UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Desarrollo de un sistema de recomendación de canciones para el aprendizaje de idiomas.

Trabajo Fin de Máster

Máster Universitario en Ingeniería y Tecnología de Sistemas Software

AUTOR/A: Dolgov, Viktor

Tutor/a: Molina Marco, Antonio

Cotutor/a: Ferri Ramírez, César

CURSO ACADÉMICO: 2023/2024

Agradecimientos

Quiero expresar mis sinceros agradecimientos a ValgrAI (Valencian Graduate School and Research Network for Artificial Intelligence) y a la Generalitat Valenciana por el apoyo económico que me brindaron. Es un gran honor y una oportunidad para mí que me ayudó a realizar mis ambiciones educativas. Gracias a su apoyo financiero, tuve la oportunidad de profundizar en la educación y la investigación que me interesan y contribuir al desarrollo de mi potencial.

Además, esto no solo tuvo un impacto positivo en mi carrera académica, sino que también me dio la confianza de que los esfuerzos y las aspiraciones de conocimiento y el desarrollo de las investigaciones científicas son valorados y apoyados. Me inspira y me motiva a seguir obteniendo mejores resultados.

Asimismo me gustaría dar las gracias a mis tutores Antonio Molina Marco y César Ferri Ramírez por su toda su ayuda y paciencia durante todo el proceso del TFM.

Resumen

Este trabajo fin de master se centra en el desarrollo de un sistema de recomendación de canciones para ayudar a las personas que aprenden idiomas extranjeros como el inglés y el español. Aprender lenguas extranjeras puede ser difícil, sin embargo elegir el material de estudio adecuado facilita mucho el proceso. Este sistema se basa en el principio del conocimiento, que utiliza datos de expertos e información sobre un área temática para proporcionar recomendaciones. Su idea principal es favorecer a las personas a estudiar un idioma de manera más efectiva eligiendo materiales (letras) que se ajusten a su nivel de conocimiento y que sean más apropiados para mejorar un aspecto específico del lenguaje (vocabulario o fonética), lo que facilita el aprendizaje de idiomas. Uno de los resultados de este trabajo es la creación del prototipo del clasificador de canciones con una interfaz de usuario intuitiva y cómoda de usar, que permite al usuario encontrar fácilmente los materiales de interés para su estudio. El prototipo cuenta con alrededor de 100.000 letras de canciones en inglés y español y dispone de varios filtros de búsqueda que ayudan a los usuarios a encontrar canciones por géneros, artistas, niveles de dificultad, etc. En general, el sistema desarrollado tiene el potencial de mejorar el proceso de aprendizaje de idiomas, ya que el uso de sus canciones favoritas hace que él sea más divertido y permite avanzar más rápido. Entre otras cosas, los resultados del estudio se pueden utilizar para otras tareas relacionadas con el análisis de textos en idiomas extranjeros.

Palabras claves

Sistemas de recomendación, Ingeniería del Lenguaje Natural, ILN, aprendizaje automático, AA, análisis de texto.

Abstract

This master's thesis is devoted to the development of a recommendation system of songs to help people learning foreign languages such as English and Spanish. Study of foreign languages can be difficult, but choosing the right educational materials can greatly facilitate this process. This system is based on the principle of knowledge, which uses expert data and information about the subject area to provide recommendations. Its main idea is to help people learn a language more effectively by choosing materials (lyrics) that match their knowledge level and that are better suited to improve a certain aspect of the language (vocabulary or phonetics), which contributes to an easier language learning process. One of the results of this work is the creation of the prototype of the song classifier with an intuitive and user-friendly interface that allows the user to easily find the materials they are interested in for study. The prototype contains about 100,000 lyrics in English and Spanish and provides various search filters that help users find songs by genre, artist, difficulty levels, etc. In general, the developed system has the potential to improve the language learning process, since the use of favorite compositions makes it more exciting and allows to move forward faster. Among other things, the results obtained in the framework of the investigation can be used for other tasks related to the analysis of texts in foreign languages.

Key words

Recommendation systems, Natural language processing, NLP, Machine Learning, ML, text analysis.

CONTENIDO

Capítulo 1. INTRODUCCIÓN

1.1. Relevancia del estudio

En la actualidad, el proceso de aprendizaje de un idioma extranjero por parte de las personas se acompaña cada vez más con la necesidad de determinar la complejidad del material estudiado. El texto es una unidad clave de material de idioma extranjero estudiado, ya que su contenido de comunicación y de lenguaje determina la actividad educativa y cognitiva de los estudiantes de un idioma extranjero.

A menudo, la dificultad (oral o escrita) de la percepción de las personas del texto estudiado se debe a un concepto psicolingüístico complejo, un fenómeno asociado con una gran cantidad de propiedades objetivas y subjetivas del texto en sí (volumen del texto, longitud de las construcciones sintácticas, tema, construcción estructural del texto, saturación informativa), así como la experiencia lingüística y no lingüística del receptor, su nivel de dominio de un idioma extranjero, el volumen de su vocabulario, etc.

Hay muchas maneras de hacer que este proceso sea más divertido y efectivo. Una de esas formas es escuchar a las canciones. Las canciones son una excelente manera de aprender o facilitar el aprendizaje de un idioma por varias razones. En primer lugar, generalmente se recuerdan más fácilmente que otros tipos de textos. Esto se debe a que contienen elementos repetitivos como la melodía, el ritmo y la rima. En segundo lugar, las canciones suelen asociarse con ciertas emociones y experiencias, lo que ayuda a recordarlas mejor. En tercer lugar, pueden usar para aprender diferentes aspectos del lenguaje, como la pronunciación, el vocabulario y la gramática.

En este sentido, el desarrollo de un sistema de recomendación para determinar la complejidad de las letras de canciones en inglés y español parece bastante relevante. Puede ayudar al estudiante a adaptar el texto a su propio nivel de conocimiento, proporcionar un sistema de aprendizaje más flexible, lo que en última instancia conducirá a un mayor rendimiento en el dominio de idiomas extranjeros y, por lo tanto, determina la relevancia de nuestro estudio.

1.2. Motivación

Los principales motivos para llevar a cabo este estudio pueden ser personales, sociales y profesionales. Entre los motivos personales, vale la pena destacar mi constante interés por el aprendizaje de idiomas extranjeros. Creo que el conocimiento de un idioma extranjero abre nuevas oportunidades para la comunicación y el autodesarrollo. Quería saber cómo el uso de canciones puede ayudar en el aprendizaje de idiomas.

Además, este estudio tiene un potencial significado práctico para mi futura carrera profesional. Los conocimientos adquiridos y las habilidades en el análisis de datos, el aprendizaje automático y la creación de aplicaciones pueden ser podrían potencialmente aplicados en mi trabajo, lo que me convertiría en un profesional más competente y demandado.

Entre otras cosas, es importante señalar que este estudio puede tener un impacto positivo en la sociedad, en particular en el campo de la educación. Por lo tanto, la realización de este estudio se debió a una combinación de motivos personales, sociales y profesionales.

1.3. Objetivos

El objetivo principal de este estudio es desarrollar un sistema de recomendaciones para la selección de canciones, teniendo en cuenta la complejidad de sus letras, para estudiantes de inglés y español. El objetivo del estudio fue establecer los siguientes subobjetivos:

1. Hacer una revisión teórica de la literatura científica contemporánea dedicada a la investigación de los niveles de complejidad de los textos en inglés y español, resaltar los enfoques para definir la complejidad de los textos de canciones en inglés y español como fenómenos lingüísticos especiales;

2. Identificar las características más relevantes y determinar el nivel de dificultad de las letras;

3. Llevar a cabo la agrupación de las letras de las canciones por dificultad en función de las características identificadas;

4. Desarrollar un prototipo del clasificador de canciones para crear una lista basada en diferentes niveles de dificultad de las letras de las canciones.

1.4. Metodología

Se pretende examinar la literatura relevante y las fuentes dedicadas al tema del análisis de la complejidad de los textos con el fin de determinar las variables más relevantes para clasificar la complejidad de letras. También es necesario identificar las herramientas más adecuadas tanto para procesar y extraer la información de las letras de las canciones, como para la agrupación y la posterior creación del prototipo. Simultáneamente el material práctico del estudio fue una base de letras de canciones en inglés y español recogidas por una muestra sólida.

1.5. Estructura

En el **primer** capítulo se presenta la información sobre los objetivos, la metodología, la motivación de la investigación, asimismo el estado actual de las tecnologías.

En el **segundo** capítulo se presenta una visión general de la investigación moderna sobre el tema estudiado, una descripción de los enfoques actuales en la literatura científica para determinar la complejidad de las letras, se consideran las peculiaridades de resaltar las características de complejidad de las letras de canciones en inglés y español.

En el **tercer** capítulo se presentan para el desarrollo de un sistema de recomendación, la descripción de las tecnologías, de los algoritmos y del *dataset* aplicados en la investigación

En el **cuarto** capítulo se presenta el análisis de canciones en inglés.

En el **quinto** capítulo se presenta el análisis de canciones en español.

En el **sexto** capítulo se trata del desarrollo del prototipo del clasificador de canciones.

En el **séptimo** capítulo se resumen los resultados del estudio y se formulan conclusiones.

En el **octavo** capítulo se presenta la lista de referencias.

En los **anexos** se puede encontrar información adicional

1.6. Estado actual de las tecnologías

Actualmente se pueden distinguir las siguientes aplicaciones en el campo del aprendizaje de idiomas a través de canciones o podcasts.

Duolingo es una aplicación popular de aprendizaje de idiomas que incluye una sección de "Duolingo Radio". Esta sección invita al usuario a escuchar podcasts en inglés y español.

Spotify proporciona una variedad de listas de reproducción para aprender idiomas, incluyendo inglés y español. Por regla general, se trata de historias en un idioma extranjero discernible por nivel.

LyricsTraining es una plataforma en línea donde puedes aprender el idioma completando las palabras que faltan en las letras de las canciones mientras ves el videoclip de la misma canción(modo karaoke). Se puede elegir canciones de diferente complejidad y en diferentes idiomas. El sistema se adapta automáticamente a su nivel y ofrece tareas más difíciles a medida que avanza.

Yabla es una plataforma para el aprendizaje de lenguas extranjeras mediante vídeo incluyendo vídeos musicales. Proporciona letras con subtítulos y ejercicios de comprensión auditiva que pueden ayudar a mejorar las habilidades para aprender inglés y español.

Todas las aplicaciones mencionadas están diseñadas para aprender idiomas extranjeros, sin embargo, se pueden destacar las siguientes deficiencias. Las plataformas como LyricsTraining o

Yabla tienden a centrarse más en la enseñanza que en las recomendaciones. Al mismo tiempo para una mejor asimilación es deseable interactuar con estas aplicaciones en formato de escritorio. A menudo estas aplicaciones carecen de las letras de canciones y la cantidad de grabaciones de audio es bastante limitada. Finalmente no siempre está claro cómo determinan el nivel de una canción en un recurso u otro.

La novedad científica es que en el presente trabajo se desarrolla un prototipo del clasificador de canciones con un agrupamiento de los niveles de dificultad de las letras en dos idiomas extranjeros: inglés y español, en base a datos científicos, con más canciones y con sus letras. También en el marco del sistema de recomendaciones, se espera desarrollar diferentes direcciones de aprendizaje de idiomas (léxica, fonética, etc.).

El valor práctico del trabajo es que los resultados obtenidos durante el estudio, así como el prototipo en sí pueden ser utilizados en cierta medida por los profesores de las instituciones de educación media y superior. Además, el prototipo desarrollado por nosotros se puede utilizar no solo para aprender idiomas extranjeros, sino también para servir como soporte para el desarrollo de otros sistemas de recomendación basados en los enfoques propuestos en ella.

Capítulo 2. ENFOQUES TEÓRICOS PARA DETERMINAR LA COMPLEJIDAD DE LOS TEXTOS EN INGLÉS Y ESPAÑOL

2.1. Revisión de la literatura científica y la investigación sobre el tema estudiado

El tema de identificar los niveles de complejidad de los textos en diferentes áreas ha sido ampliamente cubierto en la literatura científica moderna, donde los autores representan las diferentes formas en que se puede determinar la complejidad de un texto. Nuestro trabajo explora los niveles de dificultad de los textos en inglés y español, por lo que a continuación se proporcionará una revisión de la literatura sobre la evaluación de la complejidad de los textos en estos idiomas.

En lingüística, un número bastante grande de especialistas prestó atención al desarrollo de métodos para evaluar la complejidad de los textos en inglés. En la etapa inicial, los problemas de complejidad de los textos fueron abordados por científicos lingüistas, quienes desarrollaron sus métodos para identificar los niveles de complejidad de la literatura didáctica, los libros de texto escolares y los manuales. Así, desde los años 20 del siglo XX, los científicos han hecho numerosos intentos de implementar métodos para evaluar numéricamente la complejidad del texto con la ayuda de índices de lecturabilidad. Al principio, estos índices se aplicaban solo al idioma inglés, más tarde también se evaluaron los textos en otros idiomas europeos.

Como señala O. N. Lyashevskaya, el concepto intuitivo de la complejidad/facilidad de lectura del texto y la velocidad de lectura y comprensión del texto resultante en la lingüística del siglo XX se presentó en forma de índices de legibilidad. Su base es una serie de supuestos siguientes:

1. Las oraciones cortas son más fáciles de leer en comparación con las largas;

2. Las palabras largas dificultan el proceso de lectura del texto;

3. El lector ralentiza la lectura al encontrar palabras desconocidas para él o palabras con baja frecuencia de uso, etc. [Lyashevskaya, 2015].

Concepto de legibilidad. Los científicos estadounidenses E. Dale y J. Chall ya en 1949 formularon el concepto de legibilidad de la siguiente manera: "La legibilidad es la suma total (incluidas todas las interacciones) de todos los elementos dentro de un fragmento dado de material impreso que influyen en el éxito de un grupo de lectores que tratan con ese material impreso. El éxito representa el grado en que los lectores entienden el texto, lo leen a una velocidad óptima y lo encuentran interesante" [Dale, Chall cita de Sibanda, 2013, P. 7].

Como se puede ver, la formulación de E. Dale y J. Chall define los indicadores de legibilidad de los textos como la capacidad de estos últimos para ser entendidos, leídos a la velocidad de lectura normal, así como para despertar el interés de los lectores. El propósito del

texto es comunicar al lector el significado que debe comprender. De no ser así, el objetivo del texto no puede considerarse alcanzado. Según G. Abaji, solo lo que los lectores pueden entender podría interesarles. Es importante que los lectores, al leer el texto, puedan hacerlo a una velocidad óptima (es decir, cómoda) para ellos. Si el texto es demasiado complejo, la persona se ve obligada a leerlo lentamente, lo que impone mayores demandas a su memoria, lo que a su vez puede comprometer la comprensión de lectura. Así, los textos muy complejos afectan negativamente su percepción y comprensión por parte del lector, lo que lleva a este último a la frustración y, probablemente, afecta negativamente su estado de ánimo y preparación mental [Abadzi, 2008].

2.1.1. Historia de los estudios de evaluación de la complejidad de los textos en inglés

En 1923, los lingüistas estadounidenses B. Lively y C. Pressey desarrollaron el primer índice de legibilidad [Oborneva, 2006]. En 1935, W. S. Gray y W. W. Leary realizaron su investigación sobre la legibilidad, a partir de un grupo de 82 factores potenciales, destacando los 44 factores principales de legibilidad y superación de la complejidad del texto (estos son factores como la longitud de la palabra, la longitud de la oración, la proporción de oraciones explícitas, etc.).

A principios de la década de 1940, el científico estadounidense de origen austriaco Rudolf Flesch desarrolló una fórmula con la que determinar el nivel de "legibilidad" de un texto. La fórmula De R. Flesch se muestra a continuación en la *Figura 1*:

$$FRE = 206{,}835 - 1{,}015\frac{\text{total words}}{\text{total sentences}} - 84{,}6\frac{\text{total syllables}}{\text{total words}}$$

Fig. 1 Fórmula De Flesch

Dónde: FRE significa facilidad de lectura;
Total words - número total de palabras en el texto
Total sentences - número total de oraciones
Total syllables - número total de sílabas

Esta fórmula se ha generalizado y sigue siendo popular hoy en día. R. Flesch, en su libro The Art of Plain Talk (1946), identifica siete posiciones según las cuales se puede obtener un texto comprensible y fácil de entender. Por lo tanto, la evaluación de la "legibilidad" del texto en Flesch se realiza en una escala de 1 a 100 puntos. Cuanto mayor sea la puntuación del texto, más fácil será de leer. La *Tabla 1*, que se encuentra a continuación, muestra la relación de estos puntajes con el nivel de educación del lector.

Evaluación de la facilidad de lectura de R. Flesch (de 1 a 100 puntos)	Nivel de educación de R. Flesch (número de clases terminadas) (de 1 a 100 puntos)
0–30	Graduado universitario
30–50	Estudiante universitario
50–60	Estudiante de secundaria
60–70	Estudiante de 8º a 9º grado
70–80	Estudiante de séptimo grado
80–90	Estudiante de sexto grado
90-100	Estudiante de quinto grado

Tabla 1. Relación de la evaluación de la facilidad de lectura en R. Flesch con el nivel de educación del lector

Fórmula de R. Flesch es fácil de usar, esta propiedad asegura su popularidad. Sin embargo, esta fórmula, junto con otras que implican el mismo procedimiento de implementación, a menudo ha sido criticada por tener en cuenta solo dos variables, aunque hay muchos otros aspectos que afectan la legibilidad de los textos.

Otros estudios en esta área continuaron en el período de posguerra, en 1947-1958 y en 1967-1976.

En 1948, los investigadores estadounidenses E. Dale y J. Chall, basados en la fórmula de legibilidad de R. Flesh, desarrollaron su fórmula de legibilidad, que se publicó por primera vez en su trabajo conjunto "Formula for Predicting Readability" (fórmula para Predecir la Legibilidad) en 1948 y se actualizó en 1995 en el artículo "Readability Revisited: the New Dale-Chall readability Formula" (Legibilidad revisada: la nueva fórmula de legibilidad Dale-Chall), donde había una lista de palabras que se amplió a 3.000 palabras familiares.

La fórmula (o prueba) de Dale-Chall fue que utilizaron una lista de 763 palabras que el 80% de los estudiantes de cuarto grado conocían como los tokens yes (sí), no (no), etc., para determinar qué palabras causaban dificultades a los estudiantes:

0.1579 (palabras complejas ÷ (dividido por) palabras × (multiplicado por) 100) + 0.0496 (palabras ÷ (dividido por) oraciones)

Por ejemplo, si el porcentaje de palabras compuestas es superior al 5%, debe agregar 3.6365 al puntaje original para obtener el puntaje ajustado, de lo contrario, el puntaje ajustado será igual al puntaje original. Las palabras difíciles, según Dale y J. Chall, son todas unidades léxicas que faltan en la lista de palabras. Sin embargo, como señalan los autores, debe tenerse en cuenta que la lista de palabras presenta las formas principales de sustantivos y verbos. Por lo tanto, es

necesario agregar el plural correcto de los sustantivos, así como ajustar las formas temporales de los verbos: pasado, presente, futuro, etc. [Chall, 1995].

En 1952, Robert Gunning, un empresario estadounidense y editor de periódicos y libros de texto, desarrolló la fórmula de legibilidad (*the Gunning Fog Index*), que se aplica a la evaluación de la complejidad de los materiales escritos escritos en inglés. Al igual que la fórmula de R. Flesch, la fórmula de R. Gunning se basa en tener en cuenta dos aspectos: es la complejidad de las palabras y la longitud de la oración. Esta fórmula implica que las oraciones cortas que usan lenguaje sencillo son más fáciles de leer que las oraciones largas que usan expresiones complejas [DuBay, 2004]. Según la opinión de R., las palabras compuestas son palabras que contienen más de dos sílabas. La fórmula de R. Gunning se escribe de la siguiente manera:

Nivel de aprendizaje = 0.4 (ASL + PHW)

Donde ASL = longitud media de la oración (es decir, el número de palabras dividido por el número de oraciones);

PHW = porcentaje de palabras compuestas.

Se utilizó el siguiente indicador de referencia: 5 puntos – el texto es legible (fácil), 10 puntos – el texto tiene elementos de complejidad, 15 puntos – el texto es bastante difícil y 20 puntos – el texto es muy difícil (se lee con gran dificultad). Sin embargo, esta fórmula ha sido criticada por la comunidad lingüística. El punto principal de la crítica fue que la fórmula no tenía en cuenta un hecho importante, a saber, que no todas las palabras compuestas por varias sílabas pueden ser difíciles de percibir por el lector. Además, hay una serie de palabras que tienen una o dos sílabas, pero que causan dificultad al receptor para leerlas, ya que casi no se usan en el habla cotidiana. Por lo tanto, palabras tan cortas y desconocidas para el lector pueden dificultar la lectura de oraciones, lo que, en opinión de L. Sibanda, es una debilidad de la fórmula de R. Gunning [Sibanda, 2013, P. 15].

También hay que mencionar a investigadores en el campo de la evaluación de la complejidad de los textos en inglés, como R. D. Powers, W. A. Sumner y B. E. Kearl, quienes en 1958 publicaron su fórmula de legibilidad en el artículo "A Recalculation of Four Adult Readability Formulas" ("Recalculación de cuatro fórmulas De legibilidad para adultos") (revista Journal of Educational Psychology).

La fórmula de Powers-Sumner-Kearl es la siguiente:

Reading age = 0.0778 (average sentence length) + 0.0455 (number of syllables) + 2.7971

donde *Reading age* – edad del lector;

average sentence length - longitud media de la oración;

number of syllables - número de sílabas.

Cabe agregar que la fórmula de Powers-Sumner-Kearl se desarrolló para evaluar la complejidad de los textos destinados a niños de 7 a 10 años.

En 1968, el lingüista estadounidense y profesor de inglés Edward Fry desarrolló las pruebas de legibilidad de las pruebas de inglés, que se basaron en gráficos. En el artículo titulado "A readability Formula That Saves Time", publicado en la revista Journal of Reading (1968), se presentó una prueba que, mediante un gráfico adjunto, determinaba la legibilidad de textos destinados a estudiantes de secundaria. Los resultados de las pruebas fueron materiales educativos de las escuelas primarias y secundarias, así como datos de otras fórmulas para evaluar la complejidad de los textos.

El gráfico de legibilidad de E. Fry se presenta en la *Figura 2*.

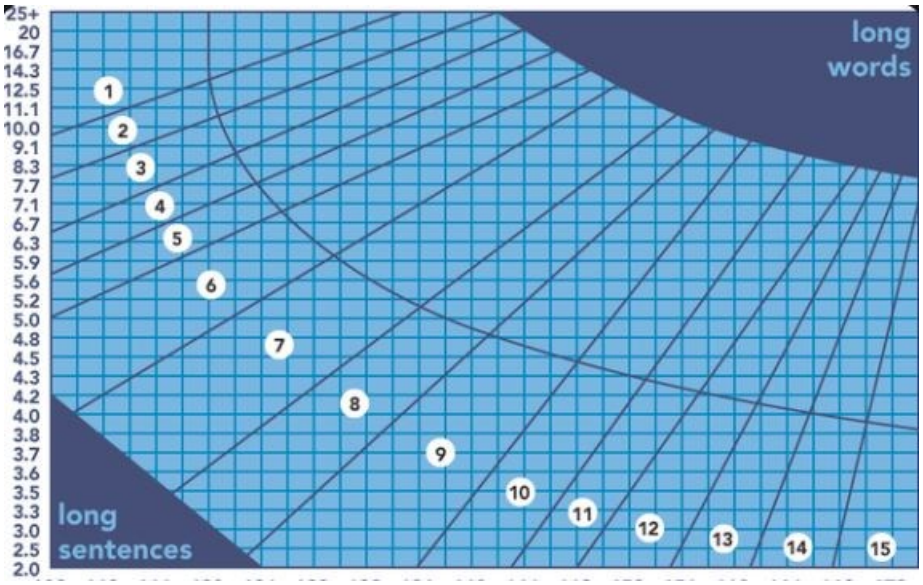


Fig. 2. Gráfico de legibilidad de Fry

El nivel de dificultad de lectura con el gráfico de Fry se calcula a partir del número promedio de oraciones (eje vertical Y) y sílabas (eje horizontal X) por 100 palabras. Estos valores medios se muestran en el gráfico. Por lo tanto, la intersección del número promedio de oraciones y el número promedio de sílabas determina el nivel de dificultad para Leer el contenido del texto.

En 1969, G. Harry McLaughlin, en su obra SMOG Grading – A New Readability Formula (1969), propuso un índice de "SMOG" de legibilidad, mediante el cual calculaba la edad del lector de texto (en prosa) a partir de la raíz cuadrada de la fracción de palabras polisílabas presentes en el texto (McLaughlin, 1969). Según G. Harry McLaughlin, la fórmula de legibilidad es una ecuación matemática que se puede obtener mediante el análisis de regresión:

Legibilidad = a + b (longitud de las palabras) + c (longitud de la oración), donde a, B y c son constantes.

Como señala G. Harry McLaughlin, como resultado de este análisis, encontramos la ecuación que mejor expresa la relación entre dos variables. "La primera variable es la evaluación de la dificultad a la que se enfrentan las personas al leer un texto dado; la segunda es la determinación de las características lingüísticas de ese texto. Esta fórmula se puede utilizar para predecir la dificultad de lectura debido a las características lingüísticas de otros textos" [McLaughlin, 1969, P. 640].

En los años 70, en Occidente, los sociólogos y psicólogos comenzaron a desarrollar formas de evaluar la complejidad de los textos, quienes señalaron la interdependencia existente entre la complejidad de los textos y las características psicológicas y de edad de una persona.

En 1975, J. Peter Kincaid, en colaboración con otros especialistas, desarrolló el llamado "Nivel de evaluación de lectura de Flesch–Kincaid" (*Flesch–Kincaid Grade Level*). El desarrollo de este Nivel (o, como se llama de otra manera, la Prueba) se llevó a cabo bajo un contrato con la Marina de los Estados Unidos (bajo la dirección de Kincaid) y se refería al campo de la educación de alta tecnología (por ejemplo, la creación y entrega de información técnica en forma electrónica), por lo que la utilidad de la fórmula de legibilidad de Flesch-Kincaid en términos de manuales informáticos para la edición de pruebas fue muy alta. En 1978, esta fórmula fue utilizada por primera vez por el ejército de la Armada de los Estados Unidos para determinar la complejidad de los manuales técnicos y manuales, después de lo cual se introdujo en el estándar militar de los Estados Unidos.

La fórmula de evaluación de lectura de Flesch-Kincaid se presenta a continuación en la *Figura 3*.

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

Fig. 3. Fórmula de Flesch-Kincaid

Como se puede ver, en los primeros paréntesis, el número total de palabras en el texto (*total words*) se divide por el número total de oraciones (*total sentences*). En el segundo paréntesis, el número total de sílabas de todas las palabras del texto (*total syllables*) se divide por el número total de todas las palabras presentes en el texto evaluado (*total words*).

El resultado será un número correspondiente al nivel de educación. Por ejemplo, 7.5 es el resultado del nivel de lectura de un estudiante de séptimo grado, 8.0 es el resultado del nivel de lectura de un estudiante de octavo grado.

También en los años 70 del siglo pasado, se propusieron índices de legibilidad para otras lenguas europeas, incluido el ruso, y se dio una razón para el método de su construcción [Mikk, 1974].

En 1977, el científico estadounidense de tecnología de la información Donald Knuth publicó un artículo en la revista SIGACT News titulado "Complejidad de canciones" (Complexity of songs) (reimpreso en 1984 en la revista mensual de la asociación de ingeniería informática en 1984), en el que intentó por primera vez evaluar la complejidad de un cuerpo seleccionado de canciones en inglés, así como canciones en otros idiomas europeos.

El tema principal del artículo de D. Knuth es la idea de la evolución de las canciones populares que está asociada con la transición de baladas largas y significativas a letras con repetibilidad frecuente de las mismas palabras y contenido con el significativo insignificante (o inexistente). D. Knuth en el trabajo señala que una serie de canciones pueden alcanzar el nivel de dificultad expresado por la fórmula O (log *N*), donde *N* es el número de palabras en una canción. El científico afirma que en las canciones más antiguas, el concepto de coro fue inventado por los autores debido al deseo de reducir la complejidad espacial de las canciones, lo que se convierte en un factor importante cuando existe la necesidad de memorizar un gran número de canciones [Knut, 1984]. Así, el lema (afirmación) en el artículo establece que si la longitud de la canción se representa como N, la inclusión del estribillo en la canción conduce a una reducción de la complejidad de la canción a cN, donde el coeficiente es $c < 1$.

Las canciones con complejidad O (logN), que D. Knuth atribuye a una clase de canciones formalmente llamada "m botellas de cerveza en la pared", son de menor importancia y un motivo más ligero. Así, la canción "99 botellas de cerveza" (*99 Bottles of Beer*) de mediados de los años 50 del siglo pasado, ampliamente conocida en Estados Unidos y Canadá por su ligero motivo y su falta de contenido serio (cuenta regresiva de botellas de cerveza caídas de la pared), dio nombre a esta clase de canciones. La complejidad de las letras de canciones de esta clase en el artículo de D. Knuth está representada por la siguiente fórmula:

$V_k = T_kBW,$

'If one of those bottles should happen to fall,'

$T_{k-1}B\ W,$

donde

V_k = *estrofas de la canción;*

B = botellas de cerveza;

W = on the wall (en la pared);

T_k = *entero.*

Finalmente, según D. Knuth, a lo largo del siglo XX, el progreso científico y tecnológico ha llevado a la necesidad de un uso aún menor de la memoria humana, lo que lleva a la aparición de una clase de canciones de longitud arbitraria, con complejidad espacial expresada por la fórmula O(1), definida por la siguiente relación recurrente:

That's the way (título de la canción estadounidense)

V_k =*U I like it, U*

U = uh huh, uh huh;

Para todas las *K (*estrofas *)* [Knut, 1984, P. 346].

Es decir, como se puede ver, en esta canción solo hay una línea *I like it (Me gustas)*, antes de la cual y después de la cual va el estribillo, que consta de solo dos y tres sílabas *uh huh.*

Sobre la base de las ideas y fórmulas de complejidad de las canciones desarrolladas por D. Knut, el científico estadounidense y profesor del departamento de matemáticas y tecnología informática Chavey Darr en 1996 en Canciones y análisis de algoritmos (*Songs and the analysis of algorithms*) (1996) presentó su conjunto de métodos y algoritmos para determinar la complejidad de las letras de canciones en inglés sobre la base de contar las sílabas de todas las palabras utilizadas en una letra de canción, expresándolo con la siguiente fórmula: $Tn = Tn-1 + f(n)$, donde Tn es es el número de sílabas en la canción; f(n) es el número de palabras.

Como señala el propio autor, "la definición en términos de sílabas evita algunos problemas relacionados con las cualidades métricas de las canciones, implica que las canciones se pueden analizar sin tocar la melodía y nos permite ser más precisos con respecto a las constantes en el análisis (si lo queremos)" [Darrah, 1996].

Desde principios de la década de 2000, ha habido un creciente interés en la comunidad lingüística en evaluar la complejidad y la legibilidad de los textos (inglés). Además de desarrollar fórmulas para evaluar la complejidad de los textos educativos para estudiantes de escuelas y colegios, los científicos también han hecho esfuerzos prometedores para desarrollar fórmulas de legibilidad para textos de "receptores adultos". Como señala M. M. Nevdah en este período (la década de 2000), los factores de complejidad de los textos en Inglés se consideran a nivel macro, ya que el público adulto no tiene dificultades para percibir un grupo de palabras o una oración completa [Nevdah, 2012].

2.1.2. Historia de los estudios de evaluación de la complejidad de los textos en español

La historia de la investigación en el campo de la complejidad de los textos en España se remonta a los años 50 del siglo XX, especialmente después de la publicación de la obra De R. Flesch "The Art of Plain Talk" (el Arte de la conversación simple), mencionada anteriormente.

En 1959, el investigador español Juan Fernández Huerta, en su trabajo "Medidas simples de legibilidad", presentó su fórmula de legibilidad para textos en español, cuyo desarrollo se basó en la fórmula de R. Flesch descrita anteriormente.

La fórmula de Fernández Huerta se presenta a continuación:

Dificultad: = 206.84 – (0.6 × número total de sílabas) - 1.02 × número total de palabras

El cálculo de la legibilidad se basa en el número de sílabas y la longitud de la oración en una muestra de texto de las primeras 100 palabras. Los resultados de la prueba se fijan en las puntuaciones de 0 a 100, donde el número más pequeño corresponde al texto de mayor complejidad, como se muestra a continuación en la *Tabla 2.*

Lecturabilidad	Nivel	Grado escolar
90-100	Muy fácil	Estudiantes de 4º grado
80-90	Fácil	Estudiantes de 5º grado
70-80	Comparativamente fácil	Estudiantes de 6 º grado
60-70	Normal (adultos)	Estudiantes de 7º y 8º grado
50-60	Bastante difícil	Preparación para estudio en la Universidad
30-50	Difícil	Cursos de Preparación
0-30	Muy difícil	Universidad (Especialización)

Tabla 2 Relación entre la Puntuación de la facilidad de lectura De J. Fernández-huerta y el nivel de educación del lector

Más tarde, en la década de 1980, científicos como B. Gilliam, S. C. Peña y L. Mountain lograron adaptar el gráfico de Fry mencionado anteriormente al español con la adición de un factor de corrección.

También en 1982, J.F. Fountain-Chambers, en su libro "Readability of primary-grade Spanish reading books: a correlational study of the Spaulding Coco et al", presenta una adaptación de la fórmula de la legibilidad de Spaulding al sistema de la lengua española. Esta fórmula se basa en el uso de palabras con un recuento de su densidad de presencia en el texto, así como en el porcentaje de palabras que no aparecen entre las palabras más utilizadas en español y la complejidad de la oración, que se mide por el promedio de la longitud de la oración en el texto de muestra [Fountain-Chambers, 1982]. El resultado de esta fórmula es un índice de nivel de dificultad para el cual se proporcionan las categorías de dificultad correspondientes.

La fórmula De J.F. Fountain-Chambers estimaciones de la complejidad del texto en español se dan a continuación en la *Figura 4.*

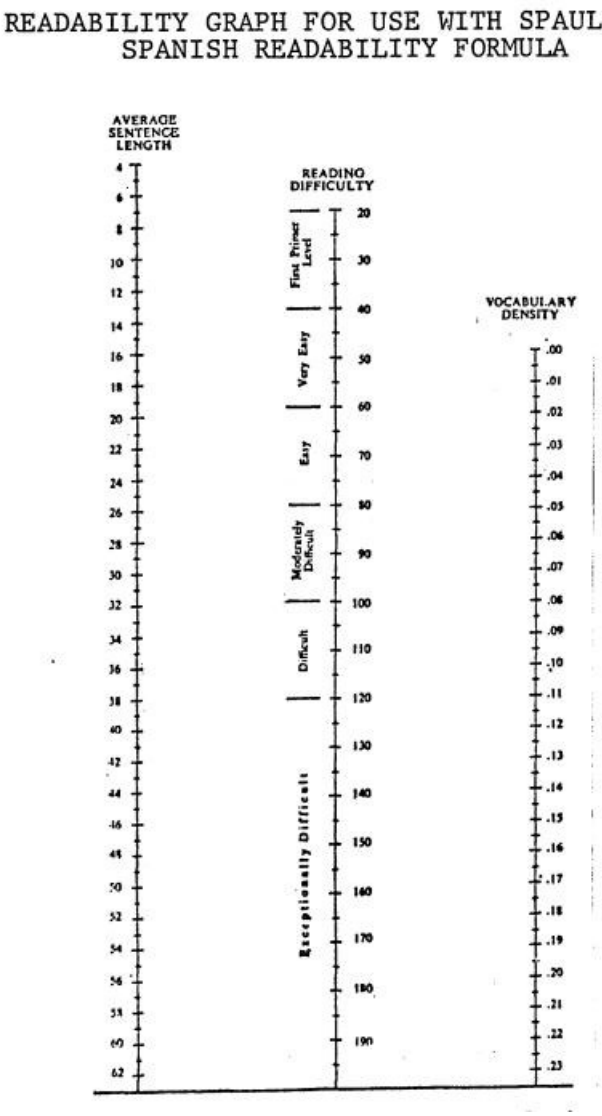


Fig. 4. Fórmula (gráfico) de Fountain-Chambers para estimar la complejidad del texto en español (basado en la fórmula de Spaulding)

Como puede ver, a la izquierda de la figura se muestra una escala que indica la longitud promedio de la oración. La escala en el centro del gráfico muestra las puntuaciones de dificultad de lectura, indicando el nivel de dificultad del texto. La tercera escala representa la densidad de las palabras utilizadas en el texto evaluado.

El análisis del gráfico de Fountain-Chambers permite concluir que cuanto más larga es la oración y la densidad de las palabras, mayor es la complejidad del texto y su legibilidad.

Por lo tanto, sobre la base de todo lo anterior, se puede concluir que el problema de evaluar la complejidad de los textos en Inglés y español ha atraído la atención de científicos de diferentes campos del conocimiento, no solo para identificar las características lingüísticas de la estructura de los textos escritos en dichos idiomas en los niveles micro y macro (palabra – oración - texto),

sino también para encontrar formas de facilitar el proceso de lectura y aprendizaje de estudiantes y estudiantes en escuelas e instituciones de educación superior.

También cabe señalar que, a pesar de la gran cantidad de fórmulas de legibilidad desarrolladas para textos en Inglés, su disponibilidad con respecto a los textos en español es bastante limitada, lo que indica la necesidad de adaptar estas fórmulas a las especificidades del español en el futuro.

2.2. Enfoques para determinar la complejidad de los textos

Sobre la base de la revisión teórica de la historia de la creación de fórmulas de legibilidad realizada anteriormente, se puede concluir que los científicos al desarrollar métodos para analizar la complejidad del texto recurrieron principalmente al uso de dos tipos de sistemas, exactamente: 1) sistemas cuantitativos de complejidad de texto (*quantitative systems of text complexity*) y 2) sistemas cualitativos de complejidad de texto (*qualitative systems of text complexity*).

Sistema cuantitativo de evaluación de la complejidad del texto.

La cuantificación de la complejidad del texto, como se ha indicado anteriormente, incluye la consideración de los siguientes indicadores:

1. Longitud de las oraciones;
2. Longitud de las palabras;
3. Cantidad de las oraciones;
4. Cantidad de las palabras;
5. Cantidad de las sílabas
6. El número de diferentes palabras complejas;
7. Porcentaje de palabras únicas;

Actualmente, para cuantificar la complejidad de los textos en inglés se utilizan programas de procesamiento automático del lenguaje natural como TextInspector, Compleat Lexical Tutor [Ziganshina, 2020, P. 389], etc., que se distinguen por un amplio conjunto de métricas calculables de texto, la visibilidad de los datos obtenidos y una base lingüística sustancial basada en el cuerpo nacional británico y el cuerpo del inglés americano moderno.

Sistema de evaluación cualitativa de la complejidad del texto.

Como señala Ch. R. Ziganshina: "Los científicos pudieron demostrar que la complejidad de los textos se debe en gran medida no tanto a las discrepancias cuantitativas como cualitativas

en los textos, como, por ejemplo, la coherencia del texto, su lógica, narrativa, etc." [Ziganshina, 2020, P. 389].

El análisis cualitativo de la complejidad del texto implica tener en cuenta los siguientes parámetros:

1. Tipo de texto (artístico, científico, etc.);
2. Narrativa del texto;
3. Construcción sintáctica (simplicidad/complejidad);
4. Especificidad de las unidades léxicas;
5. Coherencia referencial;
6. Coherencia “profunda”.

La implementación de una evaluación cualitativa de la complejidad de los textos en inglés en este momento se realiza a través de una plataforma online como Coh-Metrix [Ziganshina, 2020, P. 389], que determina el valor de cada parámetro en una escala de 0 a 100 puntos.

Así, el parámetro de "narrativa" indica la presencia en el texto de una trama, así como la presencia en el texto de un vocabulario coloquial, a menudo utilizado en la vida cotidiana, que el lector conoce bien.

La evaluación de la construcción sintáctica del texto implica la presencia de un grupo de características, como el número total de palabras en una oración, el número de tokens utilizados en una oración antes del verbo, el número de palabras de alta frecuencia para el uso, construcciones sintácticas familiares para el lector. Cuanto mayor sea este indicador, más simple será la estructura sintáctica de la oración.

La especificidad de los tokens presentes en el texto evaluado implica su capacidad para visualizar fácilmente en la mente del receptor imágenes mentales que son fáciles de entender y procesar. Como escribe Ch. R. Ziganshina, por ejemplo, la concreción de las palabras en el texto científico estudiado es "muy alta (99%), mientras que en el texto artístico la concreción de las unidades léxicas es aproximadamente 65% " [Ziganshina, 2020, P. 390].

La coherencia referencial se basa en repeticiones léxicas y sustituciones de nombres, en el papel de estos últimos a menudo actúan pronombres. La alta coherencia del texto se debe a la presencia en este último de palabras e ideas, cuya repetición frecuente en las oraciones y en el texto en general crea "cadenas" temáticas claras que unen el texto para el lector.

La coherencia “profunda” del texto se actualiza a través del vocabulario, mediante el cual las oraciones establecen relaciones causales, espaciales, temporales y otros tipos entre las palabras [Solnyshkina, 2015]. Un texto de baja coherencia suele ser más difícil de entender, ya que carece de una parte de las relaciones lógicas que el receptor tiene que reconstruir.

También cabe señalar que, además de los dos principales parámetros lingüísticos objetivos mencionados anteriormente para determinar la complejidad de los textos, hay otras clases de características lingüísticas que también tienen una influencia significativa en la precisión de la evaluación de la legibilidad del texto. Estos parámetros se consideran a continuación:

1. Indicios léxicos.

Esta clase de características incluye palabras comúnmente utilizadas, palabras raramente utilizadas (libros, ciencias, etc.), jerga, blasfemia, préstamos léxicos de otros idiomas, etc. Las palabras pueden ser polisémicas, monosilábicas/polisilábicas, pertenecen a cualquier parte específica del discurso (sustantivo, adjetivo, verbo, etc.). Estas características pueden distinguirse en porcentaje entre sí, por ejemplo, el número de unidades léxicas de varios valores en relación con el número total de palabras en el texto.

2. Indicios gramaticales.

Esta clase de características incluye:

1) En los sustantivos:

- tipo;
- número;
- persona;
- caso;
- declinación.

2) En los adjetivos:

- categoría (cualitativa, relativa, posesiva);
- grado de comparación (positivo, comparativo, excelente).

3) En el verbo:

- sistema de tiempos (presente, pasado, futuro);
- conjugación;
- reflexivo/no reflexivo.

Estos signos también pueden tener expresión en proporción.

3. Indicios semánticos.

Esta clase de características incluye parámetros de abstracción/concreción del significado de las palabras, es decir:

- nombres con el significado de los objetos físicos presentes en la realidad objetiva (casa, mesa, sofá, etc.);

- objetos concretos pero virtuales (sitio web, etc.);
- conceptos abstractos, incluidos los términos (existencia, gnoseología, etc.);
- sustancia (espíritu, materia, etc.).

Además, según I. Y. Misernov y L. A. Grashenko [Misernov, Grashenko, 2015], también hay una serie de factores subjetivos para evaluar la complejidad de los textos. Estos son factores como:

- Contenido informativo del texto. Este factor está relacionado con la repetibilidad en el texto de la misma información. Si dentro del texto hay una repetición frecuente de las mismas palabras, giros lingüísticos, frases, esto reduce el contenido informativo del texto.
- Factores relacionados con la personalidad del lector. Por lo tanto, la complejidad de la percepción del texto puede deberse a la profesión del lector, sus intereses, estilo de vida, características psicológicas, etc.;
- Factores relacionados con la edad del lector. Por ejemplo, si a un niño se le da un libro "adulto", no podrá evaluar de manera adecuada y completa la información que se le ofrece en el libro debido a la falta suficiente de una base intelectual e informativa para esto. Del mismo modo, si un adulto se compromete a estudiar un cuento de hadas para niños, entonces su informatividad para él, debido a su experiencia y conocimiento, es probable que sea cero.

Sobre la base del examen anterior de los enfoques para determinar la complejidad de los textos, en nuestra labor se destacará un conjunto de características que permitirán evaluar la complejidad de los textos en español e inglés.

2.3. Complejidad de las canciones en inglés

Hoy en día las tendencias para aprender inglés a través de las canciones son cada vez más relevantes. Recursos de Internet de canciones modernas como lyricsmode.com, englishclub.com, Crazylink, Lyricsgaps, etc. tienen como objetivo enseñar a todos los interesados los conceptos básicos del idioma inglés a través de letras de canciones en inglés distribuidas en diferentes niveles de dificultad. La base clave para esto fue el sistema lingüístico CEFR/MCER (marco común europeo de referencia para las lenguas), que implica seis niveles de dificultad para dominar el idioma inglés: A1 – principiante, A2 – elemental, B1 – umbral (intermedio), B2 – intermedio avanzado, C1 – avanzado, C2 – profesional.

A continuación, echemos un vistazo a las características gramaticales y léxicas de las canciones en inglés de cada uno de estos niveles.

Nivel gramatical.

En términos de gramática, los recursos anteriores se centran en enseñar al usuario las diferentes formas temporales del verbo, ya que en inglés tienen un sistema muy diversificado.

En los niveles A1 y A2, los estudiantes deben dominar las formas temporales del verbo inglés como *Present Simple* (presente simple): *Tears **stream** down your face when you lose something you cannot **replace*** (sitio web learnenglish-online.com); *Past Simple* (pasado simple): *Yesterday, all my troubles **seemed** so far away* (sitio web learnenglish-online.com); *Future Simple* (futuro simple): ***I'll write** home every day.*

En los niveles B1 y B2, a los estudiantes se les ofrece escuchar canciones que usan formas temporales más complejas (compuestas) de verbos, como *Present Continuous* (Estructura Sujeto + BE + Verbo (ING)), por ejemplo*: I **am working**; He **is talking** on the phone; They **are sleeping***; *Past Continuous* (Estructura Sujeto + *Was/Were* + Verbo (ING)), por ejemplo*: I **was dancing**; She **was reading**; They **were swimming***; Future Continuous (Estructura Sujeto + Verbo *will* + Verbo be + Verbo (ING)), por ejemplo: *I **will be waiting**; she **will be singing**; they **will be dancing***.

También aquí se consideran construcciones utilizando los tiempos *Present Perfect, Past Perfect, Future Perfect.*

En los niveles C1, C2 se estudian construcciones condicionales complejas basadas en canciones en inglés, por ejemplo:

Zero Conditional.

If it snows, we stay inside - Si nieva, nos quedaremos en casa;

First Conditional.

If you go to the party, you will meet my boyfriend – Si vas a la fiesta, conocerás a mi novio.

Second Conditional.

If I had a million dollars - Si tuviera un millón de dólares.

Third Conditional.

You should have said no – Deberías haber dicho que no.

Nivel léxico.

En los niveles A1, A2, las canciones presentan un vocabulario simple, a menudo utilizado en la vida cotidiana y cotidiana (*happy* – feliz, *sad* – triste, *friends* – amigos, *dreams* – sueños, etc.). El estribillo contiene muchas repeticiones léxicas, por ejemplo: *They said, "**Come** sail away, **come** sail away, **come** sail away with me, lads"*, etc.

En los niveles B1, B2 en las canciones en inglés, junto con el vocabulario simple, se observa el uso de un mayor número de palabras menos frecuentes para la comunicación cotidiana (*angels* – ángeles, *starship* – nave espacial, etc.), y también hay un uso frecuente de frases

(generalmente con dos componentes) en forma de sustantivo + sustantivo en caso indirecto (usando la preposición *of*), por ejemplo*: song of hope* – cancíon de esperanza, *pot of gold* – olla de oro. También hay frases sustantivas (de dos componentes), en las que un sustantivo es el principal, el segundo es dependiente, por ejemplo: *childhood* ***friends*** – **amigos** (principal) de la infancia (dependiente).

En los niveles C1, C2 en las canciones en inglés se observan los siguientes fenómenos léxicos:

1) omitir el sonido vocal en el medio de la palabra, que es característico del vocabulario del estilo coloquial: *Wond'ring* - en lugar de *Wondering (maravillarse);*

2) uso frecuente de la forma coloquial de los verbos:

a*) to go (to going): if we're really ever* ***gonna*** *get that far* - en lugar de *going*;

b) to want: *Hey, boy, I really* ***wanna*** *be wit' you*

c) abreviatura del verbo to be (ser), to have (tener) en forma de presente: *Do you know* ***there's*** *something wrong?* - en lugar de ***there is*** y como verbos auxiliares en el predicado compuesto: ***I've felt*** *it all along* - en lugar de ***I have felt****;*

d) también hay abreviaturas de sílabas en palabras, por ejemplo: *'Cause*-en lugar de *Because*.

e) omitiendo el fonema 'g' al final de la palabra: *You just waitin 'on the traffic jam to finish, girl; It's gettin' hot, crack a window, air it out; You're so amazin'.*

También en las canciones de estos dos niveles se observa el uso de una cantidad significativa de vocabulario asociado con la designación de diferentes eventos históricos, culturales, religiosos y fenómenos, por ejemplo: *How did they fit two on* ***the Ark****?* (El sagrario - descrito en la Biblia); *The razor that cut* ***van Gogh's*** *ear?* (Van Gogh - artista holandés) ; T*o grace* ***Hampden Park***? (Hempden Park es el estadio nacional de fútbol de Escocia), lo que requiere que el estudiante tenga un amplio conocimiento no solo del idioma Inglés, sino también de otras áreas de conocimiento.

Nivel fonético.

En los sitios y plataformas de Internet anteriores no se presentan los niveles de dificultad de las canciones en inglés en términos de fonética, por lo tanto, durante el análisis fonético independiente de las letras de canciones distribuidas en los 6 niveles de complejidad del sistema CEFR/MCER, destacamos las siguientes características.

En los niveles A1 - B2, la fonética de las letras de las canciones está estandarizada. En los niveles C1 – C2 las letras de las canciones fueron detectadas:

1. Combinaciones arbitrarias de fonemas vocales entre sí, así como vocales y consonantes para dar al texto una mayor expresividad emocional, por ejemplo, en la canción "What's my name" de Rihanna:

I need a boy to take it over

Lookin' for a guy to put in work, uh

Oh, a-whoa-a-oh

Oh, a-whoa-a-oh

Oh, na, na

What's my name? [Rihanna, "What's my name"].

2. Elongación objetivo de la vocal del fonema en el medio de la palabra, por ejemplo:

a) *Every door your enter I will **le-et** you in*

b) *I swear you got me Losin' my **mi-i-i-i-i-i-ind*** [Rihanna, "What's my name"].

Sobre la base de lo anterior en este párrafo, se puede concluir que, a medida que los niveles de las letras se vuelven más complejos, la forma estandarizada del inglés se queda cada vez más atrás, dando paso a varias desviaciones lingüísticas de las normas léxicas y fonéticas.

2.4. Complejidad de las canciones en español

En la actualidad, hay un gran número de sitios y aplicaciones que dan la oportunidad de aprender español a través de canciones. Estos son recursos de Internet como tlcdenia.es, speakasap.com, hispania-valencia.com, etc.

Plataformas de Internet tlcdenia.es y hispania-valencia.com. al igual que con las canciones en inglés discutidas anteriormente, basan sus enfoques para aprender español en el marco lingüístico CEFR/MCER (marco común europeo de referencia para las lenguas), que asume, como se mencionó, seis niveles de dificultad de dominio del idioma.

Teniendo en cuenta el sistema CEFR en plataformas educativas de Internet tlcdenia.es y hispania-valencia.com se presentan 6 niveles de aprendizaje de español a través de canciones: niveles A1, A2, B1, B2, C1, C2.

Considere a continuación las características gramaticales y léxicas de las canciones en español de cada uno de los niveles mencionados.

Nivel gramatical.

A nivel A1 (sitio web tlcdenia.es) se invita al estudiante a familiarizarse con las formas temporales del verbo español, como el presente de indicativo (presente), el pretérito perfecto

(pasado), el gerundio, ejemplos de uso del verbo gustar con sustantivos e indefinidos (sitios web tlcdenia.es, hispania-valencia.com).

En el nivel A2 el estudiante está invitado a aprender formas verbales temporales, como las perífrasis de futuro: querer + infinitivo (perífrasis de verbo gustar, en futuro), estudio de las características del infinitivo español y el imperfecto.

En el nivel B1 al estudiante se ofrece un conjunto de canciones en las que hay verbos en las forma de subjuntivo (el subjuntivo) y de imperativo (el imperativo), así como oraciones temporales, imperativo con una partícula negativa no (el imperativo negativo), verbo en forma de tiempo futuro (el futuro).

En el nivel B2, se invita a los lectores a familiarizarse con construcciones complejas en español, como las oraciones condicionales en forma de tiempo presente y pasado (condicional simple y compuesto y el pretérito pluscuamperfecto de subjuntivo), así como en las canciones de este nivel se usan nombres adjetivos y expresiones idiomáticas españolas.

Los niveles C1, C2 están diseñados para familiarizar a los estudiantes con expresiones idiomáticas españolas (géneros: música pop, rap).

Nivel léxico.

En el nivel A1 los recursos de Internet anteriores consideran la pluralidad de la palabra española gustar (1) gustar; 2) querer; 3) amar), así como el vocabulario español simple utilizado en la vida cotidiana y cotidiana (*yo te quiero; el atardecer; luna; Felicidad; romántico; casualidad*, etc.).

En el nivel A2 en las letras de las canciones españolas se observa el uso frecuente de frases (sustantivo + adjetivo), por ejemplo.: *el murmullo de sus silencios, los juguetes rotos, los amantes locos, los zapatos de charol*, etc.

En el nivel B1, las canciones españolas utilizan un vocabulario cotidiano, a menudo enmarcado en repeticiones léxico-sintácticas (***Si tienes*** *un hondo penar, piensa en mí.* ***Si tienes*** *ganas de llorar, piensa en mí;* ***Cuando*** *llores, también piensa en mí,* ***Cuando*** *quieras quitarme la vida).*

En el nivel B2, en las canciones se observa el uso de un cierto porcentaje de palabras complejas que no se usan a menudo en la vida cotidiana (críptico –misterioso, lapso – omisión, clandestino – secreto, atragantar – ahogarse, desatinar – perder la razón, fallar, una monja).

El nivel C1 incluye canciones con vocabulario cuya comprensión depende del grado de conocimiento general de fondo del estudiante. En este nivel se utilizan:

1) palabras de la realidad que denotan objetos y fenómenos culturalmente marcados de España: duro, peso (monedas españolas), etc.;

2) frases estables (modismos): *un cuento chino* (literalmente - falso);

3) nombres propios*: Almodóvar* (Pedro Almodóvar - director de cine español);

4) el uso de sufijos con un significado diminutivo en las palabras, por ejemplo: *una mordid**ita** - mordida; la orill**ita** - orilla*, etc.

En el nivel C2, en las canciones españolas hay un gran número de realidades, modismos, etc. españoles y latinoamericanos, cuya comprensión del significado solo es posible con una preparación lingüística muy seria del estudiante.

Nivel fonético.

Al analizar el perfil fonético de los cantantes más populares de la música pop española y latinoamericana moderna (como Daddy Yankee, Bad Bunny, Ozuna, etc.), se puede encontrar un gran número de variedades dialectales en la realización de los fonemas en comparación con la forma estandarizada de español (español neutral). Es la versión unificada la que constituye el estándar educativo en la enseñanza del español como lengua extranjera [Skachkova, 2021].

Está claro que en los niveles A1 a B2 en las canciones españolas solo se pueden presentar textos en forma estandarizada (en fonética, incluida), pero en los niveles C1, C2, cuando los estudiantes ya tienen suficiente conocimiento de los conceptos básicos del idioma estudiado, se pueden introducir textos con desviaciones dialectales para un conocimiento más amplio de las características de las variedades lingüísticas y el desarrollo de una perspectiva general. Para estos niveles (C1, C2), se pueden ofrecer textos de autores españoles (y latinoamericanos) contemporáneos ampliamente conocidos, con la característica bastante común de omitir el fonema s en el medio de la palabra y al final, como en la canción "Problema"de Daddy Yankee:

Ella siempre e' el tema

(En lugar de **es - (ella) es** – verbo de presente indicativo, 3 persona, singular)

Tú ere' un problema, problema

(En lugar de **eres - (tú) eres** – verbo de presente indicativo, 2 persona, singular)

Cómo me daña el sistema

Le dio hasta abajo y se le vio el gistro (Ey)

Lo má' cabrón que mis ojo' han visto… [Yankee, “Problema”].

(En lugar de **ojos,** plural)

Sobre la base de lo anterior en este párrafo, se puede concluir que, a medida que los niveles de las letras se vuelven más complejos, la estructura estandarizada de la lengua española en términos fonéticos está dando paso cada vez más a diferentes desviaciones.

Así, existen 2 enfoques principales para determinar la complejidad de los textos, a saber, cuantitativo y cualitativo. También se pueden utilizar características léxicas, gramaticales y semánticas para analizar la complejidad de los textos. En la investigación vamos a utilizar las características objetivas de las letras de las canciones, ya que son fáciles de identificar, y elegir entre un gran número de ellos los más necesarios. Al mismo tiempo, será necesario utilizar cuidadosamente fórmulas de legibilidad, ya que las letras de las canciones no pertenecen a áreas de conocimiento altamente especializadas y, por lo tanto, no contienen una gran cantidad de vocabulario profesional. Por lo tanto, las magnitudes simples, por ejemplo, la longitud promedio de las palabras, son adecuadas.

En el marco del sistema de recomendación se prevé la utilización de los siguientes indicadores cuantitativos:

- número total de palabras en una canción
- longitud media de la palabra
- número de palabras únicas (no repetidas)
- frecuencia de las palabras
- número de palabras comunes

Además de estos indicadores, también se tendrá en cuenta el número y la proporción de nombres, adjetivos y verbos, ya que estas partes del discurso tienen la mayor carga semántica.

Hay que añadir que hay una dificultad para calcular estas métricas, que se debe a la estructura de las letras de las canciones. Muy a menudo los compositores no ponen todos los signos de puntuación necesarios en las letras. Por ejemplo, como se muestra en la *Figura 5*, el texto carece de signos de puntuación (puntos) al final de las oraciones, lo que hace imposible estimar la longitud de tales oraciones.

[Intro]
Oh yeah

[Verse 1]
Maybe I don't really wanna know
How your garden grows
'Cause I just wanna fly
Lately, did you ever feel the pain
In the morning rain
As it soaks you to the bone?

Fig. 5. Letra de la canción "Oasis" - "Live Forever" genius.com

Además, no en todas partes en las canciones hay una división de las letras en estrofas, estribillos, etc. Por lo tanto, no se considerán como la longitud de la oración y el número de palabras en la oración, por lo que se decidió considerar la letra de cada canción como una oración separada.

También se considera obligatorio el uso de características léxicas para construir un sistema de recomendación:

- conocimiento y prevalencia de las palabras entre los hablantes nativos
- presencia y proporción de jerga y blasfemia
- número de significados semánticos de diferentes partes del discurso (sustantivo, adjetivo, verbo)
- número de fonemas, sílabas, morfemas y consonantes

Capítulo 3. DESARROLLO DE UN SISTEMA DE RECOMENDACIONES DE CANCIONES PARA EL APRENDIZAJE DE IDIOMAS

3.1. Enfoques para el desarrollo de un sistema de recomendación

Los sistemas de recomendación son métodos y herramientas de software que ofrecen a los usuarios los objetos más interesantes para ellos. El sistema de recomendación para la selección de canciones para aprender un idioma extranjero ofrece a los usuarios del servicio de música objetos del catálogo de música, distribuidos por su nivel de complejidad lingüística. Hoy en día, hay una serie de sistemas de recomendación de desarrollo para la selección de canciones y contenido musical teniendo en cuenta el interés de los usuarios. Los consideremos a continuación.

Se destacan los siguientes dos enfoques efectivos para construir sistemas de recomendación [D. S. Romashov, 2016]:

1. Filtrado colaborativo (método para hacer predicciones en sistemas de recomendación basados en las preferencias conocidas de los usuarios). La idea principal es encontrar usuarios o productos similares y usar sus datos para generar recomendaciones. Este método se usa ampliamente en tiendas en línea, servicios de transmisión de música y vídeo, así como en otras aplicaciones donde es importante ofrecer a los usuarios contenido que les interese.

2. Recomendaciones basadas en contenido (*content-based*). Los pasos básicos en este sistema son analizar el contenido de los elementos y crear un conjunto de sus criterios (géneros, etiquetas, palabras), averiguar qué criterios le gustan al usuario, comparar estos datos y obtener recomendaciones. Los criterios combinan usuarios y objetos en un solo sistema de coordenadas, y aquí todo es simple: si el punto del usuario y el objeto está cerca, es probable que el objeto atraiga al usuario.

También existen sistemas híbridos que son una combinación de estos dos enfoques. Por lo tanto, las recomendaciones basadas en contenido recopilan información sobre todos los usuarios, donde se presentan sus preferencias. Además, a cada objeto del catálogo de música que se puede recomendar a los usuarios se le atribuyen características que lo caracterizan (por ejemplo, canción, álbum, género, nivel de dificultad de la canción, etc.). Al filtrar conjuntamente, se recomiendan al usuario los objetos en los que otros usuarios con preferencias similares han expresado interés.

En su trabajo A. V. Melnikova describe las diversas etapas del desarrollo del sistema de recomendación entre los cuales se encuentran:

1. Selección de características del conjunto de letras de canciones (frecuencia media (densidad) de palabras en el texto, longitud media y mediana de las palabras en el texto, número total de unidades léxicas en el texto, número de tokens comunes en el texto, proporción de nombres sustantivos, adjetivos, adverbios, pronombres, verbos, numerales en el léxico del texto);

2. Agrupación de un conjunto de letras de canciones, teniendo en cuenta sus niveles de complejidad (validación de los algoritmos *DBSCAN y K-means* con el objetivo de detectar "para un conjunto de objetos X, un conjunto de marcadores (identificadores) del grupo Y" [Melnikova, 2020, P. 21-22]. Dado que las letras de las canciones en este caso se presentan en forma de ciertos rasgos, con la ayuda de estos algoritmos, se debe determinar el nivel de su complejidad.

Una interesante revisión teórica de los sistemas de recomendación en general se presenta en el trabajo de los investigadores italianos modernos Francesco Ricci, Lior Rokach y Bracha Shapira "Recommender Systems: Introduction and Challenges" (Sistemas de recomendación: introducción y problemas). En este artículo los autores consideran las siguientes características principales de los sistemas de recomendación de un objeto en particular:

1. Funciones de los sistemas de recomendación de música para mejorar la satisfacción del usuario. Un sistema de recomendación bien diseñado también puede mejorar la interacción del usuario con el sitio o la aplicación. El usuario encontrará recomendaciones interesantes para sí mismo, relevantes y, con una interacción humana correctamente organizada con una computadora, disfrutará del uso de dicho sistema.

2. Datos y fuentes de información. Esta característica del sistema de recomendación consta de tres componentes clave: objetos, usuarios y transacciones.

Objetos. Los objetos son aquellos elementos que están sujetos a recomendación. El valor de un objeto puede ser positivo si es útil para el usuario (interesante para él), o negativo si el objeto no cumple con las expectativas del usuario y este último tomó una decisión incorrecta al seleccionar el objeto dado.

Usuarios. Los usuarios de los sistemas de recomendación pueden tener una amplia variedad de propósitos y características. Para recomendaciones personalizadas e interacciones persona-computadora, los sistemas de recomendación utilizan la más amplia gama de información sobre los usuarios. Esta información se puede estructurar de varias maneras, y la elección de la información para modelar un sistema de recomendación está condicionada por el método de recomendación.

Transacciones. Las transacciones, según los autores del artículo, son datos del registro virtual de visitas del usuario, que almacenan información importante generada durante la interacción persona-computadora. Esta información es importante para construir el algoritmo de generación de recomendaciones utilizado por el sistema de recomendación. Por ejemplo, el registro de transacciones puede contener una referencia a un objeto seleccionado por el usuario y una descripción del contexto (por ejemplo, el objetivo/solicitud del usuario) para esa recomendación en particular [Ricci, Rokach, Shapira, 2015].

Adicionalmente, los autores del estudio entre los principales métodos utilizados en el desarrollo de sistemas de recomendación, además de los mencionados, destacan tales como:

- Método demográfico (*demographic*). Este tipo de sistema de recomendación ofrece objetos basados en los datos demográficos del perfil de usuario, ya que se supone que se deben crear diferentes recomendaciones para diferentes nichos demográficos;
- Método basado en el conocimiento (*knowledge-based*). Los sistemas de recomendación que se basan en la recopilación de información sobre las preferencias de los usuarios ofrecen objetos basados en el conocimiento específico (información) sobre qué propiedades específicas del objeto de interés corresponden a las necesidades y preferencias de los usuarios. Y también en qué medida este objeto es útil o necesario para el usuario final. Este método se basa en precedentes. En los sistemas de referencia que utilizan este método, la función de similitud evalúa en qué medida las necesidades del usuario (descripción del problema) se ajustan a las recomendaciones (solución del problema) [Ricci, Rokach, Shapira, 2015].

Sobre la base de todo lo anterior en este párrafo, se puede concluir que el desarrollo de sistemas de recomendación considerados y la descripción de su estructura y funcionamiento tienen una base informativa seria tanto para futuras simulaciones de sistemas de recomendación, incluido el campo de la enseñanza de idiomas extranjeros, como para mejorar y optimizar aún más los algoritmos básicos para construir tales sistemas.

Con respecto al sistema de recomendación proyectado, se considera que su base será un método basado en el conocimiento (*knowledge-based*), ya que utiliza los conocimientos adquiridos de alguna manera para obtener recomendaciones y, en la mayoría de los casos, estos conocimientos se agregan manualmente. Este conocimiento puede incluir datos como las relaciones entre los elementos.

Además, después de estudiar los materiales sobre los sistemas de recomendación de A. V. Melnikova, parece correcto incluir en el análisis la complejidad de características como la proporción de nombres, adjetivos y verbos.

El sistema de recomendación en sí se pretende crear de tal manera que el usuario tendrá la capacidad de seleccionar el idioma (inglés/español), la selección del elemento para el entrenamiento (vocabulario/fonética), la selección de la dificultad del idioma (se planea utilizar 3 o 6 niveles de dificultad). Además, el usuario podrá seleccionar canciones con un nivel de dificultad similar y/o dentro de un género específico.

3.2 Tecnologías usadas

Colab es una plataforma web para la ejecución colaborativa de código desarrollada por Google. Está diseñado para científicos, investigadores y desarrolladores que desean colaborar en proyectos en ciencia de datos, aprendizaje automático y otros campos. Con Colab, los usuarios pueden crear, editar y ejecutar Jupiter Notebook en línea, así como compartirlos con otros. La plataforma admite la mayoría de las bibliotecas y paquetes populares de ciencia de datos y aprendizaje automático, lo que permite a los usuarios integrarlos fácilmente en sus proyectos.

Python es un lenguaje de programación orientado a objetos, interpretado y de alto nivel. Admite paradigmas de programación como programación estructural, programación funcional, programación orientada a objetos y programación generalizada. Python tiene una sintaxis simple y legible que permite a los desarrolladores escribir código de manera rápida y eficiente. También tiene una gran biblioteca estándar que incluye módulos para trabajar con archivos, cadenas, bases de datos, redes y más.

Pandas es una biblioteca en Python diseñada para trabajar con datos en aplicaciones científicas e industriales. Proporciona un conjunto de herramientas para analizar y procesar datos, como leer y escribir datos en varios formatos, manipular datos, agrupar y agregar datos, visualizar datos y más. Pandas es una de las bibliotecas más populares y ampliamente utilizadas para trabajar con datos en Python y se usa a menudo en ciencia de datos, aprendizaje automático y análisis de datos.

Numpy es una biblioteca para el lenguaje de programación Python que proporciona herramientas para trabajar con matrices multidimensionales (matrices). Permite realizar diversas operaciones sobre matrices, como sumar, multiplicar, transponer, calcular la suma de elementos, entre otras. Numpy es una de las principales bibliotecas para trabajar con datos en Python y se usa ampliamente en investigación científica, ingeniería y otras áreas donde se requiere trabajar con matrices de datos.

Nltk es un conjunto de herramientas y bibliotecas para trabajar con lenguajes naturales. Se utiliza para el procesamiento de textos, clasificación de documentos, extracción de información y otras tareas, y proporciona una amplia gama de funciones para el análisis y procesamiento de textos, que incluyen tokenización, stemming, análisis, lematización, clasificación y más.

Spacy es una biblioteca abierta para el procesamiento del lenguaje natural (NLP) en Python. Está diseñado para proporcionar un procesamiento rápido y eficiente de datos de texto en idiomas naturales. Space proporciona herramientas para realizar diversas tareas de NLP, como analizar texto, detectar entidades con nombre, identificar partes del habla, resaltar frases, etc.

Sklearn es una poderosa biblioteca para el aprendizaje automático en Python. Proporciona muchos algoritmos de aprendizaje automático como clasificación, regresión, agrupación, etc. Sklearn es una biblioteca muy popular entre los científicos de datos y los investigadores, ya que es fácil de usar y tiene muchas características.

Plotly Es una biblioteca de Python para la visualización de datos. Proporciona una amplia gama de características para crear gráficos y gráficos interactivos y de alta calidad. Plotly se puede utilizar para crear diferentes tipos de gráficos.

PyCharm es un entorno de desarrollo integrado (*IDE*) para el lenguaje de programación Python. PyCharm proporciona varias herramientas de desarrollo de Python, como depuración, análisis de código, soporte para *VCS* (sistema de control de versiones) y más.

Kivy es un entorno gratuito y abierto para crear aplicaciones multiplataforma en Python. Le permite desarrollar aplicaciones para Android, iOS, Windows, Linux y otras plataformas sin tener que escribir código para cada plataforma por separado. Kivy utiliza el lenguaje Python para escribir código de aplicación.

3.3 Algoritmos de agrupación

Hay una gran cantidad de algoritmos de agrupamiento con diferentes enfoques de partición. En su trabajo A. V. Melnikova aconseja recurrir al uso de métodos de agrupamiento como *K-means* y *DBSCAN* para crear un sistema de recomendación.

El algoritmo *K-means* (también conocido como algoritmo de Lloyd) es un método de aprendizaje automático no supervisado utilizado para agrupar datos. Ayuda a agrupar objetos similares (por ejemplo, puntos de datos) en clústeres de tal manera que los objetos dentro de un clúster sean más similares entre sí que los objetos de otros clústeres. Cada grupo es un grupo de puntos, y el objetivo del algoritmo de *K-means* es minimizar la distancia cuadrática media entre los puntos dentro de un solo grupo. También existe otra variación de *K-means* que se llama el algoritmo de Elkan, que de hecho es la optimización alternativa de algoritmo de Lloyd, que utiliza la desigualdad triangular para evitar muchos cálculos de distancia en la distribución de puntos en grupos y para aumentar la velocidad.

El algoritmo *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*) es un algoritmo de agrupamiento de datos que se basa en la densidad de puntos en el espacio. *DBSCAN* identifica los grupos como regiones de alta densidad de puntos separadas por regiones de menor densidad. Permite la detección de clústeres de forma libre y puede detectar automáticamente ruidos en los datos.

En este caso la dificultad de elegir un algoritmo es el tamaño del conjunto de datos y establecer el número exacto de clústeres, por lo que dado que no hay forma de establecer el número de clústeres en *DBSCAN*, los algoritmos de *K-means* se eligieron para el estudio.

3.4 Descripción del dataset

Para seleccionar los datos con el fin de agrupar, necesitaba encontrar un conjunto de letras de canciones que cumpliera con los siguientes requisitos por lo menos:

- Tener la letra completa de la canción;
- Nombre de la canción;
- Artista;
- Especifica el género al que pertenece la canción (para recomendar al usuario en función de sus preferencias)

El resultado de la búsqueda fue un *dataset* con 5 millones de canciones en diferentes idiomas en el sitio kaggle.com. Este *dataset* incluye más de 3 millones de canciones en inglés y más de 270.000 en español, y los datos en sí son tomados del sitio web https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information. Los datos de ejemplo y los características del conjunto de letras se pueden ver más adelante en las *Tablas 3 y 4*.

	title	tag	artist	views	lyrics	language
0	Rap God	rap	Eminem	17575634	[Intro]\n"Look, I was gonna go easy on you not...	en
1	WAP	rap	Cardi B	16003444	[Intro: Cardi B, Al "T" McLaran & Megan Thee S...	en
2	Shape of You	pop	Ed Sheeran	14569727	[Verse 1]\nThe club isn't the best place to fi...	en
3	HUMBLE.	rap	Kendrick Lamar	11181199	[Intro]\nNobody pray for me\nIt been that day ...	en
4	The Hills	rb	The Weeknd	9291775	[Intro]\nYeah\nYeah\nYeah\n\n[Verse 1]\nYour m...	en

Tabla 3. Los datos de ejemplo

Característica	Conjunto de canciones en inglés	Conjunto de canciones en español
Nombre del conjunto de letras de canciones	Genius Song Lyrics	Genius Song Lyrics
Número de artistas	429060	40432
Número de letras	3316240	272110
Número de canciones	3374198	275432
Número de géneros	6	6

Tabla 4. Características de conjuntos de letras en inglés y español

El principal problema al usar esta matriz de datos fue su volumen y, en consecuencia, el tiempo empleado y la cantidad de uso de *RAM*. A este respecto se adoptaron varias decisiones que facilitaron la labor futura.

En primer lugar, como ha demostrado la práctica, es mejor descargar datos para su uso posterior en Colab a través de Google drive en formato de archivo, en lugar de directamente desde el disco duro de la computadora.

En segundo lugar, las canciones en inglés y español se dividieron en diferentes archivos.

En tercer lugar, el número de canciones en inglés se tomó casi igual al número de canciones en español, por lo que se seleccionaron las canciones en inglés más escuchadas. En total se seleccionaron 50.000 canciones en ambos idiomas.

A continuación se eliminaron las columnas adicionales que no se utilizaron para el análisis de complejidad. Además, para facilitar su uso, las letras de las canciones se pusieron en minúsculas, se eliminaron las divisiones en estrofas, estribillos y otra información de referencia, como el artista de la estrofa (un rasgo característico si la canción es interpretada por más de 1 cantante), y también se eliminó la puntuación, que, dada su originalidad y libertad en las letras, solo interfirió. Todo el proceso mencionado se puede ver en las *Figuras 6, 7 y 8*.

lyrics
[Verse 1: Bradley Cooper] Tell me somethin', girl Are you happy in this modern world? Or do you need more? Is there somethin' else you're searchin' for? [Refrain: Bradley Cooper] I'm fallin' In all the good times I find myself longing for change And, in the bad times, I fear myself [Verse 2: Lady Gaga] Tell me something, boy Aren't you tired tryna fill that void? Or do you need more? Ain't it hard keepin' it so hardcore? [Refrain: Lady Gaga] I'm falling In all the good times I find myself longing for change And, in the bad times, I fear myself [Chorus: Lady Gaga] I'm off the deep end, watch as I dive in I'll never meet the ground Crash through the surface where they can't hurt us We're far from the shallow now [Post-Chorus: Lady Gaga & Bradley Cooper] In the sha-ha, sha-hallow In the sha-ha, sha-la-la-la-low In the sha-ha, sha-hallow We're far from the shallow now [Bridge: Lady Gaga] Oh, ha, ah, ha Oh-ah, ha [Chorus: Lady Gaga] I'm off the deep end, watch as I dive in I'll never meet the ground Crash through the surface where they can't hurt us We're far from the shallow now [Post-Chorus: Lady Gaga & Bradley Cooper] In the sha-ha, shallow In the sha-ha, sha-la-la-la-low In the sha-ha, shallow We're far from the shallow now

Fig. 6. Letra original de la canción "Lady Gaga & Bradley Cooper" - "Shallow"

lyrics
tell me somethin girl are you happy in this modern world or do you need more is there somethin else youre searchin for im fallin in all the good times i find myself longing for change and in the bad times i fear myself tell me something boy arent you tired tryna fill that void or do you need more aint it hard keepin it so hardcore im falling in all the good times i find myself longing for change and in the bad times i fear myself im off the deep end watch as i dive in ill never meet the ground crash through the surface where they cant hurt us were far from the shallow now in the shaha shahallow in the shaha shalalalalow in the shaha shahallow were far from the shallow now oh ha ah ha ohah ha im off the deep end watch as i dive in ill never meet the ground crash through the surface where they cant hurt us were far from the shallow now in the shaha shallow in the shaha shalalalalow in the shaha shallow were far from the shallow now

Fig. 7. Letra procesada de la canción "Lady Gaga & Bradley Cooper" - "Shallow"

List of lemmatized lyrics
['tell', 'me', 'somethin', 'girl', 'be', 'you', 'happy', 'in', 'this', 'modern', 'world', 'or', 'do', 'you', 'need', 'more', 'be', 'there', 'somethin', 'else', 'youre', 'searchin', 'for', 'im', 'fallin', 'in', 'all', 'the', 'good', 'time', 'i', 'find', 'myself', 'long', 'for', 'change', 'and', 'in', 'the', 'bad', 'time', 'i', 'fear', 'myself', 'tell', 'me', 'something', 'boy', 'arent', 'you', 'tire', 'tryna', 'fill', 'that', 'void', 'or', 'do', 'you', 'need', 'more', 'aint', 'it', 'hard', 'keepin', 'it', 'so', 'hardcore', 'im', 'fall', 'in', 'all', 'the', 'good', 'time', 'i', 'find', 'myself', 'long', 'for', 'change', 'and', 'in', 'the', 'bad', 'time', 'i', 'fear', 'myself', 'im', 'off', 'the', 'deep', 'end', 'watch', 'a', 'i', 'dive', 'in', 'ill', 'never', 'meet', 'the', 'ground', 'crash', 'through', 'the', 'surface', 'where', 'they', 'cant', 'hurt', 'u', 'be', 'far', 'from', 'the', 'shallow', 'now', 'in', 'the', 'shaha', 'shahallow', 'in', 'the', 'shaha', 'shalalalalow', 'in', 'the', 'shaha', 'shahallow', 'be', 'far', 'from', 'the', 'shallow', 'now', 'oh', 'ha', 'ah', 'ha', 'ohah', 'ha', 'im', 'off', 'the', 'deep', 'end', 'watch', 'a', 'i', 'dive', 'in', 'ill', 'never', 'meet', 'the', 'ground', 'crash', 'through', 'the', 'surface', 'where', 'they', 'cant', 'hurt', 'u', 'be', 'far', 'from', 'the', 'shallow', 'now', 'in', 'the', 'shaha', 'shallow', 'in', 'the', 'shaha', 'shalalalalow', 'in', 'the', 'shaha', 'shallow', 'be', 'far', 'from', 'the', 'shallow', 'now']

Fig. 8. Letra procesada y lematizada de la canción "Lady Gaga & Bradley Cooper" - "Shallow"

Capítulo 4. ANÁLISIS DE CANCIONES EN INGLÉS

4.1 Características de las canciones en ingles

Como ha demostrado el análisis de la investigación y la literatura con respecto a la definición de la complejidad del texto, en el mundo de hoy no existe un sistema o fórmula coherente única para evaluar la complejidad de al menos los textos para leer en inglés. Cada técnica considerada tiene sus pros y sus contras, sus partidarios y sus críticos. En base a esto, y tomando nota de la especificidad de las letras, como características de las letras, identificamos las siguientes características después de la lematización utilizando la biblioteca nltk en Python:

1. Número y proporción de sustantivos, adjetivos y verbos en la letra. Hay un gran número de partes diferentes del habla en inglés, por lo que para simplificar la tarea, se han dividido en grupos separados. También se tuvo en cuenta el número promedio de significados posibles de cada sustantivo, adjetivo y verbo.

Para calcular estas características, se utilizó la función *pos_tag* de la biblioteca nltk utilizando el conjunto de etiquetas universal. La letra procesada de la canción (en forma de cadena) se sirve a la entrada de la función, que la divide en palabras individuales con un espacio y a partir de ellas se crea una lista de palabras individuales con la indicación de la parte del discurso.

A continuación se aplica a la lista de palabras resultante lematización para obtener la forma normal de la palabra (función *WordNetLemmatizer* de nltk) y la función *synsets* de la base de datos léxica *wordnet* para obtener el número de significados de la palabra. Después de eso, se calculan las proporciones necesarias para cada canción individual, dividiendo el número de cada parte del discurso por el número total de palabras de la canción *(noun ratio, verb ratio*, etc.) y dividiendo el número de significados de las palabras(dependiendo de la parte del discurso) por el número total de palabras de parte específica del discurso en la canción (*average noun meaning, average adj meaning, average verb meaning*). Se supone que cuantos más significados tenga una palabra, más difícil será percibir y comprender el texto. Los gráficos pueden ser vistos en la *Figura 9*.

2. Vocabulario de la canción. El número total de unidades léxicas (palabras) en el texto de la canción con el cálculo de las palabras únicas presentes en el texto, ya que se puede suponer que cuantas más palabras únicas haya en el texto, más difícil será entenderlo (*Figura 10*).

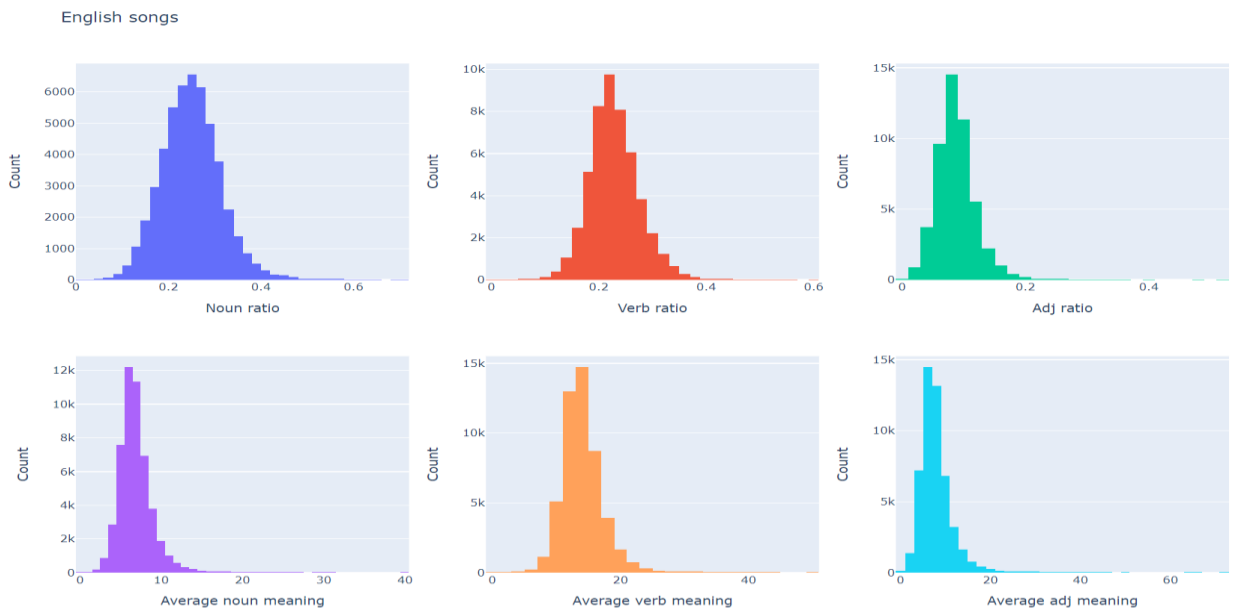


Fig. 9. Gráficos de proporciones de nombres, verbos, adjetivos y de sus números promedios de significados

3. Frecuencia media (densidad) de las palabras en el texto. En una canción a menudo se realiza una repetición múltiple de las mismas palabras. Algunas canciones también pueden repetir las mismas líneas muchas veces. Por lo tanto, se calcula el número promedio de la suma de la frecuencia de cada palabra en el texto (*Figura 10*).

4. Longitud media de las palabras en las letras de las canciones (*Figura 10*).

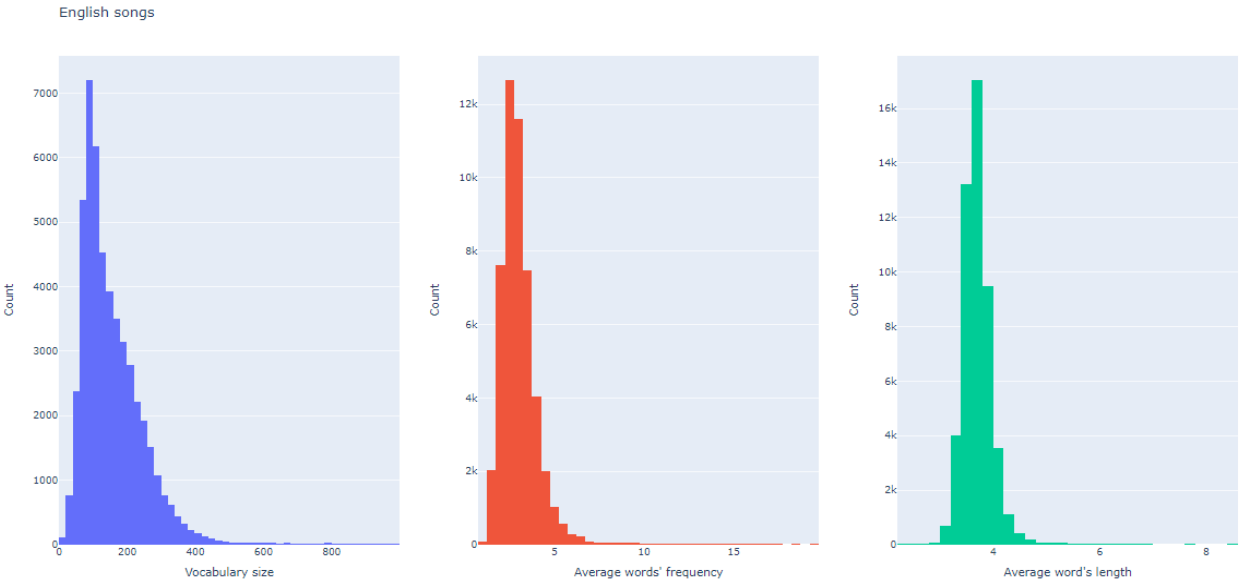


Fig. 10. Gráficos de cantidad de las palabras únicas, frecuencia media de las palabras en el texto y longitud media de las palabras

Para contar el número de palabras conocidas y de jerga, se decidió eliminar las *stopwords*. Para hacer esto, se utilizaron listas de *stopwords* de bibliotecas como nltk, spacy, gensim, sklearn, y también se creó su propia lista de *stopwords*, que, de hecho, incluye interjecciones, pronombres personales y posesivos, preposiciones, verbos auxiliares. Como punto importante, vale la pena

señalar que la lista de *stopwords* resultante es un conjunto, no una lista, ya que el conjunto es una estructura de datos hash, a diferencia de una lista, y permite búsquedas mucho más rápidas.

5. Prevalencia de palabras (*prevalences ratio*) y nivel de reconocimiento de palabras (*knowledge ratio*). Para ello, se utilizaron los resultados del estudio de Marc Brysbaert, Paweł Mandera, Samantha F. McCormick y Emmanuel Keuleers, que son una lista de casi 62.000 lemas en inglés que indica la prevalencia de cada palabra y el grado de reconocimiento (tiene valores de 0 a 1.0) entre los hablantes nativos [Brysbaert, Mandera, McCormick, Keuleers, 2019]. Prevalencia se refiere al porcentaje de personas que indican que conocen la palabra. En la práctica los porcentajes se transforman en valor z que se llama estimación estándar y que da una idea de lo lejos que está de la media de los datos. Se calculan las proporciones promedias de prevalencia y de nivel de reconocimiento a saber los valores de los indicadores mencionados de las palabras en la lista de investigación se calculan y se dividen por el número total de palabras en la canción.

6. La proporción total de palabras de jerga (*slang ratio*) en la letra de la canción, que se calcula mediante la búsqueda de la palabra en el diccionario de vocabulario callejero tomado del sitio web https://www.urbandictionary.com/ e incluye más de 100.000 palabras de jerga con la interpretación de su significado y posterior división por el número total de palabras en la canción.

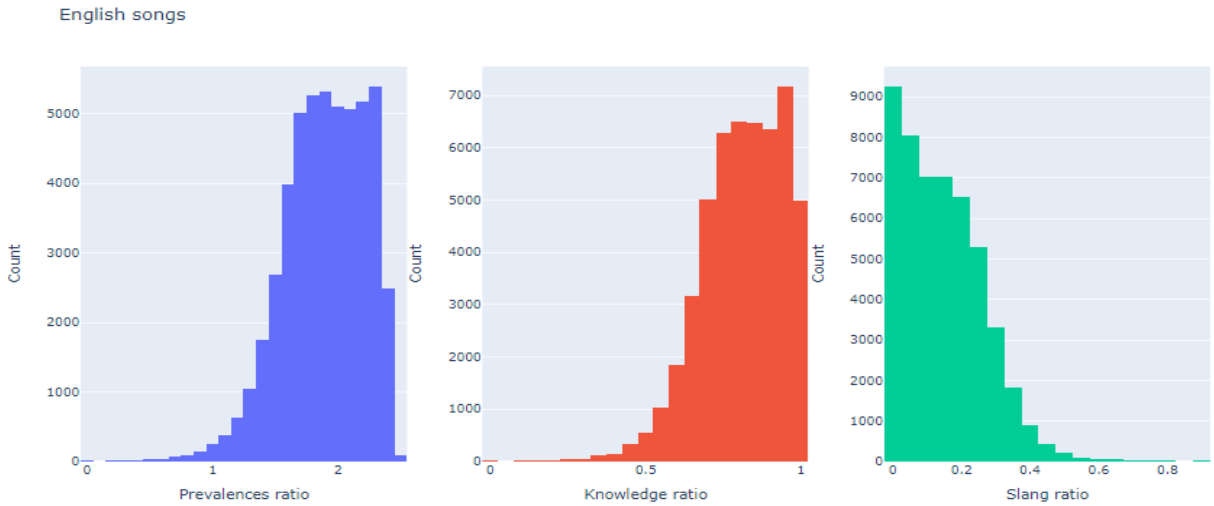


Fig. 11. Gráficos de la prevalencia de las palabras, el conocimiento entre los hablantes nativos y la proporción total de palabras de jerga

7. Como se mencionó anteriormente, para determinar el nivel de inglés, se utiliza el sistema CEFR, que incluye 6 niveles de dificultad: A1, A2, B1, B2, C1, C2. Cada nivel posterior difiere del anterior en el grado de complejidad del material lingüístico. Además, cada nivel fija su mínimo léxico, que debe ser dominado en un nivel específico por una persona que aprende inglés. Esta característica léxica también se puede utilizar para identificar la complejidad de las letras de las canciones. Además, se reveló que las *stopwrods* están principalmente presentes en los niveles A1 y A2.

8. Proporción de fonemas, sílabas y morfemas en palabras, suponiendo que cuanto mayor sea el número de fonemas, sílabas y morfemas en una palabra, más difícil será. Para el recuento de estas características también se utilizaron los resultados del estudio Marc Brysbaert, Paweł Mandera, Samantha F. McCormick y Emmanuel Keuleers, que contienen datos para más de 25.000 palabras que indican las características requeridas [Brysbaert, Mandera, McCormick, Keuleers, 2019]. Además, se utilizó el siguiente algoritmo para determinar la complejidad en fonética [https://www.geeksforgeeks.org/calculate-difficulty-sentence/]. Una palabra se considera difícil si tiene 4 consonantes consecutivas o el número de consonantes es mayor que el número de vocales. De lo contrario, las palabras son simples. La dificultad de la oración (en nuestro caso de la canción) se define como 5* (número de palabras difíciles) + 3* (número de palabras simples).

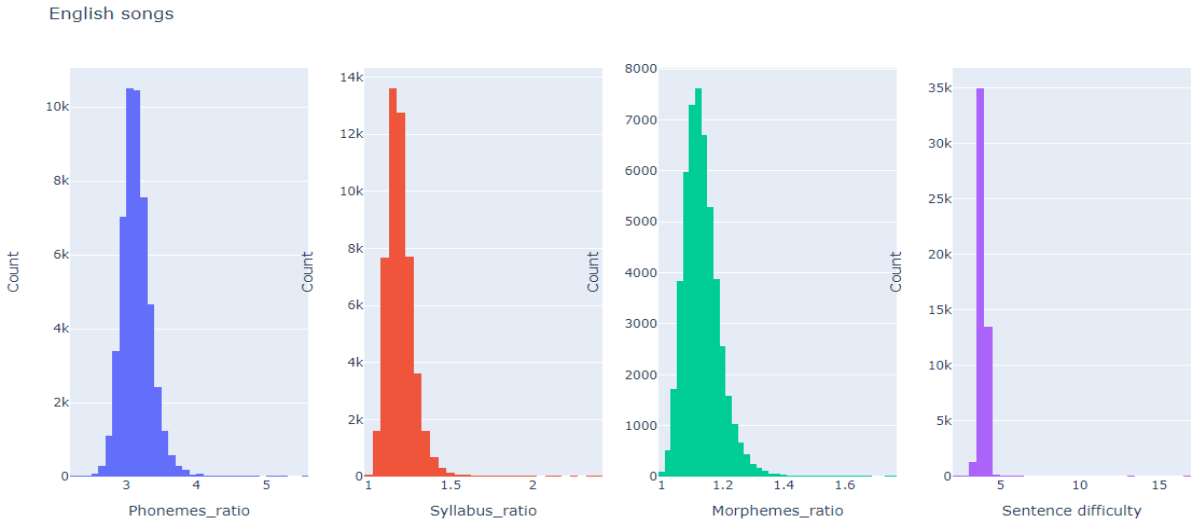


Fig. 12. Gráficos de número total de palabras de uso común, nivel promedio de prevalencia y proporción total de palabras de jerga

Mediante la construcción de la matriz de correlación (se puede ver en Anexo I) se identificaron las siguientes dependencias:

1. Cuanto mayor sea el número de palabras en la letra de la canción, mayor será el número de sustantivos, verbos y adjetivos con diferentes significados (correlación 0.96 - 0.98).

2. El número de palabras únicas en una canción depende en gran medida del número total de palabras en una canción (correlación 0.88).

3. Existe una relación obvia entre la prevalencia de una palabra y su reconocimiento para un hablante nativo (correlación 0.89).

4. Las otras características no tienen dependencias lineales explícitas entre sí.

4.2. Agrupación del conjunto de letras en ingles por dificultad

La tarea de agrupar es encontrar un conjunto de identificadores de clúster (etiquetas) Y para un conjunto de objetos X. Es necesario determinar la complejidad de las letras de las canciones expresadas como características diferentes. Dado que en la clasificación CEFR hay 6 niveles de dominio del inglés y del español, se decidió elegir este número como el número de grupos para el trabajo futuro. Teniendo en cuenta que a menudo los clústeres pueden superponerse y no tener límites claros, y considerando que la diferencia entre los niveles A1 y A2, B1 y B2, C1 y C2 frecuentemente no es muy significativa, también se decidió considerar el ejemplo de los 3 clústeres.

Antes de la agrupación, se eliminaron las columnas adicionales que contenían información de texto en lugar de numérica. Además, a la luz del hecho de que la complejidad del texto puede manifestarse de diferentes maneras (en vocabulario, en fonética, etc.), los signos se dividieron en dos casos: para identificar la complejidad léxica y la complejidad fonética.

Dificultad del vocabulario ingles

Al principio, se realizó un agrupamiento sin normalización de los datos con las características como el número de palabras por canción y el vocabulario de la canción, que, según muchas fuentes, se consideran entre las más importantes para analizar la complejidad de las letras. Se realizaron las divisiones en 3 y 6 grupos que se presentan a continuación en las *Tablas 5 y 6*.

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	25807.0	267.491766	20.0	205.00	274.0	338.00	443.0	25807.0	96.178401	5.0	73.00	93.0	116.00	281.0
1	196.0	3285.964286	2077.0	2655.75	3232.5	3846.50	4998.0	196.0	962.408163	461.0	742.75	884.0	1120.00	2017.0
2	26.0	12431.384615	9800.0	11522.25	12456.5	13516.25	16707.0	26.0	2027.884615	1228.0	1535.00	1994.5	2215.75	3672.0
3	110.0	6686.945455	5018.0	5616.75	6469.0	7718.25	9485.0	110.0	1394.845455	678.0	1130.50	1330.5	1603.75	2967.0
4	5398.0	907.713783	687.0	776.00	845.0	959.00	2133.0	5398.0	310.749907	16.0	252.00	297.0	353.00	981.0
5	18463.0	544.766993	368.0	468.00	534.0	616.00	781.0	18463.0	187.848400	7.0	150.00	185.0	223.00	441.0

Tabla 5. División en 6 grupos usando el número de palabras por canción y el vocabulario de la canción sin normalización

Las tablas contienen tal información como el número de clústeres, el número de canciones en el clúster, el valor medio de la característica en la que se basó la separación, así como sus valores mínimos y máximos y los valores de cuartiles.

Se puede observar que la agrupación no permitió distinguir 3 o 6 grupos completos y claros de la complejidad de canciones.

Esto se debe a que en ambas configuraciones vemos una intersección significativa de la característica del vocabulario de la canción en varios clusters a lo largo de los diferentes cuartiles.

Por ejemplo, los clusteres (*levels*) 0, 4 y 5 (en la configuración de 6 clústeres) se superponen en intervalos (resaltados con azul en la *Tabla 5*):

- [5;281] (0 *level*)
- [16;981] (4 *level*)
- [7;441] (5 *level*)

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	49688.0	440.834608	20.0	270.0	397.0	571.0	2272.0	49688.0	153.796188	5.0	90.0	132.0	201.0	981.0
1	249.0	4157.200803	2261.0	3034.0	3816.0	5233.0	7029.0	249.0	1091.369478	461.0	817.0	1014.0	1267.0	2967.0
2	63.0	9913.142857	7146.0	7990.5	8815.0	11849.5	16707.0	63.0	1729.571429	1048.0	1346.0	1603.0	2006.5	3672.0

Tabla 6. División en 3 grupos usando el número de palabras por canción y el vocabulario de la canción sin normalización

A continuación se normalizaron los datos y se eliminaron las enormes letras que contenían más de 2000 palabras (349 canciones), seguidas de la agrupación en la configuración anterior. Los resultados se presentan en las *Tablas 7 y 8*.

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	11115.0	0.240075	0.142785	0.216709	0.238987	0.262278	0.422785	11115.0	0.168831	0.002049	0.145492	0.168033	0.192623	0.297131
1	8050.0	0.325747	0.216203	0.297722	0.324051	0.351392	0.538734	8050.0	0.235413	0.040984	0.207992	0.235656	0.263320	0.387295
2	11783.0	0.086126	0.000000	0.067342	0.090127	0.108354	0.156962	11783.0	0.069877	0.000000	0.055328	0.070697	0.085041	0.145492
3	513.0	0.685357	0.475443	0.592405	0.655190	0.763038	1.000000	513.0	0.490191	0.221311	0.426230	0.474385	0.537910	1.000000
4	14824.0	0.162590	0.097722	0.141266	0.160506	0.181772	0.290633	14824.0	0.109353	0.001025	0.090164	0.106557	0.127049	0.204918
5	3366.0	0.439992	0.286582	0.400506	0.433924	0.474430	0.666835	3366.0	0.321618	0.030738	0.283811	0.319672	0.357582	0.588115

Tabla 7. División en 6 grupos usando el número de palabras por canción y el vocabulario de la canción con normalización

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	17938.0	0.261768	0.142785	0.224304	0.257215	0.296203	0.483544	17938.0	0.185215	0.002049	0.152664	0.183402	0.216189	0.354508
1	6045.0	0.433225	0.262785	0.368608	0.407089	0.463797	1.000000	6045.0	0.315173	0.030738	0.262295	0.300205	0.349385	1.000000
2	25668.0	0.125991	0.000000	0.093671	0.128608	0.160506	0.258228	25668.0	0.090481	0.000000	0.069672	0.089139	0.110656	0.204918

Tabla 8. División en 3 grupos usando el número de palabras por canción y el vocabulario de la canción con normalización

Ahora hay una distinción bastante clara entre los clústeres, tanto en la configuración con 3 clústeres (resaltados con azul en la *Tabla 8*) como en la configuración con 6 clústeres, debido a que los cuartiles de las clases se cruzan significativamente menos que antes de la normalización.

En total, durante el proceso de recopilación de información, se formaron más de 30 características de canciones que podrían usarse potencialmente para analizar la complejidad. Sin

embargo, algunas de las características son similares en su significado, por lo que fue necesario realizar una serie de lanzamientos experimentales para identificar las características que tienen menos y más impacto en la asignación de un objeto a un grupo en particular.

Como el número total de palabras y el tamaño del vocabulario son características importantes, parece evidente el uso de características como la tasa de prevalencia de palabras y el nivel de reconocimiento entre los hablantes nativos en el análisis. Los resultados se reflejan en las *Tablas 9 y 10*.

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	12748.0	0.202834	0.000000	0.147848	0.201013	0.254684	0.593924	12748.0	0.142728	0.000000	0.100154	0.140369	0.181352	0.610656
1	12558.0	0.356385	0.168101	0.286582	0.336203	0.401519	1.000000	12558.0	0.261022	0.011270	0.205943	0.248975	0.298924	1.000000
2	24345.0	0.143239	0.000000	0.098228	0.138228	0.183797	0.483544	24345.0	0.100746	0.001025	0.072746	0.094262	0.122951	0.299180
	Prevalences ratio							Knowledge ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	12748.0	0.595747	0.000000	0.548524	0.613472	0.663142	0.741633	12748.0	0.657671	0.000000	0.608470	0.676917	0.728703	0.880226
1	12558.0	0.707727	0.353017	0.664062	0.708972	0.751950	0.923774	12558.0	0.776807	0.395479	0.732039	0.779247	0.822858	0.985274
2	24345.0	0.846601	0.640914	0.797545	0.852745	0.899078	1.000000	24345.0	0.915252	0.754461	0.867392	0.924105	0.968617	1.000000

Tabla 9. División en 3 grupos usando el número de palabras por canción, el vocabulario de la canción, la proporción de prevalencia de palabras y el nivel de reconocimiento entre los hablantes nativos con normalización

De acuerdo con los resultados, se observa una división bastante buena en 3 grupos: Por ejemplo, se puede ver que en relación con las características como nivel de prevalencia y reconocimiento de palabras, los grupos se superponen ligeramente:

El 75% de los datos del clúster 0 tiene reconocimiento de palabras inferior a 0.72, mientras que el 75% del clúster 1 tiene valores superiores a 0.73 y el 75% del clúster 2 tiene valores superiores a 0.86, que a su vez es mayor que el 75% de los datos del clúster 1 (resaltados con naranja en la *Tabla 9*). Se pueden observar indicadores similares con respecto al nivel de prevalencia, ya que estos indicadores están relacionados.

En cuanto a los indicadores como el número de palabras y el vocabulario, la situación es similar. El 75% de los datos del clúster 0 tiene valores del vocabulario inferiores a 0.18, mientras que el 75% de los datos del clúster 1 superior a 0.2 y el 75% del clúster 2 inferior a 0.12 (resaltados con azul en la *Tabla 9*), y respectivamente el 75% de los datos del clúster 0 tiene número de palabras inferior a 0.25, el 75% del clúster 1 superior a 0.28 y el 75% del clúster 2 inferior a 0.18 (resaltados con verde en la *Tabla 9*).

Aquí se puede observar que hay una situación en la que el grupo tiene los niveles más altos del vocabulario y del número total de palabras y, al mismo tiempo, los niveles bastante altos de prevalencia y reconocimiento de palabras (*level* 1 en la configuración de 3 clústeres

en la *Tabla 9*). Esto puede explicarse, por ejemplo, por el hecho de que las canciones del clúster contienen muchas palabras diferentes pero simples.

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	3932.0	0.192710	0.000000	0.134177	0.187848	0.246076	0.588354	3932.0	0.129016	0.000000	0.088115	0.122951	0.165984	0.610656
1	7731.0	0.282975	0.150380	0.238481	0.275949	0.321519	0.538734	7731.0	0.203835	0.011270	0.163934	0.199795	0.241803	0.408811
2	16355.0	0.126924	0.000000	0.089873	0.124557	0.162025	0.369114	16355.0	0.088927	0.001025	0.067623	0.086066	0.107582	0.243852
3	3191.0	0.486201	0.288608	0.415443	0.457722	0.521013	1.000000	3191.0	0.356024	0.030738	0.300205	0.341189	0.393443	1.000000
4	9680.0	0.146867	0.000506	0.113418	0.149367	0.183291	0.309367	9680.0	0.107269	0.004098	0.079918	0.104508	0.132172	0.238730
5	8762.0	0.291483	0.138228	0.247089	0.289114	0.333671	0.518481	8762.0	0.209779	0.020492	0.174180	0.207992	0.243852	0.394467
	Prevalences ratio							Knowledge ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	3932.0	0.493745	0.000000	0.454771	0.512905	0.553123	0.617733	3932.0	0.549926	0.000000	0.509890	0.569148	0.611265	0.801041
1	7731.0	0.783568	0.636630	0.746022	0.780340	0.814423	0.940313	7731.0	0.853829	0.763195	0.816225	0.850793	0.885528	0.996487
2	16355.0	0.879370	0.722107	0.846449	0.882663	0.913228	1.000000	16355.0	0.947877	0.840313	0.918446	0.953413	0.981341	1.000000
3	3191.0	0.687886	0.285643	0.644491	0.690271	0.733478	0.906437	3191.0	0.757598	0.323410	0.711561	0.760431	0.805898	0.977278
4	9680.0	0.718922	0.518265	0.682458	0.722970	0.760353	0.822753	9680.0	0.786110	0.633128	0.748533	0.790234	0.829310	0.941014
5	8762.0	0.634245	0.401049	0.603015	0.638081	0.669867	0.721062	8762.0	0.699397	0.521376	0.666766	0.703121	0.737676	0.808955

Tabla 10. División en 6 grupos usando el número de palabras por canción, el vocabulario de la canción, la proporción de prevalencia de palabras y el nivel de reconocimiento entre los hablantes nativos con normalización

La división en 6 grupos resultó menos clara. Por ejemplo, los clústeres 0, 2 y 4 se superponen fuertemente en indicadores como el vocabulario y el número total de palabras (resaltados con azul en la *Tabla 10*). Del mismo modo, el clúster 1 y el clúster 5 se superponen en los mismos indicadores (resaltados con naranja en la *Tabla 10*). También se observan intersecciones significativas en los clusteres mencionados en nivel de prevalencia y reconocimiento de palabras.

Durante una serie de lanzamientos, se utilizaron dos variedades de *K-means*: el algoritmo de Lloyd y el algoritmo de Elkan. Esto se hizo con el objetivo para determinar qué algoritmo da un agrupamiento más claro. Al final, los resultados fueron muy similares, sin embargo, el algoritmo de Elkan dio una separación ligeramente mejor, por lo que solo se usó en el futuro.

Así mismo con el fin de evitar una posible dualidad y ambigüedad más grandes en los resultados en lo sucesivo, se decidió seguir agrupando en 3 grupos. Porque cuando se agrega para analizar características adicionales, es muy probable que la configuración en 6 clústeres muestre resultados aún peores.

En el siguiente enfoque, se tomaron dos características más, como el nivel de palabras de jerga en la canción y el nivel de palabras desconocidas, que no cayeron ni en la lista de 62.000 lemas en inglés [Brysbaert, Mandera, McCormick, Keuleers, 2019], ni en la lista de jerga (ver *Tabla 11*).

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	9892.0	0.223735	0.000000	0.151392	0.217722	0.286582	0.846582	9892.0	0.155435	0.000000	0.101434	0.150615	0.200820	0.859631
1	22637.0	0.138901	0.000000	0.094684	0.133165	0.176203	0.483544	22637.0	0.097691	0.001025	0.070697	0.091189	0.117828	0.360656
2	17122.0	0.303171	0.051646	0.222785	0.288101	0.362025	1.000000	17122.0	0.222000	0.019467	0.157787	0.211066	0.270492	1.000000
	Prevalences ratio							Knowledge ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	9892.0	0.563614	0.000000	0.526496	0.581493	0.620568	0.721851	9892.0	0.622947	0.000000	0.585174	0.642883	0.683220	0.801041
1	22637.0	0.853214	0.545999	0.810655	0.859749	0.902333	1.000000	22637.0	0.922364	0.677504	0.881366	0.931430	0.971375	1.000000
2	17122.0	0.712723	0.446583	0.673572	0.710160	0.749490	0.912361	17122.0	0.781404	0.582416	0.741424	0.778425	0.819039	0.977278
	Slang ratio							Non_present ratio						
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
level																
0	9892.0	0.332507	0.103500	0.0	0.282639	0.328571	0.382609	1.000000	9892.0	0.070820	0.098613	0.0	0.016097	0.037403	0.078611	1.000000
1	22637.0	0.054606	0.052405	0.0	0.000000	0.041509	0.089431	0.258824	22637.0	0.023032	0.039002	0.0	0.000000	0.006999	0.030453	0.341443
2	17122.0	0.193617	0.061162	0.0	0.152074	0.196648	0.239130	0.395753	17122.0	0.036911	0.038734	0.0	0.013741	0.027198	0.046179	0.377668

Tabla 11. División en 3 grupos usando el número de palabras por canción, el vocabulario de la canción, la proporción de prevalencia de palabras, el nivel de reconocimiento entre los hablantes nativos, el nivel de palabras de jerga en la canción y el nivel de palabras desconocidas con normalización

Al igual que en la situación anterior, hay una división bastante buena en grupos. Se puede observar que con respecto a características tales como nivel de prevalencia y reconocimiento de palabras, los grupos se superponen ligeramente:

Por ejemplo, el 75% de los datos del clúster 0 tiene reconocimiento de palabras inferior a 0.68, mientras que el 75% del clúster 2 tiene valores superiores a 0.74 y el 75% del clúster 1 tiene valores superiores a 0.88, que a su vez es mayor que el 75% de los datos del clúster 2 (resaltados con naranja en la *Tabla 11*). Se pueden observar indicadores similares con respecto al nivel de prevalencia, ya que estos indicadores están relacionados.

En cuanto al nivel de palabras de jerga, el 75% de los datos del clúster 0 tiene valores superiores a 0,28, mientras que el 75% de los datos del clúster 1 tiene menos de 0.09 y el 75% de los datos del clúster 2 tiene menos de 0.23 (resaltados con azul en la *Tabla 11*).

La situación es un poco peor en lo que respecta al vocabulario y al número total de palabras. Mientras que el 75% de los datos del clúster 1 tiene número total de palabras inferior a 0.17, el 75% de los datos del clúster 2 tiene valores superiores a 0.22, lo que tiene una intersección más significativa con el clúster 0, donde solo el 50% de los datos tiene valores inferiores a 0.21 (resaltados con verde en la *Tabla 11*). Por lo que corresponde al vocabulario el 75% de los datos

del clúster 1 tiene valores inferiores a 0.11 y el 75% de los datos del clúster 0 tiene valores superiores a 0.10, pero el 75% de los datos del clúster 2 tiene valores superiores a 0.15, que se superponen con el clúster 0 (resaltados con rojo en la *Tabla 11*).

Además hay una situación en la que el clúster 0 con la tasa de prevalencia de palabras más baja, el nivel de reconocimiento más bajo, los niveles de jerga y de palabras desconocidas más altos tiene un significado intermedio en el número total de palabras y vocabulario. Esto parece deberse al hecho de que este grupo contiene palabras complejas y poco conocidas.

Dado que el clúster 1 tiene los niveles más bajos de jerga, de palabras desconocidas, del vocabulario, del número palabras, y al mismo tiempo los niveles más altos de prevalencia y de reconocimiento de palabras, se puede asumir que corresponde a A1-A2. El clúster 0 en comparación con el clúster 2 tiene más jerga, más palabras desconocidas y el nivel más bajo de reconocimiento de palabras. El clúster 2 tiene vocabulario más grande y mayor número de palabras en comparación con el clúster 0, lo que puede explicarse por el hecho de que las canciones del clúster 2 contienen palabras más fáciles y comunes. Por lo tanto, se decidió que el clúster 0 corresponde a C1-C2 y el clúster 2 corresponde a B1-B2.

El uso de características como la longitud media de la palabra y la frecuencia media de la palabra no produjo ninguna mejora específica, e incluso a la inversa, empeoró la separación de los clústeres, ya que se superponen en casi todo el rango de valores (ver *Tabla 12*).

	Average words' frequency							Average word's length						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	9891.0	0.021296	0.000000	0.014789	0.019635	0.025510	0.224490	9891.0	0.220300	0.000000	0.191685	0.213748	0.241017	1.000000
1	22639.0	0.021187	0.000000	0.013410	0.019017	0.025994	1.000000	22639.0	0.210371	0.062288	0.181275	0.206002	0.234387	0.706029
2	17121.0	0.019684	0.002679	0.014120	0.018305	0.023446	0.379592	17121.0	0.217497	0.029825	0.194164	0.214473	0.237708	0.639937

Tabla 12. División en 3 grupos usando el número de palabras por canción, el vocabulario de la canción, la propoción de prevalencia de palabras, el nivel de conocimiento entre los hablantes nativos, el nivel de palabras de jerga en la canción, el nivel de palabras desconocidas, la longitud media de la palabra y la frecuencia media de la palabra con normalización

Como resultado, después de una serie de lanzamientos, se identificaron las características que menos afectan la atribución de un objeto a un grupo en particular(al usarlos en el análisis, la claridad de la división por clústeres no cambia), o interfieren con la atribución única de un grupo a un nivel particular de idioma(cuando se aplican la agrupación se deteriora), a saber: el número absoluto y relativo de sustantivos/verbos/adjetivos, el número promedio de significados de la palabra, la longitud promedia de la palabra, la frecuencia promedia de la palabra, el número de palabras de diferentes niveles de idioma A1-C2.

Por lo tanto, se identificaron 6 rasgos importantes: número de palabras, vocabulario, nivel de jerga, nivel de prevalencia, nivel de reconocimiento y de palabras desconocidas.

Dificultad de la fonética inglesa

A continuación, se agruparon los datos con el número de grupos iguales a 3 y usando 4 características para analizar la complejidad fonética de la canción. A partir de los resultados, se puede ver que se obtuvieron 3 grupos completos (*Tabla 13*).

Se puede observar que en el indicador de proporción de fonemas, el 75% de los datos del clúster 0 tiene valores inferior a 0.23, el 75% del clúster 2 tiene valores superiores a 0.25 y el 75% del clúster 1 tiene valores superiores a 0.32 (resaltados con azul en la *Tabla 13*).

En la proporción de sílabas el 75% de los datos del clúster 0 tiene valores inferiores a 0.11, el 75% del clúster 2 tiene valores superiores a 0.12 y el 75% del clúster 1 tiene valores superiores a 0.19 (resaltados con naranja en la *Tabla 13*).

En la proporción de morfemas, el 75% de los datos del clúster 0 tiene valores inferiores a 0.14, el 75% del clúster 2 tiene valores superiores a 0.17 y el 75% de los datos del clúster 1 tiene valores superiores a 0.26 (resaltados con verde en la *Tabla 13*), lo que en general indica una buena división de los clústeres. Suponiendo que las palabras con más fonemas, sílabas y morfemas son más propensos a corresponder a un nivel más alto de la lengua, entonces clúster 0 corresponde a A1-A2, clúster 1 corresponde a C1-C2 y clúster 2 corresponde a B1-B2.

	Phonemes_ratio							Syllabus_ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	21581.0	0.207921	0.000000	0.187429	0.211232	0.231707	0.365711	21581.0	0.094735	0.0	0.074785	0.094610	0.113672	0.257053
1	5938.0	0.364507	0.204472	0.327974	0.352774	0.386640	1.000000	5938.0	0.232481	0.0	0.194130	0.221027	0.256475	1.000000
2	22132.0	0.275608	0.147175	0.253885	0.273439	0.296092	0.463607	22132.0	0.149502	0.0	0.127386	0.147177	0.169404	0.384907
	Morphemes_ratio							Sentence difficulty						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	21581.0	0.113707	0.000000	0.089517	0.115789	0.139683	0.265460	21581.0	0.115507	0.000000	0.107650	0.116342	0.124036	0.280818
1	5938.0	0.315808	0.021164	0.268388	0.300494	0.345091	1.000000	5938.0	0.122052	0.028034	0.112870	0.122312	0.130538	1.000000
2	22132.0	0.197442	0.034934	0.170370	0.195011	0.223553	0.399151	22132.0	0.119006	0.058524	0.111817	0.119841	0.126633	0.763919

Tabla 13. División en 3 grupos usando proporción de fonemas, sílabas, morfemas y consonantes en palabras

Capítulo 5. ANÁLISIS DE CANCIONES EN ESPAÑOL

5.1 Características de las canciones en español

Entre las características de las canciones en español se tomaron casi similares a las que se destacaron en inglés. También se realizó una lematización preliminar, pero fue necesario utilizar la biblioteca pública spacy, ya que la biblioteca nltk no tiene lematizador en español. En particular, se utilizaron las siguientes características:

1. Número y proporción de nombres, adjetivos y verbos en el léxico de la letra (*noun ratio, adj ratio, verb ratio*), que se muestran en la *Figura 13*.

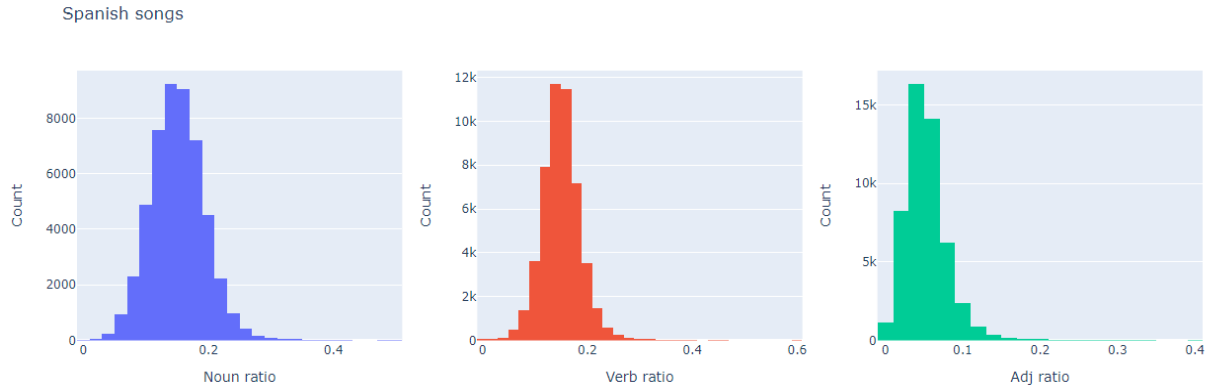


Fig. 13. Gráficos de proporciones de nombres, verbos, adjetivos

2. Número total de palabras (*word count*) en la letra de la canción, así como el vocabulario de la canción (*vocabulary size*), es decir, el número de palabras únicas.

3. Frecuencia media de las palabras y longitud media de las palabras (*average words' frequency* y *average word's length*).

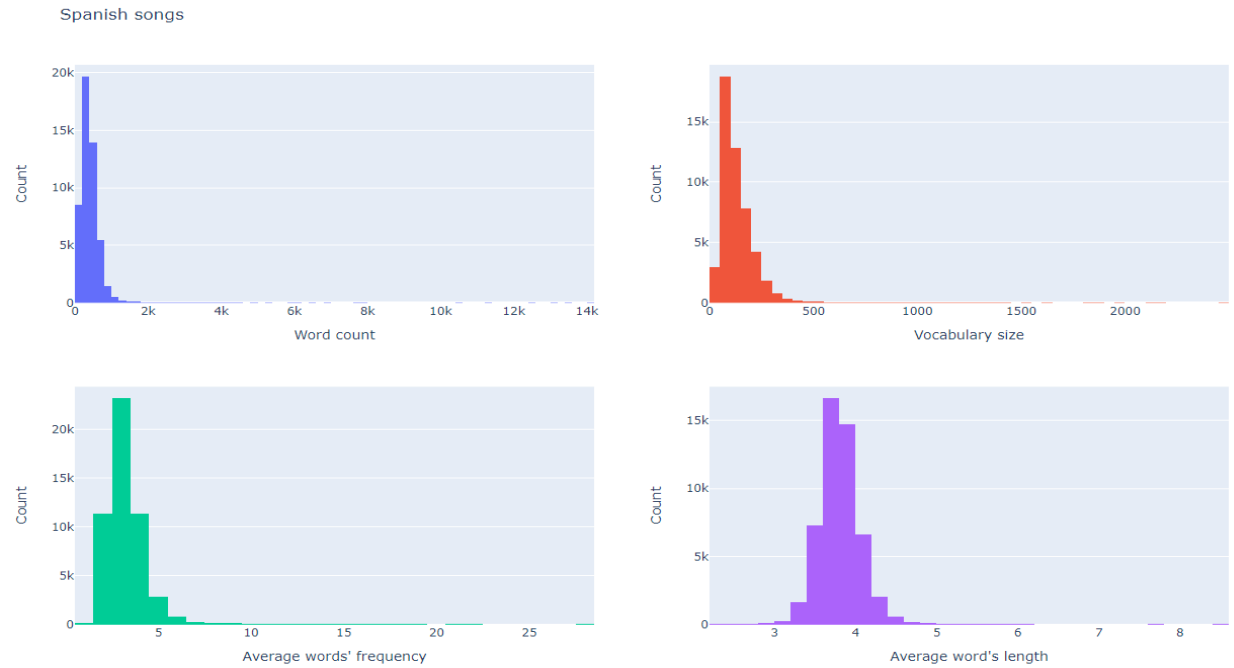


Fig. 14. Gráficos de la cantidad general de las palabras, la cantidad de las palabras únicas, frecuencia media de las palabras en el texto y longitud media de las palabras

4. También era necesario utilizar la proporción de las palabras de jerga, ya que este indicador parece ser muy importante y ha demostrado ser muy representativo en inglés. Para ello se utilizó el diccionario de Jergas Del Habla Hispana de Roxana Fitch[Fitch, 2006].

5. Número y proporción de palabras de uso común y su prevalencia y el grado de reconocimiento entre los hablantes nativos. Para ello se utilizó el artículo SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection de los autores como Jose Armando Aguasvivas, Manuel Carreiras, Marc Brysbaert, Paweł Mandera, Emmanuel Keuleers, Jon Andoni Duñabeitia[Aguasvivas, Carreiras, Brysbaert, Mandera, Keuleers, Duñabeitia, 2018]. Este artículo incluye un diccionario-base de datos con más de 44.000 palabras en español y la indicación de los niveles apropiados de reconocimiento y difusión para cada palabra [https://figshare.com/projects/SPALEX/29722].

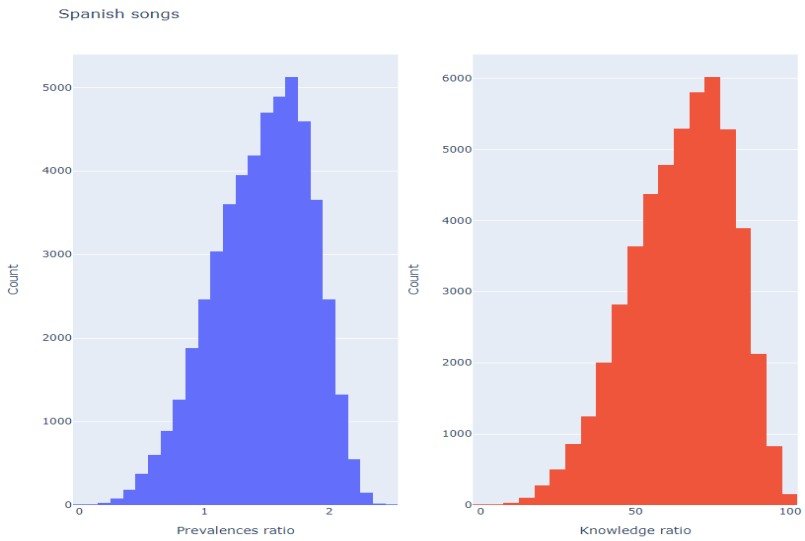


Fig. 15. Gráficos de la prevalencia de las palabras y el conocimiento entre los hablantes nativos

6. Como se mencionó anteriormente en español, al igual que en inglés, se utiliza el sistema CEFR, que incluye 6 niveles de dificultad: A1, A2, B1, B2, C1, C2, y donde cada nivel tiene su propio mínimo léxico.

Para esta información se utilizaron datos de la página web del Instituto Cervantes [https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/09_nociones_especificas_inventario_a1-a2.htm].

7. Dado que no se encontró una base de datos de palabras en español que indicara el número de fonemas, sílabas y morfemas en comparación con el inglés, se utilizó un enfoque diferente. Dado que los fonemas más difíciles de pronunciar y reconocer para los estudiantes de español son los siguientes: 'rr', 'c', 's', 'h', 'g', 'j', 'll', 'z', se hizo un recuento de estos fonemas en cada canción. Además, se utilizó un algoritmo para contar las consonantes, como en el idioma inglés. Entre otras

cosas, se encontró un programa en el sitio web del Basque Center on Cognition, Brain and Language (BSDL) [https://www.bcbl.eu/databases/espal], que le permite contar el número de sílabas y fonemas en una palabra española. Por lo tanto, ayudó a utilizar las palabras del diccionario del artículo SPALEX: Spanish Lexical Decision Database From a Massive Online Data Collection para contar fonemas y sílabas para referencia futura.

Mediante la construcción de la matriz de correlación (se puede ver en *Anexo II*) se identificaron las dependencias mismas como en el idioma inglés, a saber entre el número de palabras y los números de sustantivos, verbos y adjetivos, el número de palabras y el número de palabras únicas, la prevalencia de una palabra y su reconocimiento para un hablante nativo. Las otras características no tienen dependencias lineales explícitas entre sí.

5.2. Agrupación del conjunto de letras en español por dificultad

Dificultad del vocabulario español

Teniendo en cuenta la experiencia en el trabajo con el idioma inglés, inmediatamente se normalizaron los datos y se eliminaron enormes letras que contenían más de 1500 palabras(213 canciones). A continuación, se agruparon en 3 grupos utilizando características bien establecidas como el número de palabras, el vocabulario, la proporción de jerga, la proporción de prevalencia, el nivel de reconocimiento y la proporción de palabras desconocidas (*Tabla 14*).

	Word count							Vocabulary size						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	24059.0	0.170131	0.000000	0.115697	0.161028	0.217862	0.489175	24059.0	0.139029	0.000000	0.098616	0.129758	0.169550	0.416955
1	10984.0	0.417861	0.168471	0.323410	0.389716	0.479026	1.000000	10984.0	0.353674	0.083045	0.261246	0.333910	0.420415	1.000000
2	14744.0	0.271994	0.002706	0.186062	0.269959	0.347767	0.856563	14744.0	0.215911	0.001730	0.145329	0.207612	0.275087	0.820069
	Prevalences ratio							Knowledge ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	24059.0	0.709300	0.255669	0.646395	0.707603	0.769668	1.000000	24059.0	0.772635	0.501277	0.706006	0.771380	0.836215	1.000000
1	10984.0	0.586978	0.291114	0.533001	0.582394	0.639946	0.850785	10984.0	0.648536	0.344230	0.591834	0.641793	0.704282	0.923493
2	14744.0	0.408714	0.000000	0.352072	0.423107	0.481184	0.586929	14744.0	0.455269	0.000000	0.396746	0.472952	0.532474	0.616179
	Slang ratio							Non_present ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	24059.0	0.005055	0.0	0.0	0.0	0.000000	0.993743	24059.0	0.226264	0.000000	0.160684	0.227503	0.293750	0.435185
1	10984.0	0.011733	0.0	0.0	0.0	0.014952	0.596491	10984.0	0.349583	0.072784	0.291178	0.356136	0.407588	0.662895
2	14744.0	0.012284	0.0	0.0	0.0	0.012355	1.000000	14744.0	0.549710	0.000000	0.469243	0.530511	0.612458	1.000000

Tabla 14. División en 3 grupos usando el número de palabras, el vocabulario, la proporción de jerga, la proporción de prevalencia, el nivel de reconocimiento y la proporción de palabras desconocidas

Como resultado, se puede observar una división bastante buena en 3 grupos. Por ejemplo, el 75% de los datos del clúster 2 tiene reconocimiento de palabras inferior a 0.53, mientras que el 75% del clúster 1 tiene valores superiores a 0.59 y el 75% del clúster 0 tiene valores superiores a 0.70, que a su vez es mayor que el 75% de los datos del clúster 1 (resaltados con azul en la *Tabla 14*). Se pueden observar indicadores similares con respecto al nivel de prevalencia, ya que estos indicadores están relacionados.

En cuanto a la proporción de palabras desconocidas el 75% de los datos del clúster 0 tiene valores menos de 0.29, mientras que el 75% de los datos del clúster 1 tiene valores más de 0.29 y el 75% de los datos del clúster 2 tiene valores más de 0.46, lo que a su vez es mayor que el 75% de los datos del clúster 1 (resaltados con naranja en la *Tabla 14*).

En consideración al vocabulario, se puede observar que el 75% de los datos del clúster 0 tienen este indicador menos de 0.16, mientras que el 75% de los datos del clúster 2 tienen más de 0.14 y el 75% de los datos del clúster 1 tienen más de 0.26, que a su vez es mayor que casi el 75% de los datos del clúster 1 (resaltados con verde en la *Tabla 14*).

Al mismo tiempo, el indicador de la proporción de la jerga es poco informativo, por lo que no se usó más para el análisis. Esto se debe probablemente al hecho de que la base de datos de jerga española tiene muchas menos palabras que una base de datos de inglés similar.

Al igual que con el idioma inglés, si el nivel A1-A2 se puede rastrear explícitamente (clúster 0), debido a que tiene los indicadores más bajos como número de palabras, vocabulario, proporción de palabras desconocidas y los indicadores más altos como proporción de prevalencia y nivel de reconocimiento. Entonces la relación de los otros dos niveles no es del todo obvia. Después de todo, un nivel se destaca por un mayor número de palabras y un vocabulario más rico (clúster 1), el otro nivel se destaca por el nivel más bajo de prevalencia y reconocimiento de las palabras, así como el nivel más alto de palabras desconocidas (clúster 2).

El uso de indicadores como la proporción de sustantivos, verbos, adjetivos, frecuencia media de la palabra y longitud media de la palabra no produjo ninguna mejora, e incluso viceversa, hubo un deterioro en la claridad de la división de los límites de los grupos. Sin embargo, se tomó la decisión de probar también el impacto en la agrupación del vocabulario por niveles de CEFR.

Se ve que esto no tuvo un gran impacto, sin embargo, teniendo en cuenta la baja información del indicador de jerga para el análisis, así como el hecho de que el clúster 1 tiene una menor proporción de palabras C1 y C2 en comparación con el clúster 2, por lo que se decidió

correlacionar el clúster 2 con los niveles C1-C2, y el clúster 1 con los niveles B1-B2 (resaltados con azul en la *Tabla 15*).

	A1							A2						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	23887.0	0.050237	0.0	0.025424	0.044444	0.068627	1.000000	23887.0	0.139578	0.000000	0.089431	0.130952	0.180328	1.000000
1	15623.0	0.031304	0.0	0.017422	0.028213	0.041379	0.291667	15623.0	0.084516	0.000000	0.055276	0.079710	0.107196	0.666667
2	10277.0	0.050152	0.0	0.034247	0.047945	0.063830	0.148936	10277.0	0.125261	0.010073	0.095577	0.122222	0.152074	0.310734
	B1							B2						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	23887.0	0.192199	0.000000	0.133333	0.183673	0.241935	1.000000	23887.0	0.150210	0.0	0.088889	0.138889	0.199005	1.000000
1	15623.0	0.107976	0.000000	0.071174	0.101266	0.136364	0.545455	15623.0	0.084291	0.0	0.050553	0.078049	0.111111	0.533333
2	10277.0	0.162160	0.011561	0.123596	0.159420	0.197531	0.371134	10277.0	0.146639	0.0	0.101365	0.140351	0.186047	0.461942
	C1							C2						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	23887.0	0.072147	0.0	0.029304	0.061303	0.104575	1.00000	23887.0	0.089336	0.0	0.00000	0.074627	0.135135	1.000000
1	15623.0	0.042217	0.0	0.018018	0.036630	0.059259	0.50000	15623.0	0.050277	0.0	0.00000	0.041322	0.074627	0.714286
2	10277.0	0.077614	0.0	0.047619	0.073665	0.103060	0.25879	10277.0	0.097289	0.0	0.05291	0.088757	0.133690	0.424028

Tabla 15. División en 3 grupos usando el número de palabras, el vocabulario, la proporción de jerga, la proporción de prevalencia, el nivel de reconocimiento, la proporción de palabras desconocidas y 6 niveles de CEFR

Dificultad de la fonética española

Seguidamente se agruparon los datos con el número de grupos iguales a 3 y usando 4 características para analizar la complejidad fonética de la canción.

Se puede observar que en la proporción de fonemas, el 75% de los datos del clúster 1 tiene valores inferiores a 0.33, el 75% del clúster 2 tiene valores superiores a 0.36 y el 75% de los datos del clúster 0 tiene valores superiores a 0.44, que a su vez es mayor que prácticamente el 75% de los datos del clúster 2 (resaltados con azul en la *Tabla 16*).

En la proporción de sílabas los datos son similares.

Por lo que corresponde al número de fonemas difíciles y al número de consonantes, desafortunadamente, los datos no son informativos.

A partir de los resultados se puede asumir clúster 0 corresponde a C1-C2, clúster 1 corresponde a A1-A2 y clúster 2 corresponde a B1-B2.

	Sentence difficulty							Diff phonemes						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	7823.0	0.289929	0.000000	0.266992	0.292587	0.315344	0.474163	7823.0	0.174128	0.015218	0.147280	0.172233	0.197492	1.000000
1	16949.0	0.299574	0.119674	0.280910	0.302903	0.321105	0.481057	16949.0	0.147400	0.000000	0.126054	0.146695	0.167500	0.476191
2	25015.0	0.300574	0.120322	0.281407	0.303892	0.322374	1.000000	25015.0	0.161519	0.018578	0.140764	0.161035	0.180641	0.704499
	Phonemes_ratio							Syllabus_ratio						
	count	mean	min	25%	50%	75%	max	count	mean	min	25%	50%	75%	max
level														
0	7823.0	0.485432	0.329060	0.448867	0.471554	0.506014	1.000000	7823.0	0.489898	0.33944	0.451188	0.474609	0.511364	1.000000
1	16949.0	0.305798	0.000000	0.286134	0.312357	0.332115	0.417687	16949.0	0.302882	0.00000	0.282914	0.310258	0.330163	0.407480
2	25015.0	0.386713	0.259386	0.364316	0.385076	0.407831	0.510181	25015.0	0.385210	0.26250	0.362664	0.383523	0.406857	0.515625

Tabla 16. División en 3 grupos usando el número de fonemas, el número de sílabas, el número de fonemas difíciles y el número de consonantes.

Capítulo 6. DESARROLLO DEL PROTOTIPO

6.1 Requisitos, descripción del modelo de datos y del caso de uso

Requisitos funcionales

El prototipo desarrollado debe proporcionar a los usuarios las siguientes características:

- Selección de idioma de las opciones:
  - Eng (inglés);
  - Esp (español).
- Selección de modo de opciones:
  - vocabulario;
  - fonética.
- Selección de nivel de dificultad con posibles niveles de dificultad de fácil (*easy*) a difícil (*hard*).
- Posibilidad de buscar canciones utilizando filtros por idioma, modo, nivel de dificultad, así como buscar coincidencias en el texto para los valores de los campos o sus combinaciones para los campos de información de composición:
  - *title* (título);
  - *genre* (género;
  - *artist* (artista);
  - *year* (año).
- Muestra la composición seleccionada al hacer clic en los valores:
  - *title* (título);
  - *genre* (género;
  - *artist* (artista);
  - *year* (año);
  - texto completo de la composición con la capacidad de desplazarse por el texto (campos de *lyrics*).

Requisitos no funcionales

El prototipo debe cumplir con los siguientes requisitos:

- Debe tener una interfaz de usuario fácil de usar.
- Debe proporcionarse a los usuarios como un archivo exe ejecutable para Windows;
- Debe ser desarrollada en Python;
- La interfaz del prototipo debe crearse utilizando la biblioteca Kivy.

Descripción del modelo de datos

La tabla *music* almacena información sobre las entidades de composición e incluye campos:

- *id* (campo entero, identificador único para la composición);
- *title* (cuadro de texto, título);
- *genre* (cuadro de texto, género);
- *artist* (cuadro de texto, artista);
- *year* (campo entero, año);
- *lyrics* (cuadro de texto, texto completo de la composición);
- *level* (campo entero, clave externa para comunicarse con el nivel de dificultad en la tabla level);
- *lang* (campo entero, clave externa para comunicarse con el idioma en la tabla lang);
- *mode* (campo entero, clave externa para comunicarse con el modo en la tabla mode).

La tabla *level* almacena información sobre las entidades de nivel de complejidad e incluye campos:

- *id* (campo entero, identificador único para el nivel de dificultad);
- *name* (cuadro de texto, nombre de nivel de dificultad).

La tabla *lang* almacena información sobre los idiomas e incluye campos:

- *id* (campo entero, identificador único para el idioma);
- *name* (cuadro de texto, nombre del idioma);

La tabla *mode* almacena información de modo e incluye campos:

- *id* (campo entero, identificador único para el modo);
- *name* (cuadro de texto, nombre del modo);

Esta estructura de base de datos le permite hacer que el prototipo extensible y bastante fácil de agregar nuevos idiomas, modos, niveles de dificultad y composiciones al prototipo. Los diagramas de la base de datos y del caso de uso se muestran en las *Figuras 16 y 17*.

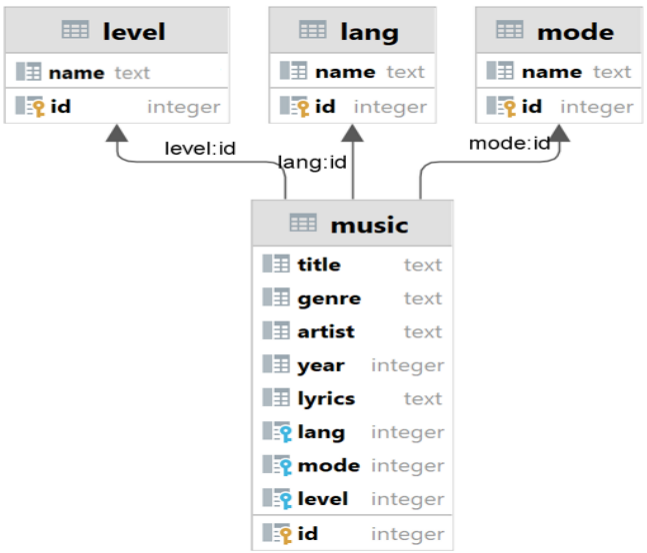


Fig. 16. Diagrama de base de datos

Caso de uso

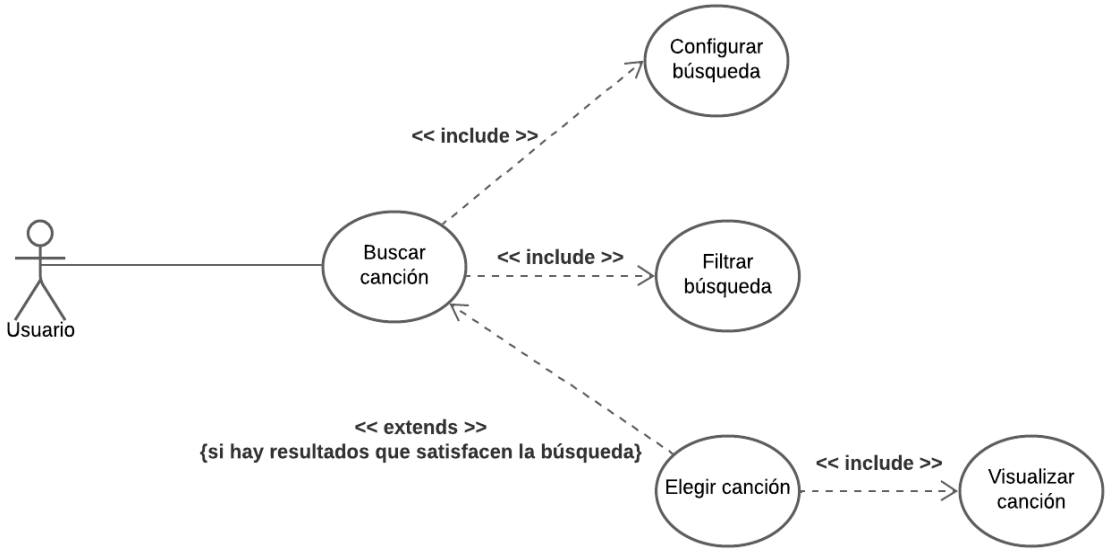


Fig. 17. Diagrama de caso de uso

Al iniciar la aplicación, el usuario selecciona el idioma de la canción, el modo y el nivel de dificultad. A continuación, la aplicación muestra una lista de canciones que cumplen con estas condiciones. Puede desplazarse por la lista de canciones con las teclas correspondientes. Al seleccionar una canción en particular, la aplicación muestra su letra. Además, el usuario puede especificar criterios de búsqueda adicionales, tales como el título, género, artista y año.

6.2 Base de datos

Después de realizar la agrupación en clústeres en el capítulo anterior se crearon archivos de excel que incluyen los datos originales y sobre los que se creó la base de datos SQLite:

- *lang* (tabla de idiomas disponibles: ingles/español);
- *mode* (tabla de modos disponibles: vocabulario/fonética);
- *level* (tabla de niveles disponibles: *easy/médium/hard*);
- *music* (tabla de canciones, cada canción usa claves externas, esta tabla está asociada con los datos de las tablas *lang*, *mode*, *level*). Esencialmente la tabla de música es la tabla principal que incluye información sobre artistas, letras, géneros, idiomas, complejidad, etc.

```
import sqlite3
filename="music_db"

con = sqlite3.connect(filename+".db")
musica.to_sql('music', con, index=False, if_exists="replace")

df = pd.read_excel('lang.xlsx')
df.to_sql('lang', con, index=False, if_exists="replace")

df = pd.read_excel('mode.xlsx')
df.to_sql('mode', con, index=False, if_exists="replace")

df = pd.read_excel('level.xlsx')
df.to_sql('level', con, index=False, if_exists="replace")

con.commit()
con.close()
```

Al mismo tiempo la aplicación accede a los datos de la base de datos y no utiliza los archivos de Excel.

6.3 Desarrollo de pantallas

La interfaz gráfica fue diseñada utilizando el framework Python Kivy. El uso del framework Kivy, que es una multiplataforma, permitirá con el desarrollo posterior del proyecto crear distribuciones para otros Sistemas Operativos (Android, Mac, Linux), mientras que no será necesario cambiar la interfaz del sistema.

En el desarrollo se utilizó la versión 2.2.1 del framework Kivy.

Se desarrollaron clases que describen las pantallas del prototipo:

- ChooseLangScreen (pantalla con menú de selección de idioma);
- ChooseModeScreen (pantalla con menú de selección de modo);

- ChooseLevelScreen (pantalla con menú de selección de nivel de dificultad);
- FilterScreen (pantalla para mostrar canciones según diferentes criterios);
- SongScreen (pantalla para mostrar el texto de la canción y la información sobre ella).

El objeto *sm* de la clase ScreenManager de kivy.uix es responsable de cambiar entre las pantallas.

En las pantallas de selección se carga información sobre los idiomas, modos y niveles disponibles desde la base de datos. Los métodos de la biblioteca *sqlite3* se utilizan para trabajar con la base de datos. Se crea un objeto *conn* para conectarse a una base de datos SQLite creada previamente, luego se crea un objeto de cursor para recuperar los datos. Para llevar a cabo el acceso a la base de datos y escribir una consulta SQL a la base de datos existe el método *execute* del objeto de cursor. A continuación, utilizando el método *fetchall*, los datos se recuperan de la base de datos como una lista de objetos Python. Una vez finalizada la conexión con la base de datos, se cierra la conexión.

Ejemplo de obtención de una lista de idiomas de una base de datos y creación de un conjunto de botones basado en la lista resultante (obtención de otros datos para el menú funciona de manera similar):

```
try:
    conn = sqlite3.connect("music_db.db")
    cursor = conn.cursor()
    cursor.execute('SELECT * from mode')
    result = cursor.fetchall()
    for r in result:
        box.add_widget(
            Button(text=r[1], on_press=lambda widget, mode_id=r[1]:
sm.get_screen("filter").set_mode(mode_id),
                   on_release=lambda x: set_screen("choose_level")))
    conn.commit()
    conn.close()
except Error as e:
    print(e)
self.add_widget(box)
```

Para crear pantallas de menú de selección de idioma, modo y nivel, se utilizó BoxLayout como contenedor principal, ya que es muy adecuado para crear menús a partir de botones de selección.

Al seleccionar el elemento deseado en las pantallas de menú utilizando el mecanismo de expresiones lambda y los métodos sobrecargados *on_press* (al hacer clic en los botones) y *on_release* (después de hacer clic en el botón), los datos recibidos se transfieren a la pantalla con los filtros y la lista de canciones.

Para hacer esto, se llama al método del objeto ScreenManager *get_screen* y luego uno de los métodos desarrollados para establecer el valor deseado del parámetro (por ejemplo, *set_mode* establece el modo):

```
box.add_widget(

    Button(text=r[1], on_press=lambda widget, mode_id=r[0]:
sm.get_screen("filter").set_mode(mode_id),on_release=lambda x:
set_screen("choose_level")))
```

El cambio entre pantallas se realiza mediante el método desarrollado *set_screen*, que utiliza la propiedad current del objeto ScreenManager:

```
def set_screen(name_screen):

    sm.current = name_screen
```

Al cargar filterScreen (en el método *on_enter*), se cargan las composiciones filtradas según las opciones de filtrado seleccionadas por el usuario y también se cargan las listas de valores disponibles para filtrar en listas DropDown. Para llenar la lista DropDown se utilizan diccionarios (la clave es id en la base de datos, el valor de texto) de datos con información sobre idiomas, modos y niveles. El método *get_data_dict* es responsable de extraer los datos de la base de datos basada en el nombre de la tabla y los proporciona como un diccionario:

```
def get_data_dict(self, table_name):
    conn = sqlite3.connect("music_db.db")
    cursor = conn.cursor()
    cursor.execute(f'SELECT * from {table_name}')
    data = cursor.fetchall()
    result = dict()
    for item in data:
        result[item[0]] = item[1]
    conn.commit()
    conn.close()
    return result
```

Los elementos en FilterScreen están dispuestos en un layout separado creado usando GridLayout (disposición tabular de los elementos):

```
self.layout = GridLayout(cols=1, spacing=12, padding=25, size_hint_y=None)
self.layout.bind(minimum_height=self.layout.setter('height'))
lang_dropdown = DropDown()
lang_dict = self.get_data_dict('lang')
lang_dict["lang"] = "all"
for key, value in lang_dict.items():
    btn = Button(text=str(value), size_hint_y=None, height=44)
    btn.bind(on_release=lambda btn, lang_id=key:
self.set_lang_selected(lang_dropdown, btn.text, lang_id))
    lang_dropdown.add_widget(btn)
lang_button = Button(text=lang_dict[self.lang], size_hint_y=None, height=44)
lang_button.bind(on_release=lang_dropdown.open)

lang_dropdown.bind(on_select=lambda instance, x: setattr(lang_button, 'text',
x))

self.layout.add_widget(lang_button)
```

El layout resultante con todos los elementos agregados a él se agrega a RecycleView que permite al usuario desplazarse cómodamente por todos los elementos en la pantalla con una lista de filtros (la pantalla creada será conveniente para usar tanto en dispositivos de escritorio con una selección de diferentes resoluciones como en dispositivos móviles). Se ha creado un GridLayout llamado SongsLayout para la lista de canciones.

Creación de un RecyclerView de raíz y agregarle un RecyclerView con parámetros:

```
root = RecycleView(size_hint=(1, 1), size=(Window.width - 50,
                                            Window.height - 50))
self.songs_layout = GridLayout(cols=1, spacing=10, size_hint_y=None)
root.add_widget(self.layout)
self.add_widget(root)
```

Para filtrar se usa una consulta SQL usando cursor.execute:

```
cursor.execute(
    f'SELECT * from music WHERE lang={self.lang} AND mode={self.mode} AND
level={self.level}'
    f' and title LIKE "%{self.title.text}%"'
    f' and genre LIKE "%{self.genre.text}%"'
    f' and artist LIKE "%{self.artist.text}%"'
    f' and year LIKE "%{self.year.text}%"')
```

La consulta creada le permite filtrar por idioma, modo y nivel, así como buscar en los campos "título", "género", "artista", "año" (el texto introducido por el usuario en los campos de búsqueda).

La información general de la canción y la letra de la canción deslizable se muestran en la pantalla SongScreen. Para crear un elemento de texto deslizable (label) con texto en el Builder (colector de la aplicación Kivy), se agregó un elemento ScrollableLabel escrito en formato kv:

```
Builder.load_string('''
<ScrollableLabel>:
    Label:
        size_hint_y: None
        height: self.texture_size[1]
        text_size: self.width, None
        text: root.text
        halign: "center"
''')
```

6.4 Características

- La aplicación se puede hacer multiplataforma sin cambiar la interfaz.
- Todas las pantallas se escalan adecuadamente.
- Los elementos en las pantallas de menú se generan en el código en base a los datos de la base de datos, cuando se agregan nuevos datos a la base de datos, estos datos pueden cambiar automáticamente el menú.
- Se agregó una pantalla con letras de canciones deslizables.

- Listas DropDown con datos de la base de datos que permiten cambiar los filtros originales.
- Consultas SQL de la base de datos.

6.5 Paquete redistribuible para Windows

Para crear una versión distribuida de la aplicación para Windows era necesario crear un archivo *exe* con la aplicación.

Para crear el archivo *exe* se utilizó la biblioteca pyinstaller. PyInstaller le permite combinar una aplicación Python y todas sus dependencias en un solo paquete. El usuario puede ejecutar una aplicación empaquetada sin instalar un intérprete de Python ni ningún módulo.

Los componentes de la biblioteca pyinstaller se instalaron en el entorno virtual del proyecto mediante un comando en el terminal:

pip install pyinstaller

Se creó la distribución de la aplicación music con la ayuda de un comando en el terminal:

pyinstaller --onefile main.py --name music

El contenido del paquete distribuido se muestra en la *Figura 18*.

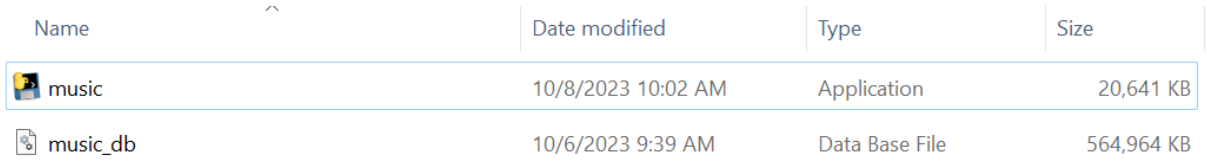


Fig. 18. El contenido del paquete distribuido

6.6 Interfaz

En las *Figuras 19 y 20* se puede ver la interfaz del prototipo. Puede elegir el idioma de la canción (inglés o español), el modo (vocabulario o fonética), el nivel de dificultad (fácil, medio y difícil, que corresponden a la división condicional A1-A2, B1-B2, C1-C2). También está disponible la búsqueda por canción, artista, género o año de lanzamiento de la canción. Debido al hecho de que esta aplicación incluye casi 100.000 canciones, por lo que también se crearon las teclas NEXT y PREV para desplazarse por la lista de canciones**,** de acuerdo con los filtros. Entre otras cosas, cada canción incluye letra completa.

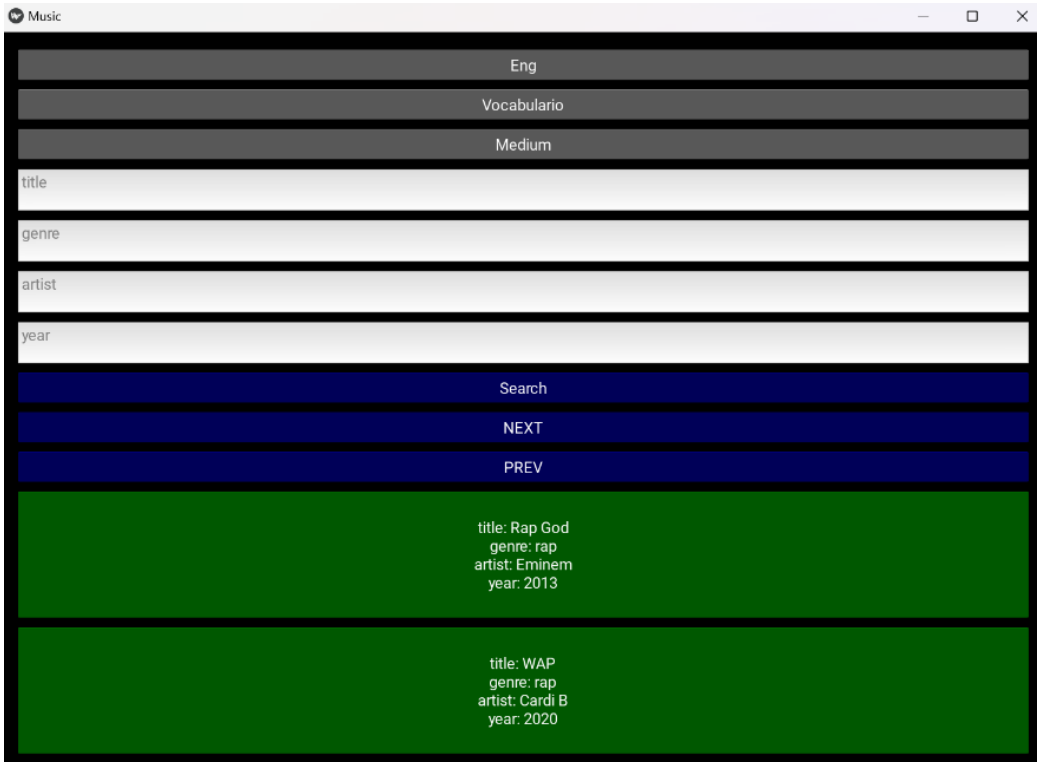


Fig. 19. Interfaz del prototipo

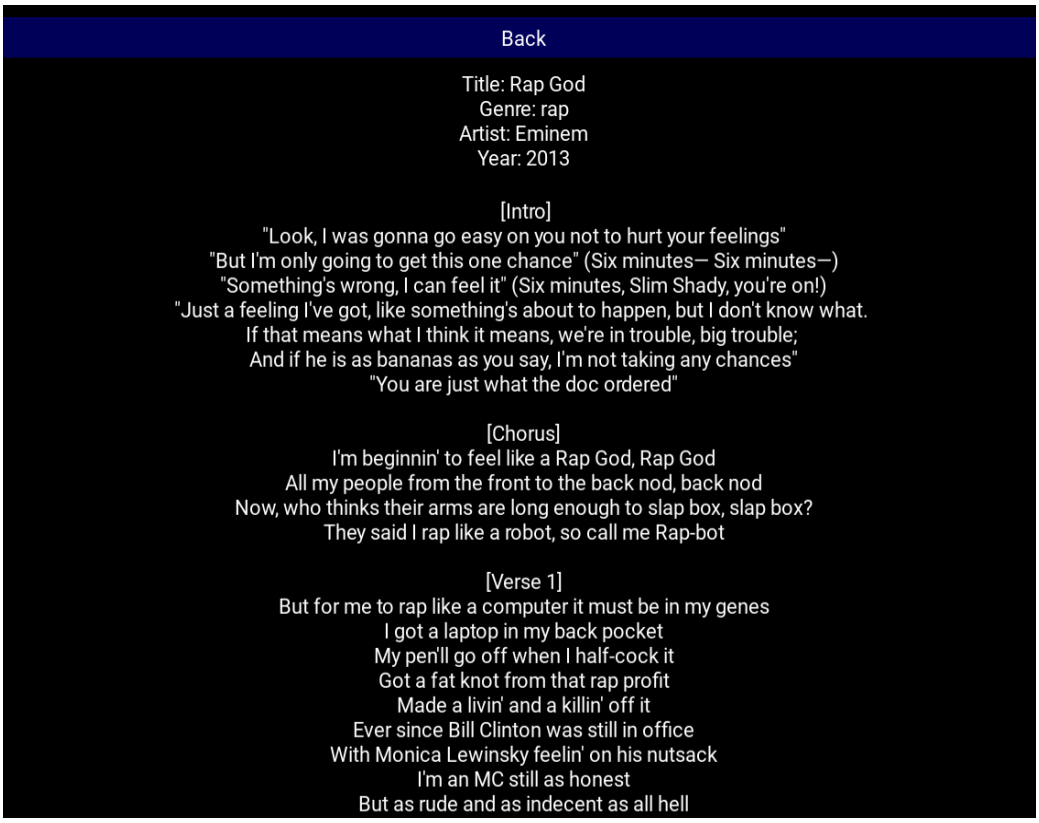


Fig. 20. Letra de la canción

Capítulo 7. CONCLUSIÓN

En el curso de esta tesis de maestría se alcanzaron los objetivos asignadas. Estudié la literatura científica dedicada a la historia de la investigación del problema de la complejidad del texto en inglés y español, examiné diferentes enfoques y fórmulas de legibilidad de los textos utilizando diferentes métricas que tienen en cuenta las características léxicas, fonéticas, etc.

El algoritmo *K-means* que se eligió para agrupar textos por complejidad permitió identificar durante la investigación las características más importantes para determinar la complejidad de las canciones en inglés y español a saber:

- número de palabras
- vocabulario
- proporción de jerga
- nivel de prevalencia de palabras
- nivel de reconocimiento de palabras
- nivel de palabras desconocidas
- número de fonemas, sílabas y morfemas en palabras

Por lo tanto, el resultado del trabajo fue un prototipo que tiene las siguientes características y oportunidades:

- Agrupación de canciones por niveles de dificultad en vocabulario y en fonética de letras en inglés;
- Agrupación de canciones por niveles de dificultad en vocabulario y en fonética de letras en español;
- Este prototipo puede ayudar en cierta medida a los profesores de inglés y español en la elaboración de recomendaciones de canciones para el aprendizaje de estos idiomas.

Con respecto a posibles trabajos futuros, en primer lugar, pueden estar relacionados con la ampliación de la funcionalidad. Por ejemplo, mediante la API, se podría enlazar con la plataforma Spotify para reproducir canciones. Además, la funcionalidad se puede mejorar agregando la capacidad de crear una lista de reproducción. Asimismo, este prototipo podría integrarse potencialmente en otras aplicaciones que son, en cierta medida, sistemas de recomendación y/o reproductores de canciones.

Al mismo tiempo, en el proceso de escribir la tesis actual, se utilizaron los conocimientos y habilidades que adquirí durante mis estudios de maestría. Esto se aplica en gran medida a las asignaturas como Ingeniería del Lenguaje Natural y *Data Science*. La primera de ellas ayudó en la primera etapa en temas como el procesamiento de letras de canciones y la posterior selección de

las características necesarias de letras. La última fue útil en la segunda etapa, cuando fue necesario analizar las características obtenidas para resaltar los clústeres.

Por supuesto, este estudio no estuvo exento de ciertas dificultades. Por ejemplo, como ha demostrado la experiencia, es extremadamente difícil distinguir 6 niveles de conocimiento del idioma en las canciones en general, y en la fonética y en el vocabulario en particular. Es por eso que se decidió concentrarse en 3 clústeres. También la dificultad fue causada por la determinación de la complejidad gramatical de las canciones. A lo que hay una serie de razones. Primero, a menudo las canciones modernas no implican el uso de construcciones gramaticales complejas de nivel C1-C2. En segundo lugar, no hay un punto de vista único sobre qué nivel de complejidad atribuir a un elemento gramatical. Por ejemplo, en el marco de la lengua española, alguien atribuye el modo de imperativo al nivel A1-A2, y alguien al nivel B1-B2. En este sentido, a determinación de la complejidad depende en gran medida del estudio realizado.

Capítulo 8. LISTA DE REFERENCIAS

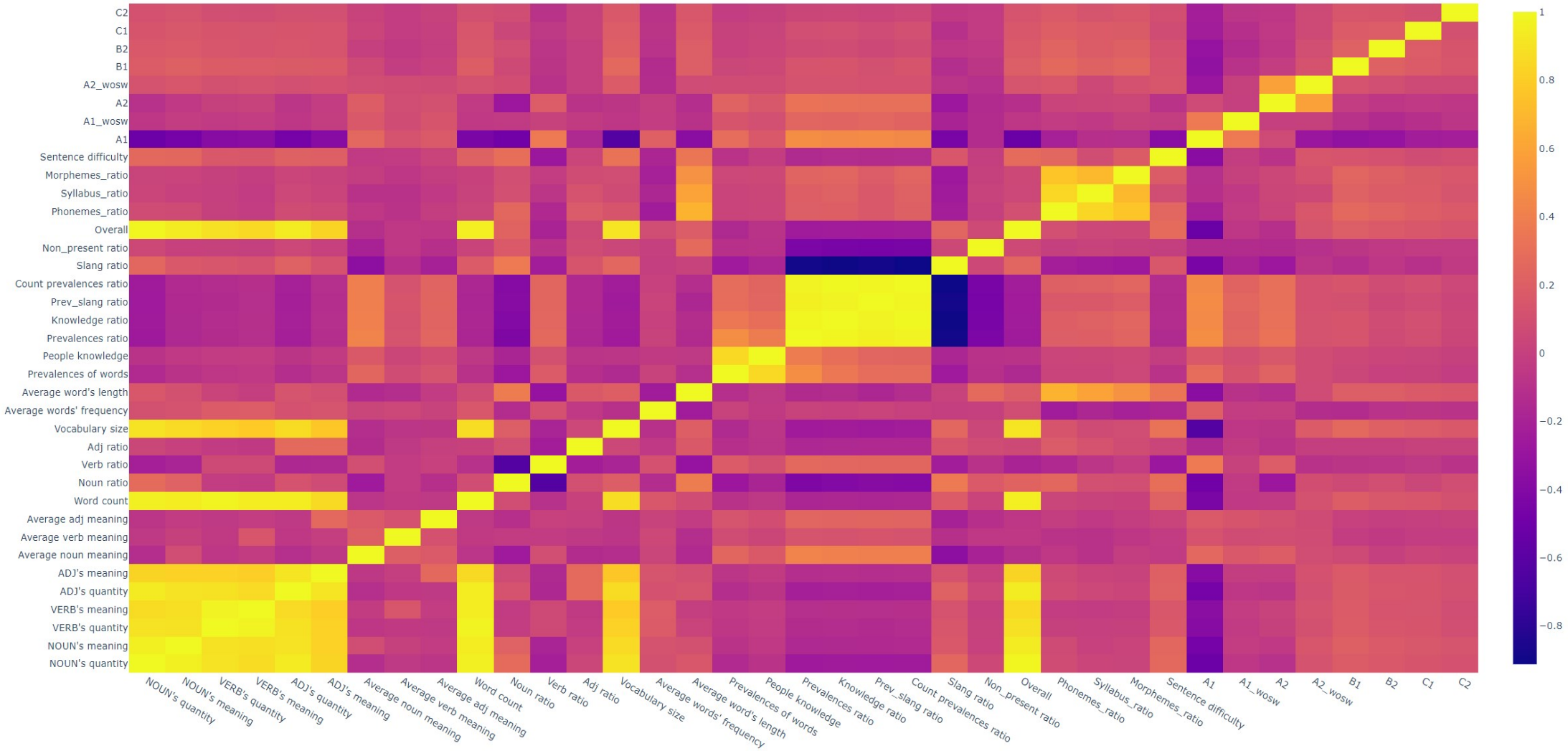
1. Abadzi, H. (2008). *Efficient learning for the poor: New insights into literacy acquisition for children.* International Review of Education.
2. Aguasvivas J.A, Carreiras M, Brysbaert M, Mandera P, Keuleers E, Duñabeitia J.A. (2018). *SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection.* Front. Psychol.
3. Brysbaert, M, Mandera, P, McCormick, S.F, Keuleers, E. (2019). *Word prevalence norms for 62,000 English lemmas*. Behavior Research Methods.
4. Chall, J.S. (1995). *Readability revisited: the new Dale-Chall readability formula*. Cambridge, Mass: Brookline Books Collection.
5. Dahl, O. (2008). *Grammatical resources and linguistic complexity. Siriono as a language without NP coordination*. Language complexity: typology, contact, change. Amsterdam.
6. Darrah, Ch. (1996). *Songs and the analysis of algorithms. In ACM SIGCSE Bulletin*. RESEARCHGATE. [Consulta: 08/07/2023]. Disponible en URL: https://www.researchgate.net/
7. DuBay, W.H. (2004). *Impact information study on the principles of readability*. PDF4PRO. [Consulta: 08/07/2023]. Disponible en URL: https://pdf4pro.com/view/principles-of-readability-impact-information-com-b5df.html
8. Fitch, R. (2006). *Diccionario de Jergas de Habla Hispana*.
9. Flesch, R. (1946). *The Art of Plain Talk. New York: Harper & brothers*.
10. Fountain-Chambers, J.F. (1982). *Readability of primary-grade Spanish reading books: a correlational study of the Spaulding Coco et al.: Readability of Spanish-Language Outcome Measures 315 Formula and the Fry Graph*. Denton (Texas): School of Library Science.
11. Gilliam, B., Peña, S. C., & Mountain, L. (1980). *The Fry Graph applied to Spanish readability*. The Reading Teacher.
12. Gray, W.S., Leary, B.W. (1935). *What makes a book readable*. Chicago: University of Chicago Press.
13. Huerta, F.J. (1959). *Medidas sencillas de lecturabilidad*. Consigna.
14. Knuth, D. (1984). *The Complexity of Songs*. Communications of the ACM.
15. Lyashevskaya, O.N. *Índices de legibilidad como medida de evaluación de la complejidad del texto* [Consulta: 07/07/2023]. Disponible en URL: https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf
16. McLaughlin, H. (1969). *SMOG grading - a new readability formula*. Journal of Reading.

17. Melnikova, A.V. (2020). *Diseño y desarrollo de un sistema de recomendación para la selección de canciones por dificultad para aprender Inglés*. Tyumen: Ed-en TSU.
18. Misernov, I.Y, Grashenko, L.A. (2015). *Análisis de métodos para evaluar la complejidad del texto*. Nuevas tecnologías de la información en sistemas automatizados.
19. Nevdah, M.M. (2012). *Mejora de la calidad de la literatura educativa*. Trabajos de la Universidad estatal de tecnología de Bielorrusia.
20. Oborneva, I.V. (2006). *Evaluación Automatizada de la complejidad de los textos educativos sobre la base de parámetros estadísticos*. M.: Ed-en ISMO de la Academia Rusa de Educación.
21. Powers, R.D, Sumner, W.A, Kearl, B.E. (1958). *A recalculation of four adult readability formulas*. Journal of Educational Psychology.
22. Powers, R.D. (1993). *Recuento de 4 fórmulas leídas*. Psicología Pedagógica.
23. Ricci, F, Rocach, L, Shapira, B. (2015). S*istemas de Recomendación: Introducción y problemas*. Springer Link.
    SPRINGERLINK. [Consulta: 27/07/2023]. Disponible en URL: https://link.springer.com/chapter/10.1007/978-1-4899-7637-6_1
24. Rihanna. (2023). *¿Cuál es mi nombre?* Lyrsense.
    LYRENSE. [Consulta: 27/07/2023]. Disponible en URL: https://lyrsense.com/rihanna/whats_my_name_r
25. Romashov, D.S. (2016). *Sistema de determinación de preferencias musicales*. San Petersburgo: Ed-en la Universidad estatal de San Petersburgo.
26. Sibanda, L. (2013). *Un estudio De caso sobre la legibilidad de dos libros de texto de Ciencias de grado 4 actualmente en uso en las escuelas de Sudáfrica*. Mahanda (Sudáfrica), Rhodes University Press.
27. Skachkova, G.A. (2021). *Características Fonéticas de la versión antillana del español en el discurso de la canción*. Colección de artículos científicos sobre los materiales de la XXX conferencia internacional científica y práctica. OTV. Editores Kormilin N.V., Shugaeva N.Yu.: Universidad Pedagógica Estatal de Chuvash (Cheboksary).
28. Solnyshkina, M.I., Kiselnikov, A.S. (2015). *Parámetros de complejidad de los textos de examen*. Boletín de la Universidad estatal de Volgogrado.
29. Yankee, D. (2023). *El problema*. LYRENSE. [Consulta: 31/07/2023]. Disponible en URL: https://lyrsense.com/daddy_yankee/problema_dy
30. Ziganshina, Chr. (2020). *Parámetros Cualitativos de la complejidad del texto (en el material de los textos artísticos y científicos populares PIRLS)*. El Mundo de la ciencia, la cultura, la educación.

31. TLCdénia. *27 canciones para aprender español*. [Consulta: 25/07/2023]. Disponible en URL: https://tlcdenia.com/es/canciones-para-aprender-espanol/
32. KAGGLE. *Genius Song Lyrics*. [Consulta: 15/07/2023]. Disponible en URL: https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information
33. MEDIUM. *Text pre-processing: Stop words removal using different libraries*. [Consulta: 15/09/2023]. Disponible en URL: https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a
34. GEEKSFORGEEKS. *Calculate the difficulty of a sentence*. [Consulta: 05/09/2023]. Disponible en URL: https://www.geeksforgeeks.org/calculate-difficulty-sentence/
35. CENTRO VIRTUAL CERVANTES. *Nociones específicas. Inventario A1-A2*. [Consulta: 18/09/2023]. Disponible en URL: https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/09_nociones_especificas_inventario_a1-a2.htm
36. BTU SFL A1 LEVEL WORDS. [Consulta: 25/07/2023]. Disponible en URL: https://depo.btu.edu.tr/dosyalar/ydyo/Dosyalar/BTU%20SFL%20A1%20LEVEL%20WORDS(1).pdf
37. EsPal. *Databases*. [Consulta: 20/09/2023]. Disponible en URL: https://www.bcbl.eu/databases/espal/

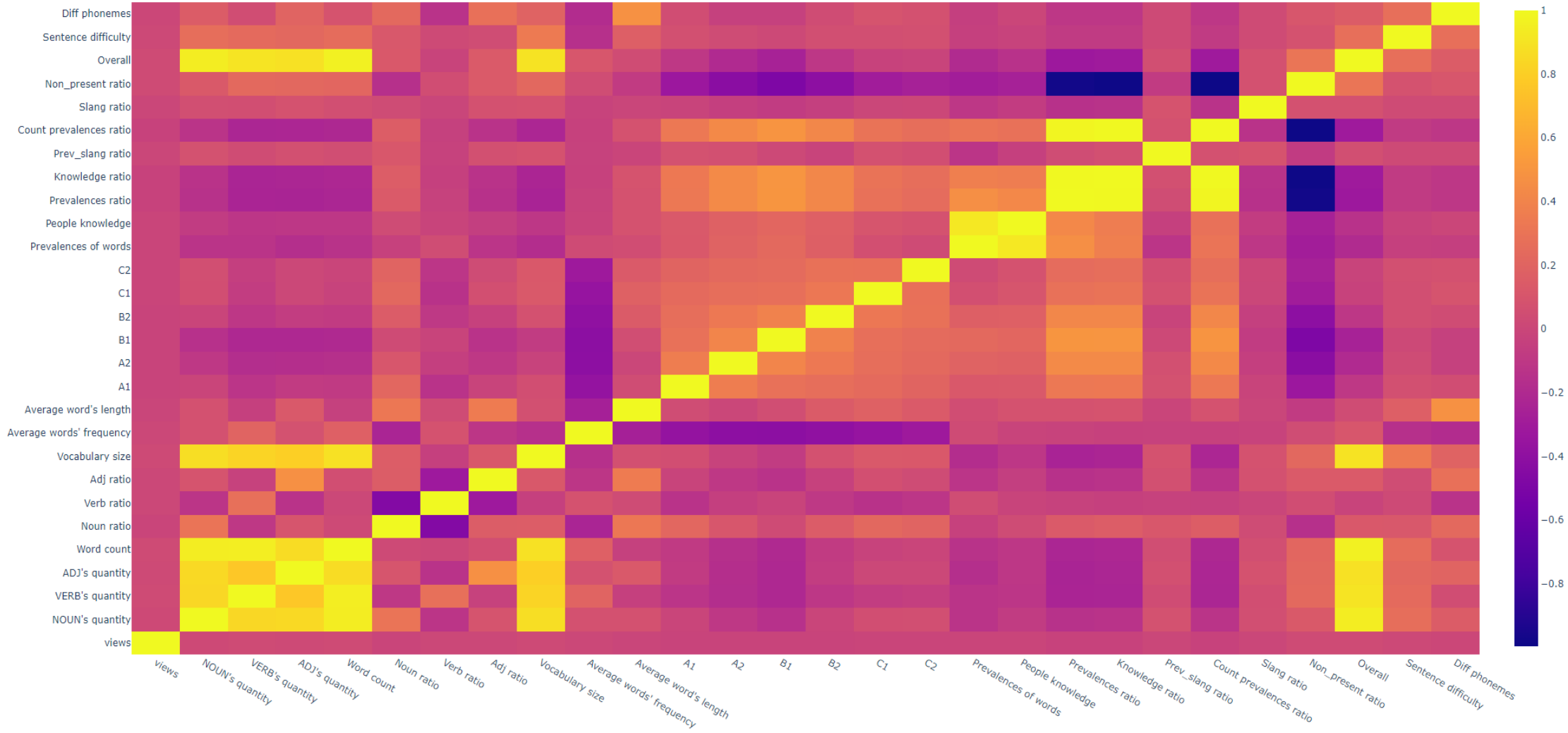
ANEXO I

Matriz de correlación(lengua inglesa)



ANEXO II

Matriz de correlación(lengua española)



ANEXO III

OBJETIVOS DE DESARROLLO SOSTENIBLE
Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Proce de**
ODS 1. **Fin de la pobreza.**				**x**
ODS 2. **Hambre cero.**				**x**
ODS 3. **Salud y bienestar.**			**x**	
ODS 4. **Educación de calidad.**		**x**		
ODS 5. **Igualdad de género.**				**x**
ODS 6. **Agua limpia y saneamiento.**				**x**
ODS 7. **Energía asequible y no contaminante.**				**x**
ODS 8. **Trabajo decente y crecimiento económico.**		**x**		
ODS 9. **Industria, innovación e infraestructuras.**				**x**
ODS 10. **Reducción de las desigualdades**.				**x**
ODS 11. **Ciudades y comunidades sostenibles**.				**x**
ODS 12. **Producción y consumo responsables.**				**x**
ODS 13. **Acción por el clima.**				**x**
ODS 14. **Vida submarina.**				**x**
ODS 15. **Vida de ecosistemas terrestres.**				**x**
ODS 16. **Paz, justicia e instituciones sólidas.**				**x**
ODS 17. **Alianzas para lograr objetivos.**				**x**

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

En mi opinión, esta tesis de maestría está relacionada con los siguientes objetivos de desarrollo sostenible de las ONU:

- **Educación de calidad**
  Dado que un sistema de recomendación para el aprendizaje de idiomas puede contribuir a este objetivo al proporcionar acceso a recursos educativos de calidad. Puede ayudar a los estudiantes a mejorar sus habilidades lingüísticas, lo que a su vez contribuye a mejorar la calidad de la educación y mejorar sus oportunidades para futuras carreras y desarrollo personal.

- **Salud y bienestar**
  El aprendizaje de idiomas tiene una relación directa con la salud y el bienestar. Las habilidades lingüísticas pueden ayudar a las personas a comprender mejor la información médica, recibir atención médica de calidad y seguir las recomendaciones de los médicos, asimismo pueden mejorar la comunicación entre los pacientes y el personal médico. Además, aprender idiomas puede ayudar a las personas a ampliar sus horizontes, desarrollar el pensamiento crítico y la creatividad. Además, aprender idiomas puede ayudar a reducir el estrés y aumentar la satisfacción con la vida.

- **Trabajo decente y crecimiento económico**
  El trabajo decente es la base del desarrollo sostenible. Proporciona a las personas medios de subsistencia, oportunidades para desarrollarse y contribuir a la sociedad. Las habilidades lingüísticas son a menudo un aspecto importante para lograr el empleo decente y el crecimiento económico. El sistema de recomendación puede ayudar a los usuarios a mejorar sus habilidades lingüísticas, lo que los hace más competitivos en él.