

# Método SPI (Stylistic Profile Index)

## Definición conceptual

El **Stylistic Profile Index (SPI)**, tal como se implementa en este script siguiendo el enfoque de Tuccinardi, es una medida de similitud diseñada para comparar dos perfiles estilísticos representados como distribuciones normalizadas de  $n$ -gramas de caracteres.

Cada texto se modeliza como una distribución de micro-rasgos (secuencias de caracteres), y el SPI cuantifica el grado de proximidad entre dichas distribuciones.

El método no compara directamente palabras, significados o estructuras sintácticas. En cambio, compara el comportamiento estadístico de patrones recurrentes de caracteres.

---

## Qué mide el SPI

Tras aplicar la normalización L1, cada texto se convierte en una distribución proporcional:

- Cada  $n$ -grama tiene un peso relativo.
- La suma total de los pesos es igual a 1.
- La longitud del texto deja de influir en la comparación.

El SPI mide el grado de solapamiento entre dos distribuciones.

De forma intuitiva:

- Si dos textos distribuyen sus hábitos estilísticos de manera muy similar, el SPI será alto.
- Si sus distribuciones divergen de forma notable, el SPI será más bajo.

El método se centra en la similitud proporcional, no en diferencias absolutas de frecuencia.

---

## Cómo funciona la comparación (explicación intuitiva)

Para cada  $n$ -grama compartido entre dos textos:

1. El script compara sus frecuencias normalizadas.
2. Evalúa cuán próximas son esas proporciones.
3. Agrega esas comparaciones locales en una puntuación global de similitud.

Por tanto, el SPI:

- Recompensa la similitud en el uso proporcional de los rasgos.
- Penaliza distribuciones desequilibradas.
- Es sensible a diferencias en el equilibrio estilístico.

Es importante destacar que los  $n$ -gramas raros no dominan el resultado, ya que la normalización impide que las frecuencias absolutas introduzcan sesgos.

---

## Por qué el SPI es adecuado para la estilometría

El SPI resulta especialmente apropiado para estudios de autoría porque:

1. Compara distribuciones estilísticas y no contenido léxico.
2. Reduce la influencia temática.
3. Captura hábitos microestructurales inconscientes.
4. Funciona bien con corpus fragmentarios.

Dado que los  $n$ -gramas de caracteres reflejan tendencias estilísticas de bajo nivel, el SPI mide eficazmente lo que puede denominarse la “firma micro-estructural” de un autor.

---

## Construcción de la línea base autoral

En este script, el SPI se utiliza en dos fases:

1. Se calculan las similitudes entre todos los textos ciertos.
  - Esto genera una distribución interna de similitud autoral.
2. Se compara el texto dudoso con el corpus cierto.
  - Su valor de similitud se evalúa respecto a la línea base interna.

Este paso es fundamental:

El método no se limita a preguntar si el texto dudoso es “similar”, sino si se comporta estadísticamente como lo hacen los textos indiscutidos del autor.

---

## Sensibilidad y limitaciones

### Sensibilidad

El SPI es sensible a:

- Hábitos estilísticos consistentes.
- Cambios de registro.
- Diferencias de género.
- Variaciones editoriales o de normalización textual.

### Limitaciones

No evalúa directamente:

- Coherencia semántica.
- Estructura retórica.
- Evolución histórica de la lengua.

Por ello, cualquier divergencia estadística debe interpretarse siempre en diálogo con el análisis filológico.

---

### **Implicación interpretativa**

Un valor alto de SPI indica que:

El texto dudoso distribuye sus micro-rasgos estilísticos de manera comparable a las obras conocidas del autor.

Un valor bajo indica que:

El texto dudoso se aparta del comportamiento distributivo típico del autor.

El análisis posterior mediante z-score determina si esa desviación es estadísticamente significativa o si entra dentro de la variabilidad esperable del autor.

## **Interpretación descriptiva**

**(antes de la normalización)**

Dos distribuciones:

1. Similitudes fragmentos–perfil Ciertos (autor consigo mismo)
2. Similitudes fragmentos–perfil Dudoso (autor con obra dudosa)

Mira:

- Media 1
- Desviación estándar 1
- Media 2
- Desviación estándar 2
- Diferencia absoluta entre medias

### **Qué significa cada cosa**

#### **Media (autor vs autor) Media 1**

Es la **cohesión estilística interna** de los Ciertos.

- Alta media + baja desviación → estilo homogéneo.
- Media más baja + mayor dispersión → obra más heterogénea.

#### **Media (autor vs obra “dudosa”) Media 2**

Mide cuánto se parece la otra obra al perfil base.

- Si está muy cerca → continuidad estilística.
- Si cae claramente por debajo → divergencia.

## Diferencia entre medias

Es la primera medida de distancia estilística entre géneros.

Pero aún no sabemos si esa diferencia es grande o pequeña en términos estadísticos.

## Paso 2: Normalización z-score

Las similitudes dependen de:

- tamaño del n-grama,
- tamaño del corpus,
- riqueza léxica,
- etc.

Por eso no basta con decir:

“la diferencia es 0.05”

Necesitamos saber:

¿esa diferencia es grande respecto a la variabilidad interna del autor?

Ahí entra el z-score.

## Paso 3: Interpretación del z-score

La formula de z-score es:

$$z = \frac{\bar{x}_{dudoso} - \mu_{autor}}{\sigma_{autor}}$$

Donde:

- $\mu_{autor}$  = media fragmentos–autor
- $\sigma_{autor}$  = desviación estándar interna
- $\bar{x}_{dudoso}$  = media fragmentos–texto “dudoso”

## Cómo leer el resultado

✓ Caso A:  $|z| < 1.96$  (no significativo al 95%)

Interpretación generalizable:

La obra “dudosa” cae dentro del rango normal de variación estilística del autor.

Conclusión metodológica:

- No rompe el perfil estilístico.
- Existe continuidad autoral detectable.
- La diferencia es compatible con variación intra-autor.

---

### ⚠ **Caso B: $|z|$ ligeramente $> 1.96$ (2–2.5)**

Interpretación:

Existe una diferencia estadísticamente significativa, pero moderada.

Lectura fina:

- Existe variación detectable.
- Pero la distancia no es extrema (p.ej.  $|z| \approx 2.1$ –2.4), no implica ruptura radical.
- Puede tratarse de especialización retórica, no cambio de identidad estilística.

Aquí conviene comparar con:

- otro autor del mismo género,
- o con una obra claramente externa.

---

### ! **Caso C: $|z| > 3$**

Interpretación:

El texto cae claramente fuera del espacio estilístico generado por la obra base.

En ese caso:

- El texto dudoso, está afectando de forma estructural el perfil.
- O bien las dos obras tienen diferencias profundas.
- O bien hay heterogeneidad editorial/transmisión/autor.

Aquí conviene revisar:

- limpieza,
- equilibrio de corpus,
- tamaño relativo.

## Ejemplo: Quintiliano

-----  
Autor conocido: quintiliano

Texto dudoso: declamationes

N-gramas: 3

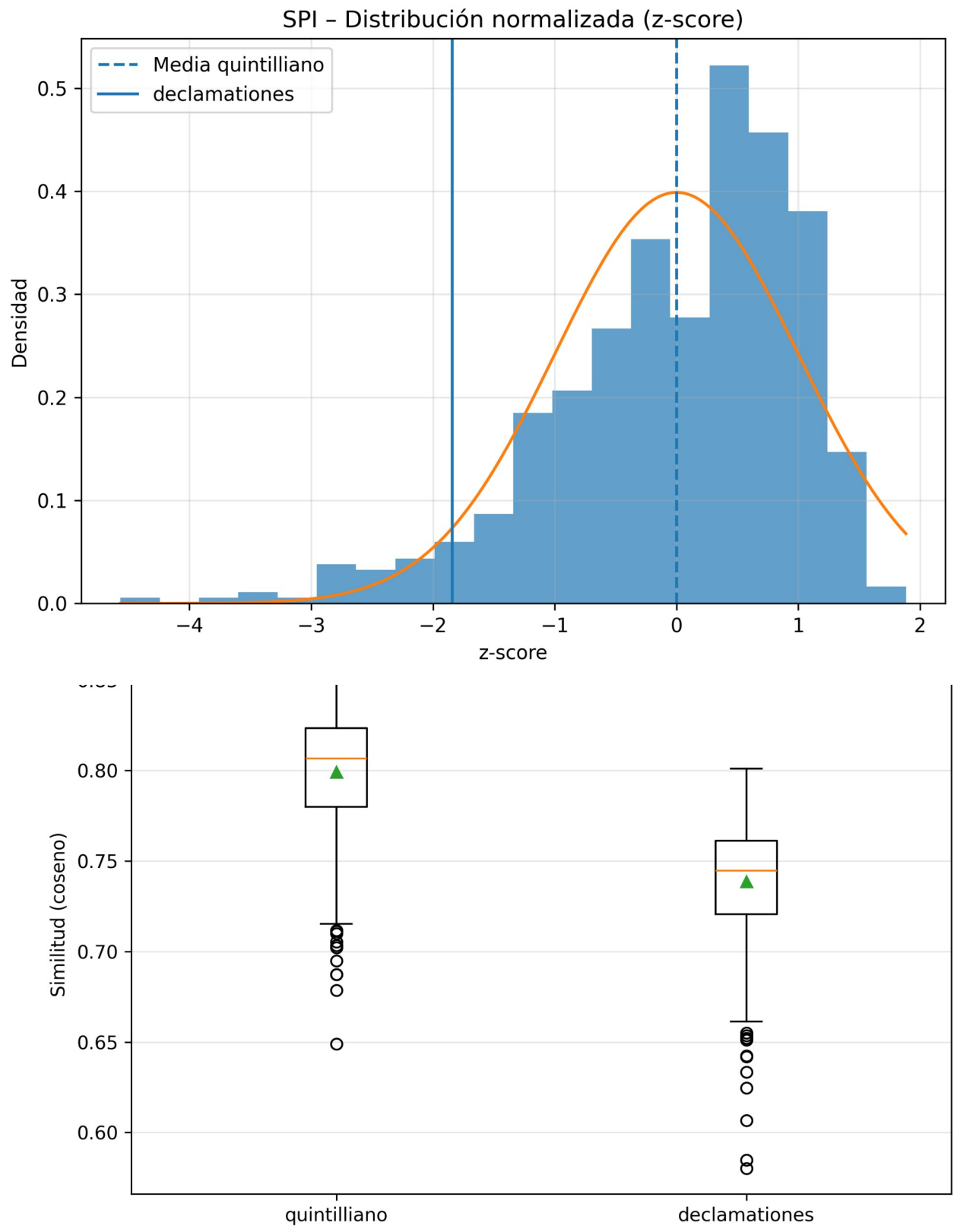
Top-k: 500

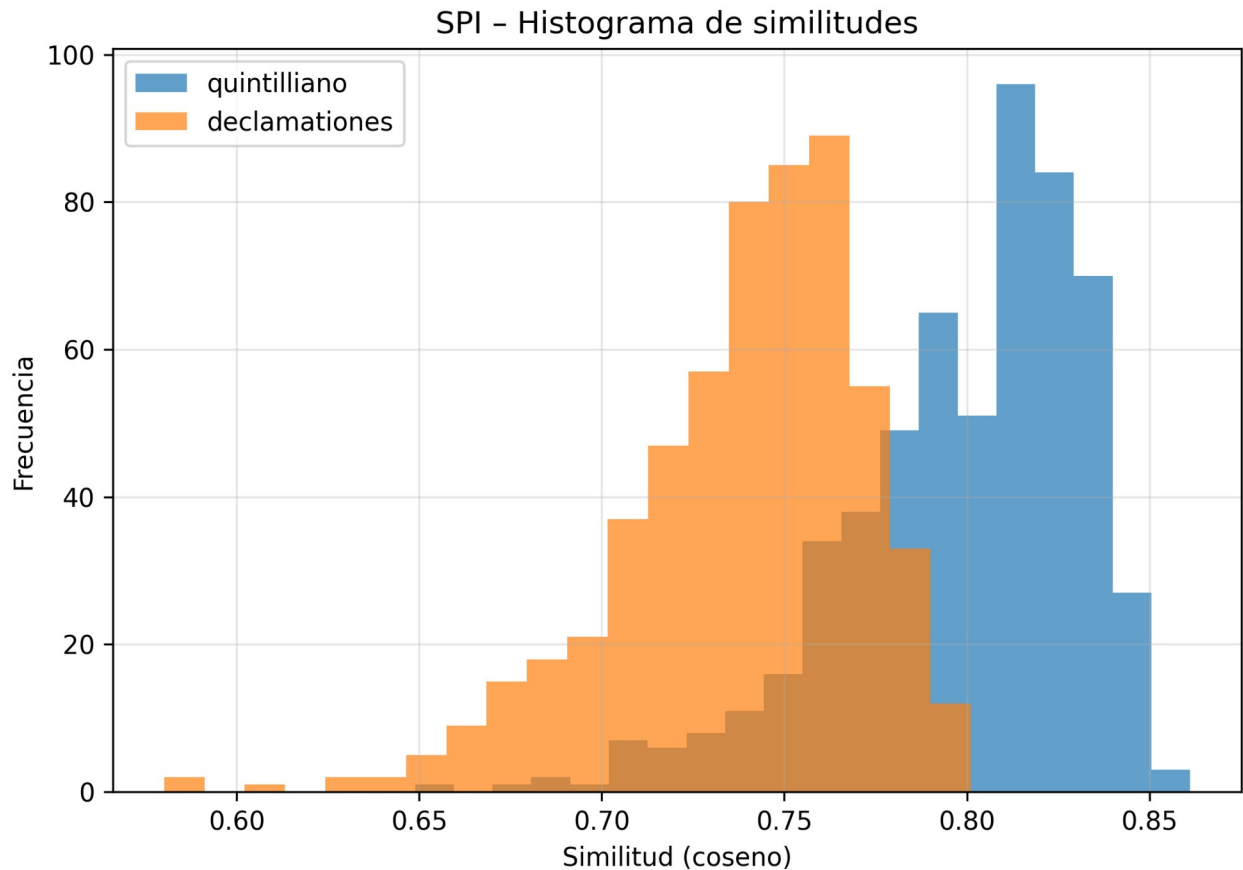
Tamaño de fragmento: 2000

RESULTADOS DESCRIPTIVOS

quintiliano\_media: 0.7992

quintilliano\_std: 0.0329  
declamaciones\_media: 0.7385  
declamaciones\_std: 0.0335  
Diferencia absoluta entre medias: 0.0606





### Interpretación descriptiva básica

#### ◆ Cohesión interna de las *Institutiones*

- Media alta: **0.8241**
- Desviación baja: **0.0297**

Esto indica:

- Fuerte coherencia estilística interna.
- Perfil muy estable.
- Baja dispersión entre fragmentos.

En términos estilométricos:

las *Institutiones* generan un **espacio estilístico compacto y bien definido**.

### Comparación con las *Declamationes*

Media: **0.7615**

Diferencia absoluta:

$$0.8241 - 0.7615 = 0.0626$$

Es decir:

Las *Declamationes* se sitúan aproximadamente **0.063 puntos por debajo** del perfil base.

Descriptivamente, hay una separación clara.

---

### Cálculo del z-score real

$$z = \frac{0.7615 - 0.8241}{0.0297}$$
$$z \approx -2.11$$

(aproximadamente  $-2.10$  /  $-2.11$ )

---

### Interpretación estadística

Con distribución normal:

- Valor crítico al 95%  $\rightarrow \pm 1.96$
- Tu resultado  $\rightarrow -2.11$

Esto implica:

- ✓ Diferencia **estadísticamente significativa** al 95%.
- ✓ El texto evaluado cae fuera del intervalo central del perfil de referencia.

### Interpretación estilométrica fina (no binaria)

Un  $z \approx -2.1$  significa:

- Está fuera del núcleo central,
- pero **no extremadamente lejos**,
- no es un valor de  $-3$  o  $-4$ ,
- no es una ruptura estructural radical.

En términos prácticos:

Existe una desviación detectable del perfil base, pero dentro de un rango compatible con continuidad autoral.

### Posible resumen textual

El análisis SPI normalizado muestra que las *Declamaciones maiores* presentan una desviación estadísticamente significativa respecto del perfil estilístico construido a partir de las *Institutiones* ( $z \approx -2.1$ ). No obstante, la magnitud de dicha desviación es moderada y no alcanza valores propios de una ruptura estructural.

### Análisis del BOXPLOT

El boxplot permite ver de un vistazo:

1. Mediana



2. Rango intercuartílico (IQR)
3. Dispersión total
4. Outliers
5. Superposición entre distribuciones

Posición central (mediana y media)

En el gráfico:

- Autor conocido → mediana  $\approx 0.82-0.83$
- Texto evaluado → mediana  $\approx 0.76-0.77$

La separación entre medianas es clara y visualmente estable.

**En general:**

- Si las medianas están muy próximas → continuidad fuerte.
- Si están separadas pero aún dentro del mismo rango visual → desplazamiento moderado.
- Si no hay solapamiento entre cajas → divergencia fuerte.

**En este caso:**

- ✓ Las cajas no se superponen plenamente.
- ✓ Existe separación clara del centro de las distribuciones.

Pero todavía no es ruptura extrema (no están en rangos totalmente disjuntos).

**Rango intercuartílico (IQR): Tamaño de las “cajas”**

El IQR del autor conocido es relativamente compacto.

Eso indica:

- Alta coherencia interna.
- Perfil estable.

El IQR del texto evaluado es similar en tamaño.

**Interpretación clave**

Aunque el centro cambia, la forma y dispersión interna son comparables.

Esto sugiere:

- No estamos ante una distribución estructuralmente distinta.
- Estamos ante un desplazamiento global.

Eso suele indicar cambio de registro/género, no cambio de identidad estilística.

**Dispersión y cola inferior**

En el texto evaluado aparecen:

- Más valores bajos.
- Algunos outliers claramente inferiores.

Esto indica:

- Mayor heterogeneidad local.
- Fragmentos especialmente alejados del perfil base.

En términos estilísticos:

El género declamatorio introduce pasajes con configuraciones formales más extremas.  
Pero el grueso de la distribución sigue siendo compacto.

## **Análisis del HISTOGRAMA SUPERPUESTO**

El histograma permite ver:

- Forma de la distribución
- Grado de solapamiento
- Si hay desplazamiento completo o parcial

### **Desplazamiento global**

La distribución azul (Institutiones) está claramente desplazada hacia la derecha.

La naranja (Declamationes) está desplazada hacia la izquierda.

Esto confirma:

- Existe diferencia sistemática.
- No es efecto de unos pocos fragmentos extremos.

El desplazamiento afecta a toda la masa de la distribución.

### **Solapamiento**

- Hay solapamiento entre aproximadamente 0.78 y 0.80.
- No son distribuciones totalmente separadas.

Para autores claramente distintos, esperaríamos:

- Mucho menor solapamiento.
- O incluso distribuciones casi disjuntas.

Aquí:

- ✓ Hay continuidad parcial.
- ✓ Comparten zona de intersección.

Eso es compatible con autor único con variación o imitación de autor o de género.

### **Forma de las distribuciones**

Ambas parecen aproximadamente:

- Unimodales (inscritas en forma “campana”)
- Sin multimodalidad evidente (No fuera de la “campana”)
- De forma relativamente normal

Eso indica:

- No hay mezcla de estilos radicalmente distintos.
- No hay “doble identidad” estilística.

### **Interpretación conceptual combinada**

Al juntar boxplot + histograma puedes aplicar esta matriz interpretativa general:

<b>Patrón observado</b>	<b>Interpretación</b>
Medianas cercanas + fuerte solapamiento	Continuidad fuerte

Patrón observado	Interpretación
Medianas separadas + solapamiento parcial	Variación genérica interna
Sin solapamiento + centros muy alejados	Divergencia estructural
Dispersión muy distinta entre grupos	Cambio en estabilidad estilística
Distribución multimodal en texto evaluado (fuera de la “campana”)	Mezcla o heterogeneidad fuerte

En este caso encaja en:

Medianas separadas + solapamiento parcial + dispersión similar.

Eso es un patrón típico de:

**Mismo autor con cambio de género.**

### Resumen del ejemplo

1. Las *Institutiones* forman un núcleo compacto y alto.
2. Las *Declamationes* están desplazadas hacia menor similitud.
3. La dispersión es comparable.
4. Hay zona común de intersección.
5. No hay ruptura estructural radical.

Interpretación sintética:

*El género declamatorio introduce un descenso sistemático en la similitud con el perfil técnico-didáctico de las Institutiones, pero sin generar una distribución independiente o estructuralmente ajena al sistema estilístico del autor.*

## 3-gramas -- 4-gramas

Diferencias estilométricas fundamentales

A) Sensibilidad estructural

3-gramas

Capturan:

- patrones morfológicos generales
- terminaciones frecuentes (-us, -um, -is...)
- combinaciones comunes de letras

Son buenos para detectar:

- hábitos ortográficos
- distribución general de morfemas
- rasgos de frecuencia global

En latín clásico, donde la morfología es muy productiva, el 3-grama tiende a reflejar el **perfil lingüístico general del autor**.

4-gramas

Capturan:

- secuencias más específicas
- fragmentos más cercanos a palabras completas

- estructuras morfosintácticas más concretas

Son más sensibles a:

- fórmulas retóricas
- colocaciones habituales
- giros recurrentes

En textos retóricos como Quintiliano, esto puede ser importante porque la retórica genera fórmulas recurrentes más largas.

---

Efecto estadístico en tu análisis Tuccinardi + SPI

Aquí es donde cambia la interpretación.

3-gramas → más estables estadísticamente

- Generan más ocurrencias
- Reducen la varianza
- Suelen producir distribuciones más suaves
- El z-score tiende a ser más estable

Resultado típico:

Las similitudes entre fragmentos del mismo autor suelen concentrarse más claramente alrededor de la media.

Si el texto dudoso cae fuera, suele ser un alejamiento bastante claro.

---

4-gramas → mayor varianza

- Hay muchos más tipos posibles
- Menor frecuencia media por n-grama
- Más dispersión en los valores de similitud

Resultado típico:

- Mayor desviación estándar
- Distribuciones más anchas
- Más sensibilidad a diferencias pequeñas

Esto puede:

- aumentar la capacidad discriminante
  - o introducir ruido si los fragmentos son cortos
- 

En términos prácticos: qué puedes esperar

Caso 1: Obras muy próximas estilísticamente

Si Instituciones y Declamaciones son estilísticamente cercanas:

- Con 3-gramas → probablemente el texto dudoso caerá dentro de la zona central.
  - Con 4-gramas → puede empezar a desviarse ligeramente si hay diferencias de registro o género.
-

Caso 2: Diferencia de género retórico

Institutiones = tratado técnico-didáctico

Declamationes = ejercicio retórico

Aquí es donde 4-gramas pueden detectar:

- fórmulas declamatorias
- patrones más enfáticos
- repeticiones retóricas

Mientras que 3-gramas pueden seguir mostrando proximidad porque la base lingüística sigue siendo quintiliana.

---

Cuándo usar cada uno

3-gramas recomendables cuando:

- Fragmentos relativamente cortos
- Corpus pequeño
- Buscas estabilidad estadística
- Quieres captar “perfil de lengua”

4-gramas recomendables cuando:

- Corpus amplio
  - Fragmentos largos
  - Buscas mayor precisión discriminante
  - Sospechas diferencias de género o registro
-