United Nations Six Way Parallel Corpora Analysis

Hello!

I am Kinan Al-Mouk

This project counts as my submission for my term project in Na-Rae Han's 2022 Data Science for Linguists course at the University of Pittsburgh.

U.N. Parallel Corpus

https://conferences.unite.un.org/uncorpus

English / Français / Español / Русский / 中文 / العربية

Background Info

Why did I choose to study the UN's Corpus?

The United Nations wants to make technology, software, and intellectual property available to everyone. Open source and free software are great tools to empower people and global collaboration.

Parallel Corpus Details

- composed of official records and other parliamentary documents of the United Nations
- produced and manually translated between 1990 and 2014
- purpose of the corpus is to allow access to multilingual language resources and facilitate research and progress in various natural language processing tasks

Corpus statistics

Statistics for pair-wise aligned documents:

	ar	en	es	fr	ru	zh
ar	–	111,241 18,539,207	113,065 18,578,118	112,605 18,281,635	111,896 18,863,363	91,345 15,595,948
en	456,552,223 512,087,009	–	123,844 21,911,121	149,741 25,805,088	133,089 23,239,280	91,028 15,886,041
es	459,383,823 593,671,507	590,672,799 678,778,068	–	125,098 21,915,504	115,921 19,993,922	91,704 15,428,381
fr	452,833,187 597,651,233	668,518,779 782,912,487	674,477,239 688,418,806	–	133,510 22,381,416	91,613 15,206,689
ru	462,021,954 491,166,055	601,002,317 569,888,234	623,230,646 513,100,827	691,062,370 557,143,420	–	92,337 16,038,721
zh	387,968,412 387,931,939	425,562,909 381,371,583	493,338,256 382,052,741	498,007,502 377,884,885	417,366,738 392,372,764	–

Research Questions

How can I use UN data to analyze linguistic features of standardized UN official languages.

How can I measure the way that standard multi language has changed in 25 years.

1

Data Processing

Pandas, NLTK, SpaCy, Juypter Notebook

main UN-Parallel-Corpora-Analysis / UN_Data_Analysis.ipynb

Uploading and Analyzing Raw Data

In [1]:

```
import nltk
import pickle
from time import time
import numpy as np
import pandas as pd
```

Reading in English File

In [2]:

```
start = time()
f = open('data/sixway/english.100k', 'r') # Reading in English File
english100 = f.read()
print("Data loaded in:", (time()-start), "seconds.")
f.close()
```

```
Data loaded in: 0.20158886909484863 seconds.
```

main UN-Parallel-Corpora-Analysis / UN_Data_Analysis.ipynb

Word Tokenizing English File

In [3]:

```
# Word Tokenization
start = time()
small_en_words = nltk.word_tokenize(english100)
print("English word tokenized in:", (time()-start), "seconds.")
small_en_words_len = len(small_en_words)
print("Word Token Count for English File:", small_en_words_len)
```

```
English word tokenized in: 18.570607900619507 seconds.
Word Token Count for English File: 3105868
```

Sentence Tokenizing English File

In [4]:

```
# Sentence Tokenization
start = time()
small_en_sents = nltk.sent_tokenize(english100)
print("English sentence tokenized in:", (time()-start), "seconds.")
small_en_sents_len = len(small_en_sents)
print("Sentence Token Count for English File:", small_en_sents_len)
```

```
English sentence tokenized in: 5.034231901168823 seconds.
Sentence Token Count for English File: 108375
```

Initial Problems in Processing

File Size

UN-Parallel-Corpora-Analysis

Size: 4,441,557,373 bytes (4.44 GB on disk) for 44 items

UNv1.0.6way.ar.zip

UNv1.0.6way.en.zip

UNv1.0.6way.es.zip

UNv1.0.6way.fr.zip

UNv1.0.6way.ru.zip

UNv1.0.6way.zh.zip

Initial Problems in Processing

NLTK

- **What I learned**: Mandarin appears to have drastically different values for every provided measure. I can understand why for most of them, but I think it may have been in part due to how the text was tokenized. For example, in the little snippet of the word tokens that I can see in `UN_Data_Analysis.ipynb` it looks like "1994年5月17日安全理事会" ("Security Council of May 17, 1994") is being treated as one token while I would've personally tokenized it as `["1994年", "5月", "17日", "安全", "理事会"]`.

Initial Problems in Processing

Time

Downloading SpaCy object for Mandarin processing

In [175…

```
start = time()
nlp = spacy.load('zh_core_web_sm')
zh_doc = nlp(mandarin_samp)
print("Mandarin document processed in:", (time()-start), "seconds.")
```

```
Mandarin document processed in: 881.6232738494873 seconds.
```

main UN-Parallel-Corpora-Analysis / SpaCyNLP.ipynb

In [61]:
```
import spacy
```

spaCy is designed specifically for production use and helps you build applications that process and “understand” large volumes of text. It can be used **to build information extraction or natural language understanding systems, or to pre-process text for deep learning**.

https://spacy.io › usage › spacy-101

spaCy 101: Everything you need to know

main UN-Parallel-Corpora-Analysis / SpaCyNLP.ipynb

In [61]:
```
import spacy
```

Downloading CAMEL Tools for Arabic processing

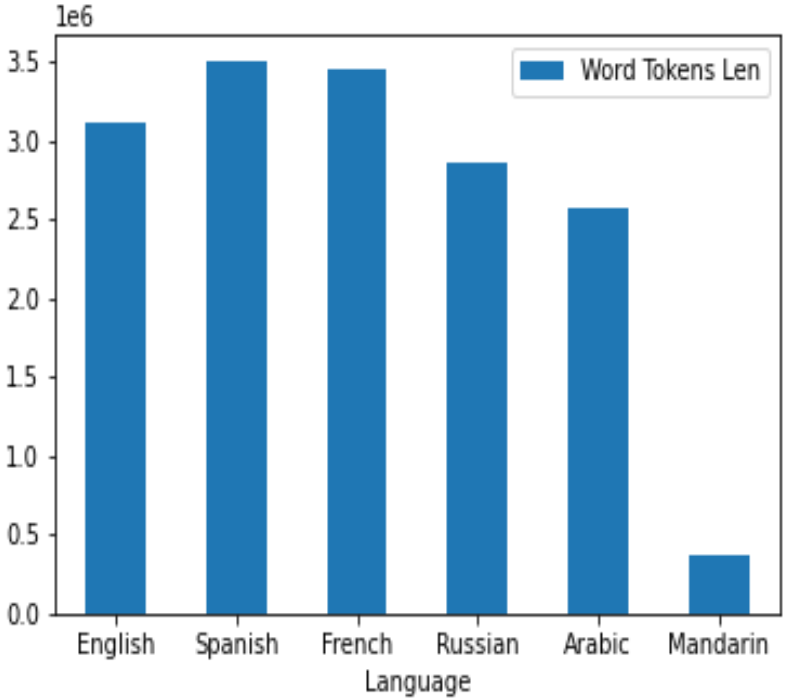
In [*]:
```
import camel_tools
```

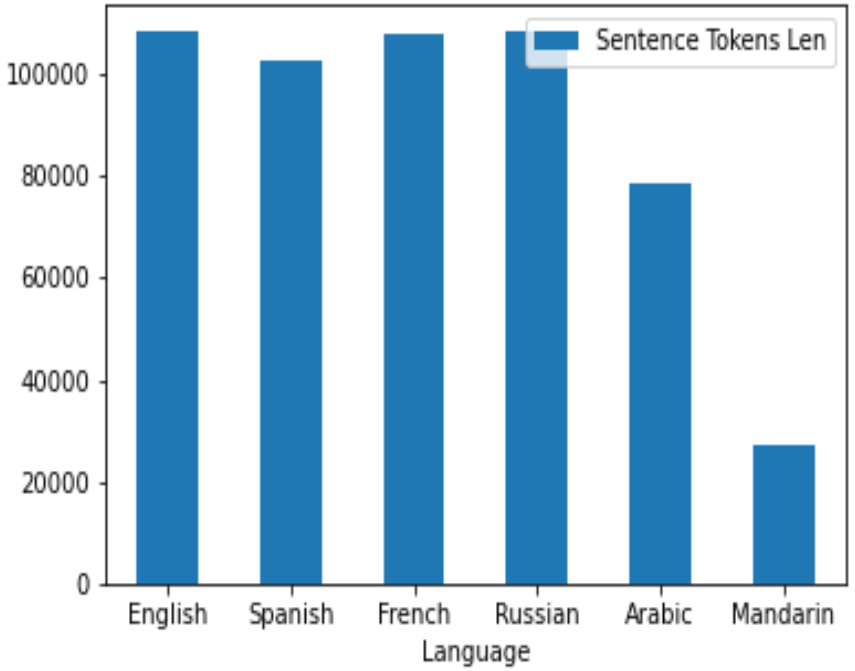
2

UN Data Analysis

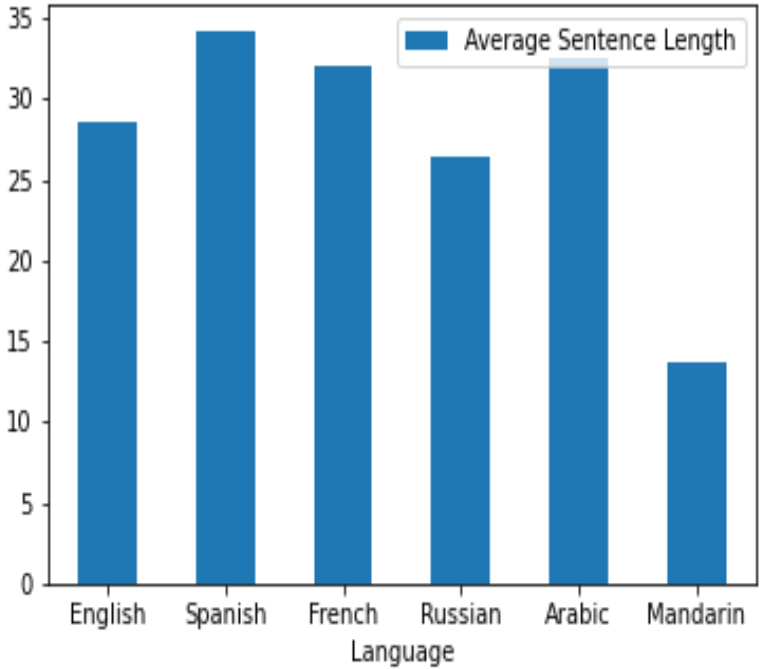
Linguistics!

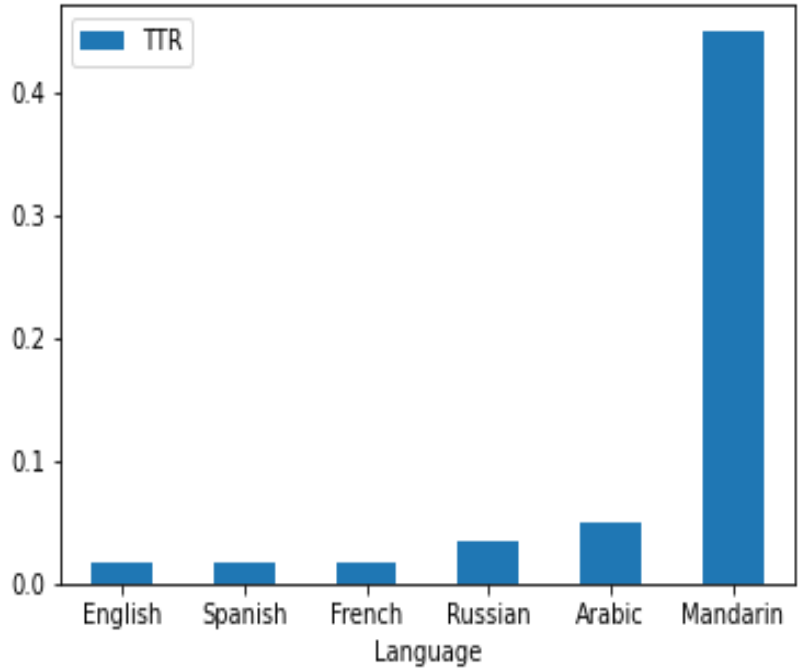
NLTK Data Analysis





NLTK Data Analysis





SpaCy Data Analysis

```
print(example_eng_sent,'\n\n',example_es_sent, '\n\n',
      example_fr_sent, '\n\n',example_ru_sent, '\n\n', example_zh_sent)
```

```
Demands that all parties to the conflict immediately cease hostilities, agree to a cease-fire, and bring an end to the mindless violence an
d carnage engulfing Rwanda;
2.

 Exige que todas las partes en el conflicto pongan fin inmediatamente a las hostilidades, convengan en una cesación del fuego y pongan térm
ino a la violencia y la carnicería insensatas en que está sumida Rwanda;
2.

 Exige que toutes les parties au conflit cessent immédiatement les hostilités, acceptent un cessez-le-feu et mettent fin à la violence et a
u carnage insensés dans lesquels est plongé le Rwanda;

 требует, чтобы все стороны в конфликте немедленно прекратили военные действия, договорились о прекращении огня и положили конец бессмысленному насилию и резне, ох
ватившим Руанду;


 要求卢旺达各当事方严格尊重联合国及其他在卢旺达服务的组织的人员和房地，并避免对从事人道主义和维持和平工作的人员进行任何恐吓或暴力行为；
12.
```

In [61]:
```
print(eng_ents[:10], '\n\nLength of English Entities:', len(eng_ents))
```

```
[['918' 'CARDINAL']
 ['1994' 'DATE']
 ['the Security Council' 'ORG']
 ['3377th' 'ORDINAL']
 ['17 May 1994' 'DATE']
 ['The Security Council' 'ORG']
 ['Rwanda' 'GPE']
 ['resolution 872' 'LAW']
 ['1993' 'DATE']
 ['5 October 1993' 'DATE']]

Length of English Entities: 12124
```

In [117…
```
fr_ents = np.column_stack((fr_ent, fr_label))
print(fr_ents[:10], '\n\nLength of French Entities:', len(fr_ents))
```

```
[['RESOLUTION 918' 'MISC']
 ['Conseil de sécurité' 'ORG']
 ['Conseil de sécurité' 'ORG']
 ['Rwanda' 'LOC']
 ['Mission des Nations Unies' 'MISC']
 ['Rwanda' 'LOC']
 ['MINUAR' 'LOC']
 ['MINUAR' 'MISC']
 ['MINUAR' 'MISC']
 ['Président' 'PER']]

Length of French Entities: 7853
```

Out [224...

	Language	Document	Word Tokens	Word Token Length	Sentence Tokens	Sentence Token Length	POS Set	POS Count	Dependency	Dependency Count	Entity Label List	Entity Label Count	Entity Count
0	English	RESOLUTION 918 (1994)\nAdopted by the Security...	[RESOLUTION, 918, (, 1994,), \n, Adopted, by,...	177808	[RESOLUTION 918 (1994)\nAdopted by the Securit...	6662	{INTJ, NOUN, X, NUM, DET, PRON, PART, ADP, VER...	18	{npadvmod, advcl, attr, acl, det, pcomp, quant...	45	{ORG, LAW, WORK_OF_ART, FAC, LOC, LANGUAGE, OR...	18	12124
1	Spanish	RESOLUCIÓN 918 (1994)\nAprobada por el Consejo...	[RESOLUCIÓN, 918, (, 1994,), \n, Aprobada, po...	176209	[RESOLUCIÓN 918 (1994)\nAprobada por el Consej...	4861	{INTJ, NOUN, NUM, DET, PRON, PART, ADP, VERB, ...	17	{expl:pass, advcl, det, acl, obl, conj, dep, c...	31	{MISC, LOC, ORG, PER}	4	9826
2	French	RESOLUTION 918 (1994)\nAdoptée par le Conseil ...	[RESOLUTION, 918, (, 1994,), \n, Adoptée, par...	181823	[RESOLUTION 918 (1994)\n, Adoptée, par le Cons...	9440	{NOUN, X, NUM, DET, PRON, ADP, VERB, PROPN, SP...	16	{aux:tense, expl:pass, advcl, det, acl, obl:ag...	36	{MISC, ORG, LOC, PER}	4	2762
3	Russian	РЕЗОЛЮЦИЯ 918 (1994),\nпринятая Советом Безопа...	[РЕЗОЛЮЦИЯ, 918, (, 1994,), ,, \n, принятая, ...	153964	[РЕЗОЛЮЦИЯ 918 (1994),\n, принятая Советом Без...	9310	{INTJ, NOUN, X, NUM, DET, PRON, PART, ADP, VER...	18	{advcl, det, acl, orphan, obl:agent, obl, conj...	40	{ORG, LOC, PER}	3	2126
4	Mandarin	第918(1994)号决议\n1994年5月17日安全理事会第3377次会议通过\n安全理事...	[第918, (, 1994,)号, 决议, \n, 1994年, 5月, 17日, 安全...	553748	[第918(1994)号决议\n1994年5月17日安全理事会第3377次会议通过\n安全理...	23363	{INTJ, NOUN, X, NUM, DET, PRON, PART, ADP, VER...	16	{advmod:dvp, name, aux:asp, amod:ordmod, mark:...	44	{ORG, LOC, LAW, WORK_OF_ART, FAC, PERCENT, ORD...	18	9476
5	Arabic	القرار ٨١٩ (١٩٩٤)\nالذي اتخذه مجلس الأمن في جلس...	null	null	null	null	null	null	null	null	null	null	null

Data is worth a thousand words

If I was able to process this Data file by file, we could make very interesting observations of political events, terminology, and linguistic features for each UN Language since 1990.

I plan to complete this with a better computer soon.

Credits

Special thanks to all the people who helped me learn the skills I needed to accomplish this

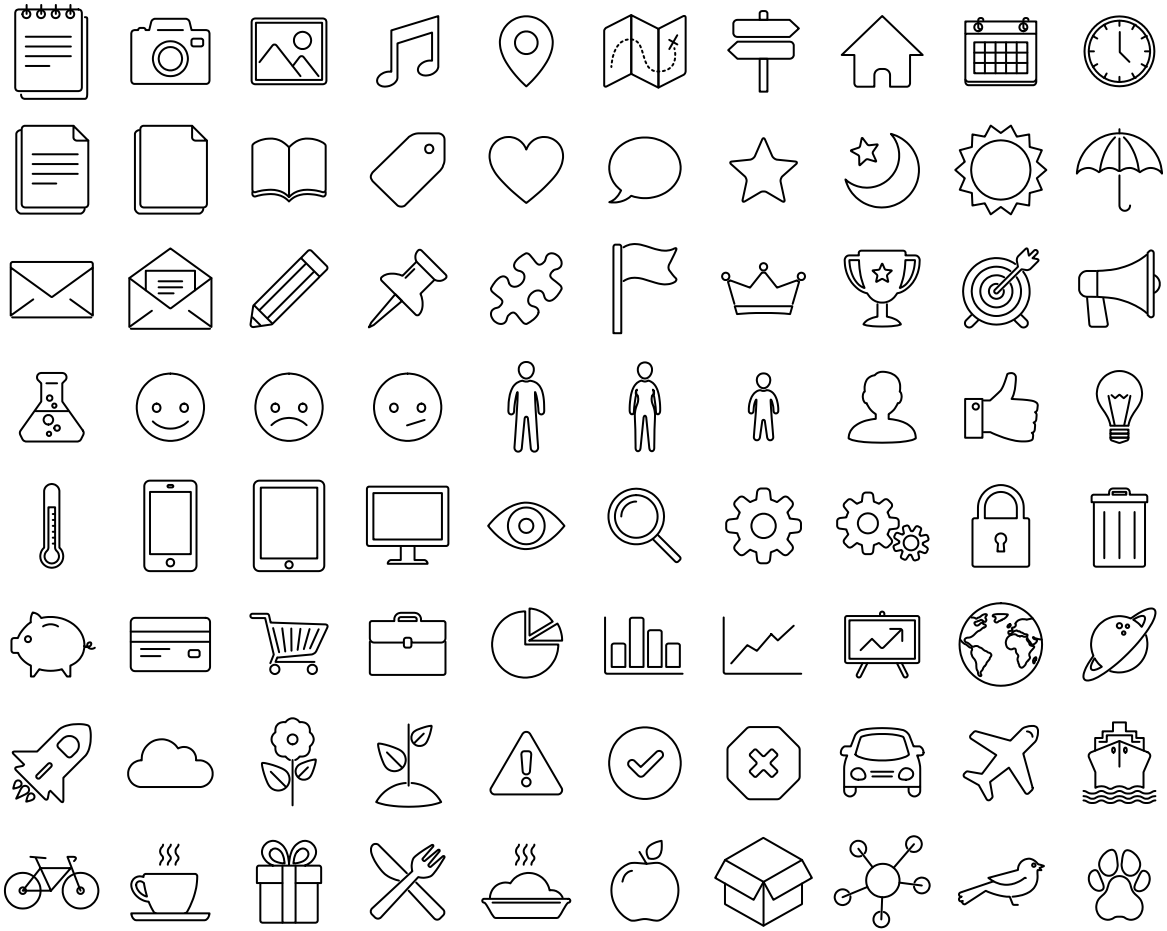
- Professor Na-Rae Han at the University of Pittsburgh
- TA's Lindsey Rojitas and Sean Steile

Thank You!

Any questions ?

You can find me at

- kim47@pitt.edu
- LinkedI



SlidesCarnival icons are editable shapes.

This means that you can:

- Resize them without losing quality.
- Change line color, width and style.

Isn't that nice? :)

Examples: