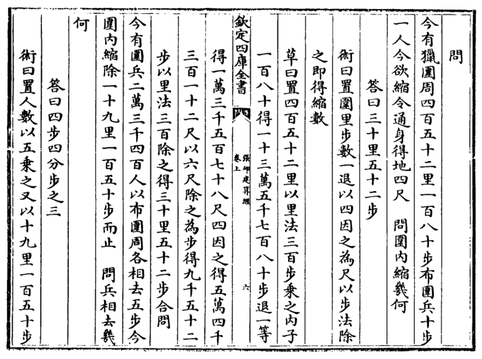
One Genre, Many Formalisms: On Formulaic Language in Pre-Modern Chinese Mathematical Texts

Florian Keßler, Diane Donner, Li Shuyi

Lehrstuhl für Sinologie mit dem Schwerpunkt Geistes- und Kulturgeschichte Chinas,
Friedrich-Alexander-Universität Erlangen-Nürnberg

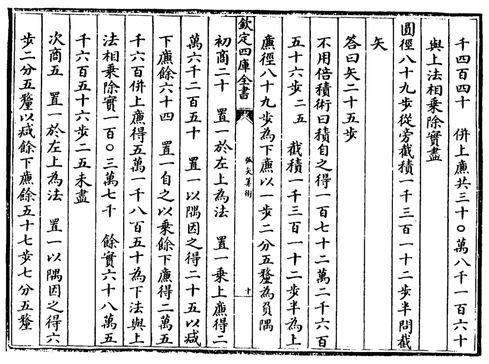
UNTERSTÜTZT VON / SUPPORTED BY
Alexander von HUMBOLDT STIFTUNG

Pre-Modern Chinese Mathematics: How formulaic was it?


欽定四庫全書
問
今有獵圍周四百五十二里一百八十步布圍兵十步一人今欲縮令通身得地四尺　問圍內縮幾何
答曰三十里五十二步
術曰置圍里步數一退以四因之為尺以步法除之即得縮數
草曰置四百五十二里以里法三百步乘之內子一百八十得一十三萬五千七百八十步退二等得一萬三千五百七十八尺以六尺除之為步得九千五十二步以里法三百除之得三十里五十二步合問
今有圍兵二萬三千四百人以布圍周各相去五步今圍內縮除一十九里一百五十步而止　問兵相去幾何
答曰四步四分步之三
術曰置人數以五乘之又以十九里一百五十步

Page from *Siku quanshu* edition of *Zhang Qiujian*

For some technical genres such as mathematics it is often assumed that their language exhibits **formulaic characteristics**. However, it is notoriously difficult to identify what properties exactly make language "formulaic". In order to nevertheless quantitatively assess the formulaicity of a corpus, Richard Forsyth has proposed a suite of computational tools titled formulib, which operates by "compiling a **formulexicon** from a corpus [...] and then us[es] **coverage** by elements of that formulexicon as an index of the degree to which a text [...] is pervaded by formulaic sequences" (Forsyth, 2021: 33, emphasis ours). In this study, we apply this suite to **pre-modern Chinese mathematical texts** (1st century CE to 18th century). These texts are particularly interesting because they were written **without modern symbolic notation systems**, which could suggest more natural, less formulaic patterns of language use.


千四百四十　併上廣共三十○萬八千一百六十步二分五釐以減餘下廣餘五十七步七分五釐
與上法相乘除實盡
圓徑八十九步從旁截積一十三百一十二步半問截
欽定四庫全書
矢
答曰矢二十五步
不用倍積術曰積自之得一百七十二萬二千六百五十六步二五　裁積一千三百一十二步半為上廣徑八十九步為下廣以一步二分五釐為員隅
初商二十　置一於左上為法　置一乘上廣得二萬六千二百五十　置一以隅因之得二十五以減下廣餘六十四　置一自之以乘餘下廣得二萬五千六百併上廣得五萬一千八百五十為下法與上法相乘除實一百○三萬七十　餘實六十八萬五千六百五十六步二五未盡
次商五　置一於左上為法　置一以隅因之得六步二分五釐以減餘下廣餘五十七步七分五釐

Page from *Siku quanshu* edition of *Hushi suanshu*

Methodology

CORPUS
- Use the "Masters Division" (*zibu* 子部) of the *Complete Library of the Four Treasuries* (*Siku quanshu* 四庫全書, compiled 1772-1782).
- Contains different genres: philosophy, arts, and different sciences (including mathematics), ...

PREPROCESSING
- Remove pre- and postface chapters
- Replace counting rods and numerals with placeholders
- Sampling to obtain balanced corpus, repeat ten times for validation
- After sampling: 11 subcategories with 430 110 tokens each (one Chinese character is one token)

EXTRACT N-GRAMS
- For each genre, compute *formulexicon*: 80 most frequent n-grams, $3 \leq n \leq 8$
- Exclude n-grams made up only of stopwords (= 30 most frequent tokens)

COMPUTE COVERAGE
- Coverage is the number of tokens of the text that occur in the text as a part of at least one of the n-grams in the *formulexicon*, divided by the total number of tokens

Example: Procedure from *Zhang Qiujian's Mathematical Classic*
(written ca. 450 CE)

置　南　北　壁　高　併之得　N　半　之得　N　尺 N　寸
place south north wall height add it obtain NUMERAL halve it obtain NUMERAL *chi* NUMERAL *cun*
"Place the heights of the southern and norther walls. Add them, obtaining a. Halve that, obtaining b *chi* and c *cun*."

又　置　長　N　尺 以　廣　N　尺 因　之得　N　尺
furthermore place length NUMERAL *chi* with width NUMERAL *chi* single.digit.multiply it obtain NUMERAL *chi*
"Furthermore, place the length, d *chi*. Multiply that with the width, e *chi*, obtaining f *chi*."

又　以　高　N　尺 N　寸　乘　之得　N　尺
furthermore with height NUMERAL *chi* NUMERAL *cun* multiply it obtain NUMERAL foot
"Furthermore, multiply it with the height, b *chi* and c *cun*, obtaining i *chi*."

以　斛 法　N　尺 N　寸 N　分　除　之得　N　斛 餘　N
with *hu* divisor NUMERAL *chi* NUMERAL *cun* NUMERAL part divide it obtain NUMERAL *hu* remain NUMERAL
"Divide it by the divisor for *hu*, j *chi*, k *cun* and l tenth-*cun*, obtaining m with a remainder of n."

Explanation

All placeholders a, b, c, ... are specific numerals in the original

Relevant n-grams in the formulexicon:

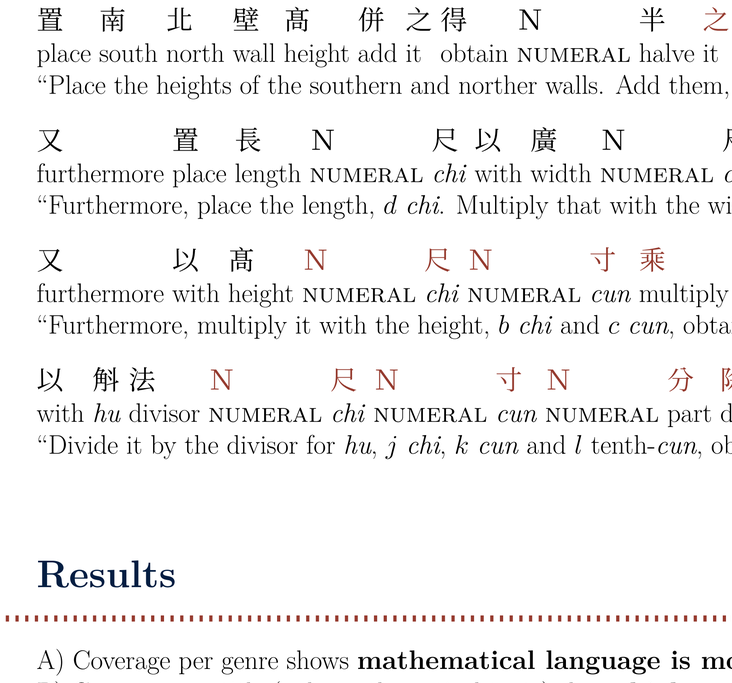
- Operations: multiplication by single digit ("multiply it, obtaining" 因之得, "multiply it, obtaining n" 因之得N), regular multiplication ("multiply it, obtaining" 乘之得, "multiply it, obtaining n" 乘之得N), division ("divide it, obtaining" 除之得, "divide it, obtaining n 除之得N)
- Quantities: "n *chi* n" N尺N, "n *chi* n *cun*" N尺N寸, "n *chi* n *cun* n" N尺N寸N, "n *chi* n *cun* n parts" N尺N寸N分
- Other: "it, obtaining n *chi*" 之得N尺, "obtaining n *chi*" 得N尺

Coverage calculation: all 30 tokens marked in color are in at least one of the n-grams in the formulexicon, so coverage $= \frac{30}{58} \approx 52\%$

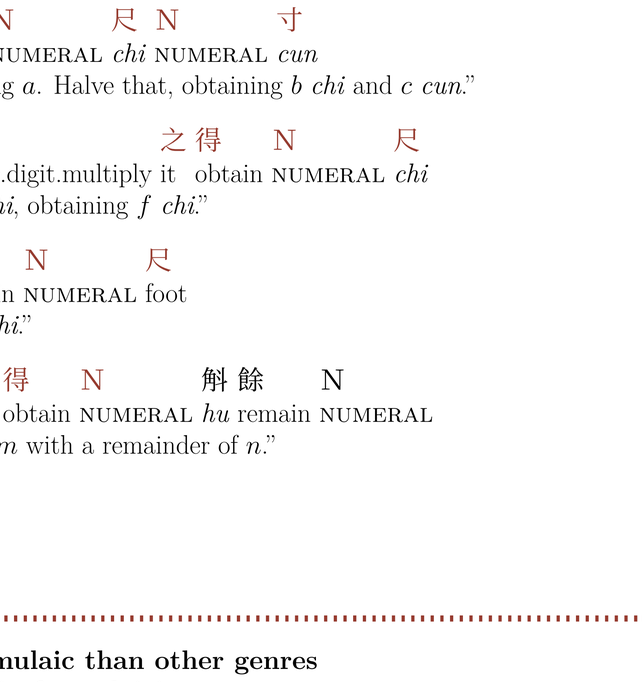
Results

A) Coverage per genre shows **mathematical language is more formulaic than other genres**
B) Coverage per title (only mathematical texts) shows **high variation in formulaicity**
C) Document frequency of n-grams in formulexicon shows **low number of formulaic sequences shared by many works**

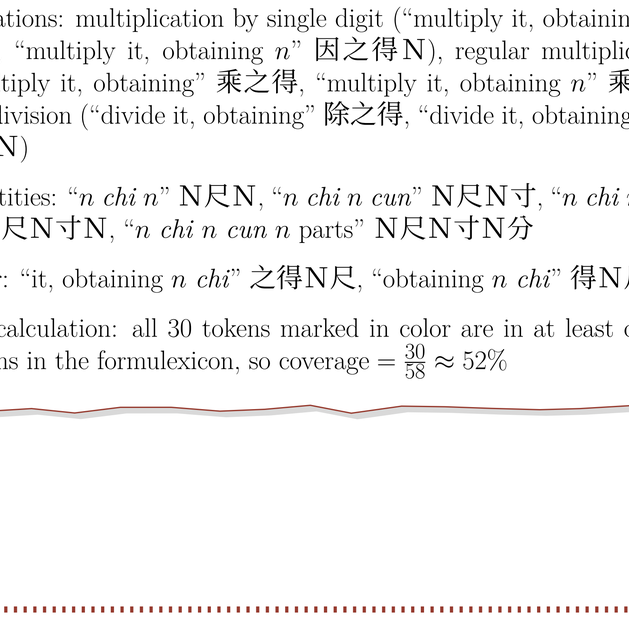
A



B



C



Key References

Forsyth, Richard. 2021. "Cascading collocations: Collocades as correlates of formulaic language". In *Formulaic language - Theories and Methods*, herausgegeben von Aleksandar Trklja und Łukasz Grabowski. Berlin: Language Science Press.
Wittern, Christian. 2016. „Kanseki Repository". CIEAS Research Report 2015 Special issue: 1–80.

Project website:

GitHub:

LaTeX TikZposter