readabilityes: Cálculo reproducible de métricas de legibilidad en español

Autores: Jesica Formoso, Sofía Ortiz, Brenda Gomez Muiño, Juan Pablo Barreyro

Palabras clave: legibilidad, español, procesamiento de texto, reproducibilidad, paquete de R

Abstract

La legibilidad es un indicador clave de la facilidad con la que un texto puede ser comprendido por su lector (Flesch, 1974). Este concepto ha sido ampliamente estudiado en el ámbito angloparlante, con aplicaciones directas en la educación, la salud, la comunicación pública y la accesibilidad. En la práctica, medir la legibilidad permite adaptar textos para que sean accesibles a audiencias con distintos niveles educativos o competencias lectoras, lo que resulta crucial para promover la comprensión y la equidad en el acceso a la información (DuBay, 2006).

En español, la evaluación de la legibilidad enfrenta limitaciones importantes. Las fórmulas más reconocidas, como las de Fernández-Huerta (Fernández Huerta,1959), INFLESZ (Barrio, 2008) y Szigriszt-Pazos (Szigriszt Pazos, 2001), requieren cálculos basados en la segmentación en sílabas y en la identificación precisa de oraciones, tareas que presentan particularidades frente al inglés debido a diferencias fonológicas y ortográficas. Entre los desafíos más frecuentes se encuentran el manejo de diptongos, triptongos, hiatos, vocales acentuadas, la “y” vocálica, grupos consonánticos que no se separan y letras con diacríticos.

En el ecosistema R existen alternativas como koRpus (que delega la silabificación en sylly y puede integrar TreeTagger) y udpipe (tokenización/segmentación y etiquetado POS mediante modelos descargables). Sin embargo, la mayoría no ofrece una silabificación robusta nativa para español ni índices específicos como Szigriszt-Pazos, INFLESZ o Gutiérrez de Polini sin dependencias externas. readabilityes implementa silabificación y segmentación orientadas al español en R, junto con métricas básicas y estos índices, lo que reduce el peso de instalación, evita descargas de modelos y mejora la reproducibilidad en flujos tidyverse.

En este contexto, presentamos readabilityes, un paquete desarrollado en R que implementa un motor de silabificación propio con reglas del español, sin dependencias externas, y funciones para tokenizar palabras y oraciones. Permite calcular métricas básicas (número de palabras, sílabas, oraciones, longitud media de palabra y de oración) e incluye directamente las fórmulas clásicas de legibilidad para el español —como Fernández-Huerta, Szigriszt-Pazos/INFLESZ y Gutiérrez de Polini—, disponibles para su uso inmediato sin necesidad de configuraciones adicionales ni instalación de software externo. El diseño de

esta herramienta prioriza la transparencia, la reproducibilidad y la integración fluida con el ecosistema tidyverse.

El paquete está dirigido a investigadores, docentes y profesionales que requieran evaluar la legibilidad de textos en español de forma automatizada y reproducible. Entre las extensiones previstas se incluyen:

- Incorporar métricas basadas en frecuencia léxica.
- Implementar detección de estructuras sintácticas complejas, como oraciones pasivas o subordinadas múltiples.
- Integrar análisis morfosintáctico mediante anotadores gramaticales como udpipe.
- Desarrollar una aplicación Shiny que permita cargar textos y descargar informes con los resultados.

Para ilustrar su uso, se presentará un ejemplo práctico en el que, a partir de una tibble con múltiples textos, se calculan automáticamente métricas e índices de legibilidad y se generan visualizaciones que permiten comparar los resultados. Este flujo de trabajo no solo automatiza el cálculo, sino que también facilita su incorporación en proyectos de análisis reproducible e informes dinámicos con Quarto o R Markdown.

Tabla 1. Ejemplo de medidas extraidas del conjunto de textos analizados con readabilityes

id_texto	N Oraciones	Promedio de palabras por oración	Promedio de sílabas por palabra	INFLESZ	INFLESZ categoría
1	1	3	1.33	121	Muy fácil
2	1	11	1.73	88.2	Muy fácil
3	1	13	2.62	30.9	Difícil
4	1	16	2.69	23.4	Difícil
5	1	22	2.41	34.8	Difícil

La Figura 1 muestra un ejemplo de salida gráfica obtenida con readabilityes, donde se comparan los índices INFLESZ de varios textos. En este caso, valores más altos indican mayor facilidad de lectura, mientras que valores más bajos reflejan mayor complejidad.

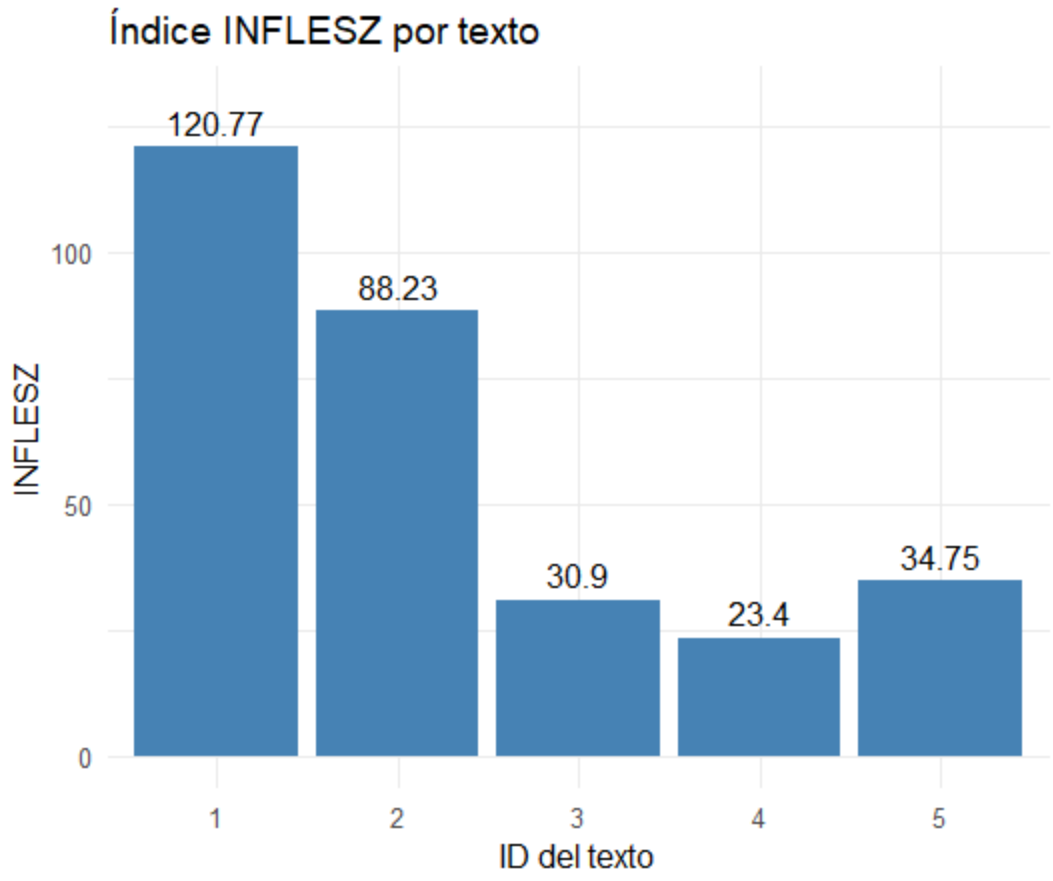

Figura 1. Comparación de índices INFLESZ en diferentes textos.

El código fuente y la documentación de readabilityes están disponibles en github. Se compartirá el link de ser aprobada la presentación para mantener el anonimato durante la revisión.

Referencias

Barrio, I. (2008). Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. An Sist Sanit Navar 2008; 31(2): 135-152.
DuBay, W.H. (2006) The Classic Readability Studies; Impact Information: Costa Mesa, CA, USA.
Fernández Huerta J.(1959). Medidas sencillas de lecturabilidad. Consigna (Revista pedagógica de la sección femenina de Falange ET y de las JONS) 1959; (214): 29-32.
Flesch, R. (1974).The Art of Readable Writing; revised and enlarged edition; Harper & Row: New York, NY, USA.
Szigriszt Pazos, F. (2001). Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad. Universidad Complutense de Madrid, Servicio de Publicaciones.