# Natural language samples

*Version 1.01. Last updated 2015-03-28.*

# Intro

**Summary:** This pack provides 52 natural language text samples in one convenient download. They are all computer friendly and contain the major languages you're looking for, so don't sweat anywhere else. ☺

It is designed to help with statistical comparisons of natural language and the Voynich Manuscript "text". This includes both types – those attempting to identify if it is written in a natural language at all, and those attempting to identify it with specific languages and features. However you can use it for anything else you want.

This pack is free to download, share and use but please read the information in section 4.

Before you do any experiments, please read this article (and its comments) for an interesting discussion on the matter. It points out potential problems you might not have thought about.

All future updates of this pack, and other resources, will be hosted at https://briancham1994.wordpress.com/voynich-resources/ .

# Contents

# 1. Rationale of selection

In light of the shortcomings of previous studies, I have compiled a large selection of linguistic samples according to the following principles:

- Select a variety of writing systems to capture the peculiarities of transcription.
- Select a variety of writing styles and genres.
- Select a large variety of languages to represent the range of statistical properties in natural languages:
    - Span many geographic areas
    - Span many time periods
    - Span many major language families
    - Span many language features
- Select languages possibly related to the Voynich Manuscript:
    - Select common languages in 15th century European manuscripts.
    - Select languages directly proposed in previous studies, marginalia readings and decipherment attempts.

# 2. Overview of selection

Spatial and temporal summaries of the language selection are below.

## 2a. Geographic spread

The languages selected cover every inhabited continent, but with a focus on Western Eurasia. The map below shows the approximate location of each language as a red dot.

## 2b. Historical spread

The languages range from the 4th millennium BCE to the present day. The samples range from the late 3rd millennium BCE to 2014 CE – a span of over four thousand years.

## 2c. Writing systems

Statistics of text may be affected by both the pecularities of the underlying information and the writing system used to record it.

This selection includes a range of types of different writing systems, rather than having one standardised representation for every language. This includes things like alphabets (e.g. Czech), abjads (e.g. Arabic), abugidas (e.g. Thai), high orthographic depth (e.g. Middle French), low orthographic depth (e.g. romanised Japanese), academic reconstructions (e.g. transliteration of Old Egyptian) and an artificially constructed system for personal use (e.g. ASLSJ transcription of American Sign Language).

# 2d. Technical details

All samples are saved as .txt files with ANSI or UTF-8 encoding. Punctuation, letter cases and some chapter/paragraph numbers have been retained, with the justification that computer code can easily be designed to ignore these if desired.

Efforts have been taken to obtain samples close to or over the length of the Voynich Manuscript (approx. 35,000 words), but please note that this is very difficult for rare or extinct languages.

Samples that used logographic or gestural systems were transcribed into an alphabetic system for ease of computer processing and comparison. The original versions of these samples have been included wherever possible.

# 3. Full list

Time and location information is approximate.

| Code | Language | | Sample |
|---|---|---|---|
| | Name | Details | |
| **N-AIN** | Ainu | **Location:** Northern Japan<br>**Time:** Unknown to present<br>**Family:** Isolate<br>**Notes:** Alphabetic. | *The Book of Common Prayer* by John Batchelor (translated 1896), with non-Ainu phrases removed.<br>4,540 words. Public domain. |
| **N-ALB** | Old Albanian | **Location:** Albania<br>**Time:** 14th century to unknown<br>**Family:** Indo-European, Albanian<br>**Notes:** Has a lot of consonants. Alphabetic. | First 104 chapters of *Meshari* by Gjon Buzuku (1555)<br>62,972 words. Public domain. |
| **N-ARA** | Arabic | **Location:** Middle East and North Africa<br>**Time:** 6th century to present<br>**Family:** Afro-Asiatic<br>**Notes:** Arabic script is an abjad, which the Voynich Manuscript's text may statistically resemble according to Reddy & Knight. Arabic's glyphs have a strong ordering system. | *Commentary on Anatomy in Avicenna's Canon* (شرح تشريح القانون لابن سينا) by Ibn al-Nafis (ابن النفيس) (13th century)[1]<br>99,065 words. Public domain. |
| **N-ARM** | Armenian (modern) | **Location:** Armenia<br>**Time:** 18th century to present<br>**Notes:** Alphabetic.<br>**Family:** Indo-European, Armenian | *Don Quixote (Դոն Կիխոտ)* by Miguel de Cervantes (1605), translated by Paul Makintsyan (Պողոս Մակինցյան) (1982)<br>54,180 words. Public domain. |
| **N-ARW** | Arawak | **Location:** Guianas and Carribean<br>**Time:** Unknown to present<br>**Family:** Arawakan<br>**Notes:** The language spoken in Cuba, which is where Christopher Columbus spent some time.<br>**Notes:** Alphabetic. | *Acts of the Apostles,* translated by Theodore Shultz (1850)<br>22,807 words. Public domain. |
| **N-ASL** | American Sign Language | **Location:** Anglo-America<br>**Time:** 19th century – present<br>**Notes:** Signed language. Represents an example of a very artificial and "engineered" notation system. | Extract from *Alice in Wonderland* by Lewis Carroll (1865), transcribed by Thomas Stone in ASLSJ notation (2014).<br>461 words. © Thomas Stone. Used with permission. |

[1] Sorry, my version of Microsoft Office 2013 can't combine Arabic alphabet text correctly.

| | | | |
|---|---|---|---|
| **N-BAQ** | Basque | **Location:** Pyrenees<br>**Time:** Unknown to present<br>**Family:** Isolate<br>**Notes:** Alphabetic. | *Gero* by Axular (1643)<br>78,344 words. Public domain. |
| **N-CUR** | Eurasian Curlew Birdsong | **Location:** Eurasia<br>**Time:** Unknown to present<br>**Family:** Animal language<br>**Notes:** Represents an example of non-human communication, and a very artificial and "engineered" notation system. | Song of a Eurasian Curlew (no date), transcribed in ABC notation.<br>166 words. Public domain. |
| **N-CZE** | Czech | **Location:** Czech Republic<br>**Time:** c. 14th century to present<br>**Family:** Indo-European, Slavic<br>**Notes:** Language spoken in Prague, the first known location of the Voynich Manuscript. Alphabetic. | "Book of Genesis" and "Book of Exodus" from the *Kralice Bible* by the Unity of the Brethren (translated 1613)<br>54,018 words. Public domain. |
| **N-EGY** | Old Egyptian | **Location:** Ancient Egypt<br>**Time:** 3rd millennium BCE<br>**Family:** Afro-Asiatic<br>**Notes:** Represents an example of an academically constructed notation system. Originally logographic and syllabic. | *Autobiography of Weni* by Weni the Elder (late 3rd millennium BCE), transliterated by Geoffrey Graham<br>1,011 words. Public domain. |
| **N-EMY** | Ch'olti' | **Location:** Guatemala<br>**Time:** Unknown to c. 17th century<br>**Family:** Mayan<br>**Notes:** Logographic or alphabetic. | *Bocabulario Grande* by Fray Francisco Morán (1695)<br>9,047 words. Public domain.<br>**Notes:** The text is a dictionary from Spanish to Ch'olti'. Represents an example of a bilingual text. |
| **N-ENM** | Middle English | **Location:** England<br>**Time:** 12th to 15th century<br>**Family:** Indo-European, Germanic<br>**Notes:** Writing system has high orthographic depth and a lot of inconsistency. Influenced by French. Alphabetic. | *The Canterbury Tales* by Geoffrey Chaucer (late 14th century)<br>150,502 words. Public domain.<br>**Notes:** Represents an example of a text with inconsistent spellings. |
| **N-FIN** | Finnish | **Location:** Finland<br>**Time:** 15th century to present<br>**Family:** Uralic<br>**Notes:** Has a lot of vowels. Alphabetic. | *Kalevala* by multiple, collated by Elias Lönnrot (1835)<br>67,443 words. Public domain.<br>**Note:** Represents an example of a poetic writing style. |

| | | | |
|---|---|---|---|
| **N-FRM** | Middle French | **Location:** France<br>**Time:** 14th to 17th century<br>**Family:** Indo-European, Romance<br>**Notes:** Postulated to be the language of the month name marginalia in the zodiac section. Writing system has high orthographic depth. Alphabetic. | *Pantagruel* by François Rabelais (1530)<br>37,017 words. Public domain. |
| **N-GEO** | Georgian | **Location:** Georgia<br>**Time:** 6th century to present<br>**Family:** Kartvelian<br>**Notes:** Alphabetic. | *The Knight in the Panther's Skin (Vepkhistqaosani/ვეფხისტყაოსანი)* by Shota Rustaveli (შოთა რუსთაველი) (12th century)<br>46,557 words. Public domain. |
| **N-GLE** | Irish (modern), a.ka. Irish Gaelic | **Location:** Ireland<br>**Time:** 13th century to present<br>**Family:** Indo-European, Celtic<br>**Notes:** Alphabetic. | *Constitution of the Irish State* by multiple (1937)<br>17,157 words. Public domain.<br>**Note:** Represents an example of a legal/formal writing style. |
| **N-GMH** | Early New High German | **Location:** Southern Germany<br>**Time:** 14th to 17th century<br>**Family:** Indo-European, Germanic<br>**Notes:** A few art history parallels may link the Voynich Manuscript to Germany. Postulated to be one of the languages in the f66r marginalia and f116v marginalia.<br>**Notes:** Alphabetic. | "Gospel of Matthew" and "Epistle to the Romans" from the *Luther Bible* by Martin Luther (translated 1522)<br>37,537 words. Public domain. |
| **N-GRC** | Ancient Greek | **Location:** Greece<br>**Time:** 9th to 4th century BCE<br>**Family:** Indo-European, Hellenic<br>**Notes:** A few have claimed parallels between the Voynich Manuscript and Greek manuscripts.<br>**Notes:** Alphabetic. | *History of Animals (Των περί τα ζώα ιστοριών)* by Aristotle (Ἀριστοτέλης) (350 BCE)<br>94,864 words. Public domain. |
| **N-GRN** | Guarani | **Location:** Paraguay, Bolivia<br>**Time:** Unknown to present<br>**Family:** Tupani<br>**Notes:** Alphabetic. | *Explanation of Catechism in the Guarani Language* by Nicolas Yapuguai (1724)<br>106,106 words. Public domain.<br>**Note:** This sample uses poor quality OCR scanned text. This has been deliberately chosen to emulate a text sample with wildly inconsistent orthography. |

| | | | |
|---|---|---|---|
| **N-HAW** | Hawaiian | **Location:** Hawaii<br>**Time:** Unknown to present<br>**Family:** Austronesian, Polynesian<br>**Notes:** Stolfi once briefly mentioned that the entropy of Voynich Manuscript text is most similar to Hawaiian. [unfortunately I cannot find the original source again] Very small alphabet.<br>**Notes:** Alphabetic. | *Ka Hoku o ka Pakipika* volume 1, numbers 1 and 2 by multiple (1861)<br>51,987 words. Public domain. |
| **N-HIT** | Hittite | **Location:** Hittite Empire (present day Turkey)<br>**Time:** 16th to 13th century BCE<br>**Family:** Indo-European, Anatolian<br>**Notes:** Represents an example of an academically constructed notation system. Originally logographic and syllabic. | *Illuyanka* (CTH 321) by unknown (c. 2nd millennium BCE)<br>1,781 words. Public domain.<br>**Notes:** Represents an example of a copy of a badly degraded text. Many consecutive lines are multiple interpretations or sources of the same original line. Line and source codes have been removed. |
| **N-IND** | Indonesian | **Location:** Insular Southeast Asia<br>**Time:** Unknown to present<br>**Family:** Austronesian, Nuclear Malayo-Polynesian<br>**Notes:** Alphabetic. | *Penalty and Passion* by Merari Siregar (1920)<br>40,296 words. Public domain. |
| **N-ITA** | Florentine, a.k.a. Florentine Italian or Tuscan Italian | **Location:** Florence<br>**Time:** 14th to 19th century<br>**Family:** Indo-European, Romance<br>**Notes:** Related to present day Italian. A few art history parallels may link the Voynich Manuscript to northern Italy. Alphabetic. | *The Divine Comedy* by Dante Alighieri (early 14th century)<br>96,316 words. Public domain. |
| **N-JPN** | Late Middle Japanese | **Location:** Japan<br>**Time:** 12th to 16th century<br>**Family:** Japonic/Isolate<br>**Notes:** Limited set of phonemes. Extremely shallow orthographic depth. Strong CVC structure. Originally logographic and syllabic script. | *Chronicles of the Authentic Lineages of the Divine Emperors (Jinnō Shōtōki/神皇正統記)* by Kitabatake Chikafusa (北畠 親房) (early 14th century)<br>69,165 words. Public domain.<br>**Notes:** Romanised with the Hepburn system, without capitalisation. |

| | | | |
|---|---|---|---|
| **N-KAA** | Karakalpak, a.k.a. Qaraqalpaq | **Location:** Uzbekistan<br>**Time:** 13th century to present<br>**Family:** Turkic<br>**Notes:** Has similar letter frequency distribution according to Jaśkiewicz. Features vowel harmony. | Collected works of Berdaq G'arg'abay ulı (19th century)<br>1,127 words. Public domain.<br>**Notes:** Uses the 1991-2009 Latin Qaraqalpaq alphabet. |
| **N-KAL** | Greenlandic | **Location:** Greenland<br>**Time:** 13th century to present<br>**Family:** Eskimo-Aleut<br>**Notes:** Polysynthetic language; words can be very long. Alphabetic. | *Universal Declaration of Human Rights* by the United Nations, translated by unknown (1998)<br>1,046 words. Public domain. |
| **N-KAN** | Middle Kannada | **Location:** Southwest India<br>**Time:** 13th to 18th century<br>**Family:** Dravidian<br>**Notes:** Has similar letter frequency distribution according to Jaśkiewicz. The script is made of half-letters that combine into compounds. | Collected *Vachana Sitya (ವಚನ ಸಾಹಿತ್ಯ)* of Sarvajna (ಸರ್ವಜ್ಞ) (16th century)<br>12,212 words. Public domain. |
| **N-KBD** | Kabardian Circassian, a.k.a. Kabardian, Kabardino-Cherkess or East Circassian | **Location:** North Caucasus<br>**Time:** Unknown to present<br>**Family:** Northwest Caucasian<br>**Notes:** Has similar letter frequency distribution according to Jaśkiewicz. The script has few vowels and many consonants, which require large multi-glyph combinations. | Various *Narts* (Нартхымэ акъыбарыхэ) by multiple (date unknown), compiled by the Kabardian Science and Research Institute (1951)<br>25,775 words. Public domain.<br>**Note:** This sample uses poor quality OCR scanned text. This has been deliberately chosen to emulate a text sample with wildly inconsistent orthography. |
| **N-KOR** | Middle Korean | **Location:** Korea<br>**Time:** 10th to 16th century<br>**Family:** Koreanic/Isolate<br>**Notes:** Alphabetic. | *Tale of Hong Gildong (□□□□)* by Heo Gyun (□□ 許筠) (c. 16th century)<br>7,158 words. Public domain. |
| **N-LAT** | Classical Latin | **Location:** Roman Empire<br>**Time:** 75 BCE to 3rd century<br>**Family:** Indo-European, Italic<br>**Notes:** Common scientific language in Medieval Europe. Features heavy inflection. Alphabetic. | *Natural History* by Pliny the Elder (c. 77-79)<br>402,702 words. Public domain. |
| **N-LIT** | Lithuanian | **Location:** Lithuania<br>**Time:** c. 11th century to present<br>**Family:** Indo-European, Baltic<br>**Notes:** Alphabetic. | *Ways of the Ancient Lithuanian* by Simonas Daukantas (1845)<br>58,674 words. Public domain. |

| | | | |
|---|---|---|---|
| **N-LZH** | Classical Chinese | **Location:** China<br>**Time:** 5th century BCE to 20th century CE<br>**Family:** Sino-Tibetan<br>**Notes:** Tonal language. Mandarin and Voynich Manuscript words have a binomial distribution, according to Stolfi. Originally logographic. | *Records of the Three Kingdoms (Sanguozhi/三國志)* volumes 1 and 2 by Chen Shou (陳壽) (3rd century)<br>45,478 words. Public domain.<br>**Notes:** Romanised with Pinyin, which captures tones. |
| **N-MNC** | Manchu | **Location:** Northeast China<br>**Time:** unknown to 19th century<br>**Family:** Tungusic<br>**Notes:** Alphabetic. | *The Fable of Mr. Dungguo and the Ungrateful Wolf (dungg'o siyanxeng jai niohe/东郭先生和狼)* by Ma Zhongzi (馬中錫) (c. 15th or 16th century)<br>586 words. Public domain.<br>**Notes:** Represented in the Latin alphabet since the native Manchu alphabet is difficult for computers to deal with. |
| **N-MON** | Classical Mongolian | **Location:** Mongolia<br>**Time:** c. 12th to 17th century<br>**Family:** Mongolic<br>**Notes:** Alphabetic. | *The Secret History of the Mongols* by unknown (13th century), transcribed by Sergei Kozin into the Latin alphabet<br>28,289 words. Public domain. |
| **N-MWP** | Kalaw Lagaw Ya, a.k.a. Western Torres Strait Language | **Location:** Australia and Torres Strait Islands<br>**Time:** Unknown to present<br>**Family:** Pama-Nyungan<br>**Notes:** Alphabetic. | *Book of Genesis*, translated by Bruce E. Waters (1981)<br>1,174 words. © Bruce E. Waters. Free to use and share for non-commercial purposes. |
| **N-NCI** | Classical Nahuatl, a.k.a. Aztec | **Location:** Central Mexico<br>**Time:** c. 7th to 16th century<br>**Family:** Uto-Aztecan<br>**Notes:** Language identified in Tucker & Talbert's decipherment attempt. Alphabetic. | *Gospel of Luke*, translated by Mariano Paz y Sanchez (1833)<br>28,389 words. Public domain. |
| **N-OCI** | Old Occitan | **Location:** France, Spain, Italy, Monaco<br>**Time:** 8th to 14th century<br>**Family:** Indo-European, Romance<br>**Notes:** Postulated to be the language of the f17r marginalia or the month names in the zodiac section. Alphabetic. | *Libre de Memorias* by Jacme Mascaron (1390)<br>18,819 words. Public domain.<br>**Notes:** Represents an example of a text with some missing and out-of-order pages. |

| | | | |
|---|---|---|---|
| **N-ORV** | Old East Slavic, a.k.a. Old Russian, Old Ukrainian or Old Belarusian | **Location:** Eastern Europe<br>**Time:** c. 10th to 17th century<br>**Family:** Indo-European, Slavic<br>**Notes:** Language identified in Stojko's decipherment attempt (as "Old Ukrainian"). Alphabetic. | *A Journey Beyond the Three Seas (Хождение за три моря)* by Afanasiy Nikitin (*Афана́сий Ники́тин*) (1472)<br>7,269 words. Public domain.<br>**Notes:** Represents an example of a travel journal, which some posit the Voynich Manuscript may be. Text contains details of Indian life and may contain a lot of foreign terms. |
| **N-OSP** | Early Modern Spanish, a.k.a. Classical Spanish, a.k.a. Castilian | **Location:** Castile<br>**Time:** 15th to 17th century<br>**Family:** Indo-European, Romance<br>**Notes:** Alphabetic. | *A Short Account of the Destruction of the Indies* by Bartolomé de las Casas (1552)<br>30,136 words. Public domain. |
| **N-PER** | New Persian, a.k.a. Farsi | **Location:** Iran<br>**Time:** 8th century to present<br>**Family:** Indo-European, Indo-Iranian<br>**Notes:** Abjad | Extract from *The Book of Kings* (Sháh Námeh/هم‌انهاش) by Ferdowsi (ی‌سودرف) (c. 1000)[2]<br>75,653 words. Public domain.<br>**Notes:** The language of the text is at the transition between Middle and New Persian. Represents an example of a poetic, column-based text. |
| **N-PML** | Sabir, a.k.a. Mediterranean Lingua Franca | **Location:** Mediterranean<br>**Time:** 10th to 19th century<br>**Family:** Creole<br>**Notes:** Represents an example of a medieval European creole. Based on Italian. Alphabetic. | Extracts from *The Bourgeois Gentleman* by Molière (1670)<br>240 words. Public domain. |
| **N-QUC** | Classical K'iche' | **Location:** Guatemala<br>**Time:** Unknown to c. 17th<br>**Family:** Mayan<br>**Notes:** Logographic or alphabetic. | *Popol Vuh* by multiple (no date), compiled and published by Francisco Ximénez (18th century)<br>28,525 words. Public domain.<br>**Notes:** Represents an example of a two-column layout. |
| **N-QUE** | Quechua a.k.a. Runa Simi | **Location:** Andes<br>**Time:** c. 14th/15th century to present<br>**Family:** Quechuan<br>**Notes:** Alphabetic. | *Inkarri* by unknown (c. 17th or 18th century)<br>3,451 words. Public domain. |

[2] Sorry, my version of Microsoft Office 2013 can't combine Arabic alphabet text correctly.

| | | | |
|---|---|---|---|
| **N-RUM** | Moldovan, a.k.a. Moldavian or Romanian | **Location:** Moldova<br>**Time:** c. 20th century to present (depending on definition of language)<br>**Family:** Indo-European, Romance<br>**Notes:** Has similar letter frequency distribution according to Jaśkiewicz. The labelling of a "Moldovan language" is derived from the study mentioned above, and does not reflect the political opinions of anybody involved. Alphabetic. | *Constitution of Transnistria (Конституция Репубричий Молдовенешть Нистрене)* as of 2000 by multiple (2000)<br>4,052 words. Public domain. |
| **N-SAN** | Classical Sanskrit | **Location:** India<br>**Time:** 2nd millennium BCE to unknown<br>**Family:** Indo-European, Indo-Iranian<br>**Notes:** No native writing system. | *Eight Chapters (Aṣṭādhyāyī/अष्टाध्यायी)* by Pāṇini (पाणिनि) (4th century BCE)<br>18,588 words. Public domain.<br>**Notes:** Sample uses Devanagari script. Represents an example of a linguistic text that discusses the rules of the language it is written in. |
| **N-SUX** | Sumerian | **Location:** Southern Mesopotamia<br>**Time:** 4th to 3rd millennium BCE<br>**Family:** Isolate<br>**Notes:** Represents an example of an academically constructed notation system. Originally logographic and syllabic script. | *Enki and the World Order* by unknown (4th millennium)<br>2,767 words. Public domain.<br>**Notes:** Represents an example of a copy of a badly degraded text. |
| **N-SWA** | Swahili | **Location:** Central and East Africa<br>**Time:** 1st millennium to present<br>**Family:** Niger-Congo<br>**Notes:** Alphabetic. | *Address to Ghanaian Parliament, 2009,* by Barack Obama (2009)<br>3,266 words. Public domain.<br>**Notes:** Represents an example of a text transcribed from an oral speech. |
| **N-THA** | Thai | **Location:** Thailand<br>**Time:** Unknown to present<br>**Family:** Tai-Kadai<br>**Notes:** Has similar letter frequency distribution according to Jaśkiewicz. Thai script is an abugida with 44 consonants and 15 vowels and 4 diacritics that combine. Tonal and analytic language. Abugida. | *Defeat of the Yuan* (ลิลิตยวนพ่าย) by unknown (c. 1475)<br>3,345 words. Public domain. |

| | | | |
|---|---|---|---|
| **N-TPI** | Tok Pisin | **Location:** New Guinea<br>**Time:** c. 19th century to present<br>**Family:** Creole<br>**Notes:** Alphabetic | *Geneva Convention* by multiple, translated by Otto Dempwolff (1914) and by *Proclamation of the British Take-Over of New Guinea* by unknown (1914)<br>552 words total. Public domain. |
| **N-TUR** | Turkish (modern) | **Location:** Turkey and Northern Cyprus<br>**Time:** 19th century to present<br>**Family:** Turkic<br>**Notes:** Features heavy agglutination and vowel harmony. Alphabetic. | *Treaty of Lausanne* by multiple (1923)<br>9,699 words. Public domain.<br>**Note:** Sample uses the modern Latin script. |
| **N-TXB** | Tocharian B | **Location:** Tarim Basin (present day Xinjiang in China)<br>**Time:** 6th to 9th century<br>**Family:** Indo-European, Tocharian<br>**Notes:** Alphabetic. | Manuscript series *A, IOL Toch, M 500.1, PK AS 1-7* (only items exclusively in Tocharian B) by unknown (6th-9th century)<br>7,903 words. Public domain.<br>**Note:** Represents an example of a text copied from many disparate sources. Sample uses academic Latin transcription. |
| **N-VIE** | Vietnamese (modern) | **Location:** Vietnam<br>**Time:** 19th century to present<br>**Family:** Austro-Asiatic<br>**Notes:** Tonal language. Alphabetic. | *The Tale of Kieu* by Nguyễn Du (1820)<br>22,848 words. Public domain.<br>**Note:** Sample uses the modern Latin script. |

# 4. Legal and authorship information

The copyright status of each sample is stated in the right hand column of the table in section 3. In nearly all cases I have found texts that are in public domain, therefore there are no restrictions on those. You are free to download, share, modify, use and abuse these files as you wish, with no permission needed.

A few cases have a different copyright status:

- N-ASL (American Sign Language) and the ASLSJ notation is copyright Thomas Stone. He has given permission to use and share this sample as long as we credit him.
- N-MWP (Kalaw Lagaw Ya) is copyright Bruce E. Waters. This sample is free to use and share for non-commercial purposes.

This sample pack was selected and compiled by Brian Cham in November 2014 with help from Thomas Stone and Luman Wang. Please direct all comments, questions, suggestions and contributions to briancham1994@gmail.com.

# 5. Works cited

List uses MLA referencing convention.


Jaśkiewicz, Grzegorz. "Analysis of Letter Frequency Distribution in the Voynich Manuscript." *Proceedings of the international workshop CS&P, 28-30 September 2011, Pułtusk, Poland*. Ed. M. Szczuka et al. Warsaw, Poland: Warsaw University of Technology. Print.

Reddy, Srvana, and Knight, Kevin. "What We Know About the Voynich Manuscript." *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011, Portland, Oregon, USA*. Ed. Zervanou, Kalliopi, and Lendvai, Piroska. Portland, Oregon, USA: Association for Computational Linguistics, 2011. Print.

Stojko, John. *Letters to God's Eye: The Voynich Manuscript for the first time deciphered and translated into English*. New York: Vantage Press, 1978.

Stolfi, Jorge. "Chinese Theory Redux: Comparing the VMS and East Asian word length distributions." *Voynich Manuscript Stuff*. Stolfi, Jorge, 19 January 2002. Web. 26 February 2014. <http://www.ic.unicamp.br/~stolfi/voynich/02-01-18-chinese-redux/ >

Tucker, Arthur O., and Talbert, Rexford H. "A Preliminary Analysis of the Botany, Zoology, and Mineralogy of the Voynich Manuscript." *Herbalgram.org* 100 (2013): 70-85. *Herbalgram.org*. Web. 6 February 2014.