Introducción al uso del paquete koRpus de R para el tratamiento de la complejidad léxica y la legibilidad en corpus lingüísticos

Nicolas Ballier, Paula Lissón, Verónica C Trujillo-González

Introducción al uso del paquete {koRpus} de R para el tratamiento de la diversidad léxica y la legibilidad en corpus lingüísticos

Nicolas Ballier[1] , Paula Lissón[1], Verónica C. Trujillo-González[2]
[1]Université Paris Diderot (USPC) – EA 3967, CLILLAC-ARP, France
[2] IATEXT (Instituto Universitario de Análisis y Aplicaciones Textuales), ULPGC, España

0. Resumen

Este poster ejemplifica la cadena de tratamiento que debe llevarse a cabo para la utilización de las métricas implementadas en {koRpus}, (Michalke, 2017) desde la importación del corpus en formato *txt*, pasando por la "tokenización" con TreeTagger, la extracción y visualización de los resultados, hasta la modelización estadística con técnicas de reducción de la dimensionalidad. Se trata de mostrar cómo el paquete {koRpus} puede ser utilizado para la investigación de la variedad léxica y la legibilidad. Estos dos conceptos son clave en la investigación de la adquisición y el aprendizaje de lenguas extranjeras, pero también en otras aplicaciones como la autoría de textos y la lingüística forense.

1. Funciones

Diversidad léxica

14 métricas de diversidad, la mayoría basadas en *type-to-token* ratio (TTR) pero con ciertas transformaciones matemáticas para obtener resultados más precisos. Ojetivo: medir la diversidad del vocabulario en un texto. Función: `lex.div()`

Legibilidad

35 métricas de legibilidad. Parámetros habituales : número de sílabas, número de frases, uso de palabras polísilabicas, uso de palabras "díficiles" (asociadas a inventorios preexistentes de palabras "sencillas" o "complejas" cf. Dale and Chall, Spache). Función: `readability()`

Hyphenation

Para el contar las sílabas automáticamente, el paquete utiliza un algorithmo de *hyphenation* (actualmente utilizado en LaTeX para la separación de las palabras). Esta función está disponible en el paquete {sylly} (Michalke, 2017). Función: `hyphen()`

2. Ejemplo: legibilidad y diversidad léxica en BROWN

- Importar el corpus utilizando tm.plugin.koRpus
- Tokenizarlo y etiquetarlo en partes del discurso (POS-tag) con TreeTagger (Schmid, 1994)
- Calcular las métricas de diversidad léxica

TTR, MSTTR, MTLD, MTLD-MA, Herdan's C (logTTR), Guiraud RootTTR, Carrol CTTR, Uber Index (U), Summer Index (S), Yule K (K), Maas a, Maas log, HDD

Tabla 1. Resultados de diversidad léxica en tres textos extraídos del corpus LOCNESS

CTTR	HD-D (vocd-D)	Herdan's C	Maas a	Maas lgV0	MATTR	MSTTR	MTLD	MTLD-MA	Root TTR	Summer	TTR	Uber index	Yule's K	vocd
7.63	34.71	0.87	0.22	4.92	0.69	0.68	71.18	74.49	10.79	0.86	0.43	21.44	130.17	85.65
9.19	35.67	0.9	0.19	5.82	0.77	0.76	134.45	140.73	13	0.9	0.54	28.62	114.35	101.9
7.19	35.24	0.88	0.21	4.88	0.72	0.71	93.91	98.01	10.17	0.87	0.47	21.75	107.57	95.04

- Calcular las métricas de legibilidad

ARI, Bormuth, Coleman (4 métricas), Coleman-Liau, Dale-Chall, Danielson-Bryan (2 métricas), Dickes-Steiwer, DRP, ELF, Farr-Jenkins-Paterson, Flesch, Flesch-Kincaid, FOG, FORCAST, Fucks, Linsear-Write, LIX, nWS1, nWS2, nWS3, nWS4, RIX, SMOG, Spache, Strain, Traenkle-Bailer (2 métricas), TRI, Tuldava, Wheeler-Smith

Tabla 2. Resultados de legibilidad en tres subpartes temáticas del corpus BROWN

Temática (genre)	ARI	Bormuth	C1	C2	C3	C4	CL	DC	DB1	DB2	DS	DRP	ELF	FJP	FLESCh	FK	FOG	FORCAST	FUCKS	LW	LIX	NWS1	NWS2	NWS3	NWS4	RIX	SMOG	Spache	STRAIN	TB1	TB2	TRI	TULDAVA	WS
reportage	11	29	47	46	46	46	45	13	8	45	62	-2779	7	53	57	10	12	10	98	13	45	6	7	7	7	5	13	4	9	31	38	-15312	5	70
editorial	11	30	48	47	48	48	46	18	8	47	62	-2923	7	56	57	10	12	10	93	13	44	6	7	7	7	5	13	4	9	32	38	-5999	5	65
reviews	11	29	46	45	45	46	44	15	8	45	57	-2780	7	52	55	11	13	10	99	13	46	7	7	7	8	5	13	5	10	31	36	-5395	5	71

3. Visualización

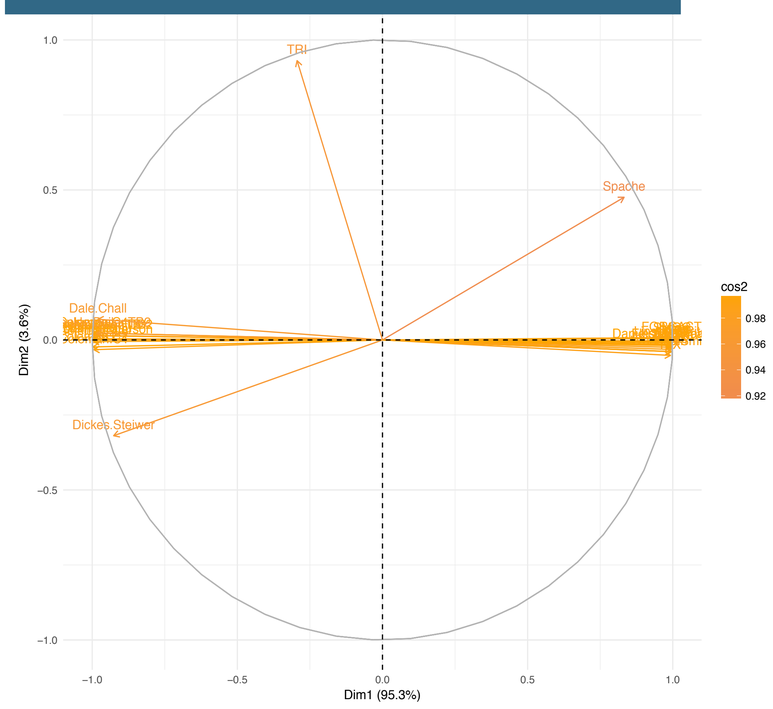

Figura 1. Mapa de las métricas de legibilidad utilizando ACP.

Conclusiones

PROS

- Fácil de usar
- Gratis, disponible para PC, Mac y Linux
- Rápido (45 métricas calculadas a la vez)
- Flexible (se adapta a varias lenguas)
- Lista de difusión y preguntas activa
- En constante desarrollo

CONTRAS

- Posibles problemas con la instalación de TreeTagger (aunque existe la alternativa *tokenizer()*)
- Funciona en línea de comanda
- Requiere el plugin tm.plugin.koRpus (posibles problemas con las versiones del paquete tm)
- El formato y los objetos creados tienen un formato propio, lo que no facilita la interoperabilidad de los resultados
- Algunas métricas todavía no han sido validadas

Contacto

Nicolas Ballier
nicolas.ballier@univ-paris-diderot.fr
Paula Lissón
paula.lisson@etu.univ-paris-diderot.fr
Verónica C. Trujillo-González
veronica.trujillo@ulpgc.es

Script del póster: https://github.com/paulalisson

Bibliografía

1. **Ballier, N., & Lissón, P.** (2017). R-based strategies for DH in English Linguistics: a case study. Presented at the Teaching NLP for Digital Humanities, workshop of the German Society for Computational Linguistics conference (GSCL 2017), Berlin, Germany
2. **McCarthy, P. M., & Jarvis, S.** (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392)
3. **Michalke, M.** (2017). Package koRpus: An R Package for Text Analysis (Version 0.10-2). Retrieved from http://reaktanz.de/?c=hacking&s=koRpus
4. **Lissón, P.** (2017). *A corpus-based study of lexical diversity and readability in EFL learners' productions: towards a critical statistical approach* (Unpublished MA thesis.). Université Paris Diderot., Paris, France
5. **Lissón, P., & Ballier, N.** (submitted). Investigating lexical progression through lexical diversity metrics and vocabulary growth curves in a corpus of French L3.
6. **Lu, X.** (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/https://doi.org/10.1075/ijcl.15.4.02lu
7. R Core Team. (2016). R: A language and environment for statistical computing. (Version 3.3.1 (2016-06-21)). Vienna, Austria.: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
8. **Schmid, H.** (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, *43*, 28.
9. **Vajjala, S., & Meurers, D.** (2012). On improving the accuracy of readability classification using insights from second language acquisition (pp. 163–173). Presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics