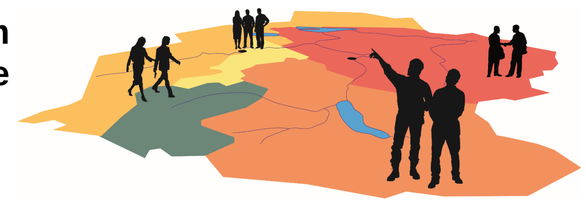
University of Zurich[UZH]

University Research Priority Program
Language and Space

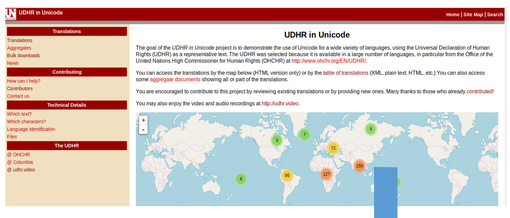


Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora

T. Samardžić (PI)[1], Ch. Bentz (PostDoc)[2], O. Sozinova (PhD student)[1]
[1]URPP Language and Space Lab, UZH, [2]URPP Language and Space Lab, UZH & DFG Center for Advanced Studies, University of Tübingen

Text Collection

Primary Sources



https://unicode.org/udhr/

2.1 Analyzed text

In this first text, Ruth Claplanhoo talks about her memories of the year 1918, when World War I ended, and she first went to boarding school.

(1) ʔiyaχaˈƛs skuˈlaʔeˈʔis.
ʔiyaχa[1] = ʾaƛ.[2] = s[3] skuˈla[4]-ʾeːʔis[5]
be.at = NOW = INDIC.1SG school-going.to
'I was going to school there.'

(2) wiʔibaqstiƛ̓itduˑ duˈwiqsubitdis.
wiʔiba[6]-aqstiƛ[7] = ʾit[8] = duː[9] duˈwiqsu[10] = (b)it[11] = dis[12]
angry -inside.MOM = PINV = 1PL father = PAST = POSS.1PL
'Our father got angry at us.'

Davidson & Werle (2010). Makah texts and analysis.

Digitalization

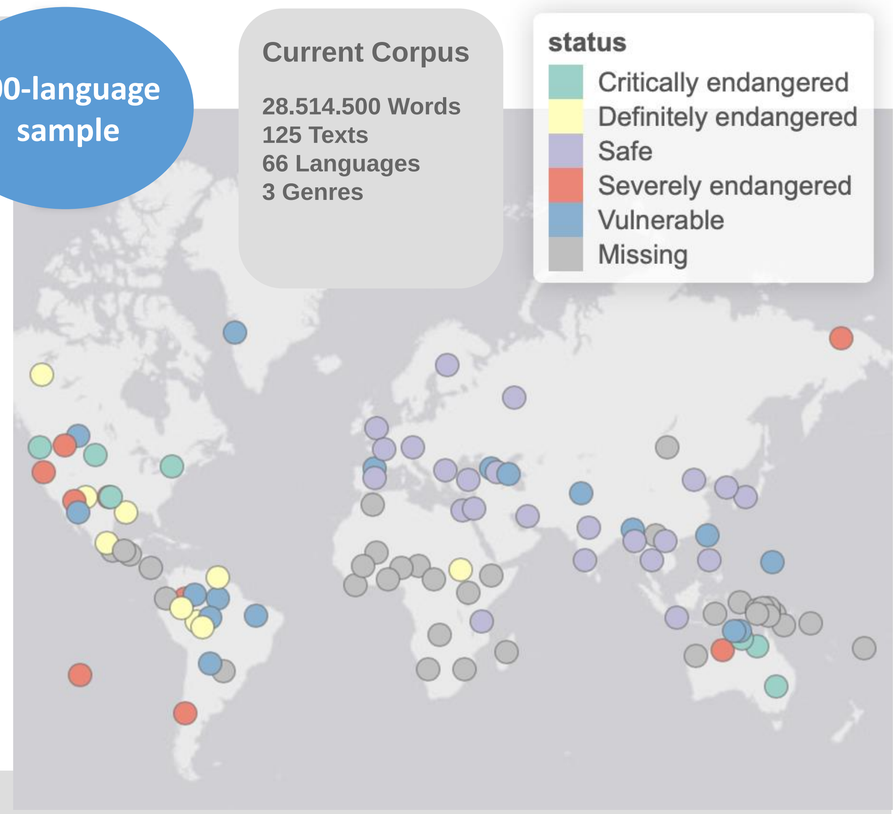
```
*rus_pro_1.txt
~/Data/1000/Texts/Raw
# language_name_wals:   Russian
# language_name_glotto: Russian
# ISO_639-3:            rus
# year_composed:        1996 - 2009
# year_published:       NA
# mode:                 written
# genre_(broad):        professional
# genre_(narrow):       official documents
# writing_system:       Cyrl

Всеобщая декларация прав человека
      Принята и провозглашена резолюцией 217 A (III) Генеральной Ассамблеи от 10
декабря 1948 года.

      ПРЕАМБУЛА
      Принимая во внимание, что признание достоинства, присущего всем членам
человеческой семьи, и равных и неотъемлемых прав их является основой свободы,
справедливости и всеобщего мира; и
      принимая во внимание, что пренебрежение и презрение к правам человека привели к
варварским актам, которые возмущают совесть человечества, и что создание такого мира, в
котором люди будут иметь свободу слова и убеждений и будут свободны от страха и нужды,
провозглашено как высокое стремление людей; и
```

```
*myh_con_1.txt
~/Data/1000/Texts/Raw
mya_nfi_1.txt | *myh_con_1.txt
# language_name_wals:   Makah
# language_name_glotto: Makah
# ISO_639-3:            myh
# year_composed:        NA
# year_published:       2010
# mode:                 spoken
# genre_(broad):        conversation
# genre_(narrow):       spontaneous speeches
# writing_system:       Latn

<line_1> ʔiyaχaˈƛs skuˈlaʔeˈʔis .
<segmentation> ʔiyaχa=ʾaƛ=s skuˈla-ʾeːʔis
<gloss> be.at=NOW=INDIC.1SG school-going.to
<translation> ' I was going to school there . '

<line_2> wiʔibaqstiƛ̓itduˑ duˈwiqsubitdis .
<segmentation> wiʔiba-aqstiƛ=ʾit=duː duˈwiqsu=(b)it=dis
<gloss> angry-inside.MOM=PINV=1PL father=PAST=POSS.1PL
<translation> ' Our father got angry at us . '
```

100-language sample

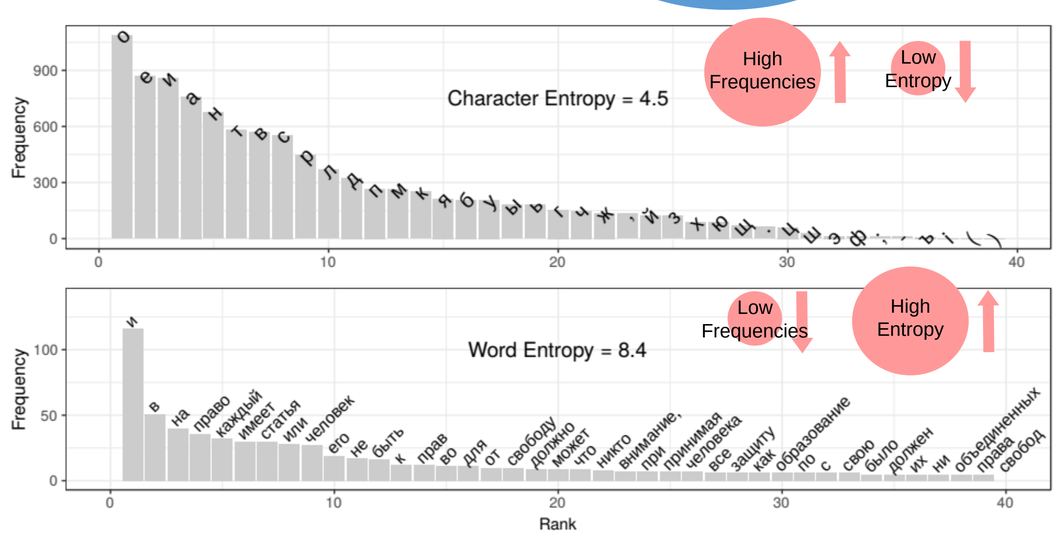
Current Corpus

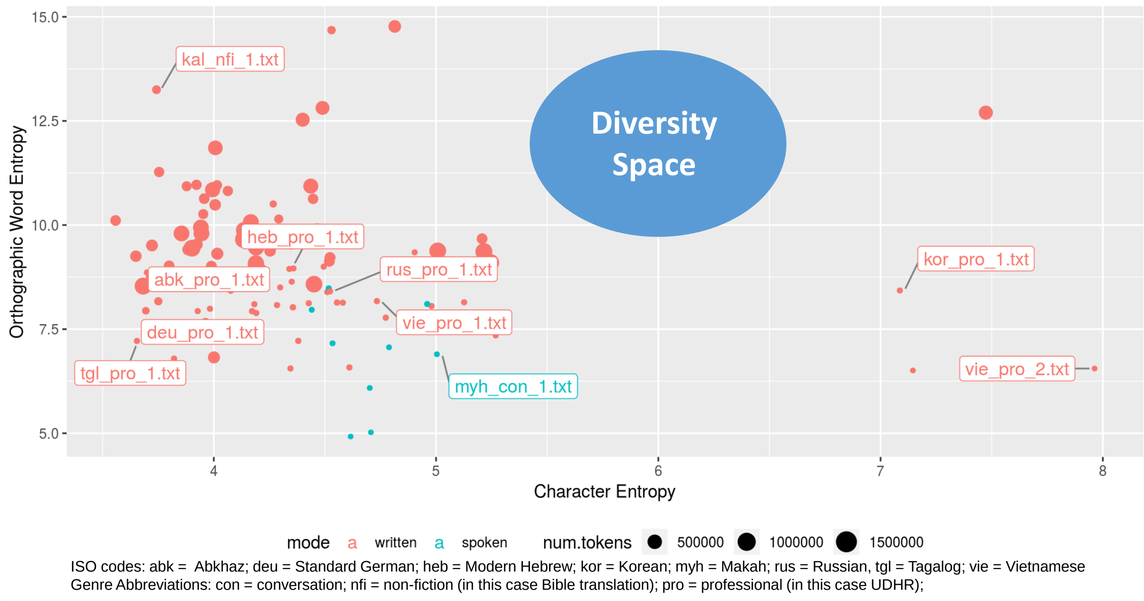
28.514.500 Words
125 Texts
66 Languages
3 Genres



Measuring Diversity

Characters & Words





ISO codes: abk = Abkhaz; deu = Standard German; heb = Modern Hebrew; kor = Korean; myh = Makah; rus = Russian, tgl = Tagalog; vie = Vietnamese
Genre Abbreviations: con = conversation; nfi = non-fiction (in this case Bible translation); pro = professional (in this case UDHR);

Explaining Diversity

Inflection & Derivation

Russian vs. German inflection 'cat'

	Singular	Plural
Nominative	koshka	koshki
Genitive	koshki	koshek
Dative	koshke	koshkam
Accusative	koshku	koshek
Instrumental	koshkoj	koshkami
Prepositional	koshke	koshkax

	Singular	Plural
Nominative	die Katze	die Katzen
Genitive	der Katze	der Katzen
Dative	der Katze	den Katzen
Accusative	die Katze	die Katzen

Russian vs. German derivation 'to see'

videt'
zavidet'
nenavidet'
voznenavidet'

predvidet'
nevzvidet'
providet'
uvidet'

sehen
absehen
hinabsehen
herabsehen
besehen
aufsehen
hinaufsehen

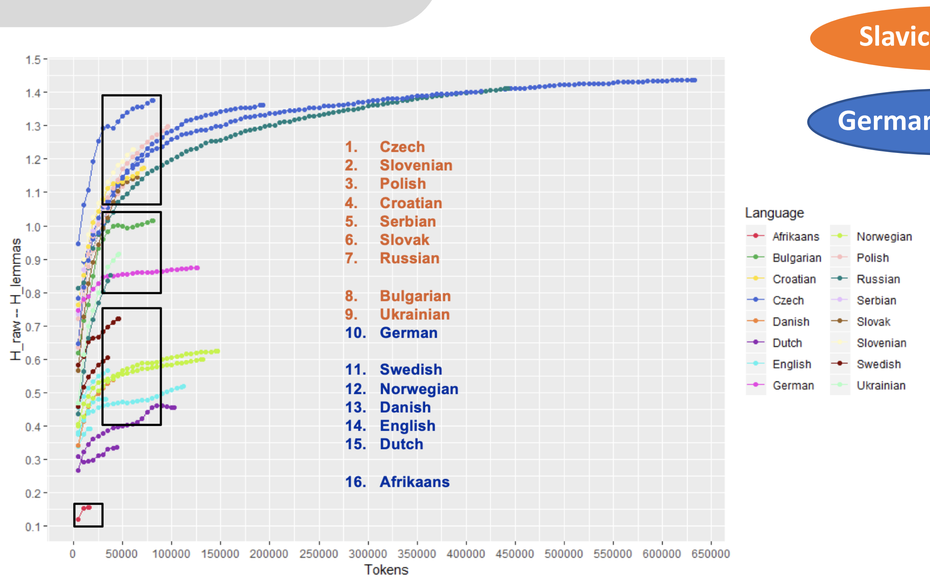
heraufsehen
wegsehen
hinwegsehen
gleichsehen
durchsehen
hindurchsehen
nachsehen

hochsehen
zurücksehen
hellsehen
umsehen
ansehen
entgegensehen
hinsehen

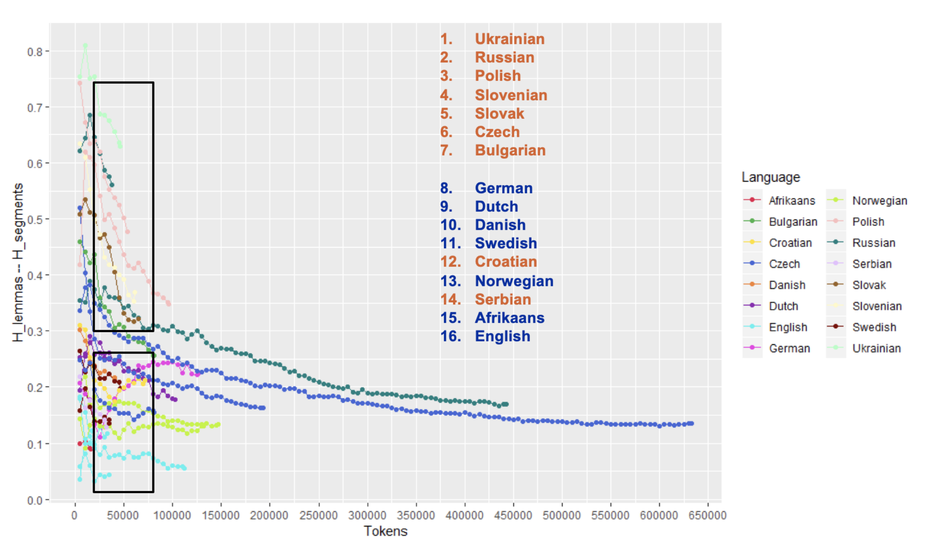
einsehen
hineinsehen
hereinsehen
reinsehen
...

Slavic languages

Germanic languages



Slavic vs. Germanic languages: inflectional complexity



Slavic vs. Germanic languages: derivational complexity