TRUNAJOD, UNA HERRAMIENTA DE ANÁLISIS AUTOMÁTICO DE TEXTOS ESCRITOS PARA OBTENER ÍNDICES DE LECTURABILIDAD

Mónica Véliz, Bernardo Riffo, Christian Soto, Daniela Campos

Proyectos Fondef IT15i10036 e IT17I0051

Trunajod

Programa computacional que analiza en forma automática textos escritos en español para extraer información sobre su lecturabilidad.

La información obtenida se expresa en una serie de índices léxicos, sintácticos y semánticos que representan la dificultad del texto.

Como base del análisis se usa un etiquetador (*Connexor*) y un diccionario de frecuencias (*Lifcach)*.

Fondef IT15i10036 e ITI710051

¿Cómo surge la herramienta?

- Proyecto Fondef D08i1179

"Desarrollo de una prueba validada y normada para el diagnóstico de la comprensión lectora en el sistema escolar chileno"

- Uno de los desafíos que planteó la investigación

¿Cómo seleccionar los textos adecuados para cada uno de los niveles escolares de la prueba *Lectum*?

La lecturabilidad y su historia

- El enfoque clásico
- El enfoque cognitivo
- Nuevas tendencias

Fondef IT15i10036 e ITI710051

Índices de lecturabilidad que se obtienen de Trunajod

Índices sintácticos

- Longitud de la O (LO)
- Longitud de la cláusula (LC)
- Índice de Subordinación (IS)
- Densidad de la frase nominal (DFN)

Fondef IT15i10036 e IT1710051

Índices de lecturabilidad que se obtienen de Trunajod

Índices léxicos:

- Densidad léxica (DeL)
- Diversidad léxica (DiL)
- Frecuencia de la palabra (FP)
- Frecuencia de palabra logarítmica FPL)

Índice semántico:

- Densidad proposicional (Dep)

Fondef TT15i10036 e TTI710051

¿Cómo trabaja Trunajod?

- Connexor Machinese Syntax

Etiqueta las palabras marcando

-- clase de palabra
-- flexión nominal y verbal
-- posición de la palabra en la O
-- dependencias sintácticas
-- la raíz de la palabra

- Lifcach

Indica la frecuencia de uso de las palabras

Fondef IT15i10036 e IT17I0051

¿Cómo trabaja *Trunajod?*

- Algoritmos básicos

Ejemplo: Índice Densidad léxica
[Sustantivos, verbos, adjetivos, adverbios]

- Reglas de ajuste

Ejemplo:
[Perífrasis verbales]

Salida de *Trunajod* : un ejemplo

Trunajod

Ingresar texto | Textos ingresados | Administrar categorías || Formateador | Manual || Salir | **fondef**

Textos ingresados

Filtrar

Nivel - todos -

Forma - todos -

Número texto - todos -

Estructura 1 - todos -

Filtrar

Fecha	Título	Nivel	Forma	Número texto	Estructura 1	Cantidades						Índices								
						P	O	V	C	PN	Prop	LO	LC	IS	DeP	DeL	DiL	DFN	FP	FPL
16:26 14/10/2014	Esperando a las nuevas hormiguitas	2	B	2		302	25	154	42	170	136	12.1	7.2	1.7	45	56	70	7.6	807	112.3
16:23 14/10/2014	El desván del duende Melodía	2	A	2		355	32	161	61	204	154	11.1	5.8	1.9	43	57	67	5.7	1283	123.2
16:18 14/10/2014	Platero	3	B	2		167	12	109	19	92	73	13.9	8.8	1.6	44	55	73	4.9	1203	95.1
16:15 14/10/2014	Comunicación engañosa	7	B	5		659	29	263	63	349	261	22.7	10.5	2.2	40	53	63	4.1	1042	89.5

Fondef IT15i10036 e ITI710051

Validación de los índices de lecturabilidad: primeros pasos

- Correlación con el gráfico de Fry, adaptado por Parodi (1986) para el español de Chile

- Correlación con la fórmula de lecturabilidad de Fernández-Huerta (1959) para el español

Fondef IT15i10036 e IT17I0051

Correlaciones entre los índices de TRUNAJOD y otros índices de lecturabilidad

	INDICES DE LECTURABILIDAD		INDICES DE TRUNAJOD								
			Indices Sintácticos				Indices Léxicos				Indice Semántico
	Indice de Fernández Huerta	Indice de Fry (Adaptado por Parodi)	LO	LC	IS	DFN	DeL	DiL	FP	FPL	DeP
Indice de Fernández-Huerta	1	,883**	,832**	,785**	,270	-,711**	-,103	-,130	-,150	-,141	,094
Indice de Fry (Adaptado por Parodi)	,883**	1	,688**	,689**	,141	-,674**	-,109	-,147	-,112	-,207	,002
LO	,832**	,688**	1	,689**	,578**	-,579**	-,434*	-,150	-,066	-,144	,002
LC	,785**	,689**	,689**	1	-,164	-,638**	-,244	-,082	-,408*	-,365	-,042
IS	,270	,141	,578**	-,164	1	-,057	-,312	-,085	,431*	,353	,099
DeP	,094	,002	,002	-,042	,099	,215	,472*	,258	,311	,335	1
DEL	-,103	-,109	-,434*	-,244	-,312	,133	1	,182	,119	,234	,472*
DiL	-,130	-,147	-,150	-,082	-,085	,383	,182	1	,187	,177	,258
DFN	-,711**	-,674**	-,579**	-,638**	-,057	1	,133	,383	,229	,339	,215
FP	-,150	-,112	-,066	-,408*	,431*	,229	,119	,187	1	,639**	,311
FPL	-,141	-,207	-,144	-,365	,353	,339	,234	,177	,639**	1	,335

Relación entre los índices de *Trunajod* y resultados obtenidos de la aplicación nacional de *Lectum 7, 5 y 3*

	B	***SE B***	***ß***		***P***
(Intercept)	37.433				
LO	-0.114	0.178	-0.644		= 0.519
LC	-0.278	0.327	-0.849		= 0.396
IS	3.857	1.818	2.121	*	> 0.05
DFN	1.524	0.158	9.640	***	> 0.001
DeL	-0.266	0.094	-2.827	**	> 0.01
DiL	-0.118	0.043	-2.756	**	> 0.01
FP	-0.005	0.001	-8.120	***	> 0.001
DeP	0.525	0.078	6.716	***	> 0.001

(*F* (8, 17126) = 55.01, *p* < .001, *R2* = .025, *R2 Ajustado* = .025)

Fondef IT15i10036 e IT1710051

Un estudio de lecturabilidad

- Estudio experimental cuyo objetivo fue establecer en qué medida la lecturabilidad de un texto depende no solo de sus propiedades estructurales (complejidad lingüística), sino también de la **naturaleza de las tareas de comprensión.**

- Evaluamos la comprensión lectora de 208 escolares de 8° Básico de 3 establecimientos educacionales de la provincia de Concepción.

Fondef TT15i10036 e TTI710051

Diseño de la investigación

- Se manipularon los textos para obtener uno de baja complejidad y otro de alta complejidad. Las diferencia en los grados de complejidad se midieron a través del Software **TRUNAJOD**.

Índices de complejidad textual

Complejidad del texto	Índices de complejidad								
	LO	LC	IS	DeP	DeL	DiL	DFN	FP	FPL
Texto baja complejidad	16.5	10.5	1,6	35	50	57	3.7	1260	124.5
Texto alta complejidad	30.9	12.6	2,5	38	51	62	3.5	944	94.4

Fuente: elaboración propia.

Fondef TT15i10036 e TTI710051

Resultados

MANOVA de los efectos del tipo de colegio y dificultad del texto sobre las tres dimensiones de la comprensión lectora de los estudiantes

Origen	Variable dependiente	Suma de cuadrados tipo III	Gl	Media cuadrática	F	p
Modelo corregido	Comp Textual	98,111[a]	3	32,704	7,302	,000
	Comp Pragmática	3,714[b]	3	1,238	5,521	,001
	Comp Crítica	,811[c]	3	,270	,769	,513
Intersección	Comp Textual	7911,131	1	7911,131	1766,384	,000
	Comp Pragmática	77,556	1	77,556	345,880	,000
	Comp Crítica	119,162	1	119,162	339,112	,000
Dificultad texto	**Comp Textual**	68,154	1	68,154	15,217	**,000****
	Comp Pragmática	2,745	1	2,745	12,243	**,001****
	Comp Crítica	,089	1	,089	,254	,615
Tipo de establecimiento	Comp Textual	27,337	2	13,669	3,052	,049
	Comp Pragmática	,813	2	,407	1,814	,166
	Comp Crítica	,733	2	,366	1,043	,354
Error	Comp Textual	913,658	204	4,479		
	Comp Pragmática	45,743	204	,224		
	Comp Crítica	71,685	204	,351		
Total	Comp Textual	8938,000	208			
	Comp Pragmática	127,000	208			
	Comp Crítica	191,000	208			

Fondef TT15i10036 e TT17I0051

Conclusiones

- La complejidad textual incide en el rendimiento lector a nivel global y en las dimensiones **comprensión textual** y **comprensión pragmática.**

- El rendimiento de los participantes en las tareas de **comprensión crítica** no resultó afectado por el nivel de lecturabilidad de los textos.

Fondef TT15i10036 e TTI710051

PROYECCIONES FUTURAS

Desarrollo de una nueva versión de TRUNAJOD, que potencie las capacidades de la versión actual.

Antes	Ahora
Análisis a nivel de palabra y oración.	Análisis a nivel textual.

Las **áreas de análisis textual** que la herramienta considerará en el examen de los textos son:

Cohesión referencial	Cohesión profunda	Narratividad	Sintaxis	Léxico
Grado en que las palabras y los conceptos se solapan o repiten de una oración a otra a lo largo del texto.	Grado en que las ideas, los eventos y la información del texto se ligan de modo coherente como un todo.	Grado en que los rasgos que caracterizan la estructura narrativa están presentes en los materiales escritos.	Examina la complejidad de las estructuras oracionales que dominan en los textos	Proporción en que aparecen rasgos tanto de la forma como del contenido de los vocablos, que resultan más complejos de procesar

La nueva versión de TRUNAJOD permitirá, mediante técnicas provenientes de la lingüística computacional e inteligencia artificial, lograr los siguientes resultados:

Un índice de lecturabilidad global

Calcula la lecturabilidad del texto calibrando la información procedente de todas las áreas de análisis e indica su adecuación a un nivel específico de escolaridad.

Cinco Índices de complejidad textual

Grado de dificultad asociado a cada una de las cinco áreas de análisis mencionadas anteriormente.

Una serie de índices específicos de complejidad

Entre ellos están los índices actuales de TRUNAJOD y los nuevos índices que se incorporarán al sistema.

Además, para promover el uso de esta herramienta en forma accesible se dispondrá de:

Una biblioteca

TRUNAJOD pondrá a disposición de sus usuarios -a modo de ejemplo- una "biblioteca", esto es, un conjunto amplio de textos escolares analizados por el sistema.

Plataforma para uso de la herramienta

Diseñada de modo que sea accesible en línea a todos los interesados a través de la web, en forma libre y mediante una interfaz fácil, expedita y amigable.

Etapas de la investigación

FASE 1:
Ajustes al sistema y generación de nuevos índices

a) Cambio a lenguaje Python.
b) Generación de nuevos índices mediante aplicación de técnicas de lingüística computacional.
c) Clasificación de la lecturabilidad de los textos del corpus.
d) Procesamiento computacional de los textos para extraer los índices de lecturabilidad.
e) Validación estadística.

FASE 2:
Implementación de TRUNAJOD en la plataforma virtual

a) Creación de la plataforma.
b) Sistema de acceso y registro.
c) Evaluación de la plataforma

“Desarrollo de una herramienta computacional para la evaluación automática de textos en el sistema escolar chileno”

(Proyecto Fondef IT17I0051)