Phonological and morphological complexity from a quantitative and typological perspective

Yoon Mi OH
Ajou University

IWMLC September 12, 2019

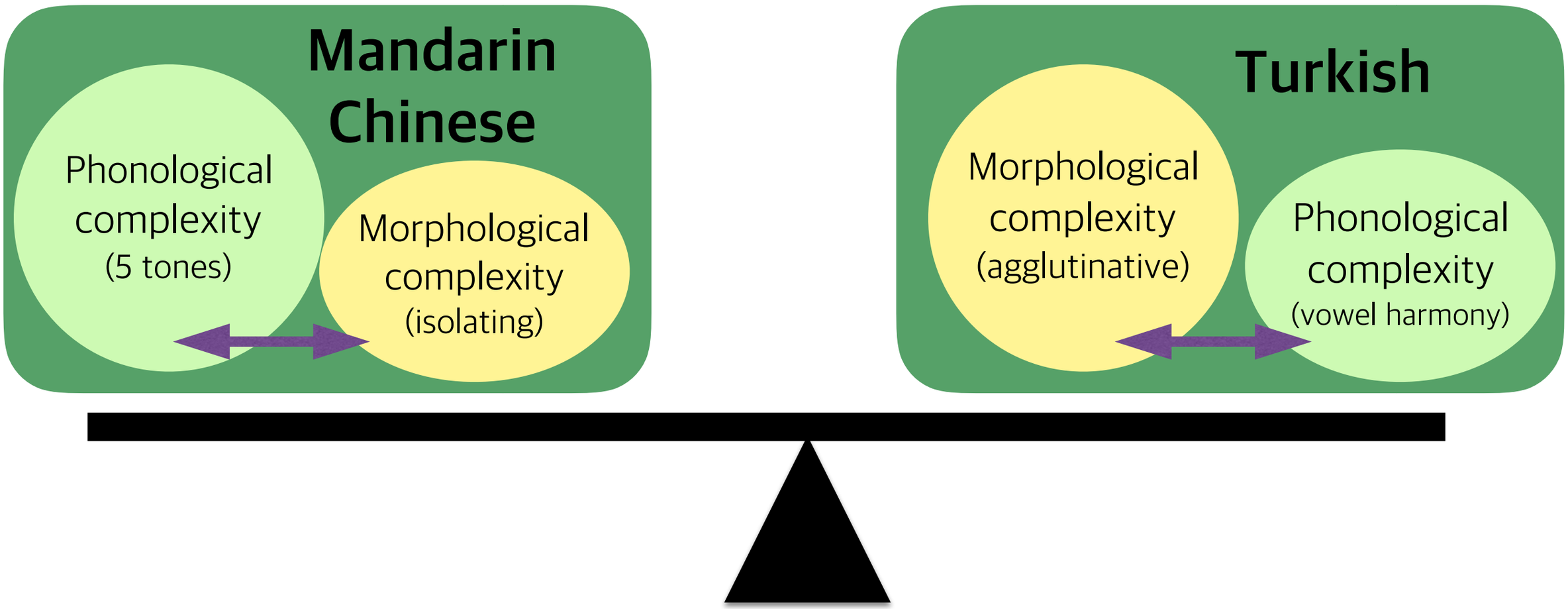
Hypothesis

Language as a **macrosystem consisting of microsystems** (i.e. linguistic modules such as morphology, phonology, semantics, and syntax)

The **equal overall complexity hypothesis**

: all languages are considered equal in terms of their overall complexity.

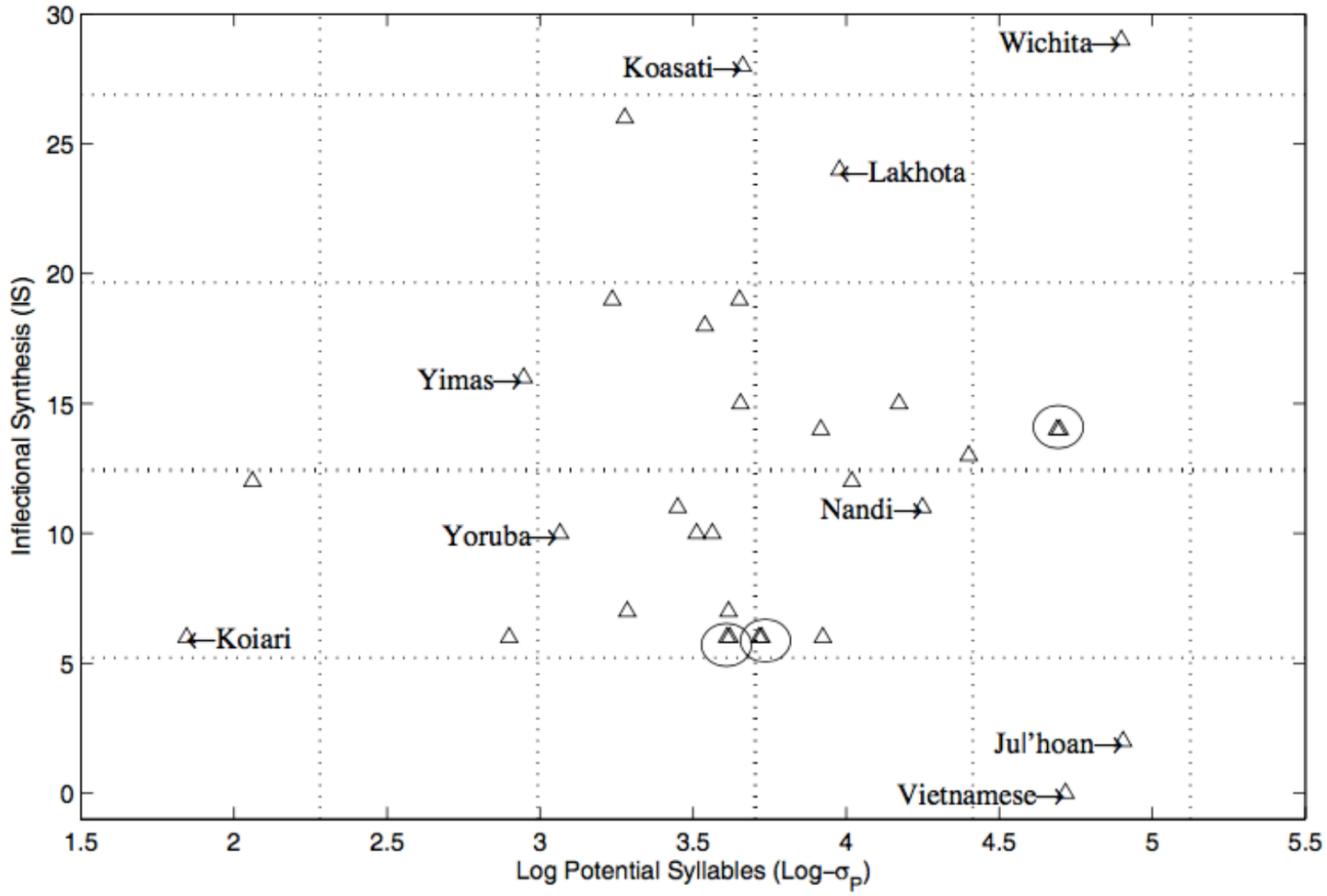
Mandarin Chinese vs. Turkish



Related works (I)

Shosted, R. (2006).
Correlating complexity: A typological approach

➡ **No significant correlation** was found between the **number of potential syllables** (log-transformed, x-axis) and the **number of verbal inflectional markers** (y-axis) in 32 languages.



(Pearson's r = **0.0704**; p-value = **0.702**; N = 32)

Related works (II)

Dahl, Ö. (2004). *The Growth and Maintenance of Linguistic Complexity*

The **distinction of linguistic complexity** accounts for the **difference between the methodologies** used to measure complexity.

1. **System complexity**: measures the "**richness**" of a system in terms of its resources. ➡ **Phonological complexity**
2. **Structural complexity**: applies to the **structure of expressions.**

➡ **Morphological complexity**

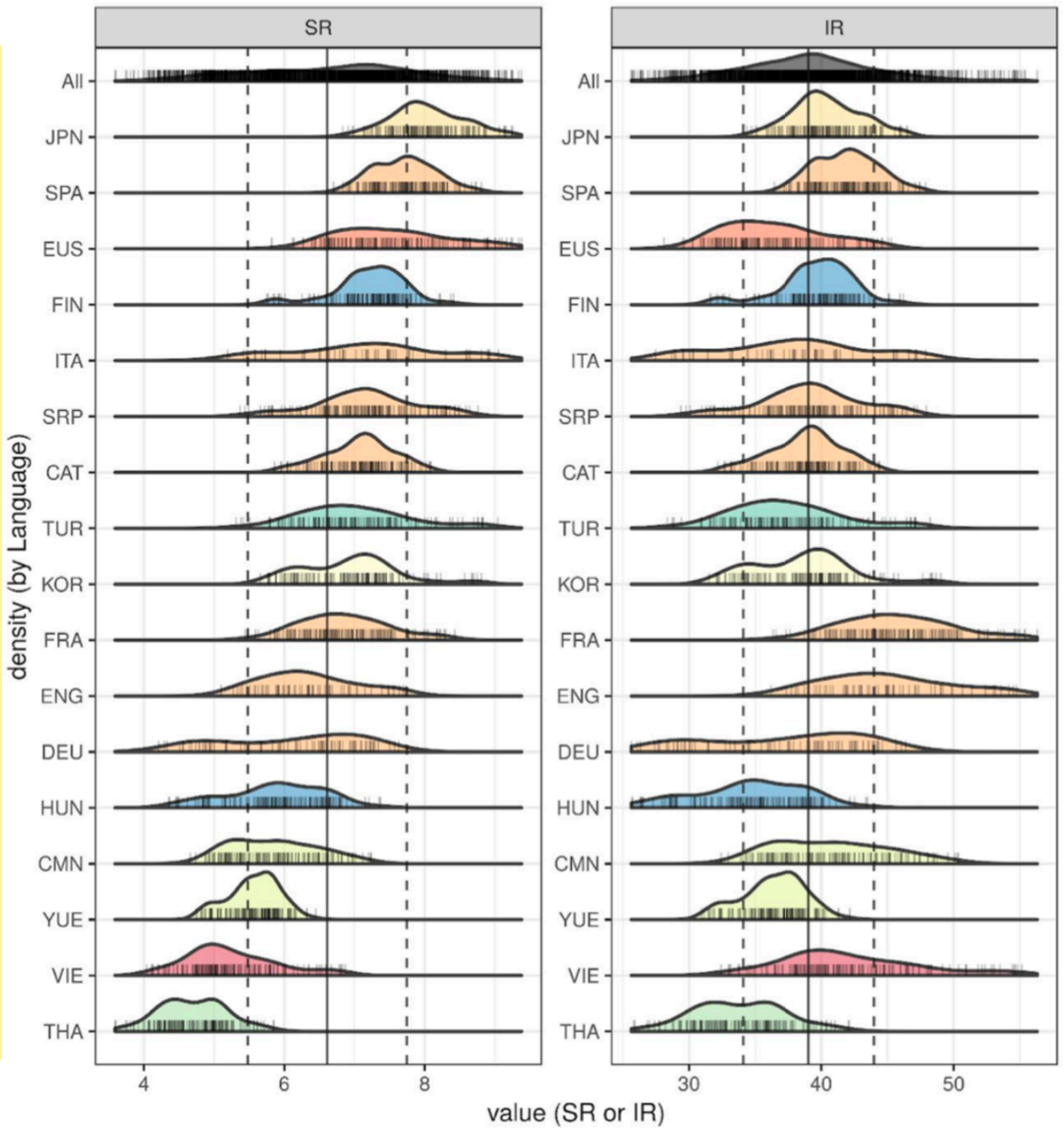
Coloma, G. (2014). *La existencia de correlación negativa entre distintos aspectos de la complejidad de los idiomas*

A negative correlation was found between **phonological complexity** and **morphological complexity** in 40 languages which was again negatively correlated with **syntactic complexity**.

Related works (III)

Coupé, Oh, Dediu & Pellegrino (2019). ***Different languages, similar efficiency: comparable information rates across the human communicative niche***

Using a large cross-linguistic corpus of 17 languages, we show that languages are more similar in **information rates** (information per second, about 39 bits per second on average) than in **information density** (information per syllable), or **speech rate** (number of syllables per second).



Objective (I)

I. **Defining linguistic parameters** to quantify linguistic complexity

➡ Methods for **quantifying linguistic complexity** differ as a function of linguistic module in question.

Phonological complexity: Bottom-up or usage-based approach

Morphological complexity: Top-down or grammatical approach

Objective (II)

2. Assessing **trade-offs** between **morphological** and **phonological complexity** by means of **multilingual parallel text corpus**

Data preprocessing

- Fully Parallelized Bible Corpus (Track A)
- Automatic phonological transcription and **syllabification**
  - in 8 languages (Basque, German, English, Finnish, French, Georgian, Russian, Spanish): automatic G2P tool (Reichel & Kisler, 2014)
  - in 2 languages (Korean, Turkish): syllabified by a program written in a bash shell script (Oh, 2015)

Parameters -Morphological complexity

Method adopted from [Lupyan & Dayle, 2010]

- **29 linguistic features** accounting for the **inflectional morphology** are chosen from **WALS**.

- **Calculation** of the score of morphological complexity: By **distinguishing between lexical (-1) and inflectional morphological coding strategies (0)**, summing the assigned values and normalizing it.

Inflectional morphology

: an effective tool for complexity reduction by simplifying the description of whole grammar

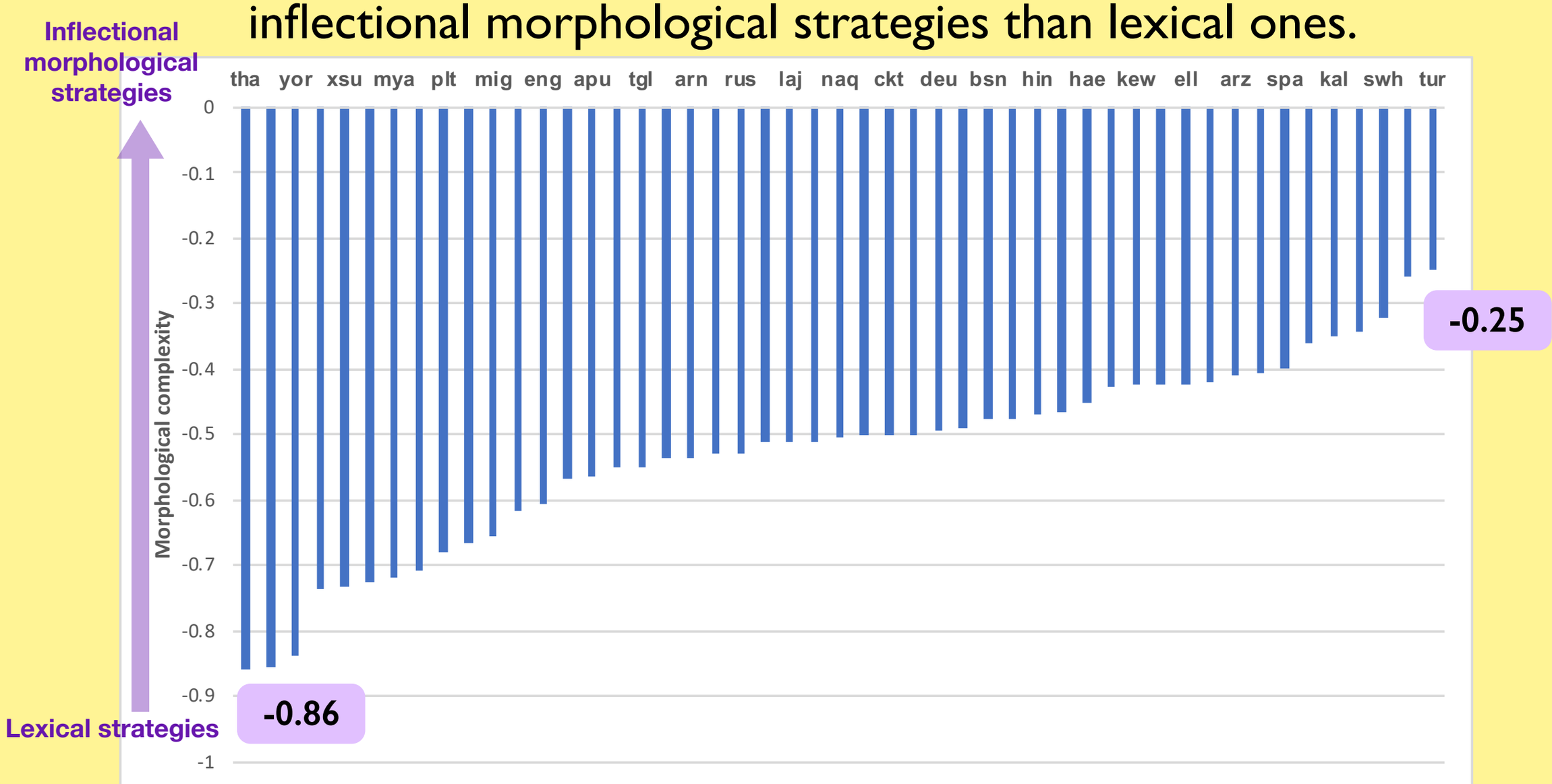
[Ackerman & Malouf, 2013]

Parameters -Linguistic features taken from WALS

Feature (WALS code)	Description
1. Fusion of selected inflectional formatives (20A)	The degree to which grammatical markers (formatives) are phonologically connected to a host word or stem
2. Prefixing vs. suffixing in inflectional morphology (26A)	The degree to which languages use prefixes or suffixes in their inflectional morphology
3. Number of cases (49A)	The number of case categories represented in a language's inflectional system
4. Case syncretism (28A)	The ways in which a single inflected form represents two or more case functions
5. Alignment of case marking of full noun phrases (98A)	The ways in which core argument noun phrases are marked to indicate which particular core argument position they occupy
6. Inflectional synthesis of the verb (22A)	The strategies of expressing grammatical categories either by individual words or by affixes attached to some other words
7. Alignment of verbal person marking (100A)	The ways in which the two arguments of the transitive verb aline with the sole argument of the intransitive verb

Parameters -Morphological complexity

If morphological complexity is closer to 0, it means languages employ more inflectional morphological strategies than lexical ones.



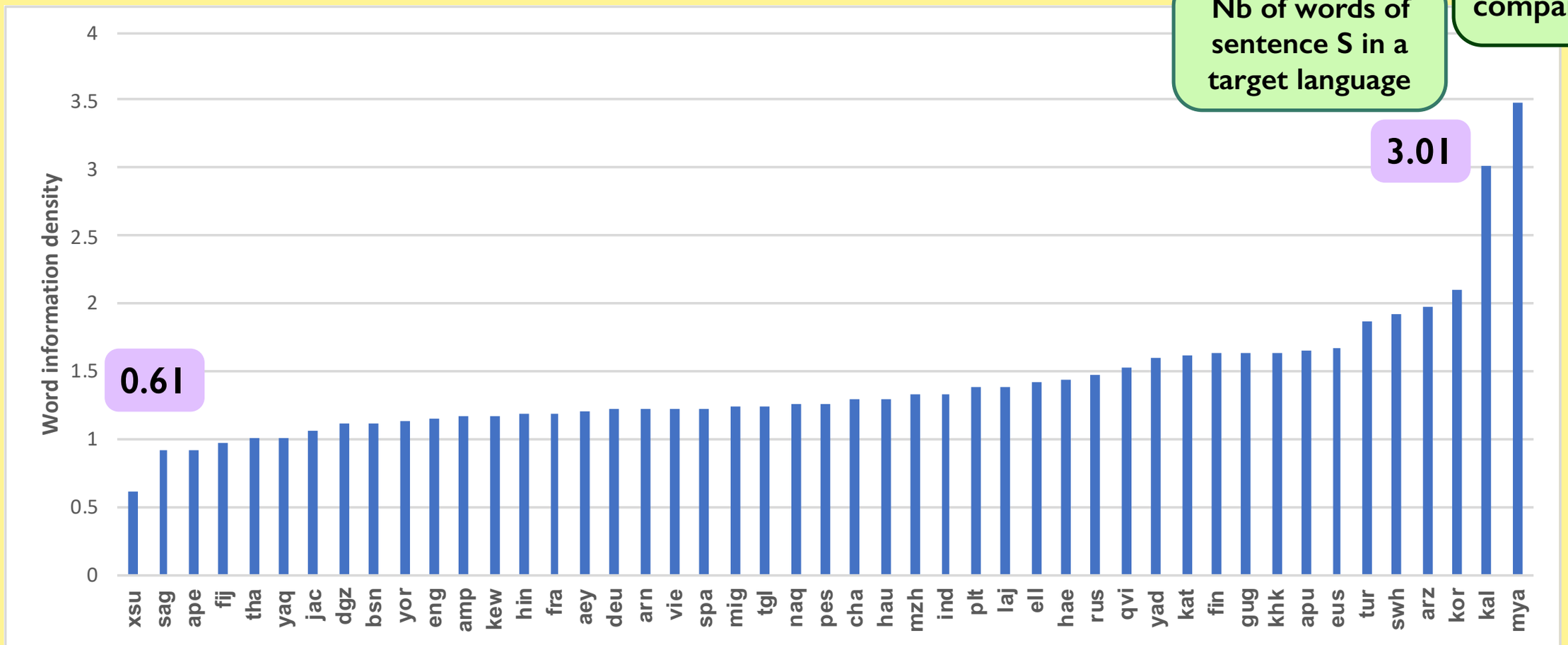
Parameters -Word Information Density

$$WID_L = \frac{1}{S}\sum_{s=1}^{S}\frac{WI_L^S}{WI_{THA}^s} = \frac{1}{S}\sum_{s=1}^{S}\frac{C_L^s}{\omega_L^s}\times\frac{\omega_{THA}^s}{C_{THA}^s} = \frac{1}{S}\sum_{s=1}^{S}\frac{\omega_{THA}^s}{\omega_L^s}$$

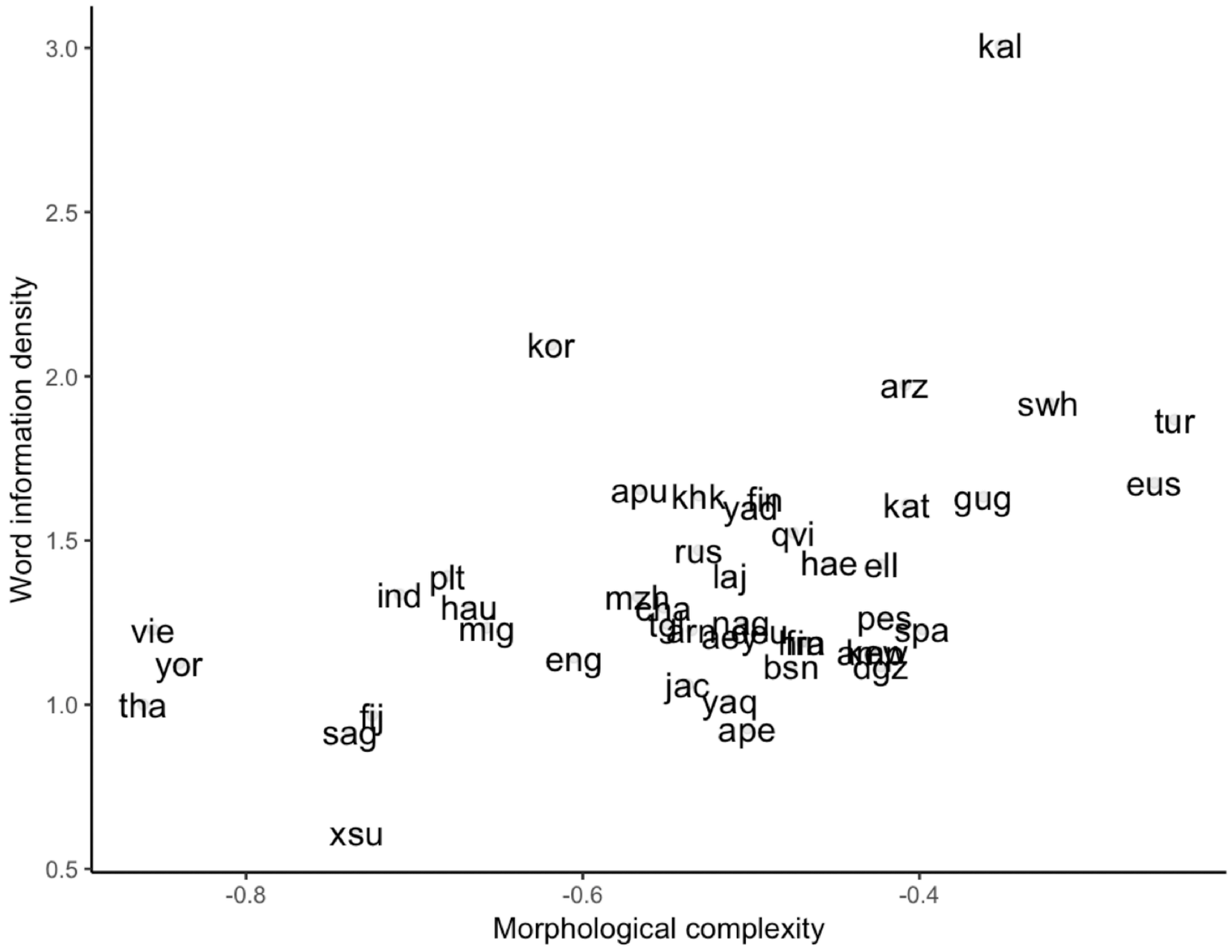
Nb of words of sentence S in THA

Pairwise comparison

Nb of words of sentence S in a target language



Results - WID & Morphological complexity



A significant correlation between WID and Morphological complexity (Pearson's r = 0.4810195***; p-value < 0.001; N = 44)

Parameters -Phonological complexity

Information theoretic measures: reduce a message into **binary arithmetic coding** (i.e. 0s and 1s) and estimate **how many bits on average** are necessary to **encode a random linguistic variable** [Goldsmith, 2000].

➡ the estimated average amount of information (in bits) contained per **syllable**

Parameters -Phonological complexity

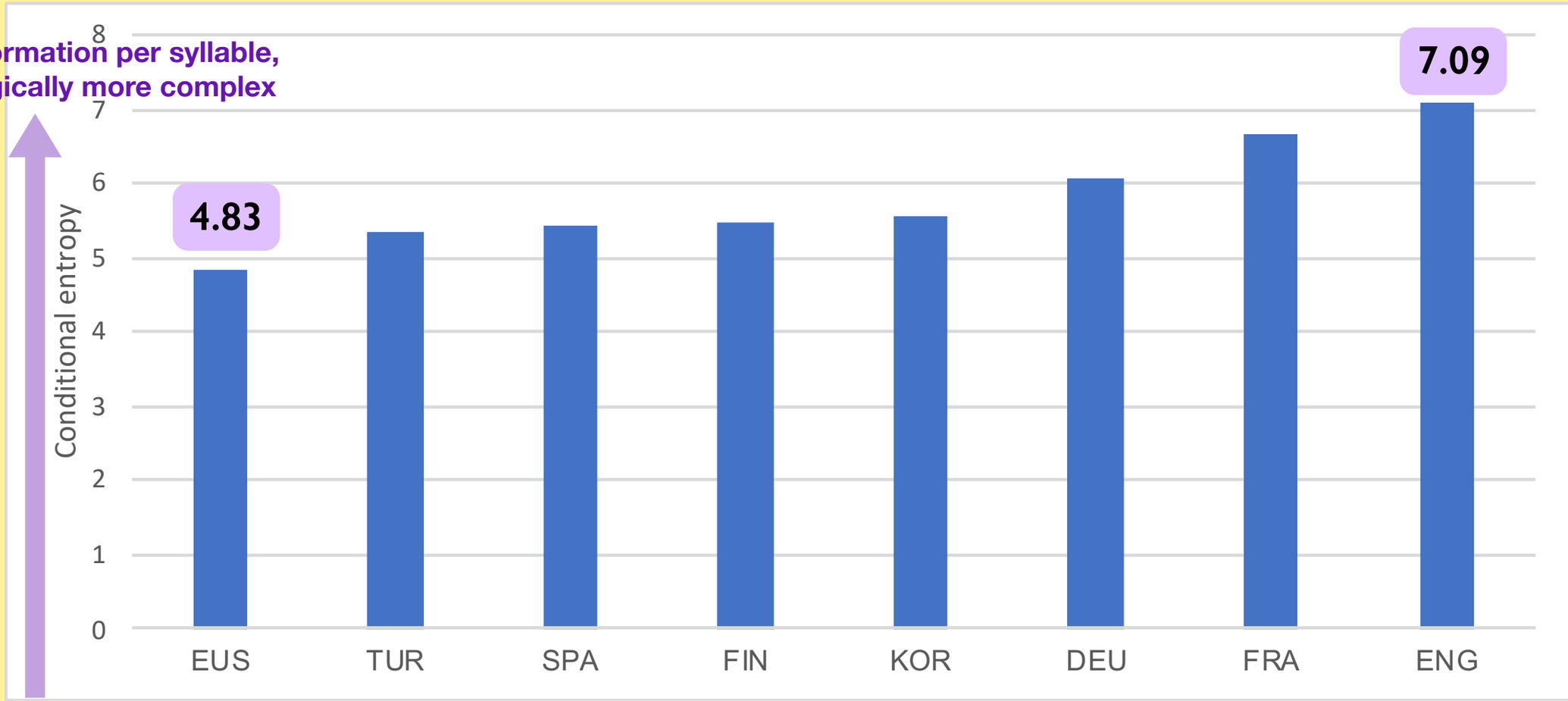
Conditional entropy: a measure of the **average amount of information** of a set of linguistic units **when the previous context** (*c*) **is known.**

$$H(X|C) = \sum_{c \in C} p(c) \cdot H(X|C = c)$$

$$= -\sum_{c \in C} p(c) \cdot \sum_{i=1}^{N_L} p(X = x_i|C = c) log_2(p(X = x_i|C = c))$$

Parameters - Phonological complexity

Languages with a tendency towards agglutination tend to encode **less information** per syllable than those with a tendency towards fusion.



Parameters -Syllable Information Density

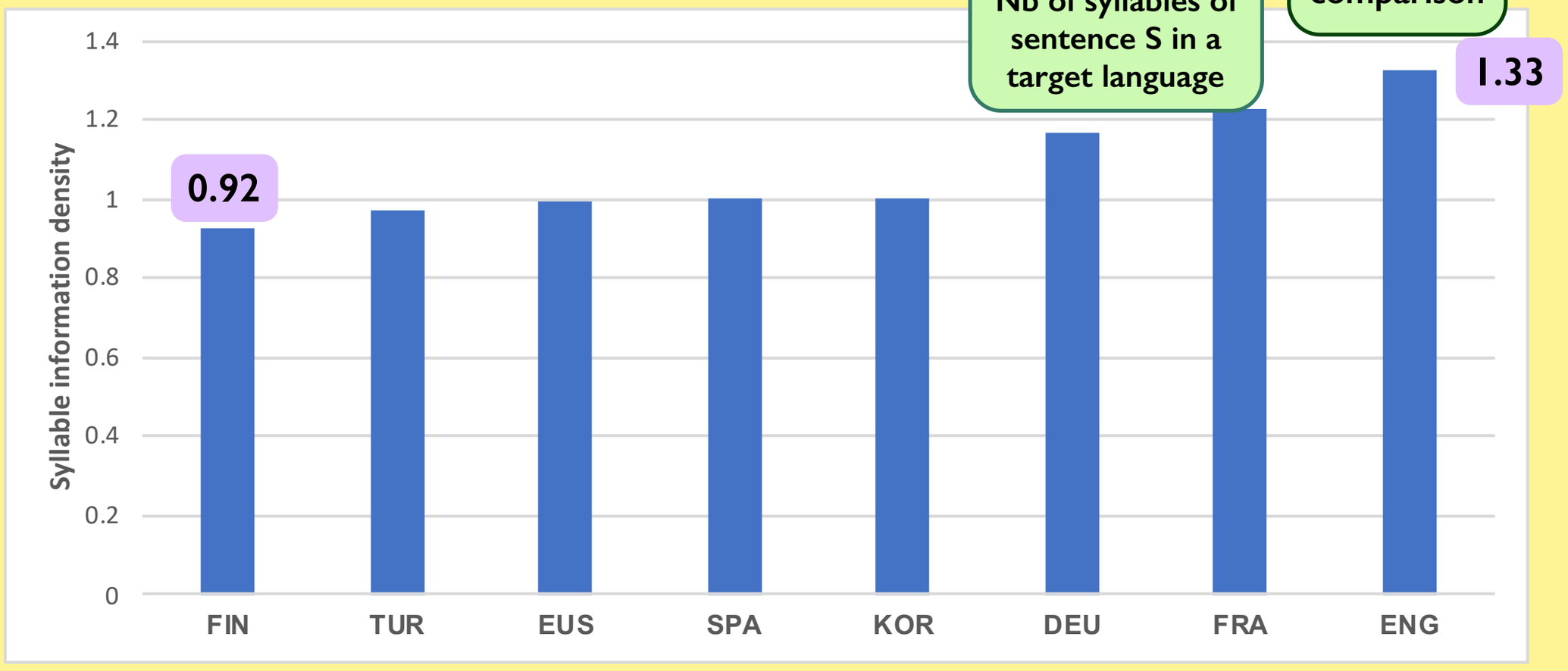
Syllable Information Density: the **average amount of information** conveyed per syllable

$$SID_L = \frac{1}{S}\sum_{s=1}^{S}\frac{SI_L^S}{SI_{KOR}^S} = \frac{1}{S}\sum_{s=1}^{S}\frac{C_L^S}{\sigma_L^S}\times\frac{\sigma_{KOR}^S}{C_{KOR}^S} = \frac{1}{S}\sum_{s=1}^{S}\frac{\sigma_{KOR}^S}{\sigma_L^S}$$

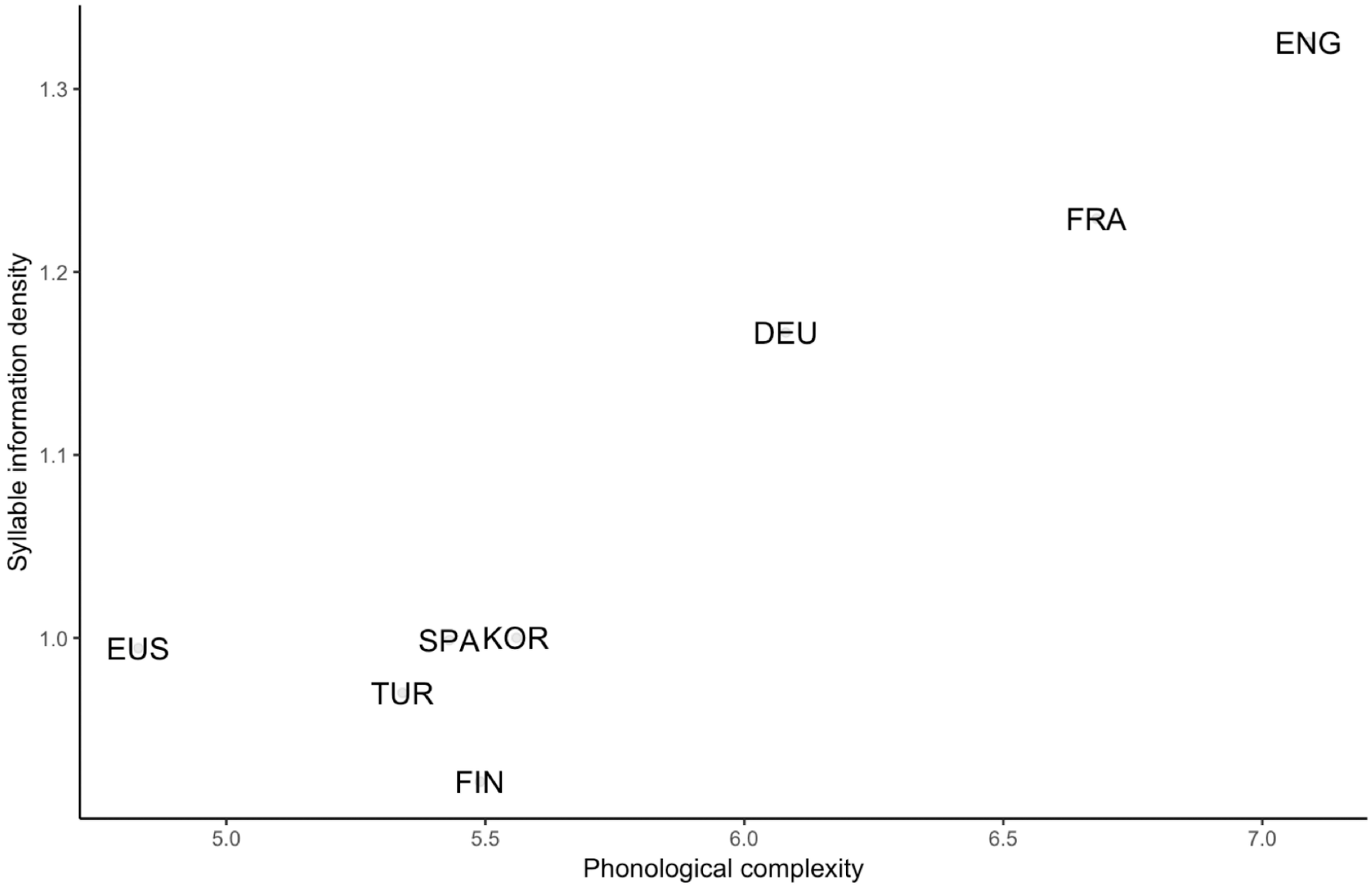
Nb of syllables of sentence S in KOR

Pairwise comparison

Nb of syllables of sentence S in a target language

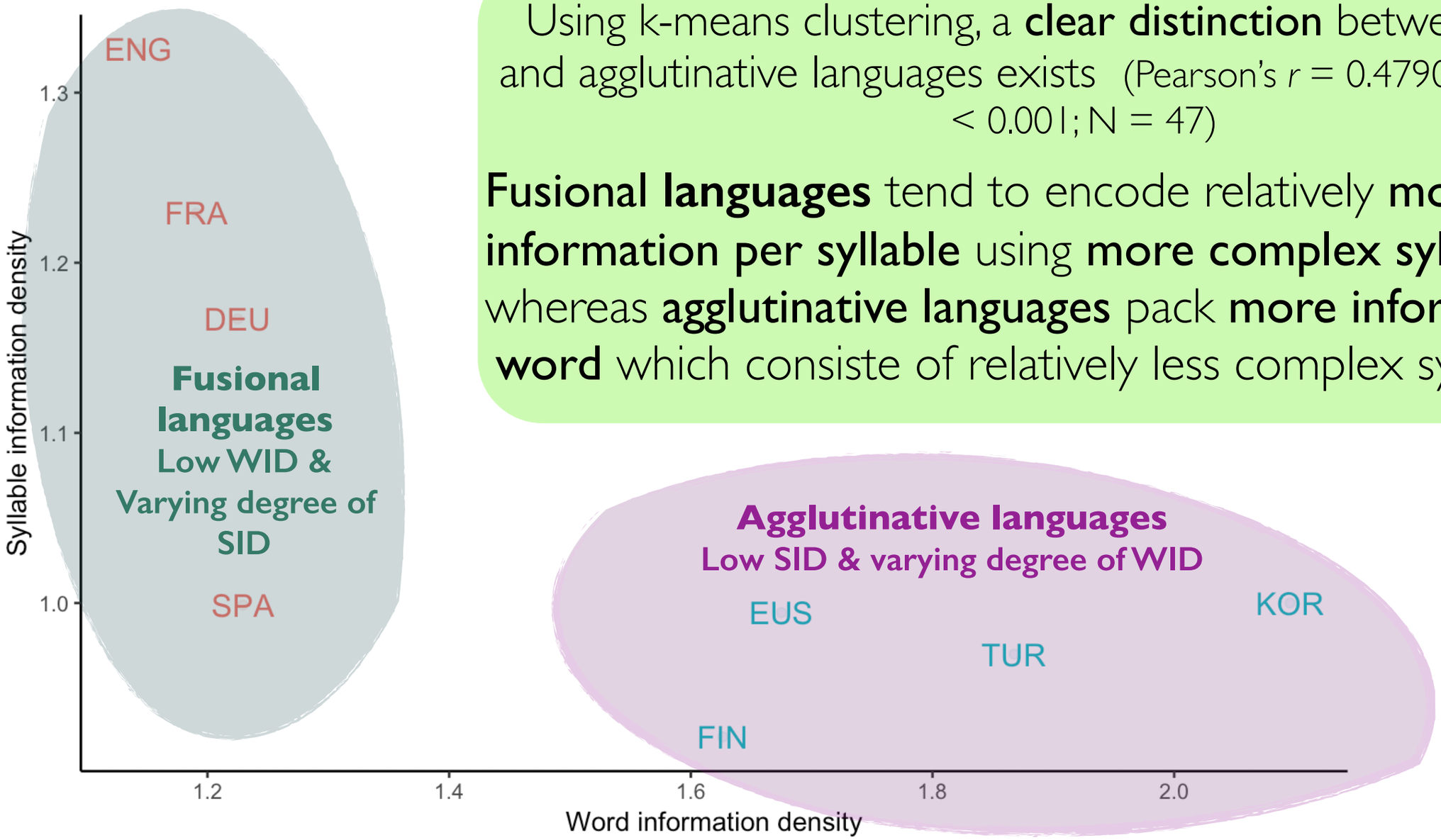


Results - SID & Phonological complexity



A highly significant correlation between SID and Phonological complexity (Pearson's r = 0.9223144***; p-value = 0.001; N = 8)

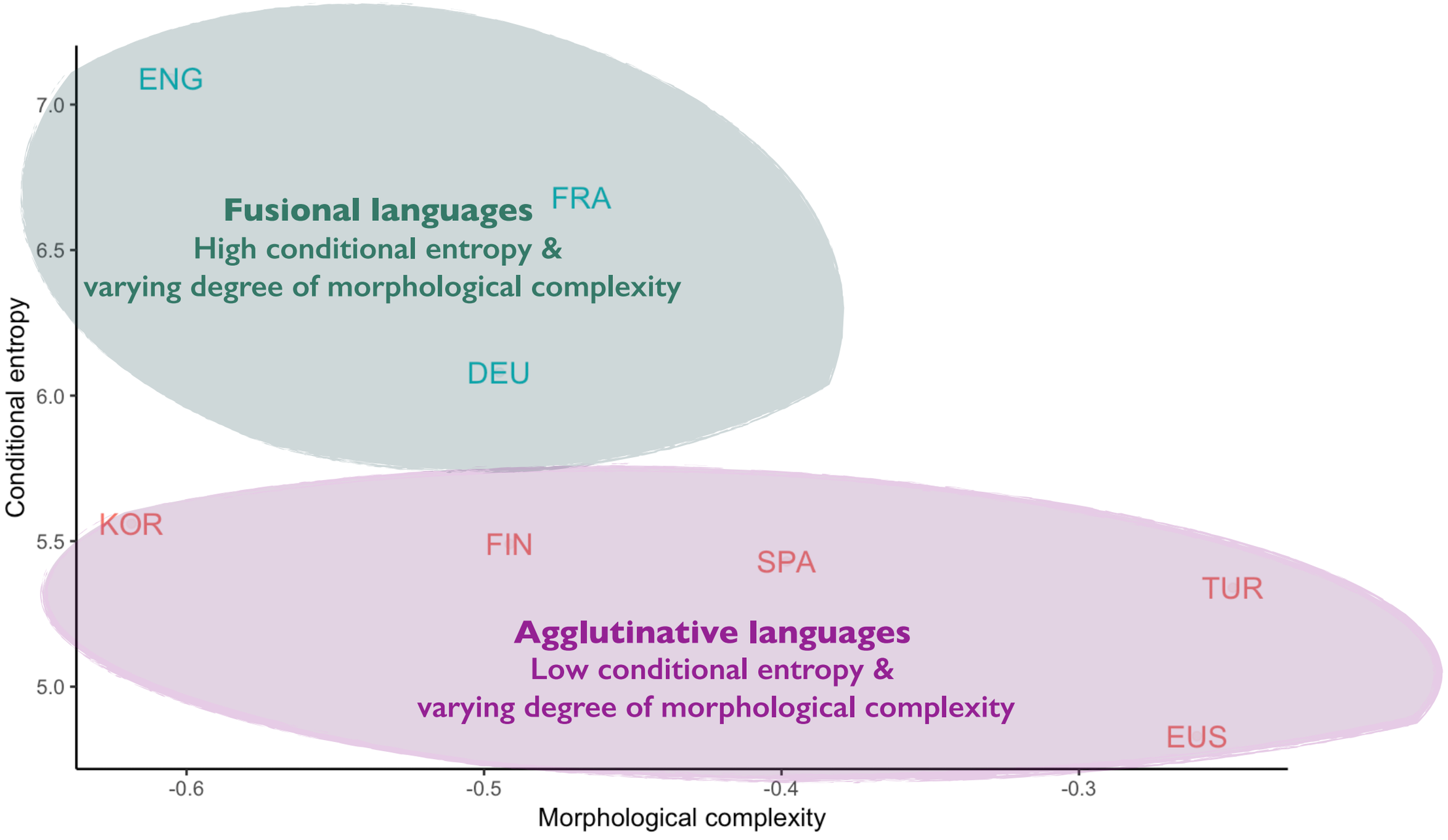
Results - WID & SID (k-means clustering)



Using k-means clustering, a **clear distinction** between fusional and agglutinative languages exists (Pearson's $r = 0.479068^{***}$; p-value < 0.001; N = 47)

Fusional languages tend to encode relatively **more information per syllable** using **more complex syllables** whereas **agglutinative languages** pack **more information per word** which consiste of relatively less complex syllables.

Results - MC & CE (k-means clustering)



Conclusion

- A trade-off between morphological complexity and phonological complexity is found using measures based on pairwise comparison.
- Both measures of information density using pairwise comparison and conditional entropy seem to capture the degree of morphophonemic alternation; agglutinative languages show a tendency toward lower conditional entropy and higher word information density whereas fusional languages exhibit the opposite tendency.

Vielen Dank!

Thank you!