Capítulo 2: análisis de los resultados

1. Unas salvedades

1.1. Los ejes verticales reales

Para analizar los resultados, primero tengo que aclarar un par de cosas. Primero quiero escribir un poco sobre los ejes verticales. La gráfica de estilo tiene un rango que va de 2.745657 a 3.300022, y la de tema uno que va de 0.4275955 a 0.5543502. Estos números pueden parecer muy pequeños, pero antes de hacer ese juicio hay que entender qué significan. Como ya describí en el capítulo anterior, lo que representan los ejes verticales en ambas gráficas es el resultado de un cálculo de distancias en un espacio multidimensional. Veamos la gráfica de estilo. En el caso del año 2020, por ejemplo, las distancias promedio entre los trabajos de titulación resultaron ser de 2.745657.[[1]](#footnote-2) ¿2.75… qué? Lo que estamos viendo cuando vemos ese 2.75 es una condensación de datos. Recordemos que lo que medimos en el capítulo pasado fueron las frecuencias relativas de las 150 palabras más comunes. Es tentador pensar que ese 2.75 es 2.75%, puesto que las frecuencias relativas se miden en porcentajes. Sin embargo, estaríamos cometiendo un error. Ese 2.75 es resultado de, dicho sencillamente, sacar la hipotenusa de un triángulo (dentro de un espacio multidimensional, como ya he explicado, pero eso es lo que está pasando cuando se usa la función "dist()" en R). El número 2.75 no representa nada que podamos ver en el mundo real: es más bien el resultado de varios procesos de simplificación de datos. Es por esto que no es acertado pensar que 2.75 es mucho ni que es muy poco. La información valiosa que sale de esta gráfica viene de comparar los datos a lo largo de los años y de observar la tendencia. Dicho de otra manera, el valor del punto (2020, 2.75) surge en relación con los demás datos, no de manera independiente. Entonces, en 2020 hubo 2.75 ¿qué?. Nada, no hubo 2.75 nada. Esto mismo aplica para la gráfica de temas. No es que en el 2011 haya habido 0.55 temas, ni que hayan sido 55%[[2]](#footnote-3) diferentes entre sí los temas de los trabajos de ese año. Para esta investigación, no es posible interpretar los valores de los ejes verticales como lo haríamos en una gráfica más común: extraemos información de la variación a lo largo del tiempo y de las tendencias.[[3]](#footnote-4)

Dicho esto, valorar los datos de manera individual puede ser posible en el futuro.[[4]](#footnote-5) Para hacerlo sería necesario tener otra base de datos que fuera representativa de algo así como la escritura promedio. Si tuviéramos la distancia que tienen dos textos promedio, tanto en estilo como en temas, podríamos ver si, para seguir con el ejemplo anterior, 2.75 es una gran distancia o si es muy poca. Claro que para llegar a esta base de datos sería necesario recopilar una gran cantidad de textos de diferentes géneros (académico, ficción, poesía, periodismo, etcétera) y llevar a cabo los cálculos que yo realicé en mi base de datos (trabajos de titulación de Letras Inglesas publicados entre el 2006 y el 2023). Una segunda opción sería realizar estos cálculos con la base de datos entera de TESIUNAM, y de esa manera sabríamos si en Letras Inglesas se escribe más homogéneamente que en la UNAM en general. En fin, en efecto es posible llegar a valorar los datos de manera individual, pero hace falta trabajo que está fuera de los límites de esta tesina.

1.2. Las fechas de publicación no indican fecha de escritura: las gráficas apuntan al profesorado más que al estudiantado

Otra cosa que hay que reconocer es que las fechas de publicación no son un indicador de cuándo fue escrito el trabajo. Bien pudo haber sido escrito en el año en el que fue publicado, o bien pudo haber sido escrito quince años antes. Mientras que no es posible conocer la fecha de escritura, algo que sí es claro es que el trabajo fue aceptado en el año de publicación o en el año anterior (pues la Forma de Examen Profesional 1 tiene vigencia de un año). Esto quiere decir que cuando ordenamos los datos por fecha de publicación, las tendencias que resultan apuntan al profesorado (quien decide cuáles trabajos se publican), más que al alumnado. Claro está que no se puede ignorar al alumnado, pero me parece mucho más acertado interpretar estas tendencias como reflejos de quienes abren las puertas que de quienes las cruzan.

Voy a profundizar un poco en por qué pienso esto. Los trabajos fueron escritos por alumnxs. Asimismo, los trabajos fueron corregidos y aceptados por profesorxs. Supongamos que yo tengo una biblioteca con libros de ficción, filosofía y ciencia. Esos libros fueron escritos por personas diferentes a mí, sin embargo, al ver la biblioteca en su totalidad, lo que emerge es mi gusto personal, no el de quienes escribieron los libros. Si elijo que en mi biblioteca no haya libros de poesía es porque a mí no me gusta la poesía, no porque a los autores que tengo en mi biblioteca no les guste la poesía. Lo mismo pasa cuando se ordenan los textos por año de publicación. Si bien quienes escribieron los trabajos fueron alumnxs, los textos están agrupados en torno al año en el que fueron aceptados, haciendo que las tendencias resultantes apunten a los gustos de quienes eligen tenerlas en su biblioteca (TESIUNAM sería el homólogo a mi biblioteca del ejemplo pasado), más que a los gustos de quienes los escribieron.

Ahora, hago esta afirmación a raíz de que me hace falta información para hacer otras. Por ejemplo, si tuviera información de más o menos cuándo se empezó a escribir cada trabajo de titulación, tendría una interpretación diferente de los datos. De la misma manera, sería valioso saber cuándo entró a la carrera el autor de cada trabajo, para así tener su año de inscripción y su año de titulación. Si resultara que en general los trabajos de titulación son publicados cuatro o cinco años después de que su autorx entró a la carrera, mi aseveración sobre quienes abren las puertas y quienes las cruzan sería muy diferente. Otra información también valiosa sería el nombre de quien asesoró el trabajo, para ver si, al ordenar los trabajos por asesorx, surgen tendencias similares o diferentes a cuando los ordenamos por año de publicación, y para ver si los trabajos asesorados por una misma persona suelen tener un estilo y temas similares o no. En fin, la falta de información me obliga a interpretar los datos de las dos gráficas como los gustos e intereses de los sínodos y no tanto del alumnado. [NO ME SIENTO TAN SEGURO DE ESTO PERO NO SE ME OCURRE NADA MÁS]

2. Interpretación de las gráficas

Con estas salvedades fuera del camino, podemos pasar a la interpretación de las gráficas. Al ver estas dos gráficas, se vuelve evidente que tienen tres etapas homólogas, que son 2006-2010, 2011-2017 y 2018-2023. En la gráfica de estilo se ve cómo en la primera etapa el estilo se diversifica un poco, luego cómo en la segunda se homogeneiza rápidamente y en la tercera cómo continúa la homogeneización pero de manera más lenta. En la de tema se ven los mismos primeros dos movimientos (aunque el primero es un poco más precipitado) y se ve cómo en el tercero vuelve a aumentar la diferencia. Es evidente que algo pasó entre el 2011 y el 2017 que hizo que tanto el estilo como los temas se homogeneizaran. A continuación voy presentar mis hipótesis respecto a qué pudo haber sido.

2.1. Estilo: Hipótesis respecto a por qué se homogeneiza

Antes de presentar mis hipótesis respecto a por qué se homogeneiza el estilo, tengo que dejar claro que no tengo datos que apoyen a ninguna. También debo decir que estas tendencias dependen de más factores de los que a mí se me pueden ocurrir. Para comprobar estas hipótesis hacen falta investigaciones futuras.

2.1.1. Autocorrector y otras herramientas que corrigen el estilo

Mi primera hipótesis (y la que intuyo que es la menos acertada y menos interesante) es que el aumento de herramientas digitales que corrigen errores de ortografía (y a veces de gramática) es un factor en la homogeneización del estilo de los trabajos de Letras Inglesas. Estoy pensando específicamente en los correctores ortográficos de MS Word y Google Docs, así como en programas más especializados como LanguageTool o QuillBot.[[5]](#footnote-6) Digo que no creo que esta hipótesis sea tan acertada porque no me parece que estos programas sean tan intrusivos como para generar un cambio así de notorio. Estos correctores gramáticos, a pesar de que sí sugieren a veces la inserción de ciertas palabras funcionales, la verdad es que son todavía muy poco confiables, por lo que no creo que el uso de estas herramientas haya sido muy popular.[[6]](#footnote-7) Si hoy no son confiables no me imagino que en el 2011 hayan sido de gran ayuda a la hora de escribir un trabajo de titulación de Letras Inglesas.

2.1.2. Aumento en el rigor del sínodo y en los seminarios de investigación y titulación.

Mi segunda hipótesis, en la cual confío un poco más, es que a lo largo del tiempo, lxs profesorxs que han sido parte de los sínodos, así como lxs que han impartido los seminarios de investigación y titulación, han ido aumentando el rigor de la escritura de los trabajos de titulación. Más que rigor respecto a los métodos de investigación (que estoy seguro que ha aumentado) me refiero a rigor en la forma de exponer argumentos. Mi hipótesis es que el aumento de rigor unifica nuestra forma de exponer argumentos, por lo que se homogeneiza el estilo.

2.1.3. Deseo de que se titule más gente: escritura formularia

Mi última hipótesis, y de la que me siento más seguro, es que dentro del Colegio de Letras Modernas, así como dentro de la UNAM entera, hay un deseo de que la cantidad de alumnxs tituladxs aumente, y que esto ha llevado a una transformación en la manera de enseñar a escribir trabajos de titulación que ha resultado en la homogeneización del estilo. La administración de la universidad tiene incentivos monetarios para aumentar las titulaciones, al igual que la del Colegio [CORROBORAR], además de que el profesorado tiene un deseo humano de que sus alumnxs lleven a término su carrera universitaria. Una estrategia para aumentar la cantidad de trabajos de titulación (y por ende de titulaciones) es desarrollar una forma de escribir formularia. Creo que consciente o inconscientemente el profesorado de Letras Inglesas ha ido desarrollando una manera de enseñar a escribir que tiene como prioridad la eficiencia y eficacia en la escritura (es decir, que el trabajo se escriba en relativamente poco tiempo y que cumpla con los estándares mínimos necesarios para que el sínodo lo apruebe).

Para comprobar esto calculé la correlación entre la cantidad anual de trabajos publicados y la distancia promedio anual del estilo (de la gráfica de estilo del primer capítulo). Usé la función de R "cor()", especificando el método Spearman[[7]](#footnote-8) para obtener una correlación de −0.179463. Esta es una correlación tan baja que básicamente indica que no existe una correlación. Sin embargo, si excluyo el año 2020 de la base de datos, la correlación aumenta a −0.3906406, que no es muy alta pero sí indica una ligera correlación.[[8]](#footnote-9) Si acoto aun más los datos para contener solamente los años del 2011 al 2017 (los años que conforman la segunda de las tres etapas que identifiqué y en la que el estilo se homogeneiza más precipitadamente) obtengo una correlación de −0.8183171, que es muy fuerte. Lo que concluyo de estas correlaciones es que, en la gran escala, la cantidad de publicaciones anuales no es un factor que esté relacionado con la similitud en el estilo, pero que, en el periodo de 2011 al 2017, existe una fuerte relación entre las dos. Si observamos la figura [["cantidad\_publicaciones\_2006\_2023.png"]] podemos notar que del 2009 al 2016 hubo un periodo en el que cada año vio las mismas o más publicaciones que el año pasado, y si lo extendemos al 2017, que, si disminuyó, disminuyó sólo por una publicación.[[9]](#footnote-10) Si vemos la cantidad de trabajos publicados en toda la historia de la licenciatura podemos observar que ese periodo de 2009 a 2017 es uno de los periodos con crecimiento más constante. Si calculamos la correlación entre la similitud del estilo y la cantidad de trabajos publicados anualmente pero del 2009 al 2017, la correlación baja (−0.5728379) pero sigue siendo significativa.[[10]](#footnote-11)

Mi conclusión es que el aumento estable en la cantidad de trabajos de titulación publicados está relacionado con una homogeneización en el estilo. Ahora, cuál sea la causa y cuál sea el efecto es algo que no puedo asegurar: no tengo manera de saber si una homogeneización en el estilo lleva a que se publiquen más trabajos de titulación o si publicar más trabajos tiene como consecuencia que se homogeneice el estilo. También es una opción que exista una tercera causa, que desconozco, que tenga como resultado tanto el aumento en publicaciones anuales como la homogeneización del estilo. Un ejemplo posible sería que unx asesorx (o un grupo de asesorxs) fueron especialmente prolíficos durante esos años y por ende aumentaron la cantidad de trabajos publicados al mismo tiempo que homogeneizaron el estilo a su gusto. El mismo efecto lo pudo haber tenido el grupo de profesorxs que impartieron el seminario de titulación durante esos años. Lo único claro es que hace falta mayor investigación al respecto.

2.2. Tema

2.2.1. Mis sentimientos iniciales respecto al tema

La verdad es que al ver la gráfica de los temas me siento más confundido que al ver la del estilo. Esto es porque es menos claro cuál es la tendencia más amplia. Es decir, sabemos que sube, baja y vuelve a subir, pero es menos evidente cómo extrapolar esta tendencia a los años previos y a los años siguientes. Veo dos opciones. La primera sería que estemos viendo una sección de una tendencia más o menos sinusoidal (que sube y baja periódicamente pero cuya línea de equilibrio siempre se mantiene), y la otra sería que estemos viendo un "bache" en una línea que va para arriba (hay una infinidad de opciones más, pero estas son las que me sugiere la curva). Ambas me son imposibles de comprobar en este momento, entonces no voy a divagar más en sus posibilidades e implicaciones.

Otra causa de confusión puede ser que los valores son aun menos concretos, pero voy a intentar hacer más concreto lo que hace el modelado de tópicos. Lo que hice en esta investigación fue dividir las palabras en funcionales y de contenido. Si la gráfica de estilo visualiza las similitudes en el uso de palabras funcionales, la de temas lo hace en el uso de palabras de contenido. ¿Por qué no simplemente calculé sus frecuencias relativas como hice con las funcionales? Porque son mucho menos frecuentes, además de que la presencia de un tema en un texto no se puede cuantificar solamente contando las palabras que consideramos que pertenecen a ese tema. Cuando se trata de cuantificar temas hay que pensar en aglomeraciones de palabras de contenido, más que en palabras de contenido individuales. [CITA DE JOCKERS HACIENDO ESTE ARGUMENTO] Antes de cualquier cuantificación, MALLET se deshace de todas las palabras funcionales de los textos para quedarse sólo con las de contenido. Después, identifica qué palabras están cerca de qué otras palabras y así identifica los temas. Las visualizaciones de los temas de las figuras [["5.png"]], [["12.png"]] y [["15.png"]] surgen a partir de que las palabras "lengua", "inglés", "alumnos", etcétera, aparecen seguido juntas, así como a su vez las palabras "Gawain", "Chaucer" y "caballero", y "cómic", "Bechdel" y "guasón" aparecen juntas. MALLET logra identificar grupos de palabras que suelen estar cerca las unas de las otras, y además identifica qué tan presente está cada grupo de palabras en cada texto. En conclusión, para interpretar la gráfica de temas hay que pensar que es lo mismo que la de estilo, pero tomando en cuenta las palabras de contenido en vez de las funcionales.

2.2.2. Correlaciones

Calculé las correlaciones entre los temas y la cantidad de trabajos publicados anualmente (como hice con el estilo), y además calculé la correlación entre el estilo y los temas. La correlación de los temas con la cantidad de trabajos publicados es de −0.06950301, es decir, inexistente. Pero, si hacemos como con el estilo y acotamos los años a que sean sólo del 2011 al 2017, la correlación se dispara a −0.7092081, que es menor a la del estilo pero sigue siendo alta. Entre el 2011 y el 2017, las tres variables —cantidad de trabajos publicados por año, distancia promedio anual del estilo y distancia promedio anual de los temas— tienen una correlación significativa. Esto es un punto a favor de la idea de que algo pasó entre el 2011 y el 2017 que hizo que aumentara la cantidad de publicaciones y se homogeneizaran tanto el estilo como los temas.

Otra cosa interesante es que, cuando calculamos la correlación entre los temas y el estilo (sin acotar los años, es decir del 2006 al 2023), obtenemos una de 0.6635707, lo que comprueba que la similitud en el estilo está ligada a la similitud en los temas, independientemente del periodo de tiempo. Cuando los temas son similares entre sí, el estilo es similar también, y cuando los temas son diversos, el estilo también.

Las dos conclusiones a las que me llevan estas correlaciones son las siguientes: algo pasó entre el 2011 y el 2017 que hizo que la cantidad de publicaciones tuviera un aumento constante, que el estilo se homogeneizara y que los temas se homogeneizaran; y, al menos cuando hablamos de trabajos de titulación de Letras Inglesas, la similitud en el estilo implica similitud en los temas (o viceversa).

Bibliografía

Jost, Steve D.. "Linear Correlation". *DePaul University*,<https://condor.depaul.edu/sjost/it223/documents/correlation.htm>, accesado el 29 de octubre del 2024.

1. Voy a redondear 2.745657 a 2.75. [↑](#footnote-ref-2)
2. Recordemos que el software de modelado de temas que usé califica los temas del 0 al 1, que se puede interpretar como de 0% al 100%, por lo que 0.55 sería lo mismo 55%. [↑](#footnote-ref-3)
3. Una gráfica común sería, por ejemplo, la gráfica de población en México que publica el INEGI cada 10 años (disponible en <https://www.inegi.org.mx/temas/estructura/>). Al ver un punto en esa gráfica, digamos el del 2020, estamos obteniendo información que por sí sola es significativa —en este caso, que en el 2020, el INEGI determinó que hay 126,014,024 personas viviendo en México—, mientras que los puntos de las gráficas de estilo y tema no son significativos por sí solos, al menos por ahora. [↑](#footnote-ref-4)
4. Cuando me refiero a los datos "individuales" me refiero tanto a la distancia promedio de un año (2020, 2.75) como a la distancia entre dos textos individuales ("Propuesta de lecciones-muestra para la enseñanza de comprensión de lectura en inglés en el marco de la ENA" y "Wide Sargasso Sea: una perspectiva diferente sobre un estereotipo femenino", 3.662696). [↑](#footnote-ref-5)
5. LanguageTool empezó en el 2003, y empezó a funcionar en español, según he podido descubrir, en el 2007 (no encuentro documentación de confianza respecto a cuándo han incluido cuáles idiomas, pero una de las ventajas del código abierto es que la historia de su desarrollo es pública, entonces pude ver que el documento más antiguo de su rama dedicada al español es del 2007. La dirección que visité es <https://github.com/languagetool-org/languagetool/blame/master/languagetool-language-modules/es/src/main/java/org/languagetool/language/Spanish.java>). Por su parte, QuillBot fue lanzado en 2017, por lo que no es posible que haya sido la razón de la homogeneización de estilo. [↑](#footnote-ref-6)
6. Será interesante ver cómo, con las innovaciones recientes en modelos extensos de lenguaje (*large language models*), evoluciona la homogeneización en el estilo. Predigo que en los próximos años va a haber una homogeneización precipitada causada por el uso de estos programas. [↑](#footnote-ref-7)
7. Para ver los detalles de esta función, visitar la documentación oficial de R: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor>. [↑](#footnote-ref-8)
8. Mi razón para quitar el año 2020 es que es un valor atípico por la cantidad tan baja de publicaciones causada por la toma de la facultad por parte de las MOFFYL y luego por el inicio de la pandemia. [↑](#footnote-ref-9)
9. Aquí está la gráfica de todas las publicaciones anuales de 1962 a 2023. [["cantidad\_publicaciones\_todo.png"]] [↑](#footnote-ref-10)
10. Para mis juicios respecto a si una correlación es significativa o no estoy siguiendo a Steve D. Jost en su guía "Linear Correlation", en donde sugiere que para las ciencias sociales una correlación de ±0.6 es una correlación significativa. [↑](#footnote-ref-11)