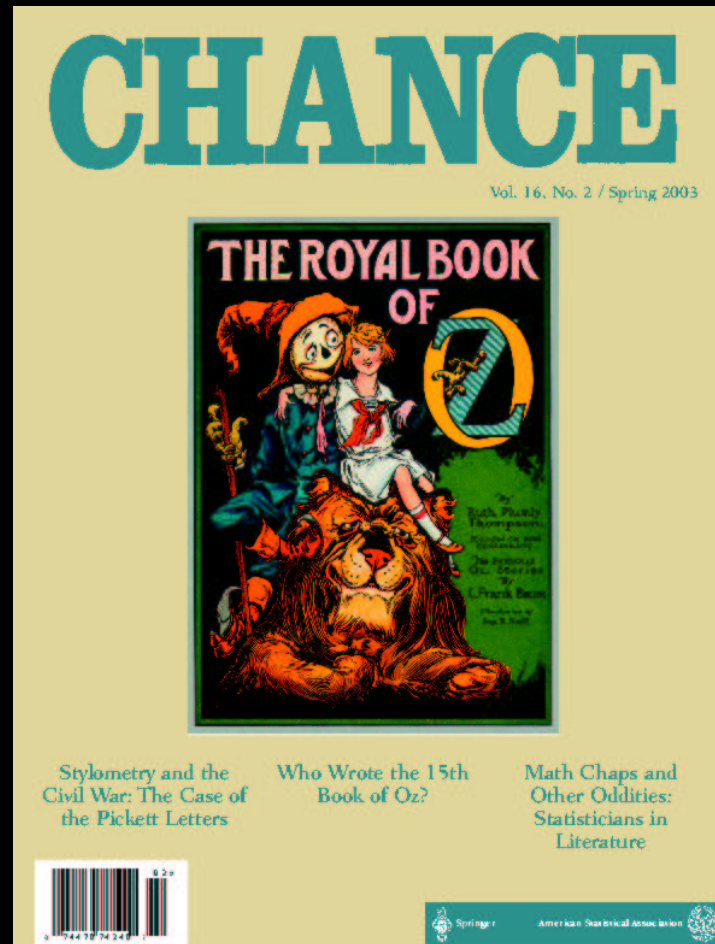




By: Ol'ga F.



'Who wrote the 15th book of Oz?' by José Binongo



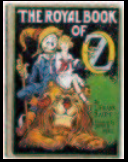
What is stylometry?...

Stylometry is the science of quantifying style. 2 sides:

- descriptive stylistics
- authorship attribution (i.e. classifying texts by author)
 - *literary texts (e.g. Oz books!)*
 - *technical writing (collaborative writing)*
 - *court evidence (forensic linguistics)*

Related classification problems:

- *by genre (including fiction vs. non)*
- *by gender (Afra!)*
- *by epoch (Victorian vs. contemporary)*




Literary background...

There are more than 40 Oz books. Most of them were written by:



- Baum books 1-14

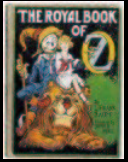
- Thompson  books 16-34

Mystery: the 15th book of Oz - 'The Royal Book of Oz'



Data

- Training Data – Oz Books
 - 14 books by Baum (Project Gutenberg)
 - 7 books by Thompson (Project Gutenberg)
 - 7 more by Thompson (scanned, proofread)
 - Total: just under 1200K words
- Validation Data
 - Baum's non-Oz works: 5 novels, 3 short stories
 - Thompson's non-Oz works: 1 short story
 - Martin Gardner's 'Visitors from Oz'
- Mystery Data: 15th book (Pr. Gutenberg) - 42K



Data Preparation

The texts were altered in the following ways:

- misspellings aimed to mimic an accent were corrected (removed if too obscure)
- portions not written in prose were removed
- contractions were expanded (ambiguous cases such as *somebody's* and *he'd* were ignored)

The average rate of occurrence of each word was then calculated (all stories with equal weight, regardless of length).



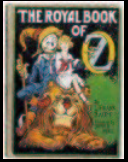
Measurements

50 most popular function words were retained.

Why function words?

- present in even of the most basic sentences
- belong to a closed class of words
- little semantic meaning, so least dependent on context
- except for auxiliary verbs and pronouns, not inflected
- not easily consciously affected

Why 50? To exclude the content-heavier ones e.g. *below*.



Measurements II

Purposely excluded, mainly because of inflections:

- auxiliary verbs
- personal pronouns

Also note that:

- no disambiguation (e.g. *to*: infinitive vs. preposition)
- contractions were expanded *after* the 50 function words were selected
- upon expansion of contractions, the corpus gained 0.5 percent of its original length



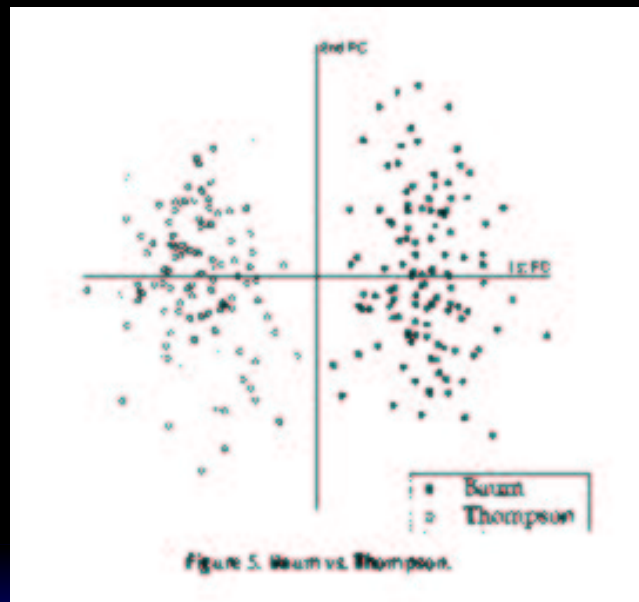
PCA (Dimensional reduction)

- *Goal* : explain as much variation as possible with as few variables as possible
- *Method* : transformation from original variables to a set of uncorrelated variables: $N\text{-}D \mapsto M\text{-}D$ where $m \ll n$
- *Why* :
 - summarize the information
 - discard useless variables
 - visual representation if $m \leq 2$
- *How* : PCs are projections of the data samples onto the eigenvectors of $\langle DD' \rangle$ with the highest eigenvalues



PCA II

- Texts \rightarrow 223 blocks of 5000 words each
- Every block \rightarrow 50 – D vector corresponding to the average occurrence rate of the 50 function words
- The variables were standardized s.t. the variance of each was reduced to unity
- In this case, $50D \mapsto 2D$, i.e. best-fitting plane



1st PC (20 percent of variation), 2nd PC (7 percent of variation)



What does that say about their styles?

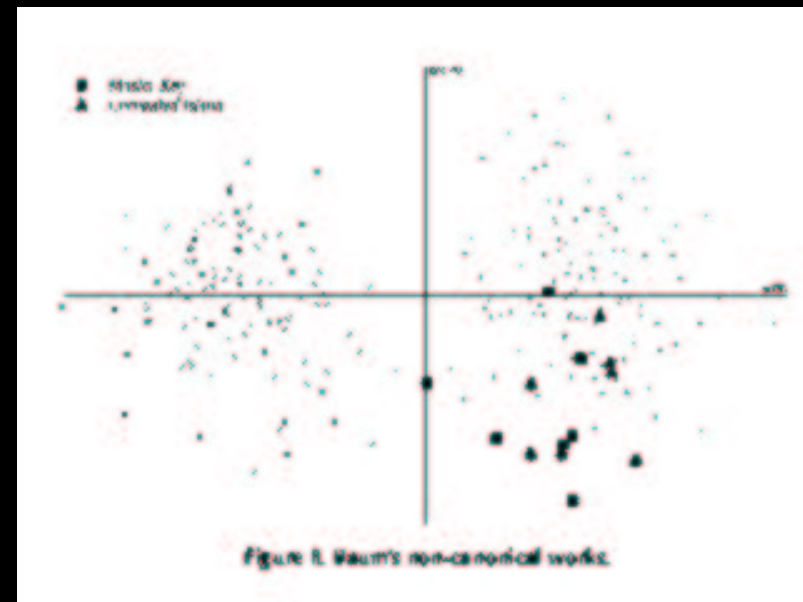
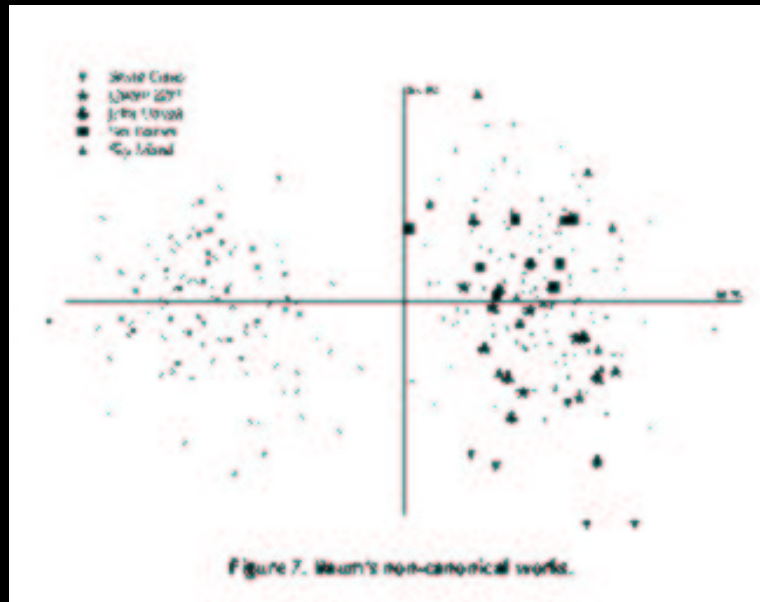
Actual difference in writing style:

- Baum uses more negative words (*but, not, no*) and *that, which*
- Thompson uses more positional words (*up, down, on, over, out, back*) - about twice as often as Baum!



Validation

Q: is good separation unique to the Oz books?
Baum's non-Oz books: Oz-related and not



Still a good separation!



Validation II

When applied to short stories, the separation wasn't quite as good.

- different genre
- improved by taking blocks of 10K words (instead of 5K)

What about a third author? (Gardner)

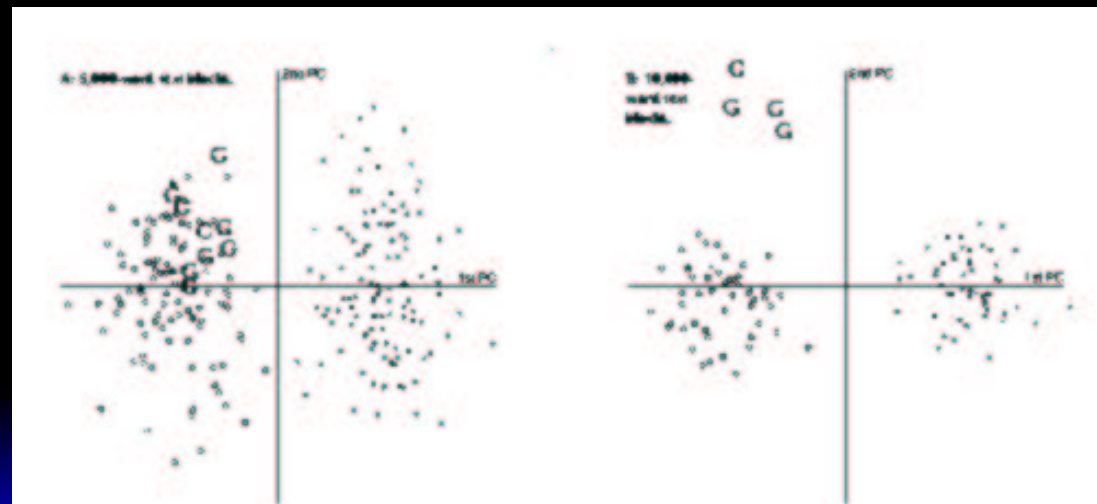


Figure 11. Gardner's Villains from Oz (1998).

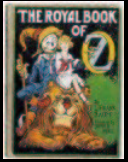


Solving the Mystery



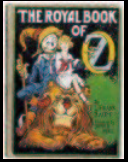
Figure 10. The Royal Book of Oz (1921).

So: 'The Royal Book of Oz' was written by *Thompson*, and 'Glinda' was indeed written by Baum (edited by his son).



So what?...

- This method allows to define an author's style not absolutely but relative to that of another author.
- Is style really what's captured? Seems so, b/c:
 - not gender
 - not time
 - not genre
 - even the characters are largely the same
- This method requires a lot of training data due to the evasive nature of style (how much data is sufficient? what does it depend on?)



So what?... II

- Incalculable degree of uncertainty about the results
- PCA doesn't depend on unverifiable statistical assumptions, unlike other methods (e.g. ?)
- Not guaranteed to work on other texts
- May help distinguish genre (e.g. Wilde's plays vs. essays)



The End

C'est tout! Thank you.