Review: of "The Linguistics of the Voynich Manuscript" by Claire Bowern and Luke Lindemann

Torsten Timm[a] and Andreas Schinner[b]

[a]T-Systems Multimedia Solutions GmbH, 01129 Dresden, Germany
[b]Institut für Experimentalphysik, Johannes Kepler Universität, A-4040 Linz, Austria

ABSTRACT
We address severe deficiencies of a recent publication by Claire L. Bowern and Luke Lindemann on the linguistics of the Voynich Manuscript [*Annual Review of Linguistics* 2021 (7): 285-308].

KEYWORDS
Voynich manuscript, scientific methodology

1. Introduction

During the last years, research on the Voynich Manuscript [see, e.g., 2], a (most likely) medieval codex written in unique "cipher" script, has drifted into a problematic direction. Even articles in peer-reviewed academic journals sometimes neglect the basic principles of scientific methodology. A recent publication [1] appears symptomatic for this questionable development.

There exist exactly two possibilities: either the Voynich Manuscript contains linguistically meaningful information, or not. Strictly speaking, one would have to more precisely define "meaningful" in this context, in case of, e.g., an irreversible writing/encoding system meaningful to the original scribe only. But this is a moot point, considering the fact that publications frequently present linguistic identification (or even partial translation) of the text. While it is perfectly legitimate for a researcher to weight evidence in favor of his/her theory, it still remains very bad scientific practice to completely ignore all arguments opposing one's own viewpoint.

Consequently, supporters of the so-called "hoax hypothesis" (an unfortunately chosen term, because of its unavoidable connotation with malicious intent) have to address at least the "language-like" features (like Zipf's laws) and, in particular, the question how a medieval scribe could maintain the internal statistical regularities of a gibberish text more or less consistently over so many pages. On the other hand, the text of the Voynich Manuscript undoubtedly shows statistical features unknown in any other language and/or writing system. It is not sufficient to simply put aside these facts by either just briefly mentioning (*"[Subsection] 1.1. Natural Language or Gibberish"*) — or, even worse, completely ignoring them.

Address correspondence to Torsten Timm. E-mail: torsten.timm@kereti.de

Important observations supporting a structured pseudo-text hypothesis do exist. For instance, the random walk analysis [6], indicating the presence of long-range correlations in the text, as well as the unusual connection between frequency, similarity and spatial vicinity of tokens [8], are very hard to explain on basis of natural language text, even under the assumption of an "exotic" phonetic encoding or shorthand notation.

While it is certainly still possible to focus on the linguistics approach, such explanations (at least in tentative form) *must* be part of any analysis meeting at least basic scientific standards. Bowern and Lindemann not only fail in this aspect, even worse, later on they write: *"However, [...], the gibberish account does not explain the higher-level document structure, which we discuss further in Section 4."* Actually, the essential results of our publication [8] are presentation and analysis of a concrete algorithm (the "self-citation process") that is able to produce key aspects of just the aforementioned text/token structure.

We wish to emphasize here that we *do not* criticize the fact that Bowern and Lindemann obviously see our work as completely irrelevant. No researcher is above the possibility of making fundamental mistakes, and/or is over-interpreting their results; thus we will always welcome any serious critical discussion of our viewpoint. However, we *do* criticize how the authors, without any discussion, simply summarize their viewpoint in the postulate: *"At this stage, we are not persuaded by any of the arguments in favor of the Voynich text being gibberish."*

2. SOME CRITICAL POINTS

From Bowern and Lindemann's most problematic arguments we are addressing the following here briefly:

- The generation of meaningless text without *"local repetitions and words from other languages"* is too difficult.

  The reported experiment with undergraduate students only proves that generating long truly random text sequences, i.e. sequences *without inherent concept*, provides an extremely difficult task. However, the equality "gibberish=random" proposed by Bowern and Lindemann is incorrect. Reusing already known text, modified following a set of simple rules, is by far easier and produces meaningless (but not *random*) text. This exactly is what the "self-citation algorithm" does. Even today nobody generates place holder texts for documents without content; instead the "lorem ipsum text" is commonly used.

- *"Following the initial observation of Currier [...], we assume that there are two 'languages' in the manuscript, labeled here for convenience as Voynich A and Voynich B."*

  It is well-known that not every page belongs to Currier A or B; there are also pages with statistical characteristics between Currier A and B. The connection between words of the two "languages" appears *much* more complicated, see our network analysis of similar words [8, p.4].

- *"Although there is some overlap, the most common vocabulary items of Voynich A and Voynich B are substantially different."*

  Actually, words typical for Currier A also exist in Currier B, but not the other way round. We interpret this in [8] as evidence for a "continuous evolution" from A to B, rather than the existence of two clearly distinguishable

languages/encoding schemes.

- Several important observations contradicting the idea that "Voynichese" words are structured like words in natural languages are not mentioned. For instance, "one of the most puzzling features of the VMS is its weak word order;" see [5]. Even Bowern and Lindemann only hypothetically assume the existence of a word morphology: *"If we assume that these chunks do in fact represent words, then we can investigate the morphology of Voynichese."*

  The authors also omit any discussion of the important, almost mathematically exact binomial word length distribution [7], as well as the fact that high-frequency tokens also tend to have higher numbers of similar word types [8, p.6]. Both effects are present (as inescapable by-product) in the text sample output of our "self-citation algorithm."

- *"The proportional frequencies of the most common words in linguistic texts are also useful for diagnosing linguistic structure."*

  Unfortunately, words like *'daiin'* or *'chedy'* actually do appear contextual, whereas so-called function words (like conjunctions, articles, etc.) normally are distributed equally over a text, because they serve to implement grammatical structures, and they usually do not possess co-occurring similar words of comparable frequency; see [8, p.6].

- *"As with the Zipfian word distribution, we find Voynichese to be well within the expected values for natural language texts and far from random gibberish."*

  Both of Zipf's laws definitely cannot serve as any evidence for linguistic structure, because obviously they are also fulfilled by gibberish created as output of a simple set of rules; see figures 6 and 7 in [8].

- *"Full reduplication, in which an entire word is repeated, is also common in Voynich. However, it is still within the realm of plausibility for natural language texts."*

  This is not the case. Bowern and Lindemann most likely used an inadequate parsing method for Wikipedia pages and therefore achieved questionable statistical results [see 3, p.15]. Their published text samples [4] still contain many repetitive structures that artificially increase word repetition counts, like traces of only in part deleted tables and hypertext markup elements.

3. Conclusions

In their summary Bowern and Lindemann state: *"The higher structure of the manuscript itself is completely consistent with natural language and is very unlikely to be manufactured."* A conclusion of such far-reaching implications should by no means be drawn without carefully weighting possible counter-arguments. In our own publication we have shown that the Voynich Manuscript text is "unusual" on various scales: long-range correlations on the bit-to-bit level ("random walk") are present, as well as a deep connection between string and spacial metrics, and token frequency. On the other hand, if some kind of morphological structure existed, then at least some characteristic relations between different word classes would be expected, rather than the linguistically extremely unusual token network uncovered by our analysis.

Again we emphasize that we do not feel offended in any way, because our work has been completely ignored, without even the slightest attempt of disproving (or at least providing concrete arguments against) it. Instead, we are worried about this practice, which has always been symptomatic for the non-scientific Voynich Manuscript research, but nowadays manages to manifest itself even in the world of serious aca-

demic publication. The Voynich Manuscript is an interdisciplinary research subject, which certainly makes it more difficult for the peer review system to detect deficiencies. Nevertheless, it is high time to counteract this fatal undermining of fundamental scientific principles; otherwise even the serious Voynich Manuscript research might sooner or later completely end up in epistemological disaster.

References

[1] Bowern, C. and L. Lindemann. 2021. The Linguistics of the Voynich Manuscript. *Annual Review of Linguistics* 2021 (7): 285-308. *doi:10.1146/annurev-linguistics-011619-030613*

[2] Kennedy, G. and R. Churchill. 2004. *The Voynich manuscript: The mysterious code that has defied interpretation for centuries.* London: Orion publishing Group Ltd.

[3] Lindemann, L and C. Bowern. 2020a. Character entropy in modern and historical texts: comparison metrics for an undeciphered manuscript. axXiv:2010.14697 [cs.CL] URL: `https://arxiv.org/abs/2010.14697`.

[4] Lindemann, L and C. Bowern. 2020b. URL: `https://github.com/chirila/Voynich-public/`

[5] Reddy, S. and K. Knight. 2011. *What we know about the Voynich manuscript. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 78-86. URL: `http://dl.acm.org/citation.cfm?id=2107647`.

[6] Schinner, A. 2007. The Voynich manuscript: Evidence of the Hoax hypothesis. *Cryptologia* 31 (2): 95-107. *doi:10.1080/01611190601133539*

[7] Stolfi, J. 2000. Voynich Manuscript Stuff. URL: `https://www.ic.unicamp.br/~stolfi/voynich/00-12-21-word-length-distr/`

[8] Timm, T. and A. Schinner. 2020. A possible generating algorithm of the Voynich manuscript. *Cryptologia* 44 (1): 1-19. *doi:10.1080/01611194.2019.1596999*