

Feature Patterns in the Voynich Manuscript

Voynich Day — **4 08** 2024
Emma May Smith

What are Feature Patterns?

- Feature patterns are cases where the features of one word co-occur with a feature in another word.
- Features might be individual glyphs or a group of glyphs, maybe combined with their position in a word. Or it could be a whole word.
- A feature might not even be in a word, but could be a linebreak or image influencing the texts adjacent to it.

Example of Feature Patterns

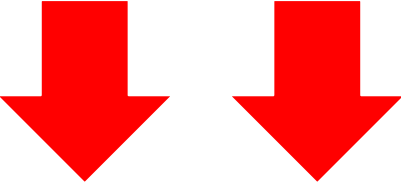
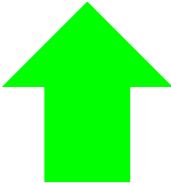
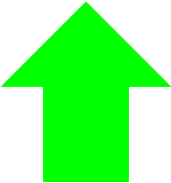
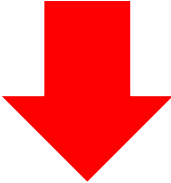
- Smith and Ponzi (2019) built on observations from Prescott Currier (1976) that words ending with certain glyphs occur more frequently immediately before words beginning with certain glyphs:

✓ oted**y.q**okaiin

✗ otai**n.k**edy

- A comparison of actual versus expected rates for these patterns allows for a division of the glyphset into two groups:
 - Strong: [q, k, t, d, r, s, n, m]
 - Weak: [o, a, y, ch, sh]
 - Ambiguous: [l]

Wordbreak Combinations Continued

		Next Word Starts	
		Strong	Weak
First Word Ends	Strong		
	Weak		

What Features to Use in This Study?

- Single glyphs and glyph combinations.
- Anywhere in a word, but also at the start and end of words.
 - How are words defined?
 - How are spaces defined?

Defining Adjacency

- Adjacency — words which occur either side of the same space.
- Linebreaks **don't** break adjacency.
- Images **don't** break adjacency.
- Paragraphs **do** break adjacency.

What Text is used?

- Only running text to maximize adjacency.
- Labels excluded as have little or no adjacency.
- Text broken down by Currier language and the hands of Davis (2020).
- Tests run on different parts: the following data based on Hand 2 Herbal, Hand 2 Quire 13, and Hand 3 Quire 20. All Language B.

How are Feature Patterns measured?

- The key thing is to know how **unusual** the relationship between two features is.
- **Z-scores** allow us to measure distance from the standard deviation as a multiple, positive or negative.
- Will work through a single example to show the calculation.

Our Example Pattern

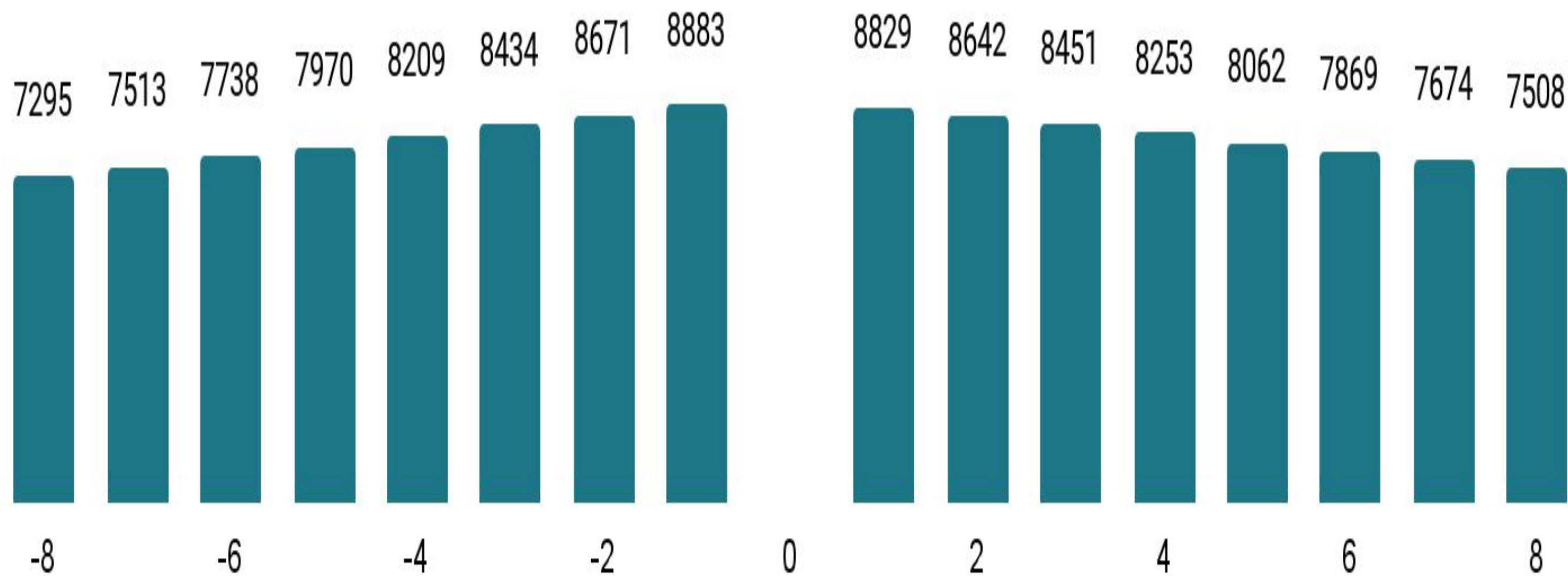
- Our example pattern:

y.q

- A word ending [y] followed immediately by a word starting with [q].
- Definitions:
 - Feature 1 is [y.]
 - Feature 2 is [.q]

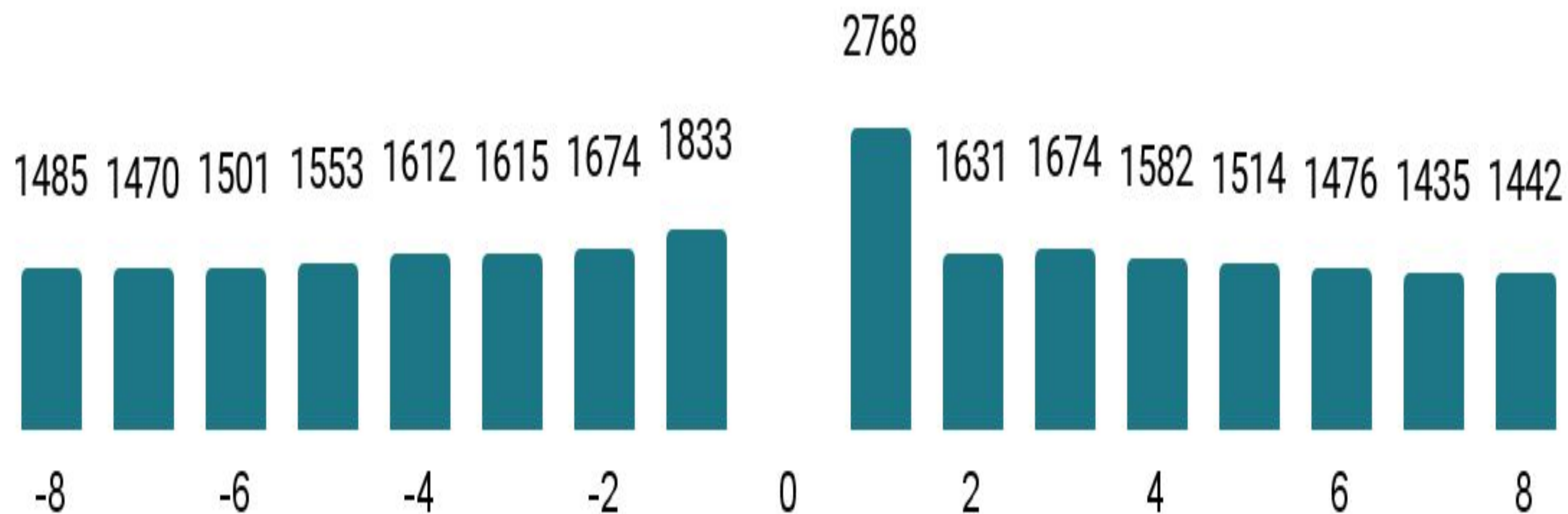
Counting the Distribution: Feature 1

- For each instance of Feature 1 [y.] count the number of times **another word** occurs eight words before or after.



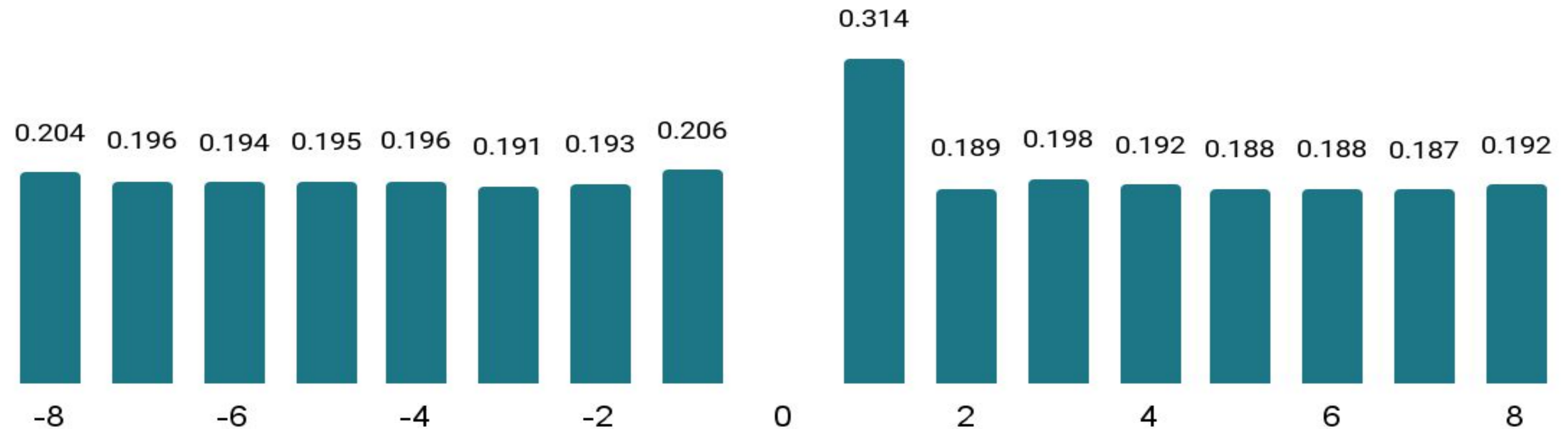
Counting the Distribution: Feature 2

- And count whenever any of the eight words either side of Feature 1 contain Feature 2 [.q].



Calculating the Likelihood

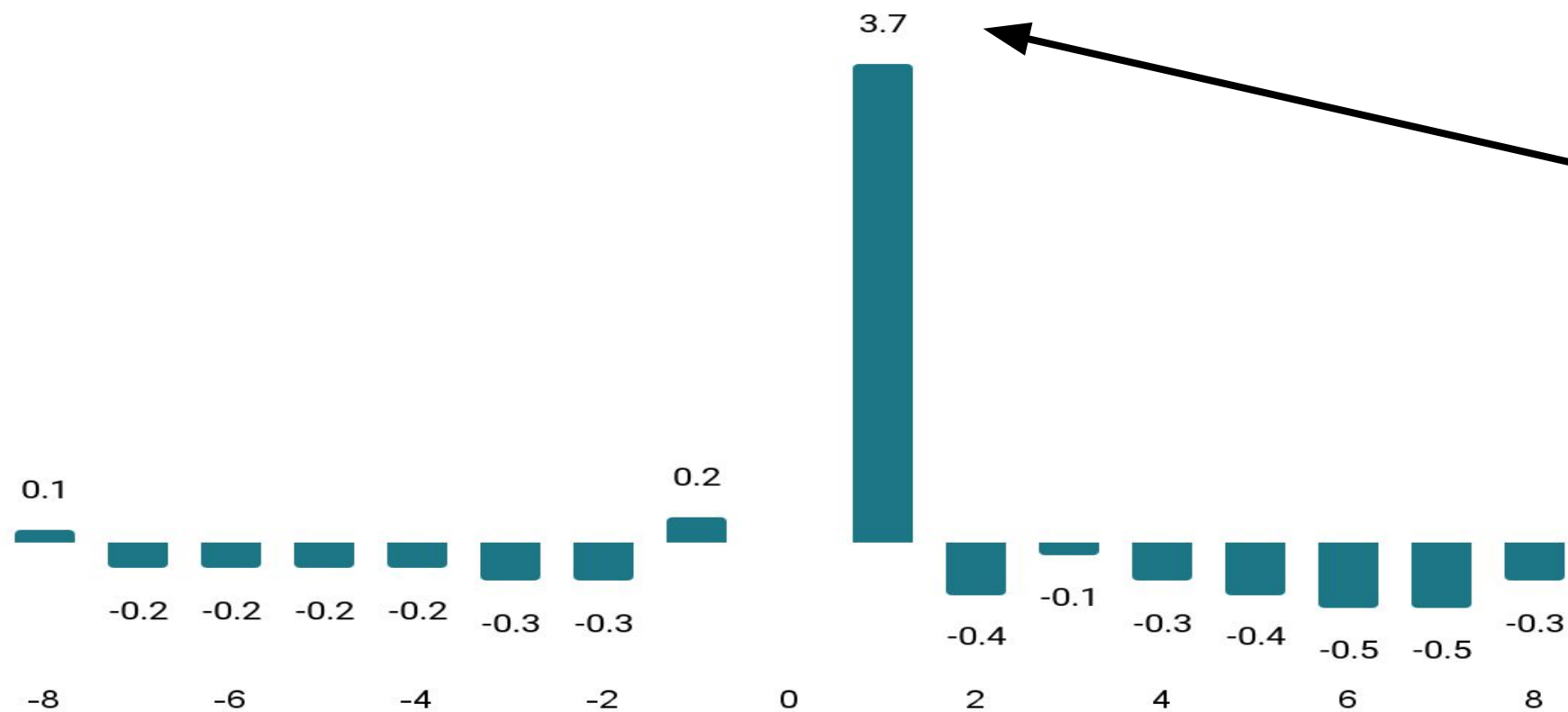
- For each position, divide number of Feature 2 by number of filled positions to obtain likelihood.



- Standard deviation calculated to be **0.0304**

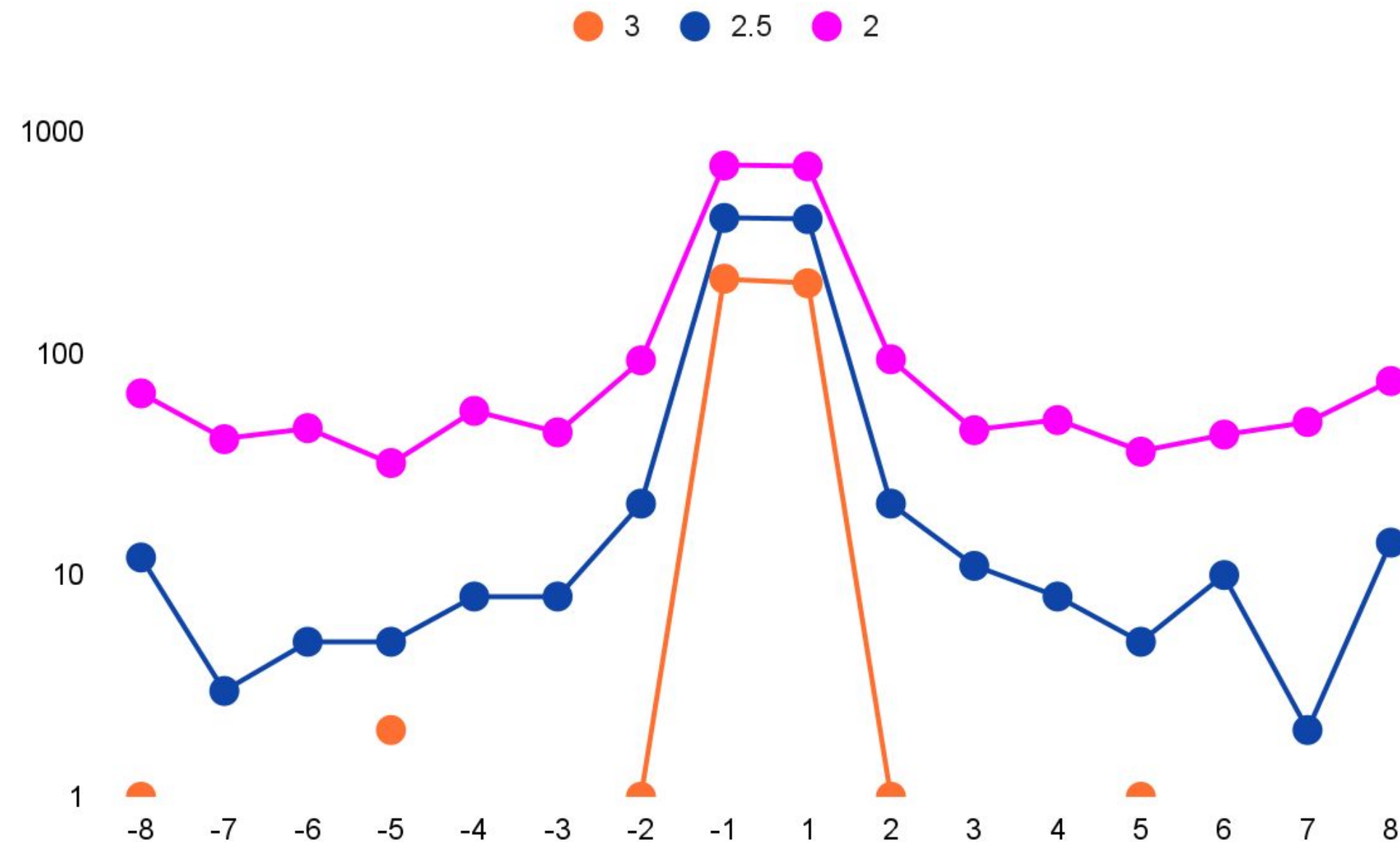
Calculating Z-scores

- Z-score for each position is calculated by:
$$(\text{Position Likelihood} - \text{Average Likelihood}) / \text{Standard Deviation}$$



Why is Adjacency Interesting?

- Adjacent positions show the most patterns at any Z-score, and account for almost all over ± 3.0 .



Let's Look at Some Results!

- Only Z-scores of ± 3.0 or greater!
- Four patterns based on position of features:

Feature 1 and **Feature 2**:

End-Start: .qokeedy**y**.**d**aiin.

Start-End: **.q**okeedy.daii**n**.

End-End: .qokeedy**y**.daii**n**.

Start-Start: **.q**okeedy.**d**aiin.

- Also Anywhere-Anywhere patterns.

Results for End–Start

- The same patterns found in Smith and Ponzi (2019):
 - The same division of glyphs was evidenced:
 - Strong: [q, k, t, d, r, s, n, m]
 - Weak: [o, a, y, ch, sh]
- The glyph [l] is strong at the start of a word, but at the end:
 - Disprefers [a] and [q].
 - Prefers [k, d, ch, sh]

New Unexpected Patterns: 1

- [dy.ka] is dispreferred at -3.0.
- Patterns with lower Z-scores suggest *any* [dy.k] is dispreferred.
- However: [ey.ka] at 3.2
- Evidence that [dy] has its own identity?

New Unexpected Patterns: 2

- [in.ai] is dispreferred at -3.0.
- However, this combination should be preferred.
- Most words starting [ai] will end with [n]. But [n.*n.] is a *weaker* dispreference pattern.
- Is this about [*aiin.aiin]?

Other Adjacency Patterns

- Nothing for Start–End above ± 3.0 .
- Most patterns for End–End and Start–Start are simply “echoes” of End–Start patterns.
 - [**in.*ckhy.*] at 3.6, [**in.*cthy.*] at 3.2, but many of the second words are complete.
 - [*.a*.a*] and [*.a*.o*] are common, but many words starting [a] end with [l], [r], or [n], which have a preference for [a] or [o], or both.
 - A similar explanation for words beginning [da] or [sa] being less common before words beginning [qo].

End–End: 1

- The End–End pattern [*l.*dy.] has a Z–score of 3.0.
- Remember from End–Start we saw that [l.] disprefers [a] and [q], but prefers [k, d, ch, sh].
- Is there a connection between these patterns?
- Certainly words starting [a] are unlikely to end [dy]. But what else?

End–End: 2

- Two End–End patterns with [s]:

[*s.*in.] at 3.2

[*s.*y.] at -3.2

- Though [*s.*iin.] has a score of 3.3, but [*s.*ain] shows no preference.
- The [*s.*y.] is mostly from [dy], but that specific pattern is at -2.7.

A Start–Start Pattern?

- The Start–Start patterns found after [sh]:
 - [.qo] at 3.7
 - [.qok] at 3.6
 - [.qot] at 3.3
- Many words beginning [sh] end with [y], which would prefer [q].
- However, these patterns with words beginning [ch] are *much* weaker, despite a similar percentage of those ending [y].

Anywhere–Anywhere Patterns

- Patterns also occur between glyphs *anywhere* in adjacent words.
- Many could be echoes of End–Start patterns due to word structure rigidity.
- The following are a selection of the more interesting.

Anywhere–Anywhere: 1

- Two patterns with [d] and gallows [k/t]:

[d ke] 3.0

[d te] 3.0

- Compare with the gallows followed by [ch]:

[d kch] 2.6

[d tch] 0.3

Anywhere–Anywhere: 2

- We can assume these are actually End–Anywhere patterns:

[n ke] -3.5

[n lch] -3.3

- Specifically [ke] in the first pattern: [n k] is -1.7. Nor does [n te] have a high Z–score.
- Likewise for the second pattern: [n] with [l], [lsh], and [lk] are weaker.

Anywhere–Anywhere: 3

- A number of patterns with [k]:
 - [e k] 3.6
 - [ee lke] 3.2
 - [che k] 3.2
 - [ke k] 3.2
 - [lche k] 3.1
- All of the first feature contain [e].
- Same patterns not seen with [t]!

What have We Learnt?

- A handful of additional adjacency patterns were identified.
- More research is needed.
- Z-scores are an interesting way of seeing relationships at a glance, but not perfect.

The End

Thanks for listening.

Happy to take questions!