to a, but nearly always as or and ol.
* l, usually preceded by a or o, is very commonly followed by d, much less commonly by t (p), with or without a space between.
* Speaking generally, each character behaves as if it has its own place in an 'order of precedence' within words; some symbols such as o and y seem to be able to occupy two functionally different places.

*Currier (1976)[[22]](#endnote-22)*

* These letters: o, d, y, s all seem to start with a 'c'-curve. [...] The forms all have counterparts starting with i: j , l , r, etc. We also have a = e + i.
* The final letters are in two series, one preceded by a and the other by o, giving a series of sixteen: n in iin iiin, l (il) (iil) (iiil), r ir (iir) (iiir), m im (iim) (iiim).  
  The ones in parentheses are very low-frequency; the others all occur with respectable frequency. In addition, these combinations of symbols which appear as finals may occur separately. A large number of unattached finals is a characteristic of 'Language' B, and not 'Language' A.
* One can pick out resemblances between Latin abbreviations and other alphabets for most symbols except for the series t , k , p , f. [...] The last two [...] appear 90-95% of the time in the first lines of paragraphs, in some 400 occurrences in one section of the manuscript.
* One might conclude that p , f are an elaborate form of t , k , with the same value. [...] But, not true! These two letters p , f are *not* the same as those two t , k , as the statistics show. The letters t , k are followed anywhere in a 'word' by e about half the time (say 750 out of a total of 1500), including initially. These two, p , f, are *never, ever, anywhere* in the manuscript, followed by e. [...] Therefore, p , f are *not* aberrant or variant forms of t , k

Currier's first observation has been noted independently by several people, and was taken up recently by Brian Cham, who developed the curve-line system[[23]](#endnote-23) out of it.

*D'Imperio (1978)[[24]](#endnote-24)*

* The split gallows seem only to occur on first lines of paragraphs, and in labels.
* The same "word" may be repeated two, three or more times
* Many "words" differ by only one character and are found in each other's vicinity
* Certain symbols occur characteristically at the beginnings, middles or ends of words, and in certain preferred sequences
* Certain symbols appear very rarely, and only on certain pages
* There are very few doublets. Primarily: e or i and occasionally also y, d and o.
* There are very few single-letter words in the running text, primarily s and y.
* Prefix-like elements are found in front of certain "words" that also occur commonly without them. Such prefix-like elements are: qo, o and y
* The symbol q almost always precedes o, connected to it by an extension of the crossbar of the 4. This combination is found almost always at word start.
* On most herbal folios, the first paragraph usually starts with t, k, p or f, usually immediately followed by ch, Sh, o, y, aiin or dy.
* Labels very rarely start with t, k, p or f . Instead, they often start with o, d, y or sometimes s or ch.

2.3 Character frequencies

There is no consolidated set of this most basic statistic due to the use of different transliteration alphabets and different transliteration sources. Several examples may be found in different sources.

One example is found in D'Imperio (1978), Fig. 28 on p.106, from several sources but none covering the entire MS text.

René Zandbergen has produced some graphics of the single character frequency distribution of the entire MS, for four different transliterations[[25]](#endnote-25).

As a very short summary, the single character frequency distribution in the most important transliteration alphabets is largely similar to that of texts in normal European languages, thought the drop in frequency appears to be marginally steeper.

2.4 Entropy

In general, the entropy is a single value computed from a frequency or probability distribution. Clearly, each frequency distribution will have its corresponding entropy value, but any particular entropy value can match an infinite number of different distributions. The entropy value is maximum in case all frequencies are equal, and is smaller in case this distribution is skewed.

The entropy of the Voynich MS text was first analysed in detail by the Yale professor William Ralph Bennett Jr. He developed the concept in many easy steps and in more detail than in the above-mentioned introductory page. He first analysed texts in common European languages and then addressed the Voynich MS text, which he transliterated using his own transliteration alphabet. He writes:

*[...] the statistical properties of the Voynich Manuscript are quite remarkable. The writing exhibits fantastically low values of the entropy per character over that found in any normal text written in any of the possible source languages (see Table 5). The values of h1 [i.e. first order entropy - RZ] are comparable to those encountered earlier in this chapter with tables of numbers. Yet the ratio h1/h2 is much more representative of European languages than of a table of numbers alone.*

His computed values are as follows[[26]](#endnote-26):

|  |  |  |
| --- | --- | --- |
| **Entropy order** | **Normal languages** | **Voynich MS** |
| **First** | 3.91 - 4.14 | 3.66 |
| **Second** | 3.01 - 3.37 | 2.22 |
| **Third** | 2.12 - 2.62 | 1.86 |

He finally identified one language with a set of similarly low entropy values, namely Hawaiian, but he also pointed out that this is not likely to be significant.

Zandbergen has re-done the calculation for first- and second-order entropy for a larger number of languages, using the text of the Universal Declaration of Human Rights. In this analysis, the space character has not been interpreted as a character, but as a separator between words. The first plot below shows the (conditional) second-order entropy plotted against the first-order entropy, for a number of modern European languages, also including the results for the Voynich MS. The Voynich MS statistics are those computed by Bennett (left-most point) and those computed by Dennis Stallings for Herbal-A and Herbal-B using the Currier alphabet. The meaning of the legend is shown in a table below the figure. It is clear that none of the languages shows a similar behaviour to the Voynich MS text.

Chart, scatter chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Code** | **Meaning** | **Languages** |
| ROM | Romance languages | Latin, French, Spanish, Italian, Portuguese, Catalan, Galician, Occitan (Auvergnan), Corsican, Friuli, Maltese |
| GER | Germanic languages | English, German, Dutch, Frisian, Afrikaans |
| SCA | Scandinavian languages | Swedish, Norwegian (Modern and Bogmål), Danish, Icelandic |
| SLA | Slavic languages | Russian, Polish, Czech, Slovak, Croatian, Bulgarian, Macedonian, Belorus, Georgian |
| GAE | Gaelic languages | Scottish Gaelic, Irish Gaelic, Breton, Manx |

The calculations have been repeated for a number of other languages from around the world. These are listed in the following table:

|  |  |  |
| --- | --- | --- |
| **Code** | **Meaning** | **Languages** |
| EUO | European, other | Albanian, Basque, Finnish, Hungarian |
| AFR | African | Ethiopian (Amharic), Swahili, Hausa, Edo, Somali, Bari |
| IEU | Indo-European | Greek, Estonian, Latvian, Lithuanian, Farsi, Hindi, Nepali, Urdu |
| DRA | Dravidian | Malayalam, Kannada |
| ASI | Asian | Turkish, Armenian, Turkmen, Kurdish, Hebrew, Arabic, Azerbaidjani, Bengali, Minjiang (a Chinese dialect, spoken vs. written), Tibetan, Mongolian, Japanese, Korean, Thai, Laotian, Burmese, Cambodian, Vietnamese, Indonesian, Tagalog, Cebuano, Hawaiian |

The result is shown in the following plot, where the points for the "European" languages have been repeated in grey.

Chart, scatter chart

Description automatically generated

Here we see a number of points among the group of Asian languages that lie in the relative vicinity of the Voynich MS text. The lowest and leftmost point is Tagalog (Philippinian). The two points to the right of this are the spoken and written version of Minjiang. These text have been written in the Latin alphabet without indication of tones. Hawaiian, the language named by Bennett, is the lowest point directly above those for the Voynich MS.

A more recent study by Luke Lindemann and Claire Bowern of Yale University[[27]](#endnote-27) covers some of the same ground, and analyses even more different languages, including historical corpora. It comes to the same conclusion and summarises:

*The character set size and frequency of characters is conventional, but the characters are combined in an extremely predictable way, as indicated by an unusual conditional character entropy that is distinctly lower than any of the 316 comparison texts. This discrepancy is not attributable to the transcription system used to encode Voynich, although decisions about the compositionality of glyph sequences can have a significant effect on entropy. Nor is it the result of conventional scholarly abbreviations of the historical period or the absence of written vowels. Rather, it is largely the result of common characters which are heavily restricted to certain positions within the word. Voynichese most closely resembles tonal languages written in the Latin script and languages with relatively limited syllabic inventories.*

An alternative method to compute entropy is the so-called 'commas' method, which has been used by Jim Reeds and later by Gabriel Landini.

Jorge Stolfi has set up a tool to visualise the number of bits of entropy per character in the following location: [>>Jorge Stolfi: where are the bits?](http://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/98-07-09-local-entropy/)

Furthermore, Zandbergen addressed the question how it is possible that the character and bigram entropy of the Voynich MS text is so much lower than that of, say, Latin, while the word entropy is similar. This is addressed at this page: [From bigram entropy to word entropy](http://www.voynich.nu/extra/wordent.html). The short summary is that, counting from the start of each word, the entropy per character in the Voynich MS starts off lower for the first two characters, but is higher for the remainder of the characters, when compared to normal languages.

While the entropy values are single values derived from a frequency distribution, more can be learned by looking at the detail of these distributions, for which [see here](http://www.voynich.nu/extra/sol_ent.html). This discussion exemplifies even better how much different the Voynich MS text is from 'regular languages'.

There is a ***critically important conclusion*** to be drawn from the first- and second-order entropy values reported by various authors. As already mentioned, the entropy values do not change when one consistently replaces characters by others, i.e. in a simple substitution cipher. This tells us something about the possible plain text of the Voynich MS.

1. It could be that the text is meaningless, i.e. there is no plain text language, and the anomalously low entropy is the result of whatever process was used to generate the strings of characters
2. If there is a plaintext that was encoded using a simple substitution, then this plaintext must have the same anomalously low entropy values. This then excludes most of the typical languages that might be assumed for a European MS of the 15th Century. In fact, no candidate plaintext language could yet be identified. Hawaiian, the one identified by Bennett, does not match for other reasons (as will become apparent in later pages). Some languages like Hebrew, the various Arabic languages, Persian, Armenian etc. have not yet been tested quantitatively.
3. If there is a plaintext in one of the known languages used in European MSs of the 15th Century, then this text must have been modified by some process changing the statistics quite drastically. This change is indeed so drastic that it is no longer possible to identify the plaintext language from the Voynich MS text, and any attempts of this nature will be invalid.

In general, and quite briefly, any attempt to translate the Voynich MS into something meaningful in Greek, Latin, English, etc. using a simple substitution must fail. As this is the first thing most people will try, we can begin to understand how the MS has resisted all translation attempts.

2.5 Vowel/consonant detection

An algorithm for detection of vowels and consonants was designed by B.V. Sukhotin, and Jacques Guy has experimented with this in the 1990's. He published a first English summary of the algorithm in Cryptologia [(see note 5)](http://www.voynich.nu/a2_char.html#n05). Results indicated that the characters that look like vowels (a, o, y) also appeared statistically like vowels, though the confidence of the result was not very high. There is also a more recent [>>internet blog entry](http://www.davidjackson.info/voynich/2015/10/11/an-examination-of-the-vowels-in-voynichese/) related to running Sukhotin's algorithm on individual pages of the MS.

Significant further work on vowel/consonant detection has been done using Hidden Markov Modelling, which is summarised in the next section.

2.6 Hidden Markov Modelling

Markov and Hidden Markov models are engineered to handle data which can be represented as ‘sequence’ of observations over time. Hidden Markov models are probabilistic frameworks where the observed data are modeled as a series of outputs generated by one of several (hidden) internal states. There are not many publications that discuss its application to the Voynich MS. Following is a summary.

**Jeffrey Krischer**  
Techniques similar to HMM processing were initially classified, and publications of its application have been restricted. D'Imperio mentions [(see note 4)](http://www.voynich.nu/a2_char.html#n04) a first application of this in a paper by Jeffrey Krischer. He also published an unclassified summary [(13)](http://www.voynich.nu/a2_char.html#n13), which does not give any details about the method and the results.

**Mary D'Imperio**  
Mary D'Imperio made her own analysis, which was published in an undated paper that was also classified, but released a few years ago [(14)](http://www.voynich.nu/a2_char.html#n14).

**Jim Reeds**  
Jim Reeds has made similar analyses, but these have not been published in full. He summarised that, while for known languages the HMM tends to separate vowels and consonants, for the Voynich MS text the analysis brings out a circular pattern of characters. This may be the same as what D'Imperio found and published.

**Reddy and Knight**  
One of the analyses reported in Reddy and Knight (2011) [(15)](http://www.voynich.nu/a2_char.html#n15) is the application of a two-state bigram HMM to the characters in the Voynich MS words. They report that normally, in alphabetic languages like English, the clusters correspond almost perfectly to vowels and consonants. They find, however, that with the Voynich MS the last character of every word is generated by one of the HMM states, and all other characters by another, i.e. as if the vowels are at the end of each word. They speculate further on this point in their paper.

**Later work**  
A recent paper by Acedo [(16)](http://www.voynich.nu/a2_char.html#n16) describes a single application of a two-state HMM to the Takehashi transliteration of the MS and tentatively concludes that this successfully shows a separation into vowels and consonants.

2.7 Other material

*2.7.1 Line-initial/final and word initial/final character properties*

Following observations are paraphrased from Currier's papers.

* In those pages where the text is presented linearly, the line is a functional entity. The following three bullets clarify this general observation
* 1: the frequency counts of characters at beginnings and endings of lines are markedly different from elsewhere. There are some characters that may not occur initially in a line. There are others whose occurrence is about one hundredth of the expected
* 2: the ends of lines seem to contain meaningless symbols: little groups of letters which don't occur anywhere else [(17)](http://www.voynich.nu/a2_char.html#n17). There is one symbol [(18)](http://www.voynich.nu/a2_char.html#n18) that occurs at the end of the last word of a line 85% of all of its occurrences: m
* 3: there is not one single case of a repeat going over the end of a line to the beginning of the next
* Skewed frequencies at beginnings of lines may be illustrated by the two letters ch and Sh. If its occurrence as an initial were random, we would expect it to occur one seventh of the time in each word position of a line. Actually, it is a very infrequent word initial at the beginning of a line, except when there is an intercalated o. This applies only to 'Language' A.
* The 'ligatures' [ cKh cTh cFh cPh ] can *never* occur as paragraph initial, and almost never line initial.

The observation of Currier that the line appears to be a functional unit was further analysed in 2012 by Elmar Vogt, for which [>>see here](http://voynichthoughts.files.wordpress.com/2012/11/the_voynich_line.pdf). One of the most obvious features he shows is that, when using the EVA alphabet, the first word tends to be on average 1 character longer than the second and following words.

*2.7.2 Location of gallows (and other) characters*

Julian Bunn highlighted the positions of the gallows characters on each folio of the MS in [>>a page at his blog](http://voynichattacks.wordpress.com/2012/06/29/page-positional-gallows-mk-ii/), in colour coded graphics. They show a peculiar vertical pattern, which may be related to the observations of Andreas Schinner in his 2007 Cryptologia paper [(19)](http://www.voynich.nu/a2_char.html#n19), which is discussed in a [later page](http://www.voynich.nu/a5_synt.html#schinner).

The following page by Sean V. Palmer gives a very visual representation of the feature that many characters have very preferential positions inside the words of the MS: [>>Voynich MS glyph position stacks](http://inamidst.com/voynich/stacks).

Analysis Section ( 3/5 ) - Word structure

3.1 Introduction

Several authors have identified structures in the composition of words in the Voynich MS. These are, roughly in chronological order:

*3.1.1 Tiltman's split in roots and suffixes*

Tiltman observed that many words in the stars or recipes section (which was the only sample he had available for a detailed analysis) were composed of two parts. He set up a table in Plate 17 of his publication, which is shown below, converted to EVA:

|  |  |
| --- | --- |
| **Roots** | **Suffixes** |
| ok   of | an   ain   aiin   aiiin |
| ot   op | ar   air   aiir   aiiir |
| qok   qof | al   ail   aiil   aiiil |
| qot   qop | or |
| ch | ol |
| Sh | ey   eey   eeey |
| d | edy   eedy   eeedy |
| s |  |

Every combination of a 'root' and a 'suffix' gives a valid word. He roughly subdivided the suffixes into three groups, depending on whether they contain a, o or e. Tiltman observes that the suffixes (which he also calls 'finals') are often found standing alone in the stars section of the MS.

Some additional observations by Tiltman related to this are:

* “Regarding the second type of suffix, some of the combinations are so rare that I have been uncertain whether to take any account of them at all. Some are very common indeed. It seems to me that each of these combinations beginning with a has its own characteristic frequency which it maintains throughout the MS and independent of context (except in case where two or more a groups are together in series, as referred to later). These a groups, e.g. ar or aiin, frequently occur attached directly to "roots", particularly ok (of), ot (op), d and s. okaiin (ofaiin), qokaiin (qofaiin) and daiin rank high among the commonest words in the MS.
* I am unable to avoid the conclusion that the occurrences of the symbol e up to three times in one form of suffix, and the symbol i up to three times in the other must have some systematic significance.”

Tiltman also presents an observation, apparently offered to him by one Peter Long, "that the a groups might represent Roman numerals. Thus aiin might represent 'iij', and ar ar al 'xxv', but this, if true would only present one with a set of numbered categories, which doesn't solve the problem. In any case, though it accounts for the properties of the commoner combinations, it produces many impossible ones."

*3.1.2 Mike Roe's generic word*

The following pattern was contributed to the Voynich MS mailing list by one of its original participants Micheal Roe. His system is represented here translated to the EVA alphabet. Each path represents a valid word, and Mike suggested that this could perhaps present evidence of grammar of the Voynich language:

+- o --+ +- r -+

o --+ +--+ +--+ +--+

| +- t -+ | +- cho -+ +- l -+ |

qo --+--+ +--+ |

| +- k -+ | +- e ---+ |

cho --+ | | | |

| +- ee --+ |

| | | |

+--+- che -+-- y ------+------>

| | | |

| +- ch --+ |

| | | |

| +- sh --+ |

| | | |

| +-------+ |

| |

| +- al ---+ |

| | | |

+--+- am ---+----------+

| |

+- ain --+

| |

+- aiin -+

Diagram, schematic

Description automatically generated

3.2 Jorge Stolfi's ground-breaking work

*3.2.1 Split into 'soft' and 'hard' characters*

Jorge Stolfi discovered a new structure in the words of the Voynich MS, by grouping all characters into 'soft' and 'hard' and showing that the vast majority of words consists of one, two or three groups, which he calls prefix, stem or midfix, and suffix. The first and last consist of 'soft' characters, and the stem or midfix of hard characters. Stolfi has since then been able to set up a more detailed (and slightly more complicated) description of the words (see 'word grammar' below). However, the simple principle behind the prefix-stem-suffix structure makes it worthwhile to look at it first.

It is explained in detail on a page on his site[[28]](#endnote-28).

It explains that the following characters are 'hard' characters and build the optional stem of a word: ch Sh t k p f e cTh cKh cPh cFh

All other characters are soft, and build the prefix and suffix. The rule states that the vast majority of Voynich words are made up as one of:

* Soft characters only (i.e. a combination of prefix and suffix)
* Prefix + stem
* Stem + suffix
* Prefix + stem + suffix

The distribution of these patterns throughout the MS, and the possible patterns for each word part should provide further interesting clues about the language of the Voynich MS.

Jorge Stolfi later found that the distribution of the 'hard' characters seems to be governed by a strict and simple rule. First of all, they can include something that has been called a pedestal or a plateau (i.e. typically the character ch) and independently of that they can include a gallows character ( t k p f ).

*3.2.2 Fine structure*

Later, Stolfi analysed a 'fine structure'[[29]](#endnote-29) of words in the Voynich MS. This is also known as the 'OKOKO' paradigm.

*3.2.3 Word grammar*

Most of the features found by Jorge Stolfi were later combined into what he calls the grammar of Voynich words. It is the most complete and most accurate breakdown of the word structure into a set of rules, and includes core, mantle, and crust characters. It is explained in great detail on this page[[30]](#endnote-30), while the formal grammar definition is here[[31]](#endnote-31). The interested reader is advised to study this page directly, as it is quite difficult to summarise.

3.3 Hidden Markov Modelling

This topic was discussed previously but its results also reflect on the word structure.

3.4 Summary

The fact that structures like the ones introduced in this page exist, tells us that the MS text is not one that was encrypted from an Indo-European plain text using the type of encryption available in the early 15th Century. Any tentative solution working along these lines will necessarily fail.

The exact word structure has not been identified definitely. This page shows several cases, and in general one may observe that the simple paradigms will 'cover' or 'explain' a smaller percentage of the word types in the MS, while the more complicated ones cover a larger percentage.

The word structure is also likely to completely explain the anomalously low entropy values of the Voynich MS text, though what is cause and what is effect is not yet fully understood.

Analysis Section ( 4/5 ) - Word statistics

4.1 Introduction

The term 'word' requires an additional introduction. The MS clearly has groups of characters separated by spaces. It has become usual to call these groups of characters 'words', but it is not certain that these also represent words in the grammatical sense. When one looks a bit more closely, it becomes evident that not all word spaces are that clear, and occasionally, spaces between individual characters are as wide as some word spaces.

A second point concerns terminology as used by different authors. The question 'How many words are there in the Voynich MS?' could lead to two different answers, namely the total number of words, or the number of different words. Reddy and Knight use in this context the terms word tokens and word types. The meaning is most easily explained by noting that the phrase:

*To Be Or Not To Be*

has six word tokens and four word types. Whenever the short term 'word' is used (e.g. quoting a source) the distinction is not relevant. Reddy and Knight report that the Voynich MS text contains 37,919 word tokens and 8114 word types[[32]](#endnote-32).

As before, this page will first list a number of observations about the Voynich MS words that may be found in the older literature, and then concentrate on word statistics analyses that have been performed later.

Due to various reasons, word statistics are likely to be less reliable than most other statistics. One the one hand, there is the uncertainty in the identification of word spaces. Secondly, there are uncertainties in the transliteration, causing that identical words may be represented differently, or different words may be transliterated the same. Finally, when considering word length, the answer will depend on the transliteration alphabet used, and the uncertainty what constitutes a single character.

4.2 Observations in printed literature at the word level

*4.2.1 D'Imperio (1978)*

* There are very few single-letter words in the running text, primarily s and y.

4.3 Are word spaces significant?

At first sight this seems an easy question, since the word spaces are clearly observable, and the word tokens delineated by them show a frequency distribution that is quite natural. Immediately, a number of frequently occurring word types like daiin or chol may be recognised.

On the other hand, it can also be observed that there tend to be a few specific characters that occur most frequently immediately before or immediately after word spaces. It is as if this were similar to the rules in the Arabic script that certain characters cannot be connected to the next character. Such spaces would not be real word spaces. In addition, some words that appear standing alone can also appear connected together. For example beside chol daiin also choldaiin occurs.

Then again, an argument in favour that the word spaces are real is formed by the labels. The label words, which are found standing alone, also occur in the running text separated by spaces, and only very rarely with a space in between. This qualitative argument still needs to be confirmed quantitatively, though.

In conclusion, one cannot be absolutely certain about this question, but the evidence tends to be in favour that the visible spaces in the MS are meant to be word (or other unit of meaning) separators.

4.4 Are words really words?

Even if the word spaces are significant, one may still doubt that each such word token really represents a complete word in plain text. It has been suggested that each written word token only represents a syllable, or even just one character. In the latter case, the MS text would either be highly verbose, or contain lots of nulls.

Even if the Voynich MS word tokens are words, there is the additional important question whether it is possible to convert the Voynich MS text to a meaningful plain text by a word for word substitution. In most proposed solutions this tends to be an unstated assumption.

4.5 Word frequencies

This again varies due to transliterations, etc.. This is closely related with the analyses of Zipf's law.

The number of word types that appear only once in the MS (so-called *hapax legomena*) is rather high (about half of all word types). A detailed comparison with plain texts in other languages should be made.

4.6 Word length distribution

Word length statistics are perhaps the least reliable of all, as explained in the introduction. One can still do the calculations, and expect different results for the word token length distribution and the word type length distribution, since the more frequent word types tend to be shorter. (This is true both for normal languages and the Voynich MS text). In this[[33]](#endnote-33) analysis Jorge Stolfi arrived at the surprising conclusion that, based on his assumptions about character and space definition, the word type length distribution is almost perfectly binomial. This is unusual for a natural language. This observation is still not properly understood or explained. Stolfi presents one potential solution here[[34]](#endnote-34).

This topic is also addressed by Reddy and Knight who suggest that this shape is very similar to that of devowelled English or Arabic, or in fact Chinese represented in Pinyin. They suggest that the Voynich MS text in this respect resembles an abjad. There is, however, the problem that devoweled languages will have entropy values that are much higher than the Voynich MS text.

4.7 Word entropy

The word token entropy for the entire MS has once been computed to be 11 bits, which would be normal for a text of this length (see here[[35]](#endnote-35)).

Reddy and Knight compute the value for combined biological and recipes section (all in Currier B language), which comprises 17,597 word tokens, as 9.666, which is again an expected value.

Here[[36]](#endnote-36), the word entropy for the first 8000 words in the recipes section is computed as 9.927, not fully consistent with the value above, but all values are the right order of magnitude for a normal text.

4.8 Zipf's laws

This topic was analysed by Gabriel Landini, who wrote a paper titled "Zipf's law in the Voynich Manuscript". This is not available on-line.

Analysis Section ( 5/5 ) - Sentences, paragraphs, sections

5.1 Introduction

As before, this page will first list a number of observations about the Voynich MS syntax or grammar that may be found in the earlier literature, and then concentrate on analyses that have been performed more recently. The latter includes among others cluster analyses detecting variation of language 'usage' across the MS and long-range correlation studies which indicate that the MS text seems to behave just like normal language.

5.2 Observations in printed literature at syntax level

*Tiltman (1967)*

(Note: Tiltman treats f as a variant form of k and p as a variant form of t. In the following, characters or sequences in parentheses represent such variant forms).

* There are [...] many examples of 2, 3, 4 or even 5 a groups strung together on end, with or without spaces between them. When this occurs, there appears to be some selective preference. For example, ar is very frequently doubled, whereas aiin, which is generally significantly commoner, is rarely found doubled. Perhaps the commonest succession of three of these groups is ar ar al.  
  al very frequently follows ar, but ar hardly ever follows al.
* Some of the commoner words, e.g. okeey, okeedy, qokeedy, okaiin, okar, okal, daiin, chedy occur twice running, occasionally three times.

*Currier (1976)*

'Unattached' finals scattered throughout language 'B' texts in considerable profusion; generally *much* less noticeable in Language 'A'.

*D'Imperio (1978)*

* The split gallows seem only to occur on first lines of paragraphs, and in labels.
* The same "word" may be repeated two, three or more times
* Many "words" differ by only one character and are found in each other's vicinity
* Certain symbols occur characteristically at the beginnings, middles or ends of words, and in certain preferred sequences

5.3 Repeating Phrases

Several authors have remarked that the MS curiously lacks repeated phrases of 2 or more words. Such repeated expressions would typically be expected, for example, in the herbal section, as may be observed in editions of medieval herbal MSs[[37]](#endnote-37). Various different mechanisms can be proposed that would lead to this phenomenon (e.g. word transposition or scrambling). It could also be an indication that the text is meaningless. Furthermore, it may be partly due to errors in the text and the fact that orthography was not yet standardised, though one may seriously doubt whether that is sufficient to explain it. This is something that could be tested numerically, by taking an edited text and introducing errors and non-standard orthography. I do not believe that this has been attempted so far. In any case, the reason for the missing long repeats is so far not understood, and it remains a key issue.

5.4 Cluster analysis of the MS text

Typically, cluster analysis is applied to the MS text after it has been split into pages or other groups, in order to analyse the correspondence between the texts of the different parts. The earliest analyses mainly tended to confirm Currier's split between A and B languages, but also demonstrated that there are additional details. This point has become more relevant in more recent publications on this topic.

This subject requires a more intensive discussion than what is available now, and not all work has been published. Initially, the following is a combination of older on-line sources and some more recent publications.

*5.4.2 Topic modelling*

* Montemurro and Zanette[[38]](#endnote-38). The publication in 2013 by Montemurro and Zanette received quite a bit of attention in the media because, contrary to the work of Rugg and Schinner, it appears to demonstrate that the Voynich MS contains a meaningful message.
* Sterneck, Polish and Bowern[[39]](#endnote-39). This 2021 study by the research group at Yale University of Claire Bowern warrants a discussion and comparison with earlier work.

5.5 Long-range correlations

*5.5.1 The Letter Serial Correlation analysis*

This technique, used by Brendan McKay and Mark Perakh[[40]](#endnote-40), has been applied to a number of plain language texts and also the Voynich MS. While the principle behind the computed statistic is not fully understood, it clearly appears that meaningful texts generate a particular curve exhibiting one minimum, which is equally observed for the Voynich MS. If the words in the text are shuffled around arbitrarily, the curve changes into a flat line.

The tentative conclusion is that the text in the Voynich MS appears to represent something meaningful.

*5.5.2 Long-range correlations by Gabriel Landini*

In a 2001 paper in Cryptologia[[41]](#endnote-41) Gabriel Landini comes to a very similar conclusion.

*5.5.3 The work of Amancio et al.*

Amancio et al. published a paper in 2013[[42]](#endnote-42), based on the results of applying 'big data' analysis techniques to the Voynich MS text. It does not concentrate only on the Voynich MS, but looks at it as an example. The main conclusion is that the text in the Voynich MS is significantly different from a scrambled version of the same text, i.e. it is not an arbitrary sequence of words.

5.6 Other analyses

*5.6.1 Andreas Schinner*

The important 2007 paper by A. Schinner[[43]](#endnote-43) deserves a thorough discussion. It treats several anomalies of the Voynich MS text, such as vertical patterns, and anomalous 'random walk' behaviour, which led the author to conclude that the text is possibly meaningless, and the result of a hoax.

Torsten Timm wrote a paper in 2014, with the most recent revision in Dec.2015[[44]](#endnote-44), in which he proposes a method by which the Voynich MS would have been created. This works along the lines of the vertical patterns mentioned above. Apart from the theory, it includes a large number of additional statistics.

*5.6.2 Word permutation analysis by Marke Fincher*

In a study looking at different statistics of word pairs, where the pair of words in question may be adjacent in the text, or separated by any number of intermediate words, Fincher noted that the text in the Voynich MS does not at all behave in the same way as plain texts in several known languages. The paper does not appear to be available online. The analysis is closely related to the problem of lack of repeating strings, and also suffers from possible inconsistent spelling and errors in the text or transliteration.

*5.6.3 Reddy and Knight*

Sravana Reddy and Kevin Knight produced a publication in 2011[[45]](#endnote-45) with a large selection of old and new statistics of the text of the Voynich MS, titled "What we know about the Voynich MS".

Some of the statistics and results of Reddy and Knight are revisited, and shown graphically, by Sean Palmer[[46]](#endnote-46).

REFERENCES

1. Neal, P., 2021. Barschius (1637) Latin | Voynich Sources. [online] Philipneal.net. Available at: <http://philipneal.net/voynichsources/barschius\_1637\_latin/> [↑](#endnote-ref-1)
2. https://collections.library.yale.edu/catalog/2002046 [↑](#endnote-ref-2)
3. Internet Archive. 2021. Transactions of the College of Physicians of Philadelphia : College of Physicians of Philadelphia : Free Download, Borrow, and Streaming : Internet Archive. [online] Available at: <https://archive.org/details/transactionsofco3431coll/page/414/mode/2up?view=theater> [↑](#endnote-ref-3)
4. 2021. [image] Available at: <https://archimede.imss.fi.it/kircher/562/large/114r.jpg> [↑](#endnote-ref-4)
5. Nast, C., 2021. The Unread: The Mystery of the Voynich Manuscript. [online] The New Yorker. Available at: <https://www.newyorker.com/books/page-turner/the-unread-the-mystery-of-the-voynich-manuscript> [↑](#endnote-ref-5)
6. Blunt, Wilfrid, and Sandra Raphael: The illustrated herbal, London, 1979 [↑](#endnote-ref-6)
7. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-7)
8. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-8)
9. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-9)
10. Newbold, William Romaine: The Cipher of Roger Bacon. Edited with foreword and notes by Roland Grubb Kent. Philadelphia: University of Pennsylvania Press, 1928. [↑](#endnote-ref-10)
11. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-11)
12. Hauer, Bradley and Grzegorz Kondrak: Decoding Anagrammed Texts Written in an Unknown Language and Script, Transactions of the Association for Computational Linguistics, vol. 4, pp. 75ï¿½86, 2016. [↑](#endnote-ref-12)
13. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-13)
14. Cappelli, A.: dizionario di abbreviature Latini ed Italiani, Milano, 1912. 529 pp., ill. [↑](#endnote-ref-14)
15. Gessman, Gustav W.: Die Geheimsymbole der Alchymie, Arzneikunde und Astrologie des Mittelalters, Berlin, Verlag von Karl Siegismund, 1922. [↑](#endnote-ref-15)
16. Tranchedino, Francesco: Diplomatische Geheimschriften - Codex Vindobonensis 2398; The Fine Art Facsimile Edition: Codices Selecti vol. XXII. Graz 1970, ed. Walter Höflechner [↑](#endnote-ref-16)
17. Voynich.nu. 2021. Codex Taurinensis C VII 15. [online] Available at: <http://www.voynich.nu/extra/mstaur.html> [↑](#endnote-ref-17)
18. Petersen, J., 2021. The Origin of the Voynich Glyphs | Voynich Portal. [online] Voynichportal.com. Available at: <http://voynichportal.com/2017/02/24/the-origin-of-the-voynich-glyphs/> [↑](#endnote-ref-18)
19. Some Voynich ideas. 2021. About month names. [online] Available at: <https://voynichanalysis.wordpress.com/2012/03/25/about-month-names/> [↑](#endnote-ref-19)
20. King, David: The Ciphers of the Monks: A Forgotten Number-notation of the Middle Ages, Stuttgart, 2001. [↑](#endnote-ref-20)
21. Tiltman, John H.: The Voynich Manuscript, The Most Mysterious Manuscript in the World. NSA Technical Journal 12 (July 1967), pp.41-85. [↑](#endnote-ref-21)
22. http://www.voynich.nu/extra/img/curr\_main.pdf [↑](#endnote-ref-22)
23. Brian Cham's personal website. 2021. Introduction to the Curve-Line System. [online] Available at: <https://briancham1994.com/2014/12/17/curve-line-system/>. [↑](#endnote-ref-23)
24. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-24)
25. http://www.voynich.nu/extra/sol\_ent.html [↑](#endnote-ref-25)
26. Bennett, William Ralph: Scientific and Engineering Problem Solving with the Computer. Englewood Cliffs: Prentice-Hall, 1976. [↑](#endnote-ref-26)
27. Lindemann, Luke and Claire Bowern: Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript. 27 October 2020. [↑](#endnote-ref-27)
28. https://www.ic.unicamp.br/~stolfi/voynich/97-11-12-pms/ [↑](#endnote-ref-28)
29. https://www.ic.unicamp.br/~stolfi/voynich/Notes/017/Note-017.html [↑](#endnote-ref-29)
30. https://www.ic.unicamp.br/~stolfi/voynich/00-06-07-word-grammar/ [↑](#endnote-ref-30)
31. D'Imperio, Mary E. : The Voynich Manuscript - an elegant enigma. Aegean Park Press, 1978. [↑](#endnote-ref-31)
32. Reddy, Sravana, and Kevin Knight: What we know about the Voynich Manuscript, in Proceedings of the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011. [↑](#endnote-ref-32)
33. https://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/00-12-21-word-length-distr/ [↑](#endnote-ref-33)
34. https://www.ic.unicamp.br/~stolfi/EXPORT/projects/voynich/02-01-18-chinese-redux/ [↑](#endnote-ref-34)
35. http://www.voynich.nu/a1\_intro.html#zipf [↑](#endnote-ref-35)
36. http://www.voynich.nu/extra/wordent.html [↑](#endnote-ref-36)
37. Segre Rutz, Vera: Il giardino magico degli alchimisti: un erbario illustrato trecentesco della Biblioteca universitaria di Pavia e la sua tradizione. Introduzione, edizione critica e commento, Milano, 2000 [↑](#endnote-ref-37)
38. Montemurro, Marcelo and Damian Zanette: Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis, PLoS ONE 8(6): e66344. doi:10.1371/journal.pone.0066344 (2013) [↑](#endnote-ref-38)
39. Sterneck, Rachel, Annie Polish and Claire Bowern: Topic modeling in the Voynich Manuscript. arXiv:2107.02858v1 [cs.CL] 6 July 2021. [↑](#endnote-ref-39)
40. http://www.talkreason.org/Mark's%20sites/Mark's%20perakm%20site/members.cox.net/marperak/Texts/index.html [↑](#endnote-ref-40)
41. Landini, Gabriel: Evidence of linguistic structure in the Voynich MS using spectral analysis, in Cryptologia Vol. 25 Issue 4, 2001, pp. 275-295. [↑](#endnote-ref-41)
42. Amancio DR, Altmann EG, Rybski D, Oliveira ON Jr, Costa LdF: Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript. PLoS ONE 8(7), 2013. [↑](#endnote-ref-42)
43. Skinner, Stephen, with an introduction by R. Prinke and R.Zandbergen: The Voynich Manuscript: The Complete Edition of the World's Most Mysterious and Esoteric Codex, Watkins, London, 2017. [↑](#endnote-ref-43)
44. Timm, T., 2014. How the Voynich Manuscript was created. arXiv preprint arXiv:1407.6639. [↑](#endnote-ref-44)
45. Reddy, Sravana, and Kevin Knight: What we know about the Voynich Manuscript, in Proceedings of the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011. [↑](#endnote-ref-45)
46. Inamidst.com. 2021. Related Voynich MS glyphs. [online] Available at: <http://inamidst.com/voynich/related> [↑](#endnote-ref-46)