Mighty Morpho-Tagging Rangers

Group 6

Amandalynne Paullada
Naomi Tachikawa
Shapiro

Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages

Shauli Ravfogel[1] **Yoav Goldberg**[1,2] **Tal Linzen**[3]

[1]Computer Science Department, Bar Ilan University

[2]Allen Institute for Artificial Intelligence

[3]Department of Cognitive Science, Johns Hopkins University

{shauli.ravfogel, yoav.goldberg}@gmail.com, tal.linzen@jhu.edu

in ***Proceedings of NAACL-HLT 2019***

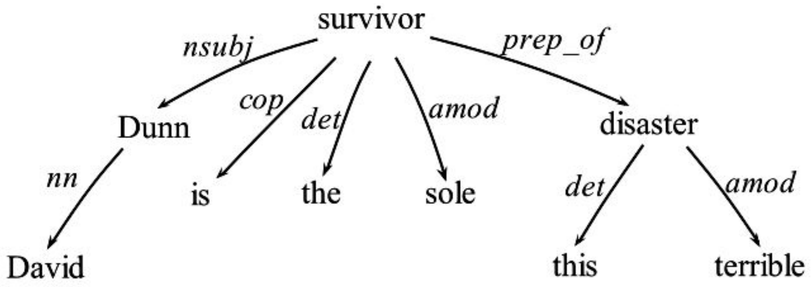
Background

- ❖ It's hard to make crosslinguistic comparisons of RNN syntactic performance (e.g., on subject-verb agreement prediction)
  - ➢ Languages differ in **multiple typological properties**
  - ➢ Cannot hold training data constant across languages

Proposal: **generate synthetic data** to devise a controlled experimental paradigm for studying the interaction of the inductive bias of a neural architecture with particular typological properties.

Setup

❖ Data: English Penn Treebank sentences converted to Universal Dependencies scheme



Example of a dependency parse tree

Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Language~~s~~

ONE

Shauli Ravfogel[1] **Yoav Goldberg**[1,2] **Tal Linzen**[3]

[1]Computer Science Department, Bar Ilan University

[2]Allen Institute for Artificial Intelligence

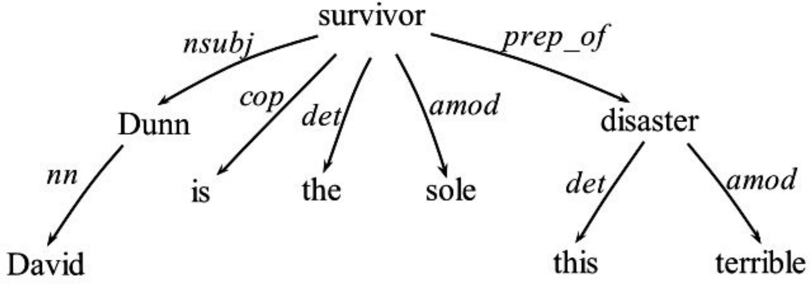
[3]Department of Cognitive Science, Johns Hopkins University

{shauli.ravfogel, yoav.goldberg}@gmail.com, tal.linzen@jhu.edu

in *Proceedings of NAACL-HLT 2019*

Setup

- Identify all verb arguments with *nsubj*, *nsubjpass*, *dobj* and record plurality (**HOW? manually?**)



Example of a dependency parse tree

Setup

- ❖ Generate synthetic data by appending novel morphemes to the verb arguments identified to inflect them for argument role and number

	Singular	Plural
Subject	-kar	-kon
Object	-kin	-ker
Indirect Object	-ken	-kre

Table 2: Case suffixes used in the experiments. Verbs are marked by a concatenation of the suffixes of their corresponding arguments.

Setup

❖ Generate synthetic data by appending novel morphemes to the verb arguments identified to inflect them for argument role and number

	Singular	Plural
Subject	-kar	-kon
Object	-kin	-ker
Indirect Object	-ken	-kre

No explanation or motivation given for how the novel morphemes were developed, nor an explicit mention that they're novel! Might length matter?

Typological properties

- Does jointly predicting object and subject plurality improve overall performance?
  - Generate data with **polypersonal agreement**
- Do RNNs have inductive biases favoring certain word orders over others?
  - Generate data with different **word orders**
- Does overt case marking influence agreement prediction?
  - Generate data with different **case marking systems**
    - unambiguous, syncretic, argument marking

Examples of synthetic data

Original		they say the broker took them out for lunch frequently . *(they, broker: subjects; say, took: verbs; them: object)*
Polypersonal agreement		they saykon the broker tookkarker them out for lunch frequently . *(kon: plural subject; kar: singular subject; ker: plural object)*
Word order variation	SVO	they say the broker took out frequently them for lunch .
	SOV	they the broker them took out frequently for lunch say .
	VOS	say took out frequently them the broker for lunch they.
	VSO	say they took out frequently the broker them for lunch .
	OSV	them the broker took out frequently for lunch they say .
	OVS	them took out frequently the broker for lunch say they *(they, broker: subjects; say, took: verbs; them: object)*
Case systems	Unambiguous	theykon saykon the brokerkar tookkarker theyker out for lunch frequently . *(kon: plural subject; kar: singular subject; ker: plural object)*
	Syncretic	theykon saykon the brokerkar tookkarkar theykar out for lunch frequently . *(kon: plural subject; kar: plural object/singular subject)*
	Argument marking	theyker sayker the brokerkin tookkerkin theyker out for lunch frequently . *(ker: plural argument; kin: singular argument)*

Task

- ❖ Predict a verb's subject and object plurality features.
  Input: synthetically-inflected sentence
  Output: one category prediction each for subject & object
    - subject: [singular, plural]
    - object: [singular, plural, none]   (if no object)

(It's NOT CLEAR in the paper WHAT the actual prediction task is / what the actual output space is. I had to look at their actual code to guess this. >:/)

Model

- Bidirectional LSTM with **randomly initialized** embeddings
  - so no influence on statistics of e.g. '-kar' & its ngrams in other data I guess
- Each word is represented as the sum of the word's embedding and its constituent character ngram (1-5) embeddings
- bi-LSTM representation of **left and right contexts of verb** fed into **two independent multilayer perceptrons**, one for subject prediction task, one for object prediction task

The prediction target (i.e., the inflected verb) is withheld during training, so what's in its place in the input??? Nothing? or a placeholder vector? -_-

Findings

- Performance was higher in subject-verb-object order (as in English) than in subject-object-verb order (as in Japanese), **suggesting that RNNs have a recency bias**
- Predicting agreement with both subject and object (polypersonal agreement) performs better than predicting each separately, **suggesting that underlying syntactic knowledge transfers** across the two tasks
- **Overt morphological case** makes **agreement prediction significantly easier**, regardless of word order.

Beyond plurality features

- ❖ No shade at number agreement!
- ❖ We're interested in predicting part-of-speech, grammatical gender, verb aspect, and more
- ❖ Control task paradigm is cool
- ❖ AP out.

Exploring BERT'S Vocabulary

Judit Ács

Introduction

- *Old news*: BERT models uses **WordPiece** (**WP**) tokenization!
  - Word pieces are *subword* tokens (e.g., "##ing")
  - WP tokenization models are data-driven:
    - *Given a training corpus, what set of D word pieces minimizes the number of tokens in the corpus?*
    - After specifying the # of desired tokens *D*, a WP model is trained to define a vocabulary of size *D* while greedily segmenting the training corpus into a minimal number of tokens (Wu et al. 2016; Schuster and Nakajima 2012)

BERT's multilingual vocabulary

- Ács (2019) focuses on BERT's *cased* multilingual WP vocabulary
  - 119,547 word pieces across 104 languages
  - Created using the top 100 Wikipedia dumps
  - WP tokenization ≠ morphological segmentation; e.g., *Elvégezhetitek*:

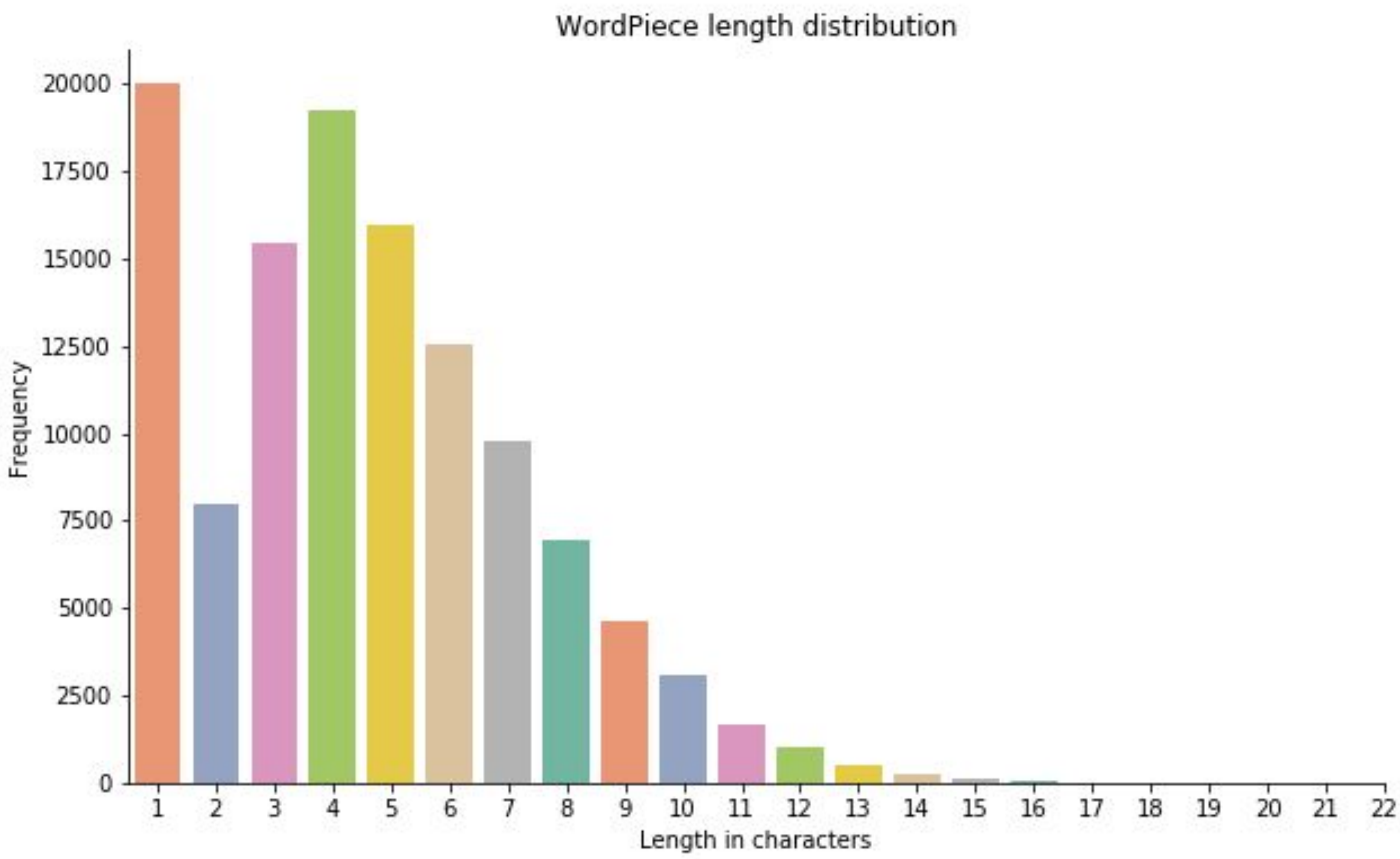
El, **végez**, **het**, **itek** (morphemes)

vs.

El, **##vé**, **##ge**, **##zhet**, **##ite**, **##k** (word pieces)

BERT's multilingual vocabulary (*cont'd*)

- 119,547 word pieces across 104 languages
- The first 106 pieces are reserved for special characters (e.g., PAD, UNK)
- 36.5% of the vocabulary are continuation pieces (e.g., "##ing")
- Every character is included as both a standalone word piece (e.g., "な") and as a continuation word piece (e.g., "##な").
  - The alphabet consists of 9,997, contributing 19,994 pieces
- The rest are multi-character word pieces of various lengths…


WordPiece length distribution
Frequency
0
2500
5000
7500
10000
12500
15000
17500
20000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Length in characters

The 20 longest word pieces

Token	Length
bewerkingsgeschiedenis	22
ուսումնասիրությունների	22
Territorialgeschichte	21
Europameisterschaften	21
huvudavrinningsområde	21
தேர்ந்தெடுக்கின்றனர்	20
Rechtswissenschaften	20
eenoogkreeftjessoort	20
Årsmedeltemperaturen	20
நிர்வகிக்கப்படுகிறது	20

Token	Length
Auseinandersetzungen	20
தொகுக்கப்பட்டுள்ளது	19
delstatshuvudstaden	19
Bevölkerungsstandes	19
Nationalsozialisten	19
Weltmeisterschaften	19
delavrinningsområde	19
bevolkingsdichtheid	19
Nationalsozialismus	19
Europameisterschaft	19

The land of Unicode

A word piece is said to *belong* to a Unicode category if all of its characters fall into that category or are digits.

Script	Sum	%
Latin	93495	78.21
ASCII	92327	77.23
CJK+kana	14932	12.49
Cyrillic	13782	11.53
CJK	13601	11.38
Indian	6545	5.47
Arabic	4873	4.08
Korean	3273	2.74
Hebrew	2482	2.08
Greek	1566	1.31
Kana	1331	1.11
Armenian	1236	1.03
Georgian	705	0.59
Misc	639	0.53
Thai	370	0.31
Myanmar	271	0.23
Tibetan	40	0.03
Mongolian	4	0.0

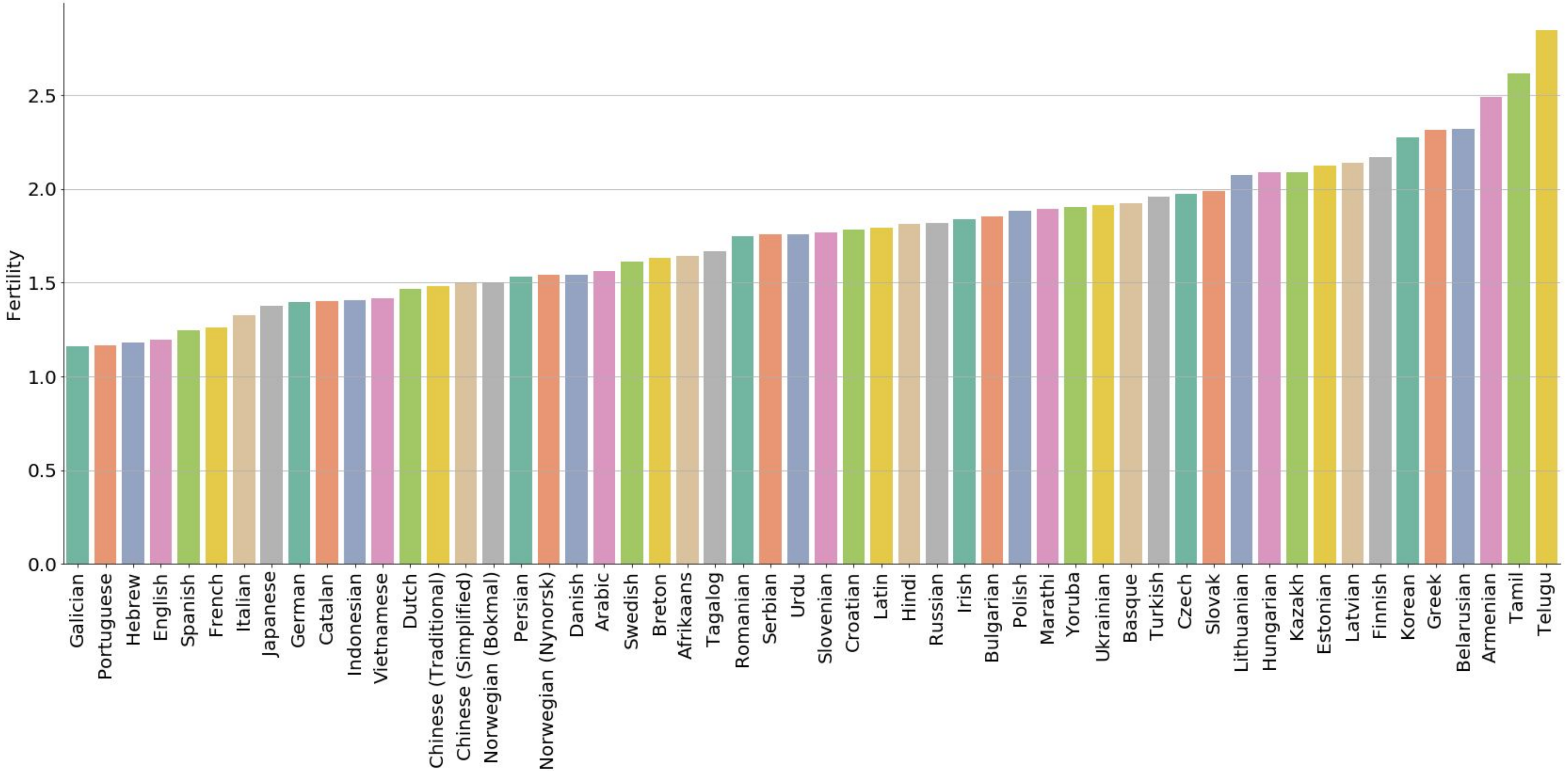
Tokenizing Universal Dependency (UD) treebanks

- UD provides treebanks for 70 languages that are annotated for morphosyntactic information, dependencies, and more
  - 54 of the languages overlap with multilingual BERT
  - *Nota bene*: UD treebanks differ in their cross-linguistic tokenization schemes
- Ács (2019) tokenized each of the 54 treebanks with HuggingFace's **`BertTokenizer`**

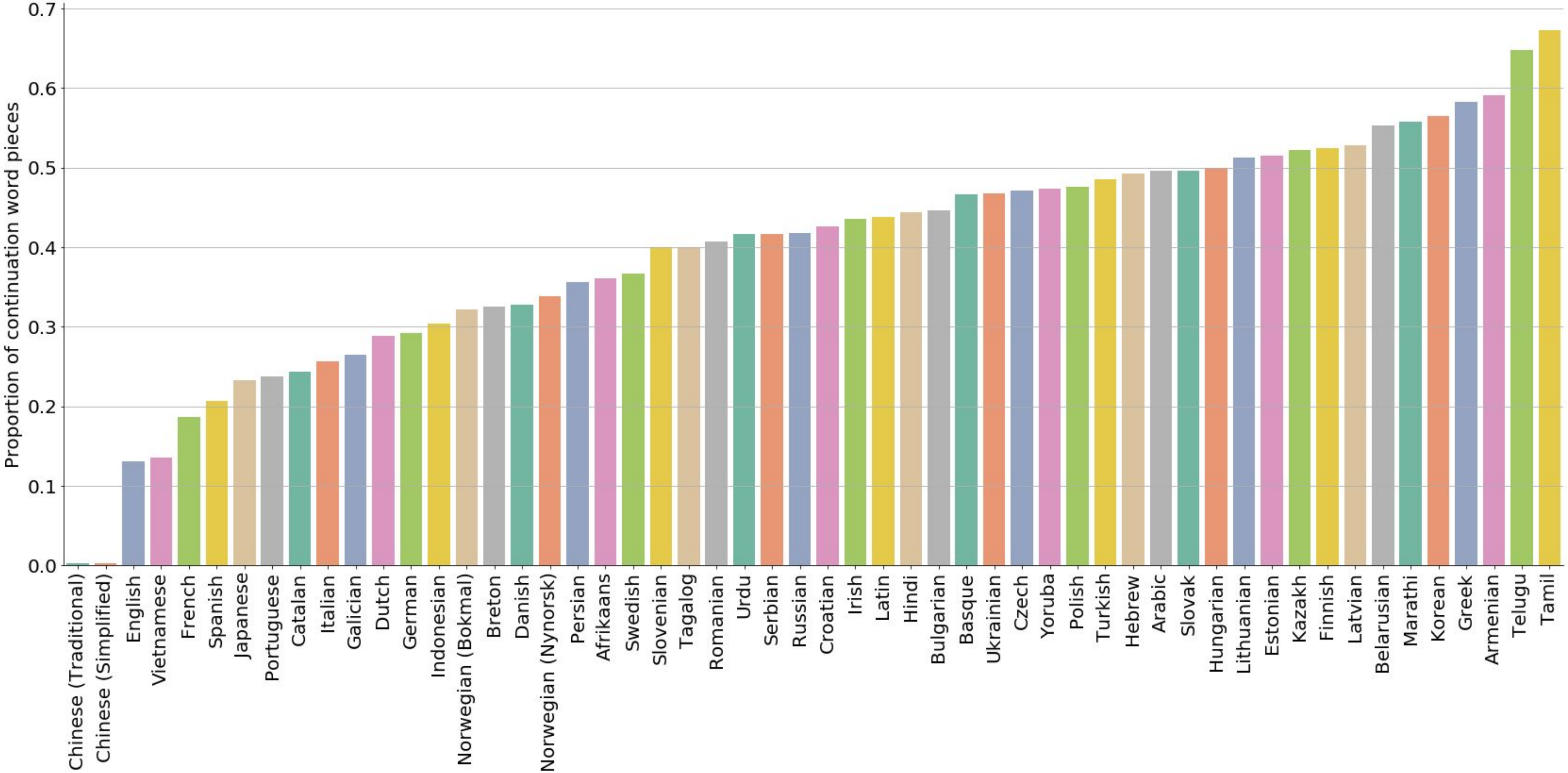
Fertility

Let ***fertility*** equal the number of word pieces corresponding to a single word-level token.

E.g., ["fail", "##ing"] has a fertility of 2.



Fertility
0.0
0.5
1.0
1.5
2.0
2.5
Galician
Portuguese
Hebrew
English
Spanish
French
Italian
Japanese
German
Catalan
Indonesian
Vietnamese
Dutch
Chinese (Traditional)
Chinese (Simplified)
Norwegian (Bokmal)
Persian
Norwegian (Nynorsk)
Danish
Arabic
Swedish
Breton
Afrikaans
Tagalog
Romanian
Serbian
Urdu
Slovenian
Croatian
Latin
Hindi
Russian
Irish
Bulgarian
Polish
Marathi
Yoruba
Ukrainian
Basque
Turkish
Czech
Slovak
Lithuanian
Hungarian
Kazakh
Estonian
Latvian
Finnish
Korean
Greek
Belarusian
Armenian
Tamil
Telugu


Proportion of continuation word pieces
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Chinese (Traditional)
Chinese (Simplified)
English
Vietnamese
French
Spanish
Japanese
Portuguese
Catalan
Italian
Galician
Dutch
German
Indonesian
Norwegian (Bokmal)
Breton
Danish
Norwegian (Nynorsk)
Persian
Afrikaans
Swedish
Slovenian
Tagalog
Romanian
Urdu
Serbian
Russian
Croatian
Irish
Latin
Hindi
Bulgarian
Basque
Ukrainian
Czech
Yoruba
Polish
Turkish
Hebrew
Arabic
Slovak
Hungarian
Lithuanian
Estonian
Kazakh
Finnish
Latvian
Belarusian
Marathi
Korean
Greek
Armenian
Telugu
Tamil

Crosslinguistic comparison of sentence and token lengths

- Ács (2019) also juxtaposes **sentences lengths** in word pieces and word-level tokens across the 54 languages:
  - juditacs.github.io/2019/02/19/bert-tokenization-stats.html (alphabetical order)
  - juditacs.github.io/assets/bert_vocab/bert_sent_len_full_fertility_sorted.png (fertility order)
- She also compares the distribution of **token lengths** across the same languages:
  - juditacs.github.io/assets/bert_vocab/bert_token_len_full.png (alphabetical order)
  - juditacs.github.io/assets/bert_vocab/bert_token_len_full_fertility_sorted.png (fertility order)

> *What are the ramifications of operating on* **word pieces?**