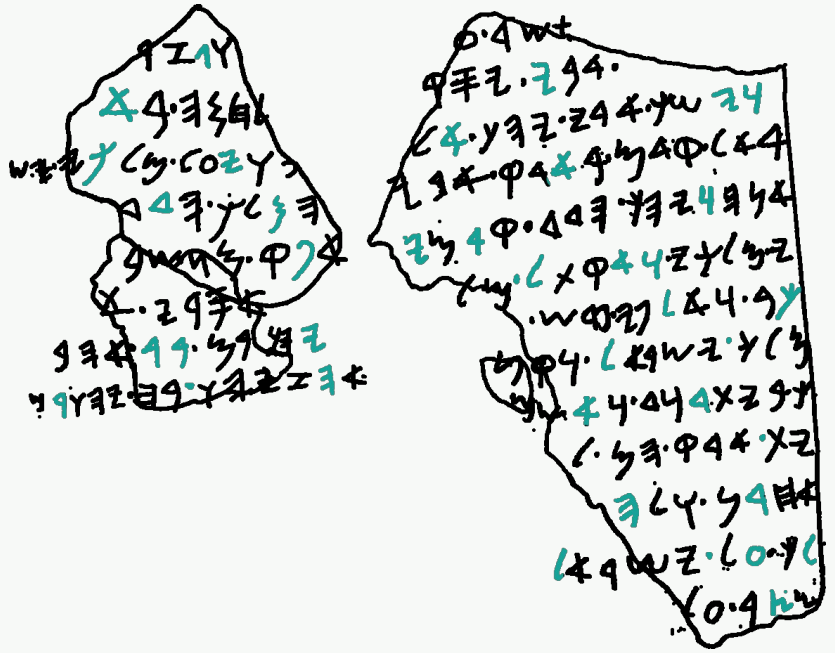
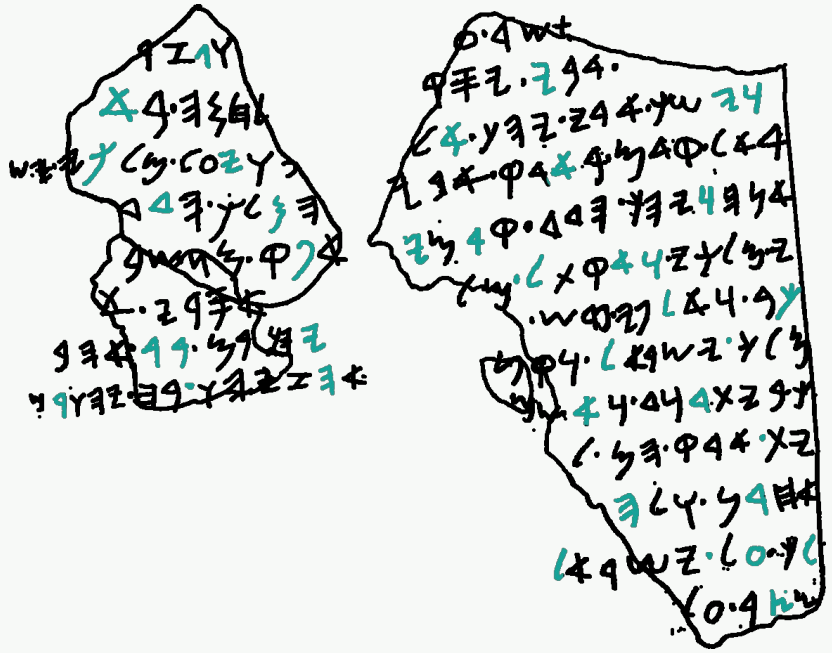


COMO COMPUTAR A COMPLEXIDADE DAS LÍNGUAS HUMANAS?

FELIPE R. SERRAS

SNAIL — 13.09.24



COMO COMPUTAR A COMPLEXIDADE DAS LÍNGUAS HUMANAS?

Parte 1

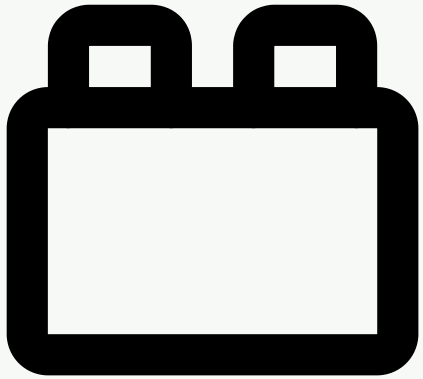
FELIPE R. SERRAS

SNAIL — 13.09.24

COMO COMPUTAR A COMPLEXIDADE DAS LÍNGUAS HUMANAS?

Um SNAIL em quatro atos

- Ato I – Dois experimentos bobos e suas consequências...
- Ato II - Uma viagem ao redor do mundo
- Ato III - A Anatomia de uma Métrica
- Ato IV - Resultados Selecionados



ATO I

DOIS EXPERIMENTOS BOBOS E SUAS CONSEQUÊNCIAS ONTOLÓGICAS E EPISTEMOLÓGICAS

EXPERIMENTO 1 – O HINO BRASILEIRO

Ouviram do Ipiranga as margens plácidas
De um povo heróico o brado retumbante,
E o sol da Liberdade, em raios fúlgidos,
Brilhou no céu da Pátria nesse instante.

[Excerto 1 – 1ª Estrofe do Hino Brasileiro]

EXPERIMENTO 1 – O HINO BRASILEIRO

As margens calmas do Ipiranga ouviram o grito alto de um povo heróico e a luz intensa da liberdade brilhou no céu do nosso país naquele momento

[Excerto 2 – 1ª Estrofe do Hino Brasileiro Reescrita]

EXPERIMENTO 2 – ~~DOIS REAIS OU~~ DUAS LÍNGUA MISTERIOSA

Kian bonegan seminarion ni partoprenas hodiaŭ, ĉu ne?

[Excerto 3 – Fatos em Língua Misteriosa A]

EXPERIMENTO 2 – ~~DOIS REAIS OU~~ DUAS LÍNGUA MISTERIOSA

Imuphi umhlangano omuhle kangaka esiya kuwo namuhla, akunjalo?

[Excerto 4 – Fatos em Língua Misteriosa B]

CONSEQUÊNCIAS ONTOLÓGICAS

[Existe uma qualidade dita complexidade, ao menos convencionalmente separável do significado do texto, que os seres humanos são capazes de diferenciar entre textos de uma mesma língua. Chamaremos essa qualidade de **Complexidade Intra-Linguística (CIL)**]

[Hipótese 1 – CIL Existe]

CONSEQUÊNCIAS ONTOLÓGICAS

[Existe uma qualidade dita complexidade, ao menos convencionalmente separável do significado do texto, que os seres humanos são capazes de diferenciar entre amostras de línguas diferentes. Chamaremos essa qualidade de **Complexidade Extra-Linguística (CEL)**]

[Hipótese 2 – CEL Existe]

HIPÓTESE DE TRABALHO

[CEL e CIL são passíveis de aproximação por dois (ou mais) aproximantes numéricos ou categóricos ~CEL e ~CIL, a partir dos quais nós tentaremos construir uma Teoria Quantitativa/Computacional da Complexidade da Linguagem Humana (não hoje)]

[Hipótese 3 – Hipótese de Trabalho]

UMA AGENDA EPISTEMOLÓGICA

(AE1) Análise fenomenológica de CEL e CIL;

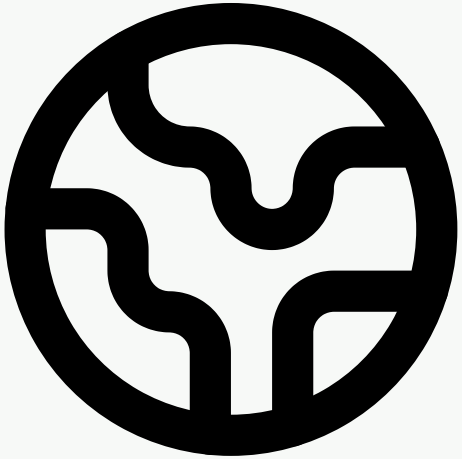
(AE2) Busca por aproximantes computáveis para CEL e CIL: ~CEL e ~CIL;

(AE3) Análise da eficiência de ~CEL e ~CIL como aproximantes de CEL e CIL;

(AE4) O que ~CEL e ~CIL nos revelam sobre CEL e CIL?

NOTAS SOBRE A AGENDA EPISTEMOLÓGICA

- Podemos entender nossa tarefa como uma instância de uma tarefa maior: Natural Language Assessment
- Existe uma dependência epistêmica entre (AE3) e (AE4)
- Vamos começar com (AE1), com um foco especial em CIL, para isso vamos fazer...



ATO II

UMA VIAGEM AO REDOR DO MUNDO

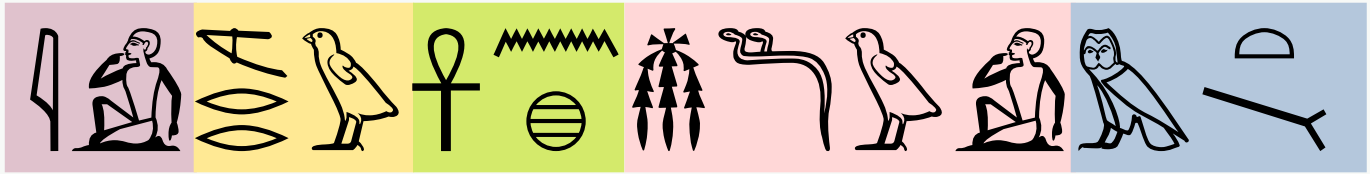
我昨天吃了苹果。

Wǒ zuótiān chī le píngguǒ.

I yesterday eat PRT-PST.PFV apple.

I ate an apple yesterday.

[Excerto 5 – Sentença em chinês/mandarim]



ỉ mrrw ꜥnḫ msḏḏw m(w)t

PTCL-VOC amar-PTCP.ACT vida odiar-PTCP.ACT morte

Oh você, aquele que ama a vida, aquele que odeia a morte

[Excerto 6 – Sentença em Egípcio Médio, de [8]]

toto yonoye kamara

Pessoa 3SG- comer -DIST.PAST.COMPL onça

A onça comeu o homem

[Excerto 7 – Sentença em Hixkaryana, de [7]]

talo	taloni	talossani
casa	Casa -POSS	Casa -POSS -LOC
casa	minha casa*	na minha casa*

[Excerto 8 (a/b/c) – Variantes da palavra casa em Finlandês]

Io mangio	Noi mangiamo
Pon.1Sg √comer -1SG	Pron.1Pl √comer -1PL
Eu como	Nós comemos

[Excerto 9 (a/b) – Conjugações de comer em Italiano]

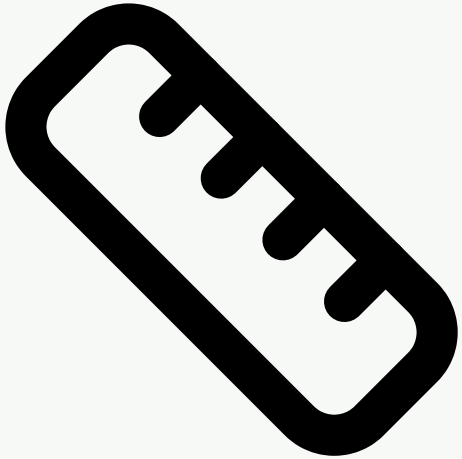
iitghesqesaghiisqaa

Itegh- -sqe- -yaghtugh- -sqe- -aa

Entrar padir-para ir pedir-para INDIC-3SG.3SG

Ele pediu para ele pedi-lo para entrar

[Excerto 10 – Sentença em Yupik Central Siberiano, de [6]]



ATO II

A ANATOMIA DE UMA MÉTRICA

CONTEXTUALIZAÇÃO

- Apresentação das métricas propostas no trabalho de Patrick Juola e Katharina Ehret e Benedikt Szmrecsanyi.
- Baseadas em teoria da informação e tipologia linguística.
- Distinguir complexidade geral e complexidades aspectuais.

COMPLEXIDADES ASPECTUAIS

- Mediriam a complexidade em diferentes níveis de análise
- Foco: morfologia, sintaxe e pragmática.
- Alternativamente: intra-palavra, inter-palavra e inter-excerto.

COMPLEXIDADES ASPECTUAIS

- A priori são métricas para a complexidade interlinguística

~CEL$_A$ ~CEL$_M$ ~CEL$_S$ ~CEL$_P$

ESTRUTURA DAS MÉTRICAS

- As métricas dependem de três mecanismos principais:

  (M1) Infômetro: Mede a informação transmitida em uma mensagem típica.

  (M2) Injetor de Ruído: Perturba a mensagem, afetando um nível específico de análise.

  (M3) Kernel: Combina os dois mecanismos anteriores em um valor coeso.

INFÔMETRO

- Usaremos algoritmos de compressão de dados
- Quantidade de Informação ~ Tamanho do texto comprimido
- Motivação psicolinguística: reconhecimento de padrões frequentes e consulta a padrões

INJETOR DE RUÍDO/DEGRADAÇÃO

- Baseado na deleção de partes da amostra.
- Deleção acontece de forma a afetar um nível específico:

  Morfologia ↔ Caracteres
  Sintaxe ↔ Palavras
  Pragmática ↔ Excertos

INJETOR DE RUÍDO/DEGRADAÇÃO - EXEMPLO

«O conceito de complexidade da linguagem é intuitivo para os seres humanos. As pessoas reconhecem textos complexos e percebem que algumas línguas são mais fáceis ou difíceis de aprender, conforme sua formação linguística prévia»

[Excerto 11a – Trecho do abstract, imaculado]

INJETOR DE RUÍDO/DEGRADAÇÃO - EXEMPLO

«O conceito de complexidade da linguagem é intuitivo para os seres humanos. As pessoas reconhecem textos complexos e percebem que algumas línguas são mais fáceis ou difíceis de aprender, conforme sua formação linguística prévia»

[Excerto 11b – Trecho do abstract, morfologicamente degradado]

INJETOR DE RUÍDO/DEGRADAÇÃO - EXEMPLO

«O conceito de complexidade da linguagem é intuitivo para os seres humanos. As pessoas reconhecem textos complexos e percebem que algumas línguas são mais fáceis ou difíceis de aprender, conforme sua formação linguística prévia»

[Excerto 11c – Trecho do abstract, sintaticamente degradado]

INJETOR DE RUÍDO/DEGRADAÇÃO - EXEMPLO

«O conceito de complexidade da linguagem é intuitivo para os seres humanos. As pessoas reconhecem textos complexos e percebem que algumas línguas são mais fáceis ou difíceis de aprender, conforme sua formação linguística prévia»

[Excerto 11d – Trecho do abstract, pragmaticamente degradado]

KERNELS

(K1) $\sim CEL_A(T) = \quad Res(|C(T)|, |T|)$

(K2) $\sim CEL_S(T) = \quad |C(TS)| \; / \; |C(T)|$

(K3) $\sim CEL_P(T) = \quad |C(TP)| \; / \; |C(T)|$

(K4) $\sim CEL_M(T) = - \; |C(TM)| \; / \; |C(T)|$

KERNELS - EXEMPLO

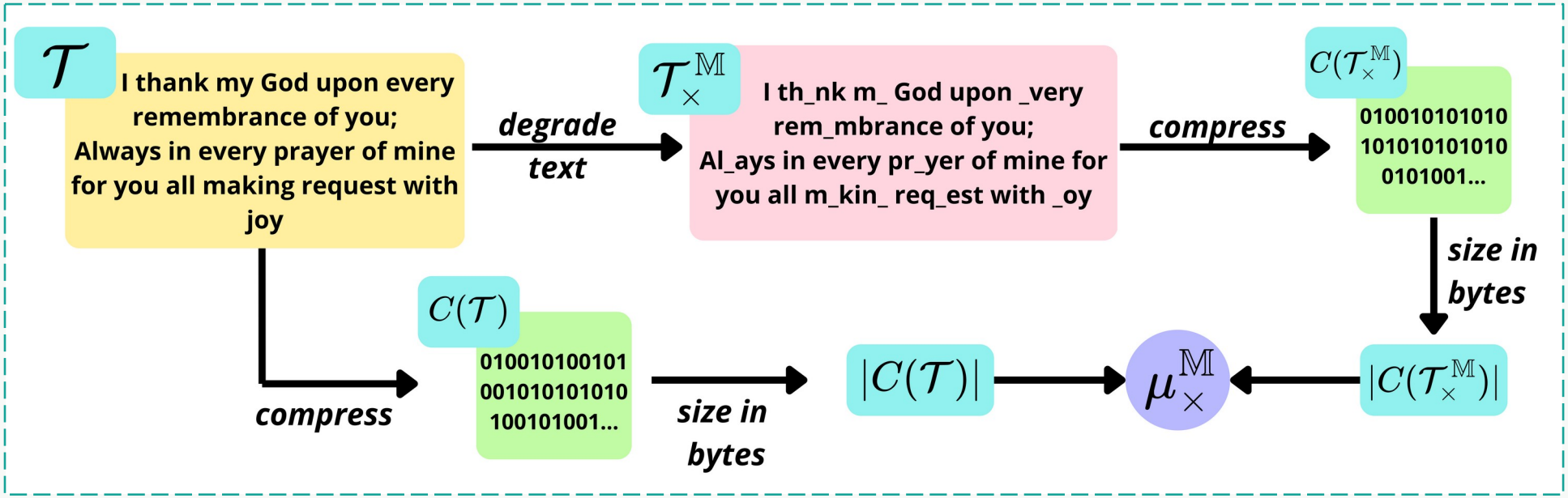
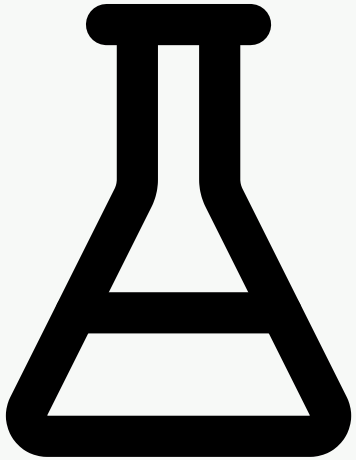


Imagem 1 – Exemplo do Cômputo de ~CELM

OBSERVAÇÕES SOBRE AS MÉTRICAS

- Esse framework permite diversas variações, mantendo-se anatomia geral das métricas
- De fato, as formas apresentadas aqui são resultado da exploração de variações em todas as componentes: infômetro, degradação, kernel



ATO III

RESULTADOS FORTUITOS

ALERTA ESTÉTICO

MUITOS DOS GRÁFICOS DESSA SEÇÃO FORAM RETIRADOS DE OUTROS TRABALHOS. POR CONSEQUÊNCIA, ELES NÃO SERÃO NECESSARIAMENTE COMPATÍVEIS COM A UNIDADE ESTÉTICA DA APRESENTAÇÃO, CONSTRUÍDA ATÉ O MOMENTO 😔

DE ONDE ESSES RESULTADOS VIERAM

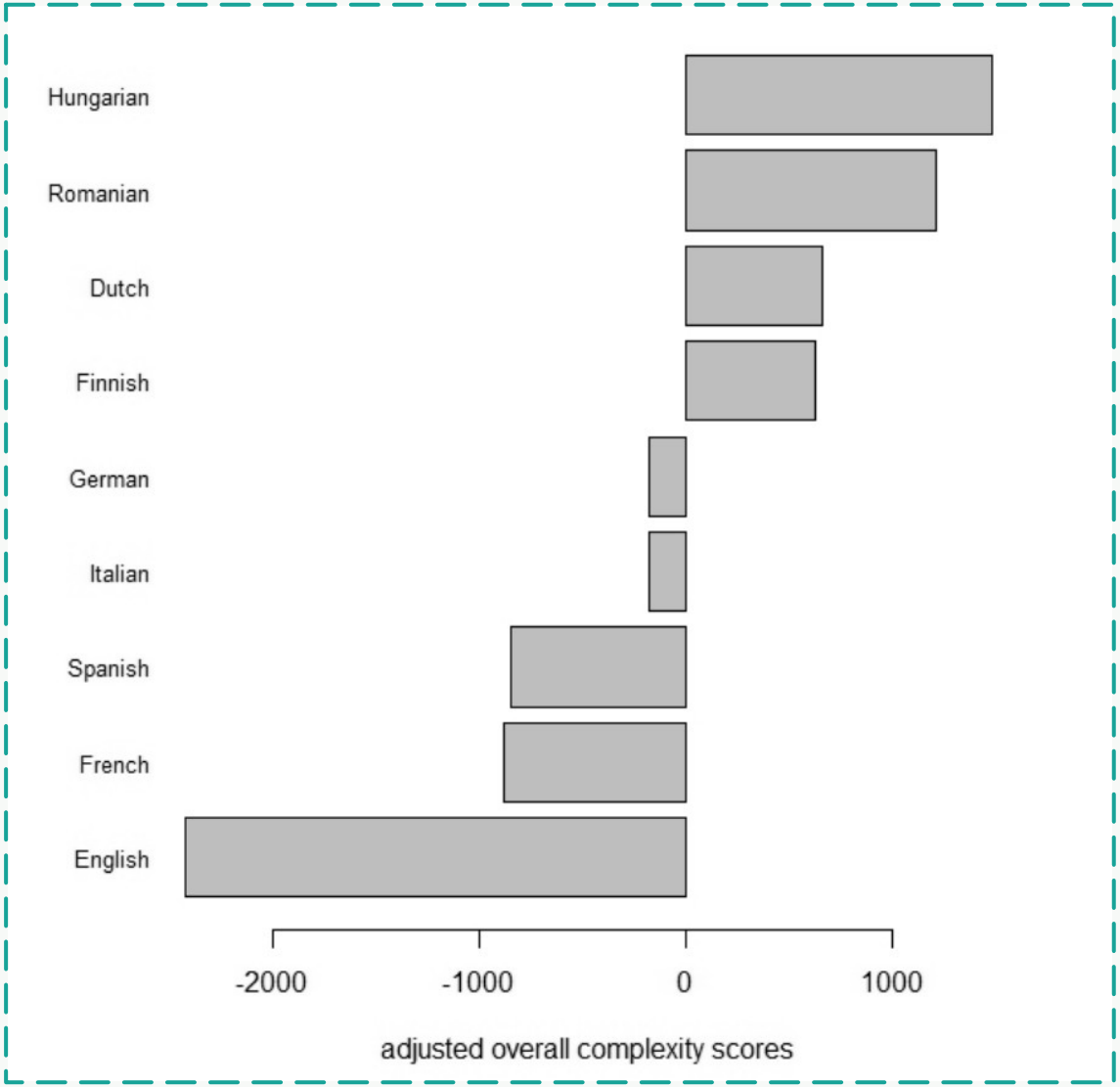
- Assessing linguistic complexity por Patrick Juola
- An information-theoretic approach to assess linguistic complexity, por Katharina Ehret and Benedikt Szmrecsanyi

INTERLÚDIO – HIPÓTESES NO CORAÇÃO DA ÁREA

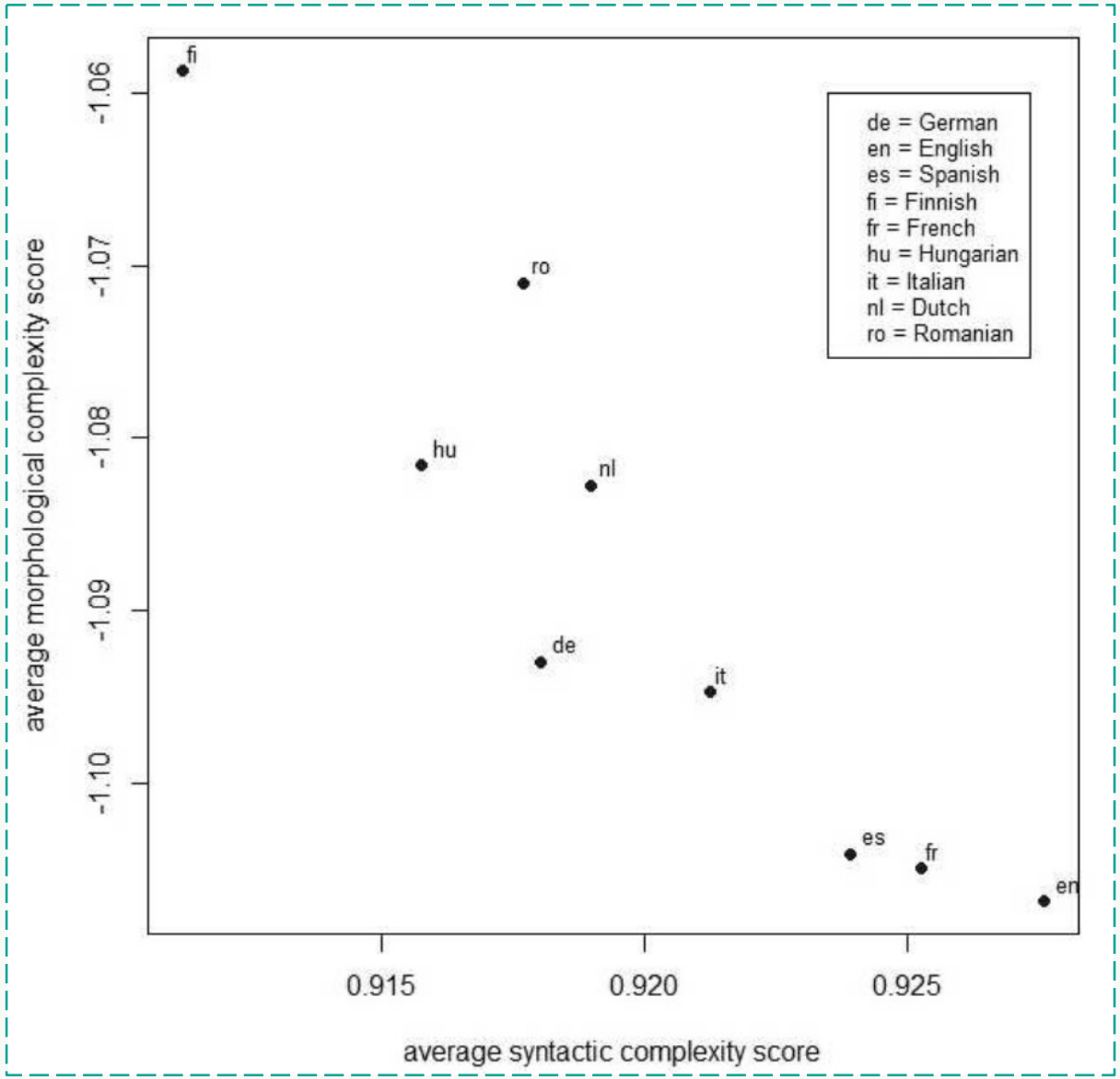
(♡1) Hipótese da Equi-complexidade

(♡2) Hipótese do *trade-off* sintático-morfológico

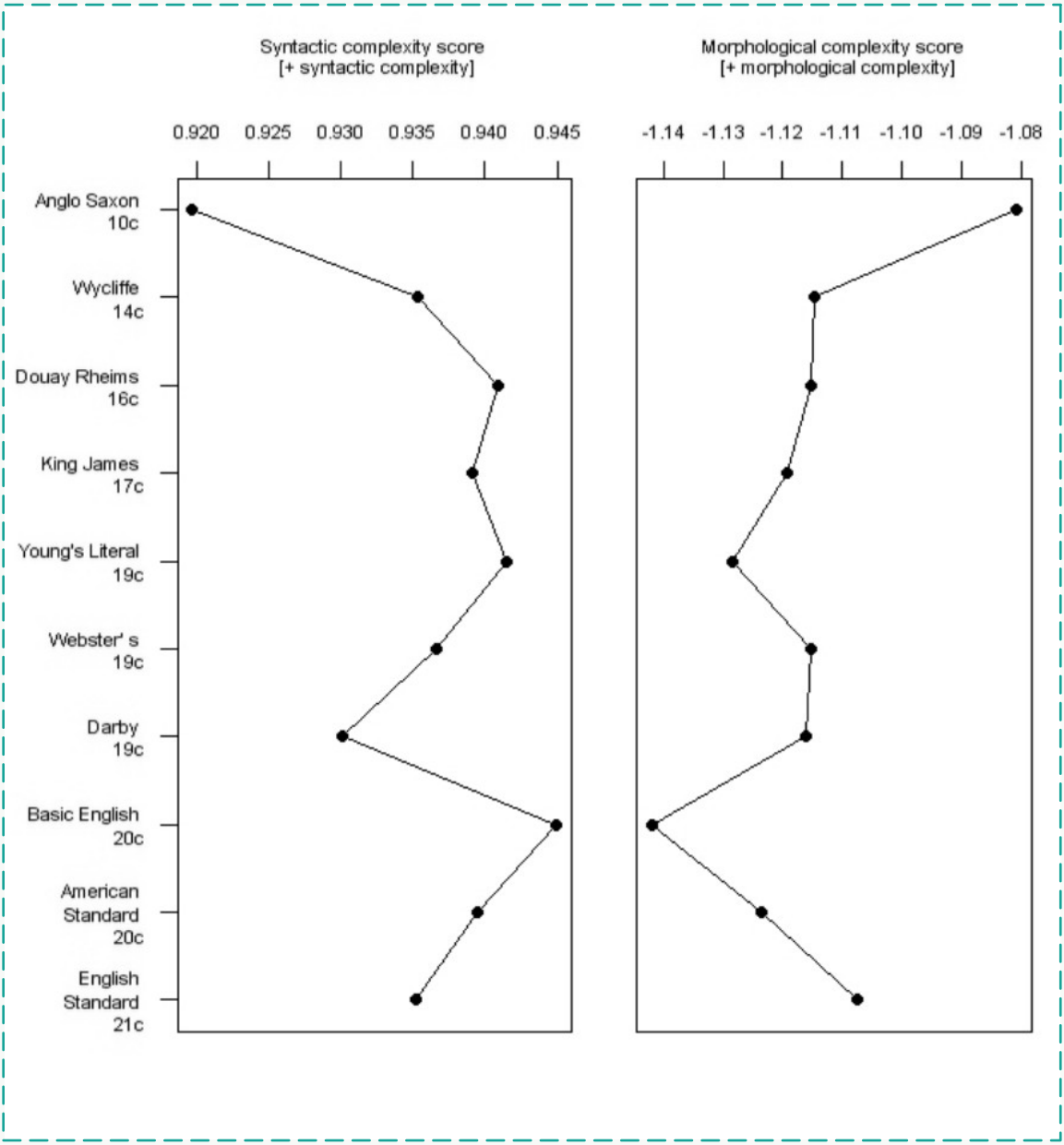
(♡3) Hipótese da simplicidade dos pidgins e das línguas crioulas


Hungarian
Romanian
Dutch
Finnish
German
Italian
Spanish
French
English
-2000
-1000
0
1000
adjusted overall complexity scores

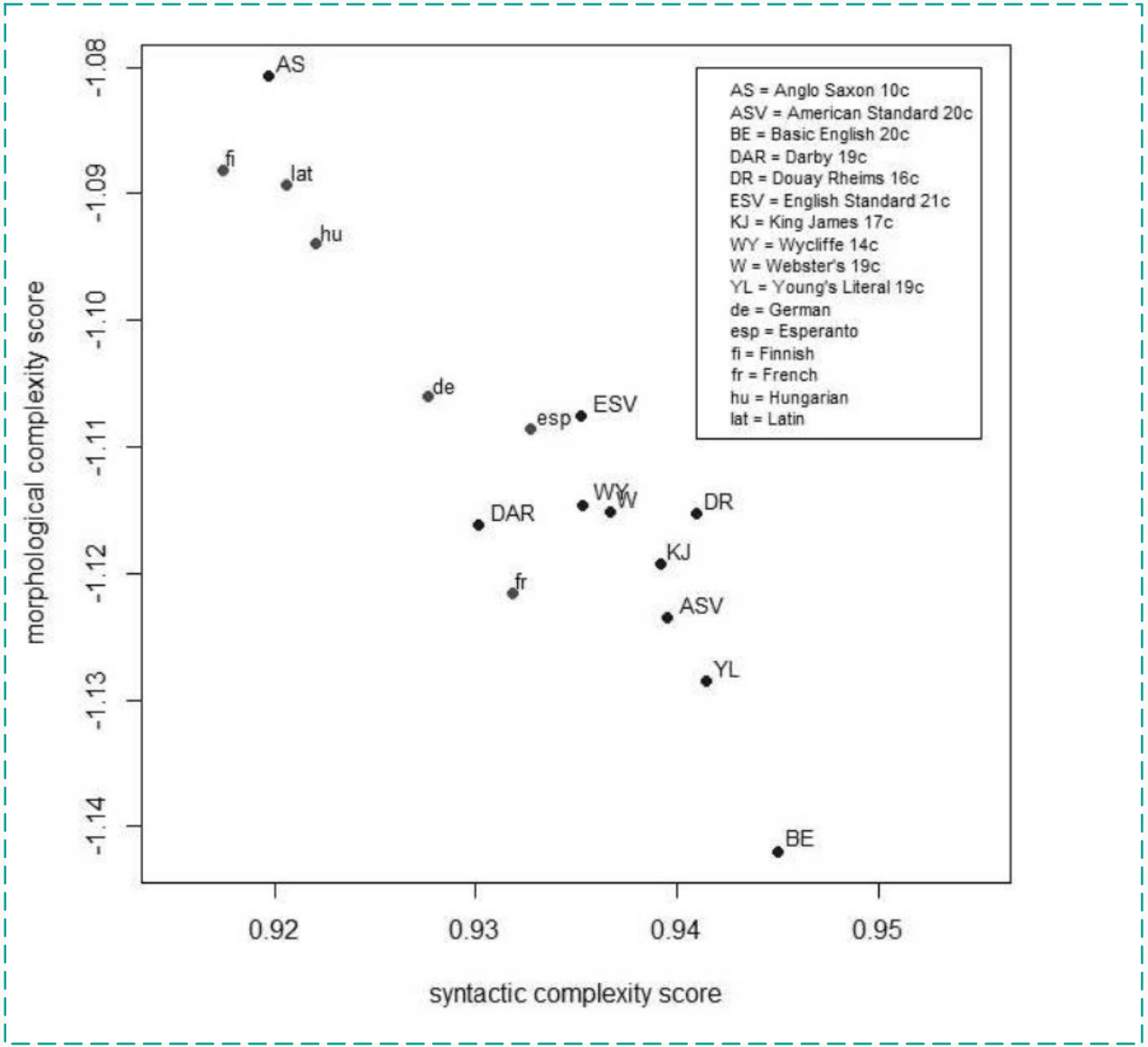
[de[2]]


de = German
en = English
es = Spanish
fi = Finnish
fr = French
hu = Hungarian
it = Italian
nl = Dutch
ro = Romanian
fi
ro
hu
nl
de
it
es
fr
en
-1.06
-1.07
-1.08
-1.09
-1.10
0.915
0.920
0.925
average morphological complexity score
average syntactic complexity score

[de[2]]


Syntactic complexity score
[+ syntactic complexity]
0.920
0.925
0.930
0.935
0.940
0.945
Morphological complexity score
[+ morphological complexity]
-1.14
-1.13
-1.12
-1.11
-1.10
-1.09
-1.08
Anglo Saxon 10c
Wycliffe 14c
Douay Rheims 16c
King James 17c
Young's Literal 19c
Webster' s 19c
Darby 19c
Basic English 20c
American Standard 20c
English Standard 21c

[de[21]]


AS = Anglo Saxon 10c
ASV = American Standard 20c
BE = Basic English 20c
DAR = Darby 19c
DR = Douay Rheims 16c
ESV = English Standard 21c
KJ = King James 17c
WY = Wycliffe 14c
W = Webster's 19c
YL = Young's Literal 19c
de = German
esp = Esperanto
fi = Finnish
fr = French
hu = Hungarian
lat = Latin
AS
fi
lat
hu
de
esp
ESV
WY
W
DR
DAR
KJ
fr
ASV
YL
BE
-1.08
-1.09
-1.10
-1.11
-1.12
-1.13
-1.14
morphological complexity score
0.92
0.93
0.94
0.95
syntactic complexity score

[de[2]]

ALERTA LINGUÍSTICO

APESAR DE MUITO RELEVANTES, ESSES RESULTADOS ESTÃO RESTRITOS MAJORITARIAMENTE A LÍNGUAS EUROPEIAS!

É AÍ QUE A GENTE ENTRA!

Analysing and Validating Language Complexity Metrics Across South American Indigenous Languages

Felipe Ribas Serras
Miguel de Mello Carpi
Matheus Castello Branco
Marcelo Finger
Institute of Mathematics and Statistics, University of São Paulo
R. do Matão, 1010 - Butantã, São Paulo - SP, Brazil, 05508-090
{frserras, miguel, matheus.castello, mfinger}@ime.usp.br

Abstract

Language complexity is an emerging concept critical for NLP and for quantitative and cognitive approaches to linguistics. In this work, we evaluate the behavior of a set of compression-based language complexity metrics when applied to a large set of native South American languages. Our goal is to validate the desirable properties of such metrics against a more diverse set of languages, guaranteeing the universality of the techniques developed on the basis of this type of theoretical artifact. Our analysis confirmed with statistical confidence most propositions about the metrics studied, affirming their robustness, despite showing less stability than when the same metrics were applied to Indo-European languages. We also observed that the trade-off between morphological and syntactic complexities is strongly related to language phylogeny.

1 Introduction

The development of means for quantifying linguistic properties is essential for cognitive approaches to computational linguistics, becoming simultaneously more challenging and useful as the property of interest is transversal to different languages and, therefore, an important clue for accessing cognitive processes behind human language. This is the case of language complexity.

The concept of language complexity, whether of an utterance or of a language as a whole, is instinctive for us. People know how to recognize when a text is written in a difficult or elaborate way and they usually recognize that certain languages are less or more complicated to learn depending on their linguistic background.

Informally, we can say that: (i) the *complexity of an utterance* encompasses the quantity and sophistication of linguistic constructs necessary to form and understand the utterance and (ii) *the complexity of a language as a whole* refers to the quantity and sophistication of communicative strategies available for the formation of such utterances in that language.

Despite a relative consensus around these intuitions, we lack established formal and quantifiable definitions of language complexity. It is difficult to find a definition that encompasses the heterogeneous range of human language manifestations, both in terms of different languages and of different levels in which meaning can be conveyed within a language.

Even in light of these challenges, it is crucial to establish rigorous, theoretically and experimentally validated definitions of language complexity. Both cognitive and non-cognitive approaches to Linguistics can significantly enhance their expressive capacity and theoretical framework. In NLP, complexity measures can be used in automatic text simplifiers, translators, domain-sensitive correctors and completers (Leal et al., 2023), but can also be integrated into the of training machine learning models, to increase performance (Sarti et al., 2021).

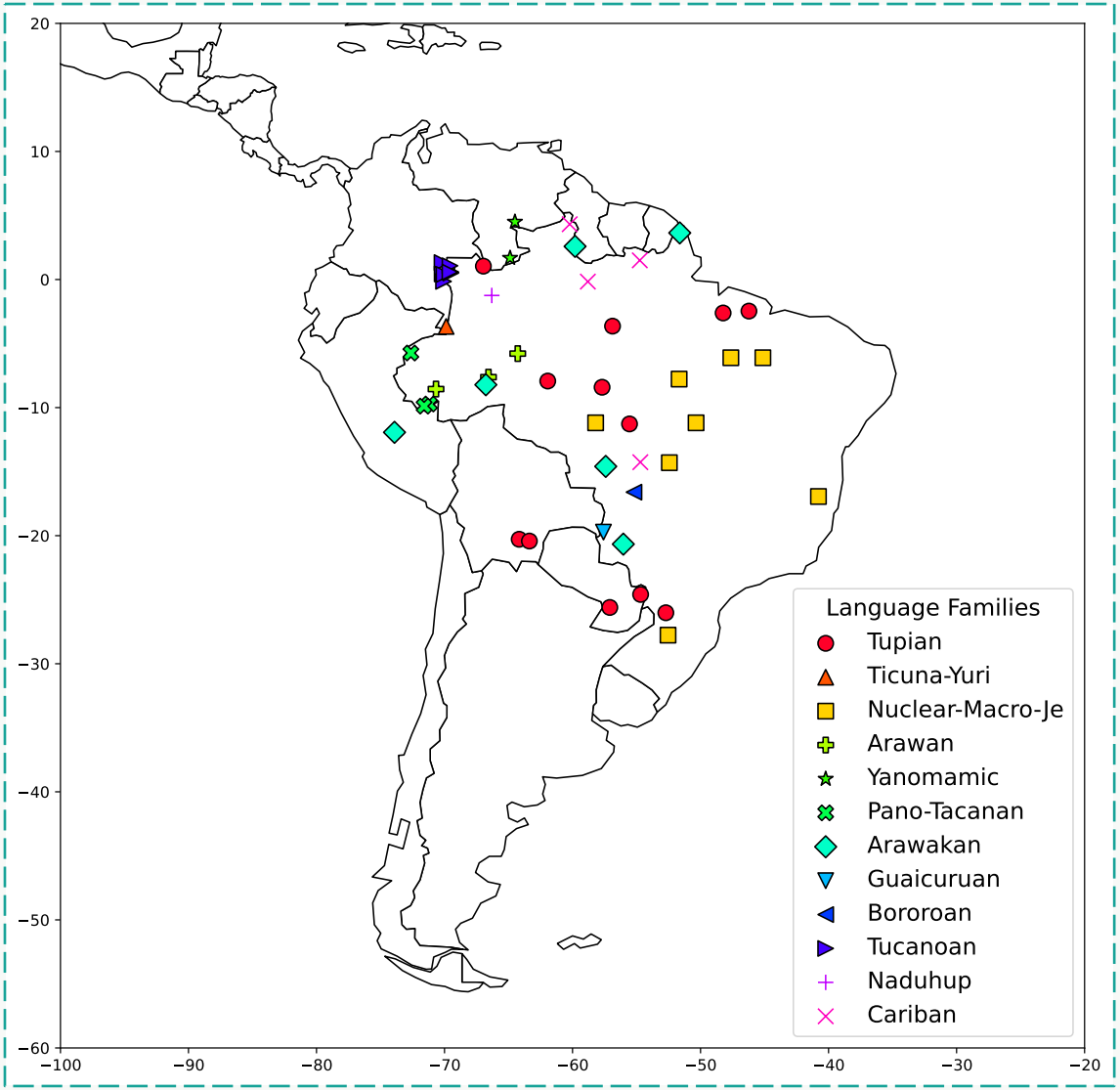
Another challenge for the construction of a robust theory for language complexity is that of inclusion: historically, the construction of tools and theories of human language has included Indo-European languages, to the detriment of other linguistic manifestations, e.g. American native languages. For a concept that aims to be transversal to different languages and provide universal insights into them and their underlying cognitive processes, as is the case with language complexity, it is necessary to include the broadest possible range of languages in its development and validation.

This inclusion is the focus of our work. Here, we examine a set of language complexity metrics derived from Information Theory, proposed in Juola (1998, 2005, 2008), and Ehret and Szmrecsanyi (2016). The authors ran several experiments with the proposed metrics, drawing on data from a sub-

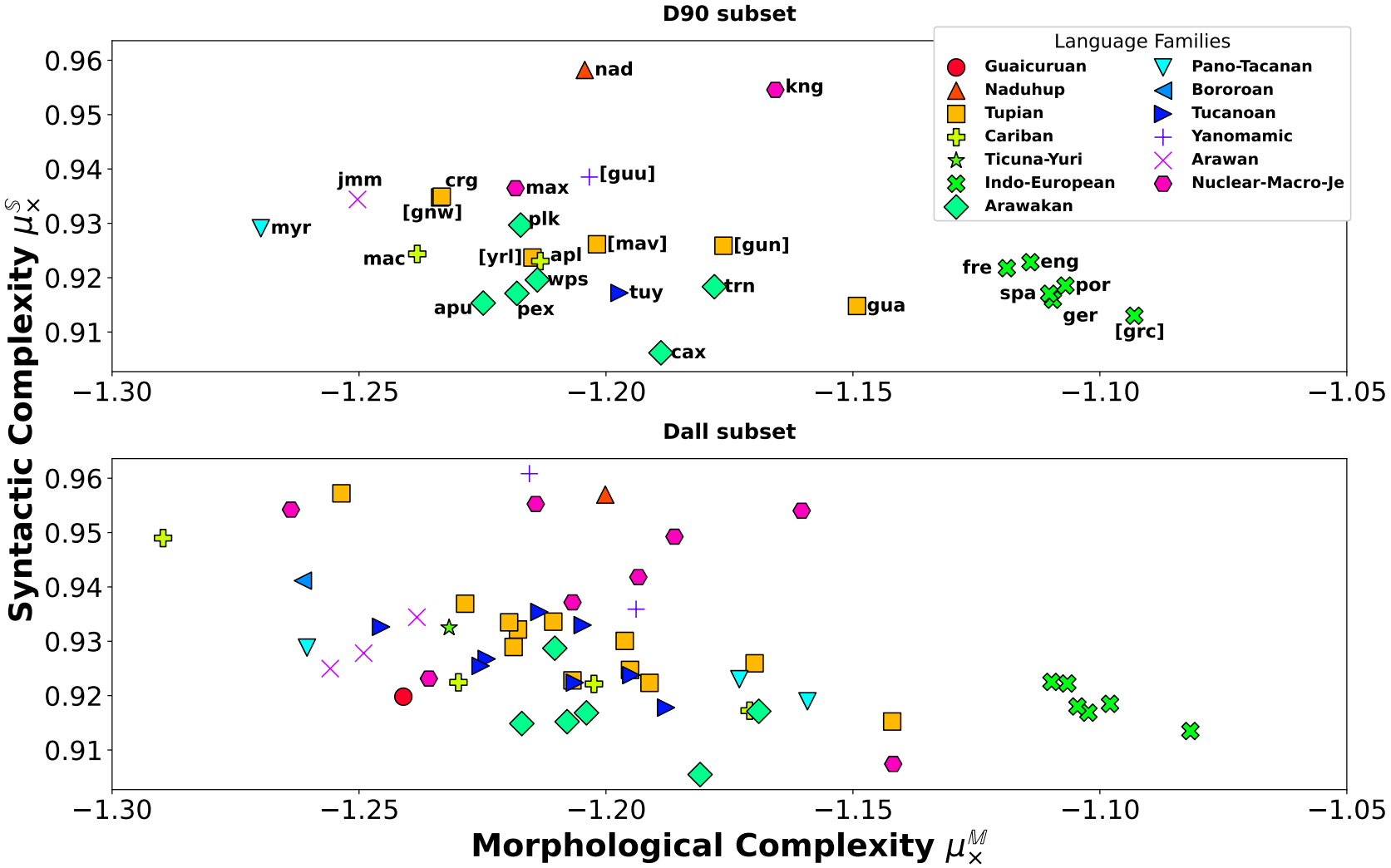
Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 152–165
August 15, 2024 ©2024 Association for Computational Linguistics

Analysing and Validating Language Complexity Metrics Across South American Indigenous Languages

Por: Eu, Miguel, Matheus e Finger


Language Families
Tupian
Ticuna-Yuri
Nuclear-Macro-Je
Arawan
Yanomamic
Pano-Tacanan
Arawakan
Guaicuruan
Bororoan
Tucanoan
Naduhup
Cariban
20
10
0
−10
−20
−30
−40
−50
−60
−100
−90
−80
−70
−60
−50
−40
−30
−20

Trade-off hypothesis ($\mathcal{H}_3$) with *gzip*

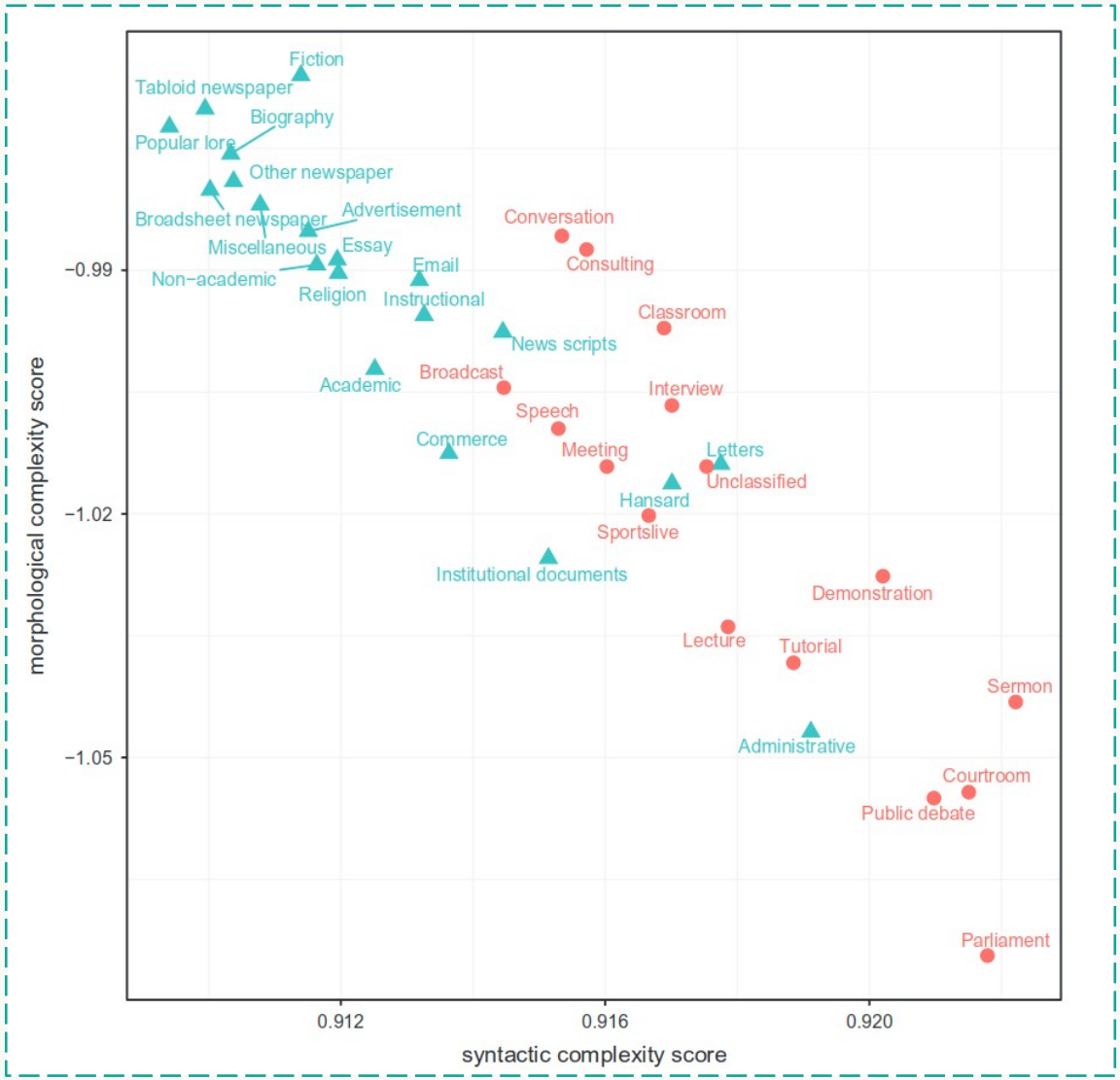


MAS E O CIL, GENTE?

👦 - “Até agora só falamos da CEL, mas você prometeu que ia falar da CIL também”

👧 - “Isso significa que a gente vai passar mais 1 hora aqui ouvindo sobre o CEL?!?”

Não, graças à Katharina Ehret em An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data


Fiction
Tabloid newspaper
Biography
Popular lore
Other newspaper
Broadsheet newspaper
Advertisement
Miscellaneous
Essay
Non−academic
Religion
Email
Instructional
Conversation
Consulting
Classroom
News scripts
Academic
Broadcast
Interview
Speech
Commerce
Meeting
Letters
Unclassified
Hansard
Sportslive
Institutional documents
Demonstration
Lecture
Tutorial
Sermon
Administrative
Courtroom
Public debate
Parliament
−0.99
−1.02
−1.05
morphological complexity score
0.912
0.916
0.920
syntactic complexity score

[de[3]]

CONSEQUÊNCIAS ONTOLÓGICAS

[CEL e CIL são a mesma coisa, ou ao menos aproximáveis pela mesma coisa]

[Hipótese 4 – Hipótese da Unicidade]

NÃO-ATO

EPÍLOGO

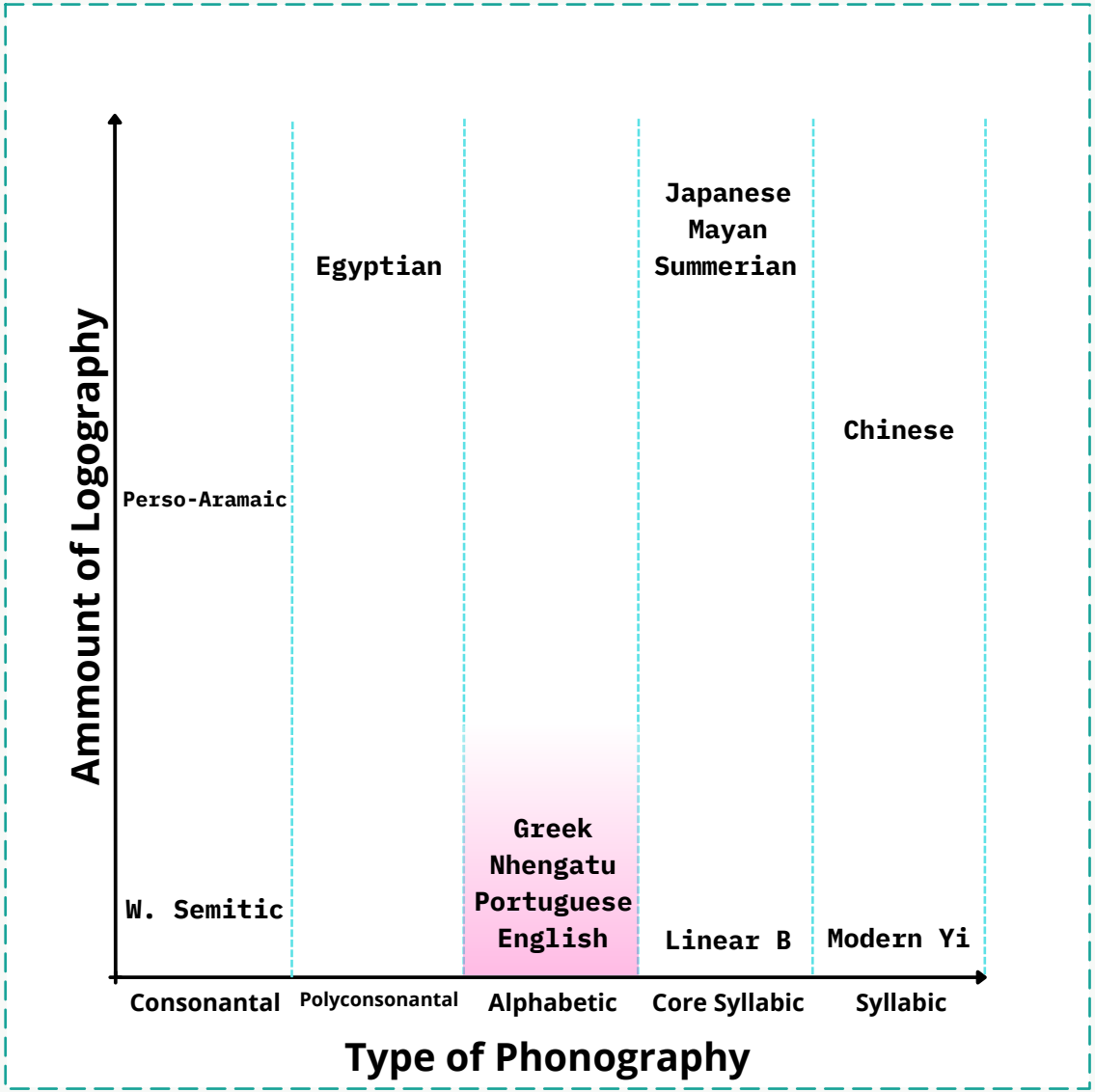
RESUMÃO

“Fiquei com preguiça de escrever esse slide, simplesmente vá passando pelos slides anteriores”

— Felipe do Passado

FUTURAS DIREÇÕES DE TRABALHO

- Validação estatística mais robusta e compreensiva sobre possíveis variações
- Análise do comportamento das métricas dadas variações de domínio de linguagem e recorte temporal em ptbr (e em outras línguas)
- Extensão e Adaptação das métricas para outros sistemas de escrita


Ammount of Logography
Perso-Aramaic
W. Semitic
Egyptian
Japanese
Mayan
Summerian
Chinese
Greek
Nhengatu
Portuguese
English
Linear B
Modern Yi
Consonantal
Polyconsonantal
Alphabetic
Core Syllabic
Syllabic
Type of Phonography

PRINCIPAIS REFERÊNCIAS

[1] Juola, P. (2008). Assessing linguistic complexity. Language Complexity: Typology, Contact, Change. John Benjamins Press, Amsterdam, Netherlands.

[2] Ehret, K. and Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. Complexity, isolation, and variation, 57, 71.

[3] Ehret, K. (2021). An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. Corpus Linguistics and Linguistic Theory, 17(2), 383-410.

[4] Serras, F., Carpi, M., Branco, M., & Finger, M. (2024, August). Analysing and Validating Language Complexity Metrics Across South American Indigenous Languages. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (pp. 152-165).

[5] Karlsson, F. (2006). Finnish as an agglutinating language. In K. Brown (Ed.), Encyclopedia of language & linguistics (second edition) (Second Edition, pp. 476–480). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00147-4

[6] de Reuse, W. J. (2006). Polysynthetic language: Central siberian yupik. In K. Brown (Ed.), Encyclopedia of language & linguistics (second edition) (Second Edition, pp. 745–748). Elsevier. https://doi.org/10.1016/B0-08-044854-2/04669-1

[7] Wikipedia contributors. (n.d.). Hixkaryana language. Wikipedia. Retrieved September 13, 2024, from https://en.wikipedia.org/wiki/Hixkaryana_language

[8] Bibliotheca Alexandrina Contributors. (n.d.) Active participles. Hieroglyphs Step by Step. Retrieved September 13, 2024, from https://www.bibalex.org/learnhieroglyphs/lesson/LessonDetails_En.aspx?l=127

CONTEÚDO EXTRA

→ Capa: Inscription from Tel Dan, drawn by User:Schreiber and originally uploaded to German Wikipedia by Benutzer:Schreiber, created by Schreiber, distributed under CC-by-sa 3.0, edited.
→ Flags from https://github.com/kapowaz/square-flags
→ Icons from https://lucide.dev/
→ Egyptian Hieroglyphs rendered using Jsesh (https://jsesh.qenherkhopeshef.org/)

AGRADECIMENTOS

→ Bruna Bazaluk
→ Miguel Carpi

Ferramentas de IA Gerativa foram utilizadas de forma assistiva na preparação desses slides.

OBRIGADO PELA ATENÇÃO ♡

CONTATO: FRSERRAS ‘AT’ IME ‘DOT’ USP ‘DOT’ BR