Complexity trade-offs in written and spoken registers

Yoon Mi Oh, Christophe Coupé, Dan Dediu,
Matt Stave, François Pellegrino

July 02, 2026

COLLEGIUM
DE LYON
Université de Lyon

Introduction (I)

In human communication, information is transmitted primarily by speaking, signing, writing, or non-verbal communication.

(1) In the written register, lexical and grammatical choices occur more systematically than in the spoken register due to time and cognitive constraints, resulting in more complex sentence structures and words.

(2) However, the redundancy of written English at the letter level (Shannon, 1951) and the semantic redundancy of spoken English at the clause level (Sivan & Tsodyks, 2025) are the same: about 50%.

What does it imply?

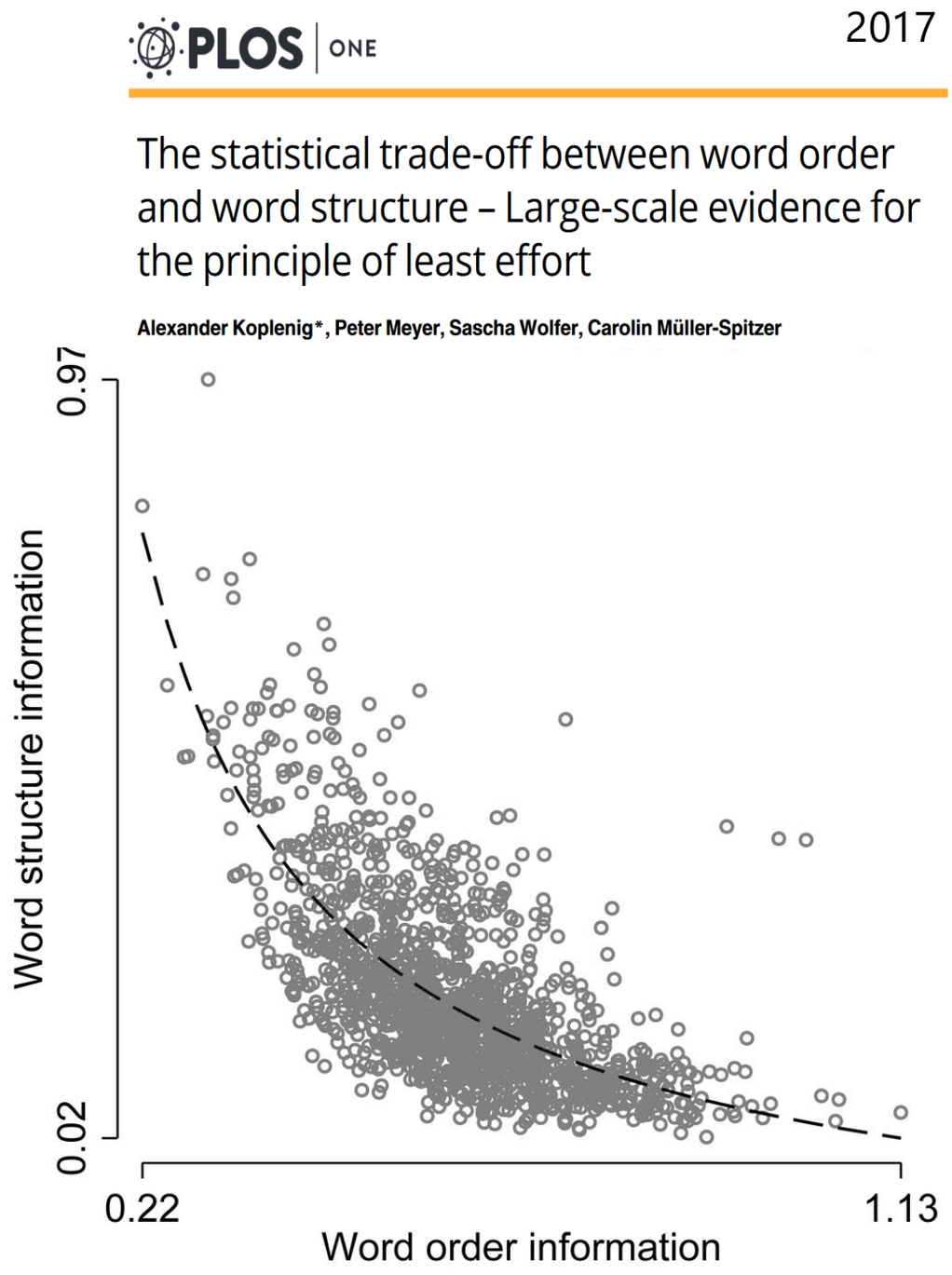
Shannon (1951). Prediction and entropy of printed English. *Bell System Technical Journal 30*(1). 50-64.
Mehrabian (1971). *Silent messages* (Vol. 8, No. 152, p. 30). Belmont, CA: Wadsworth.
Sivan & Tsodyks. (2025). Information rate of meaningful communication. *Proceedings of the National Academy of Sciences 122*(25). e2502353122.

Introduction (II)

(3) In previous studies,

cross-language comparisons are mostly based on large parallel written corpora; some convincingly show that there exists a balance between within-word and across-word information.

Does this balance hold true across both registers?



Koplenig, Meyer, Wolfer, Müller-Spitzer (2017). The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLoS ONE 12*(3): e0173614. https://doi.org/10.1371/journal.pone.0173614

Introduction (III)

Rationale: the redundancy of written English and the semantic redundancy of spoken English are the same: about 50%.

Research question 1: How do written and spoken registers regulate information transmission within and across words?

→ Do they differ in terms of complexity trade-offs across and within words?

→ Why words as a basic unit? Because they serve as a modality-neutral unit which bridges the gap between morphological (within words) and syntactic (across words) structures in both registers.

→ We expect that a consistent complexity trade-off across registers will reveal a universal cognitive constraint optimising information transmission.

Shannon (1951). Prediction and entropy of printed English. *Bell System Technical Journal 30*(1). 50-64.
Sivan & Tsodyks. (2025). Information rate of meaningful communication. *Proceedings of the National Academy of Sciences 122*(25). e2502353122.

Introduction (IV)

Rationale: the redundancy of written English and the semantic redundancy of spoken English are the same: about 50%.

Research question 2: Do written and spoken registers exhibit any difference regarding lexical complexity?

→ **Does the role of lexical complexity differ across registers with respect to syntactic and morphological complexity?**

→ **We expect that in written registers, due to lower cognitive and time constraints on word choice, lexical complexity would exhibit a more distinct role within the complexity trade-off in comparison to spoken registers.**

Shannon (1951). Prediction and entropy of printed English. *Bell System Technical Journal 30*(1). 50-64.
Sivan & Tsodyks. (2025). Information rate of meaningful communication. *Proceedings of the National Academy of Sciences 122*(25). e2502353122.

Data compilation (I)

Corpus compilation: a cross-language corpus of spoken and written registers by assembling data from 3 corpora.

Data for spoken register

1) DoReCo Language Documentation Reference Corpora **(v2.0) (Seifart et al., 2024)**

- Transcriptions of spontaneous speech in 53 areally and genealogically diverse languages.

- Mostly personal narratives, traditional narratives and descriptions.

- 38 languages supplying morphological glosses selected in this study.

https://doreco.huma-num.fr/languages

Seifart, Paschen & Stave (eds.). 2024. *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). https://doi.org/10.34847/nkl.7cbfq779.

Data compilation (II)

Corpus compilation: a cross-language corpus of spoken and written registers by assembling data from 3 corpora.

Data for written register:

2) the Parallel Bible Corpus (Mayer & Cysouw, 2014)

3) Manually collected additional Bible translations

- A subset of 31 languages from DoReCo (spoken registers) is matched with Bible translations in 14 languages from the Parallel Bible Corpus and 17 languages manually collected (written registers).

- For 10 languages, translations from close proxy languages are used after validation by colleagues.

- Except for 5 languages (Goemai, Jejuan, Fanbyak (North Ambrym), Daakie, Teop), 26 languages rely on a fully parallelised translation of the Gospel of Luke.

Mayer & Cysouw. 2014. Creating a massively parallel Bible corpus. *Oceania 135*(273). 40.

Data compilation (III)

Data for spoken register

(1) DoReCo (Seifart et al., 2024): a subset of 31 languages matched with

Data for written register

(2) the Parallel Bible Corpus for 14 languages (Mayer & Cysouw, 2014)

Language	Proxy	Fully Parallelised (Y/N)
Arapaho	-	Y
Bora	-	
Cabécar	-	
Cashinahua	-	
Dalabon	Kunwijku	
Dolgan	Yakut	
English (Southern England)	-	
Mojeño Trinitario	-	
Northern Kurdish (Kurmanji)	-	
Pnar	Khasi	
Sanzhi Dargwa	Dargwa	
Sümi	-	
Tabasaran	-	
Texistepec Popoluca	Popoluca de la Sierra	

(3) Manually collected Bible translations for 17 languages

Source / Author	Language	Proxy	Fully Parallelised (Y/N)
Wycliffe Bible Translators, Inc.	Fanbyak	North Ambrym	N
	Gorwaa	-	Y
	Northern Alta	Casiguran Agta	Y
	Ruuli	-	Y
	Teop	-	N
American Bible Society	Hoocąk	-	Y
bejalanguage.org	Beja	-	Y
Bible Society of the South Pacific	Vera'a	Mota	Y
Bulatova & Myreeva	Evenki	-	Y
Indonesian Bible Society	Yali (Apahapsili)	-	Y
Instituto Lingüístico de Verano	Movima	-	Y
Jeju Hyangto Munhwasa	Jejuan	-	N
Krifka & Taho	Daakie	-	N
Mission Evangélique Réformée Néerlandaise	Kakabe	Mogofin	Y
Nigeria Bible Translation Trust	Goemai	-	N
Seriken & eBible.org	Urum	Turkish	Y
Thieberger	Nafsan (South Efate)	-	Y

Seifart, Paschen & Stave (eds.). 2024. *Language Documentation Reference Corpus (DoReCo) 2.0.* Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). https://doi.org/10.34847/nkl.7cbfq779.
Mayer & Cysouw. 2014. Creating a massively parallel Bible corpus. *Oceania 135*(273). 40.

Methods (I)

Three measures of information

- They were chosen due to their robustness (Oh & Pellegrino, 2023).

(1) Inter-Word Information (IWI): Average amount of information across words, estimated by Kolmogorov complexity (Juola, 1998), a widely used measure of the complexity of syntactic structures by means of compression ratios for texts.

$$MeanC_L = 1 - \frac{1}{10}\sum_{i=1}^{10}\frac{CA_L^i}{CP_L^i}$$

the average compression ratio between the change in size of compressed text files before (CA_L) and after (CP_L) word order distortion by a random permutation iterated 10 times.

- **In a putative language with a totally free word-order, $MeanC_L$ is close to zero.**
- **A language with a strict word order results in a $MeanC_L$ close to one.**

Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics 5*(3). 206-213.

Methods (II)

Three measures of information

Towards robust complexity indices in linguistic typology
A corpus-based assessment
STUDIES IN LANGUAGE 2023
Yoon Mi Oh and François Pellegrino

- They were chosen due to their robustness (Oh & Pellegrino, 2023).

(1) Inter-Word Information (IWI): Average amount of information across words, estimated by Kolmogorov complexity (Juola, 1998), a widely used measure of the complexity of syntactic structures by means of compression ratios for texts.

$$IWI_L = \frac{MeanC_L}{MeanC_{ENG}}$$

a pairwise comparison between the average compression ratio in a target language ($MeanC_L$) and our reference language (English, $MeanC_{ENG}$).

- **English is used as a reference language in order to normalise the degree of flexibility of word-order across our language samples.**

Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics 5*(3). 206-213.

Methods (III)

Three measures of information

(2) Word Information Density (WID): Average amount of information within words

$$WI_L^v = \frac{S_L^v}{N_L^v}$$

the average amount of information conveyed per word, defined as the semantic content *S* of verse *v* in language *L* divided by the number of its words *N*.

- **Since a parallel corpus is used, the semantic content *S* of each verse is assumed to be equivalent between English and a target language.**

$$WID_L = \frac{1}{V}\sum_{v=1}^{V}\frac{WI_L{}^v}{WI_{ENG}^v} = \frac{1}{V}\sum_{v=1}^{V}\frac{S_L^v}{N_L^v}\times\frac{N_{ENG}^v}{S_{ENG}^v} = \frac{1}{V}\sum_{v=1}^{V}\frac{N_{ENG}^v}{N_L^v}$$

a pairwise comparison between the number of words of verse *v* in our reference language (English, N_{ENG}) and in a target language *L* (N_L).

- **English selected as a reference given all languages used in DoReCo had English translations.**
- **Using a reference eliminates the need for computing absolute average amounts of information + allows comparisons by means of relative normalisation between languages.**

Methods (IV)

Three measures of information

(3) Measure of Textual Lexical Diversity (MTLD): Lexical complexity

$$TTR_L = \frac{T_L}{N_L}$$

the degree of the lexical diversity and morphological productivity calculated for language *L* as the ratio of the number of unique word types T_L over the word token count N_L.

- **A critical limitation of TTR is that it is highly dependent on text length.**

$$MTLD_L = \frac{N_L}{2}\left(\frac{1}{F_{fwd}} + \frac{1}{F_{bwd}}\right)$$

the bi-directional (F_{fwd}, F_{bwd}) average number of word tokens N_L required to reach a standard TTR threshold of 0.72 (McCarthy & Jarvis, 2010).

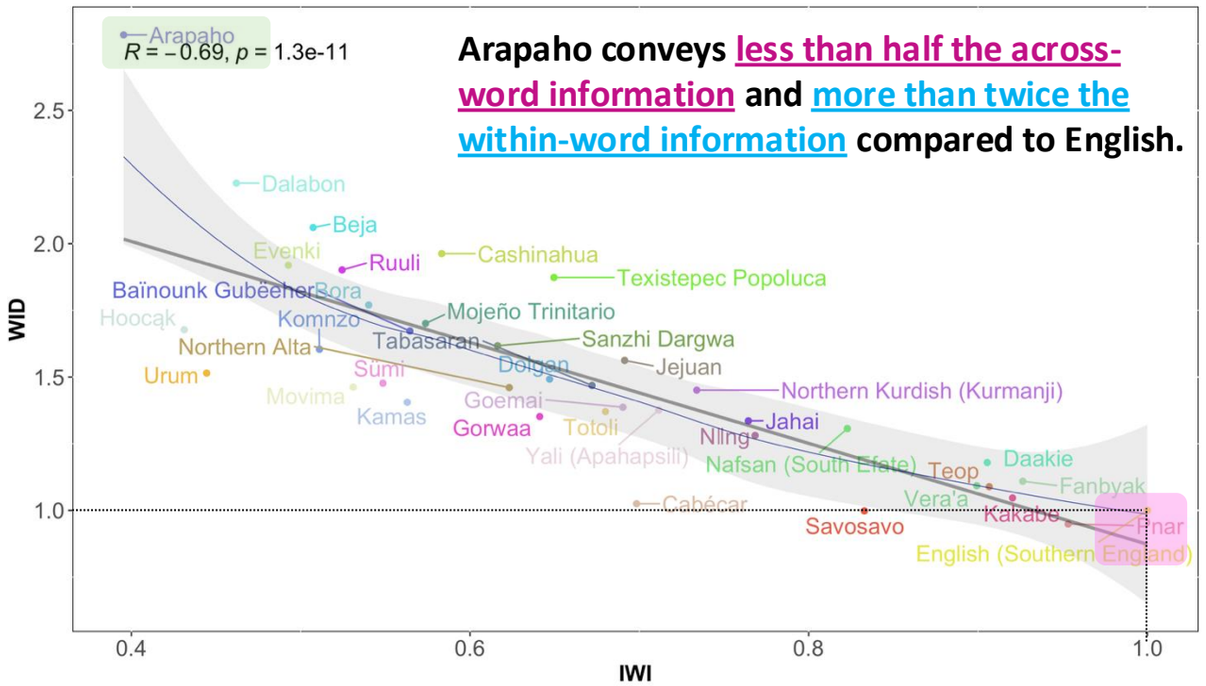
- MTLD is a text-length invariant tool for measuring complexity of lexical structures.

- In a text with high lexical diversity, TTR drops slowly, leading to fewer factors *F* and a higher MTLD.

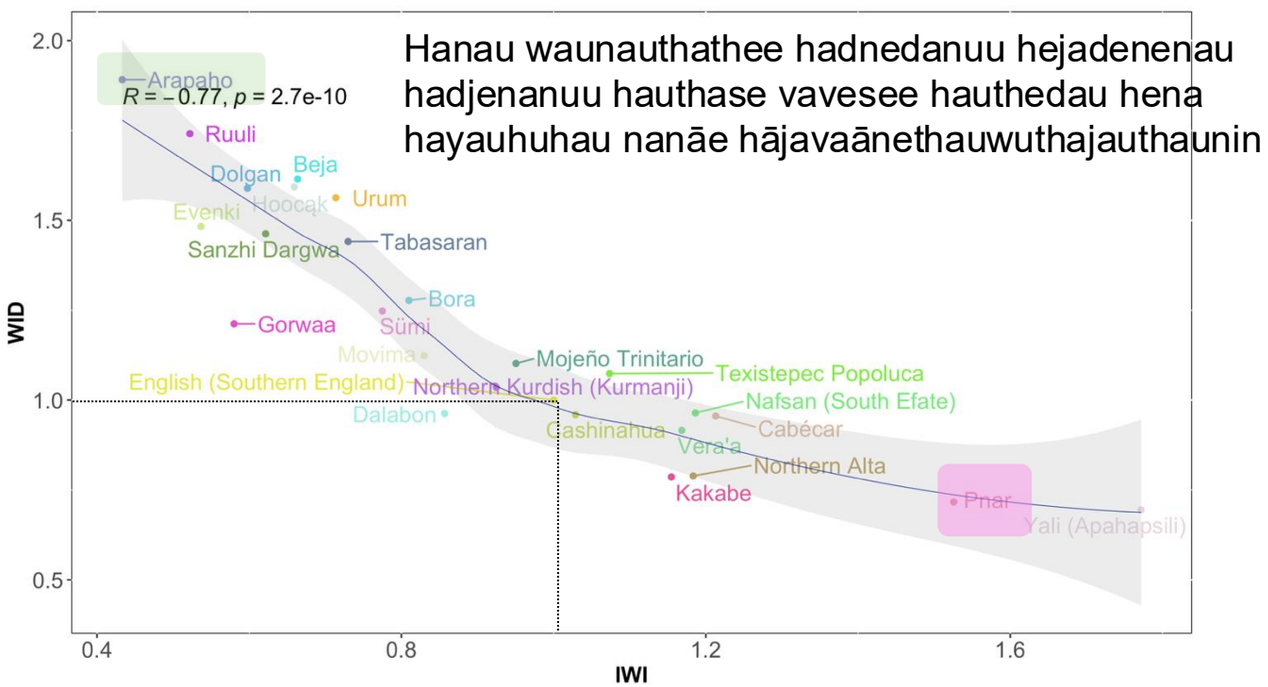
Results (I)

Research question 1: How do written and spoken registers regulate information transmission within and across words?

A trade-off between information within words (WID) and across words (IWI) exists across typologically diverse languages in both written and spoken registers



The spoken register (DoReCo Corpora, 38 languages)



The written register (Bible Corpora, 31 languages)

(i) A large range of variation in both within-word and across-word information

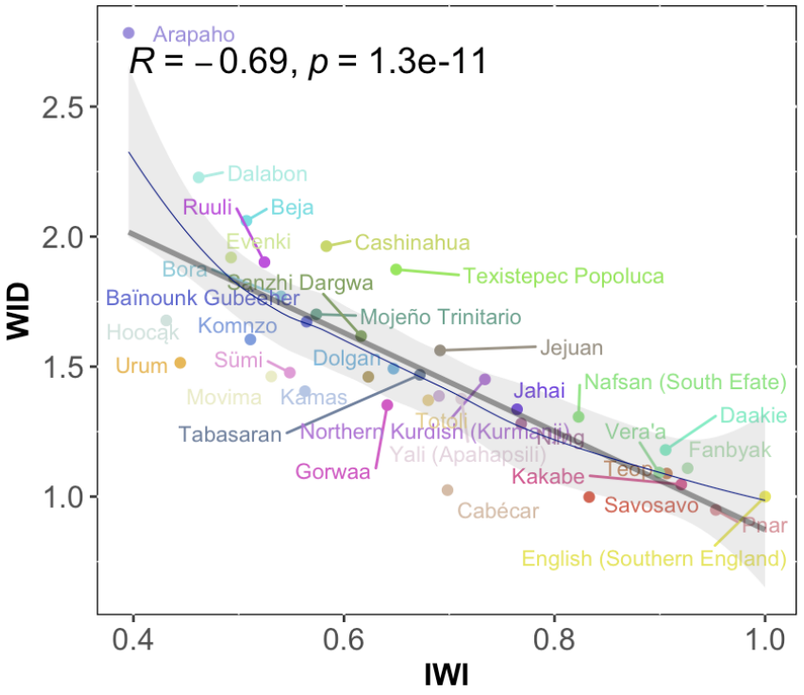
(ii) No language is characterized by low values on both dimensions

(iii) Overestimation of WID in DoReCo Corpora due to simple translations in English

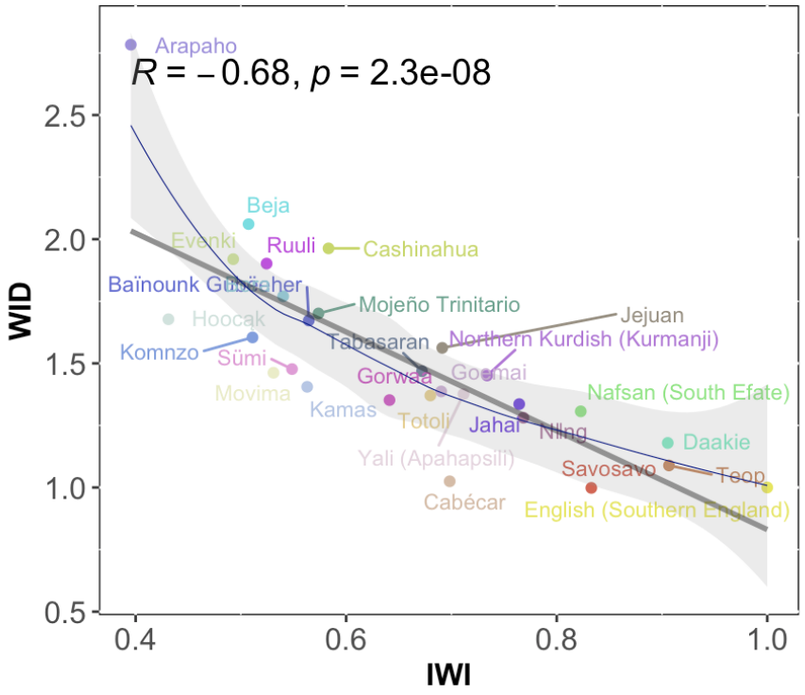
Results (II)

Assessing the robustness of the spoken register across language subsets:

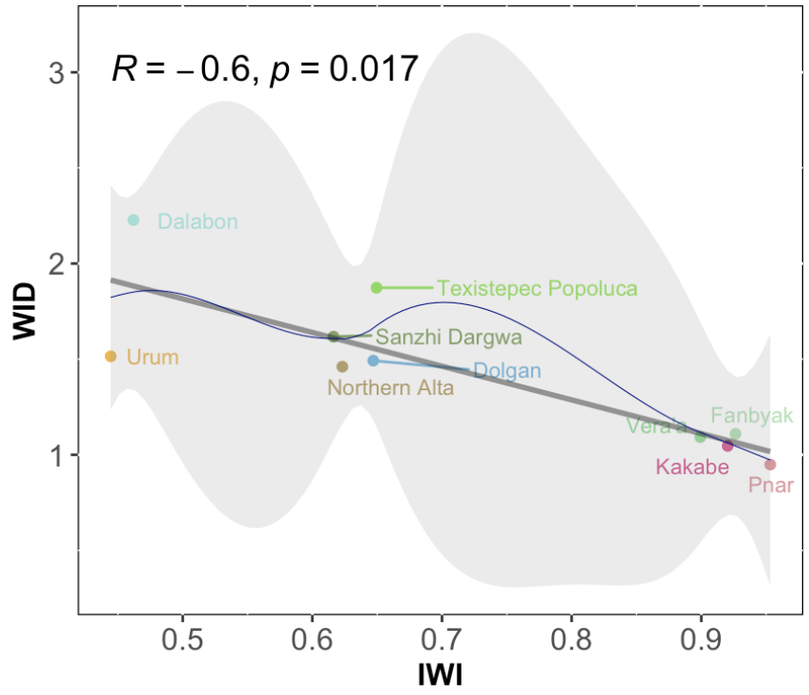
The correlations remain consistent across different language subsets.



Full DoReCo dataset (N=38)



Subset of 28 languages matched with actual Bible translations



Subset of 10 languages matched with proxy Bible translations

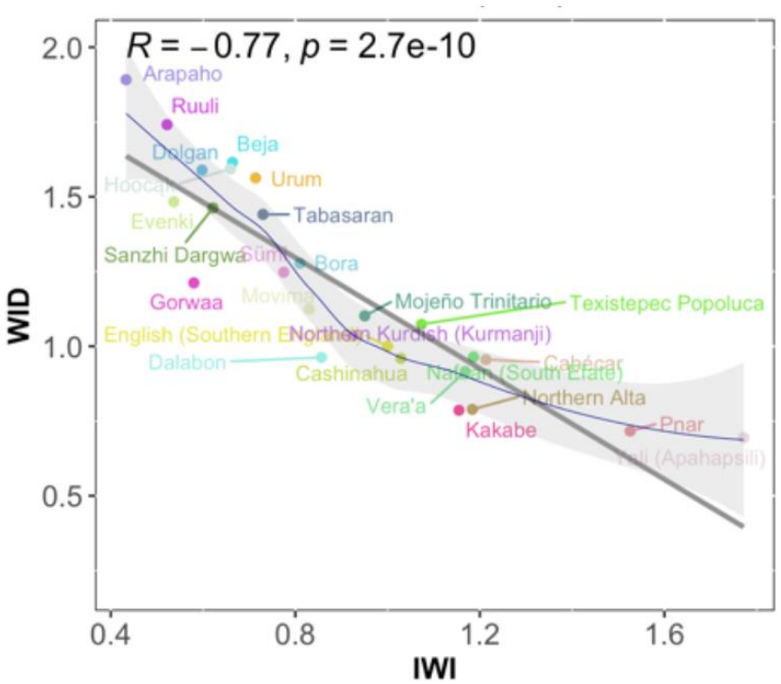
(i) Robust and highly significant negative correlations are maintained across all subsets.

(ii) The subset of 10 languages (mapped to written proxies) exhibits a stable trade-off pattern.

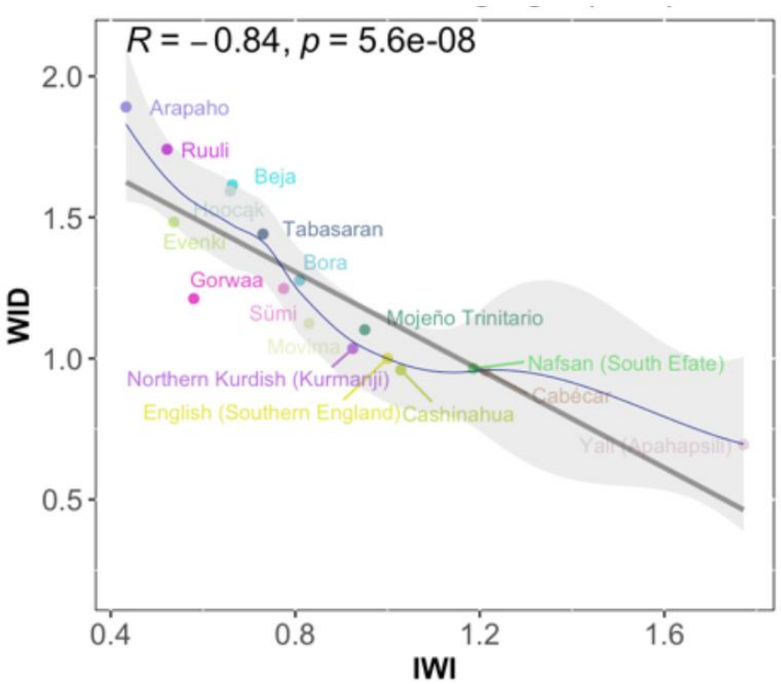
Results (III)

Assessing the robustness of the written register across language subsets:

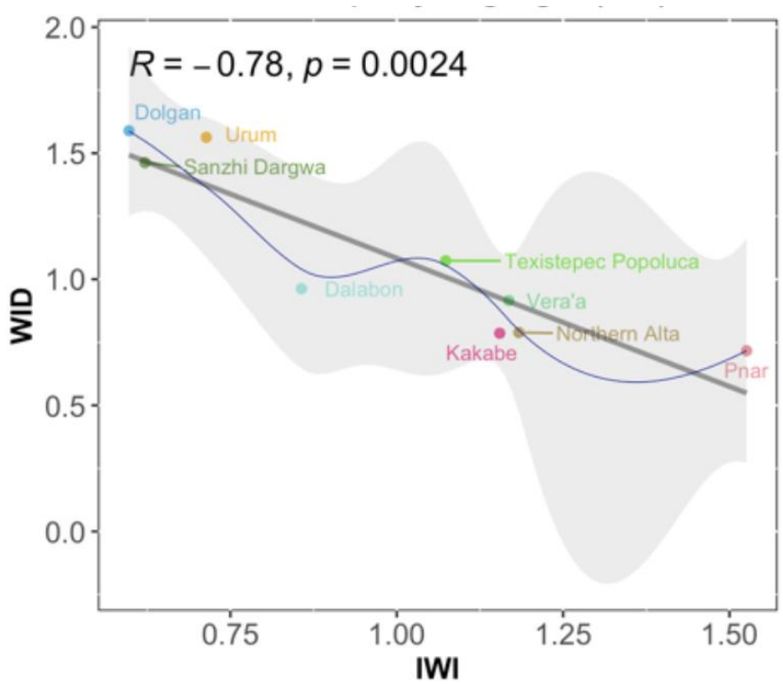
The correlations remain consistent across different language subsets.



Fully Parallelised Bible translations (N=26)



Subset of actual Bible translations (N=17)



Subset of proxy Bible translations (N=9)

(i) Robust and highly significant negative correlations are maintained across all subsets.

(ii) The subset of 9 proxy languages exhibits a stable trade-off pattern.

Results (IV)

Research question 2: **Do written and spoken registers exhibit any difference regarding lexical complexity?**

→ **Yes, an asymmetric pattern of lexical complexity exists between registers.**

Comparison of the Kendall correlations between MTLD and WID

Register	Source	Kendall's τ_b	*p*-value
Spoken	DoReCo (N=38)	$\tau_b = 0.19$	$p = 0.1$
Written	Bible (N=26)	$\tau_b = 0.56$	$p = 2.286 \times 10^{-5}$

(i) Lexical complexity does not seem to play any significant role in transmitting information in the spoken register.

(ii) It confirms a more distinct role of lexical complexity in the written register.

Discussion

(1) A strong, significant trade-off between information within words and across words in both spoken and written registers

- **They share similar cognitive mechanisms for balancing information transmission within and across words, to smooth and minimise processing effort.**

(2) Distinct role of lexical complexity between the registers

- Lexical complexity plays a distinct role in the written register by allowing writers to maximise text elaboration in terms of morphological and lexical richness.

(3) Some existing limitations

- Use of close proxy Bible translations for 10 languages

- Simple English translations used as a reference in the DoReCo corpus

- Limited genre diversity (mostly based on narratives)

Conclusion

(1) Universal cognitive mechanisms

- While languages vary in terms of their preference for morphological or syntactic structures, they share similar cognitive mechanisms for regulating information encoding within and across words.

(2) Integration of prosodic cues and temporal aspects (for the spoken register), as well as sociolinguistic variables

Thank you!

https://yoonmioh.github.io