Introduction to Corpus Linguistics Final Project

Matthieu Rouxel, David McKean

Goal :

Look at the 3 verbs 'convince', 'obligate', and 'force (v)' and find significant differences between where they appear in certain texts (context, etc.)

Find the frequency counts of the three verbs

Train a model on classification between three verbs: 'convince', 'obligate', and 'force'.

Hypothesis: Convince can appear in a human interaction. Force can be used in more cases such as animated or inanimate objects. Obligate will mostly appear on a hierarchical level.

Matthieu Rouxel

Layout
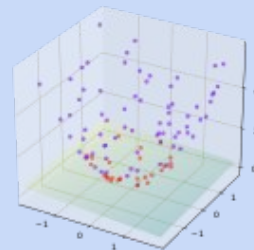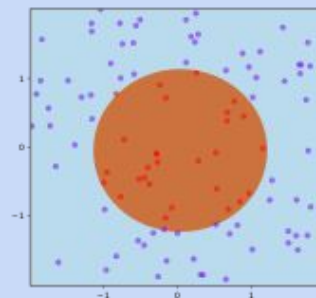
Sketch Engine - "Convince"

Here we looked at the verb convince from list A

modifiers of "convince"	objects of "convince"	subjects of "convince"	"convince" and/or ...
firmly ••• firmly convinced that	**sceptic** ••• convince sceptics	**argument** ••• convinced by this argument	**try** ••• to try and convince the
totally ••• totally convinced that	**public** ••• convince the public	**evidence** ••• convinced by the evidence	
entirely ••• not entirely convinced	**reader** ••• convince the reader	**success** ••• success convincing	
absolutely ••• I 'm absolutely convinced	**voter** ••• to convince voters that	**trouble** ••• had trouble convincing	
half ••• was half convinced	**other** ••• to convince others	**experience** ••• experience has convinced	
utterly ••• utterly convinced that	**doubter** ••• to convince the doubters	**case** ••• cases are not convincing	
finally ••• finally convinced	**colleague** ••• convince his colleagues	**problem** ••• problems convincing	
wholly ••• not wholly convinced by	**jury** ••• convince the jury of		
altogether ••• not altogether convinced	**employer** ••• convincing their employers		
thoroughly ••• thoroughly convinced that	**everyone** ••• not everyone is convinced		
apparently ••• apparently convinced	**Nizan** ••• Nizan was convinced that the		
personally ••• was personally convinced it was	**manager** ••• managers are convinced		

Sketch Engine - "Force"

Here we looked at the verb force from list A

modifiers of "force"	nouns modified by "force"	verbs with "force" as object	verbs with "force" as subject
armed •••	**majeure** •••	**join** •••	**act** •••
the armed forces	of force majeure	joined forces	forces acting on
• especially: W_hansard	**wind** •••	• especially: W_newsp_other_report	**launch** •••
• especially: W_non_ac_polit_law_edu	gale force winds	• especially: W_news_script	forces launched a
• especially: W_newsp_brdsht_nat_report	**personnel** •••	• usually: Written books and periodicals	• usually: W_non_ac_polit_law_edu
show more (1)	armed forces personnel	**peacekeepe** •••	**shape** •••
labour •••	**participation** •••	peacekeeping force	forces shaping
the labour force	labour force participation	• especially: W_non_ac_polit_law_edu	**attack** •••
• especially: W_non_ac_soc_science	**commander** •••	**drive** •••	forces attacked
• especially: W_ac_soc_science	air force commander	the driving force behind	**fight** •••
• especially: W_commerce	**ECOMOG** •••	• usually: Written books and periodicals	forces fighting
show more (1)	the regional peacekeeping force ECOMOG	**deploy** •••	**capture** •••
security •••	**F** •••	forces deployed	forces captured
the security forces	force F	**station** •••	**operate** •••
• especially: W_newsp_other_report	**constant** •••	forces stationed in	forces operating
• especially: W_newsp_other_reportage	the force constants	• especially: W_non_ac_polit_law_edu	**seize** •••
• especially: W_hansard	**pilot** •••	• usually: W_non_ac_polit_law_edu	forces seized
show more (3)	air force pilot	**exert** •••	**kill** •••
police •••	**helicopter** •••	force exerted	killed by the security forces
police force	the force helicopter	**oppose** •••	**determine** •••
• especially: W_hansard	**field** •••	opposing forces	forces determining
• especially: W_non_ac_polit_law_edu	the force field	**motivate** •••	**control** •••
• especially: W_news_script	**jet** •••	the motivating force	forces controlled
show more (3)	two Yugoslav air force jets	**withdraw** •••	**drive** •••
task •••		withdraw its forces	
task force		**arm** •••	

Sketch Engine - "Obligate"

Here we looked at the verb obligate from list A

Egrep - "Convince"


```
[maro3214@verbs2 brown]$ egrep 'convince' BROWN1_A.TXT
A14 0640 immediate expansion, he sought to convince another National League
A31 1280 Jim became convinced that an individual can do something constructive
A34 0070 later probably was inevitable and even desirable. But he was convinced
A34 0220 Mr& Kennedy had become convinced that a personal confrontation with
A34 0510 as Mr& Khrushchev is convinced that the balance of world power is
A34 1020 for Laos. Mr& Kennedy was convinced that insistence on the demand
A42 0200 The conversation apparently convinced Mr& Kennedy that the
[maro3214@verbs2 brown]$
```

Egrep - "Force"

Here we have the result from the expression: egrep "convince" BROWN1_A.TXT

```
[maro3214@verbs2 brown]$ egrep 'force' BROWN1_A.TXT
A01 0940 the audience was warned that entering a candidate for governor would force
A02 0180 would force banks, insurance firms, pipeline companies and other
A02 0200 law cannot be enforced now because it is almost impossible to locate
A02 0230 keynote when he said it would force banks to violate their contractual
A02 1870 which have forced these people to depend upon ~ADC for subsistence",
A03 1870 The rule was enforced by demand of Sen& Wayne Morse [D&,
A04 0660 and the need for strengthening conventional forces as well as the
A04 0680 "threshold", as the conventional forces strengthening is called, will
A05 0690 problem which confronts our local and state law enforcement officers
A05 0710 attorney general has advised local police that it is their duty to enforce
A05 0750 Rhode Island in recent weeks toward enforcement of the Sunday sales
A05 1000 that a regrouping of forces might allow the average voter a better pull
A05 1800 as a means of making enforcement of minor offenses more effective.
A06 0310 is being waged on state issues and I will not be forced into re-arguing
A06 0590 is an insult to the man who led our forces to victory in the greatest
A06 1130 me that the office of sheriff is best filled by a man with law enforcement
A07 0830 of Arkansas, who has been siding with Speaker Sam Rayburn's forces
A07 1780 head of the pro-Communist Pathet Lao forces. The latter
A08 0510 branches of the armed forces will take part in the parade.   Division
A08 0600 by the reserve forces of all services, five states, Puerto Rico, the
A14 0230 with the Pittsburgh Steelers.   An injury forced Skorich
A14 0680 got off the ground, but after two years it forced the existing majors
A14 1610 rules into force. #RULING FROM ON HIGH# "I can remove the ball,
A19 1030 use snow tires or chains on them, should be "strictly enforced".
A19 1940 Heating is by individual gas-fired, forced warm air systems.
A21 1630 the death penalty for captured invasion forces and suspected collaborators.
A27 0340 he said, was not due to inadequate growth forces inherent in our
A28 0150 are getting such high yields per acre that many are being forced
A34 1190 moves that greatly strengthened the pro-Communist forces. As a result,
A36 0560 can be forced to the wall in a very short time. Should these giants
A36 0710 govern the enforcers of the law on Sunday? ## THE QUESTION
A36 1440 by any forced and artificial means".   Compare this statement
A37 0330 to increase the readiness of our forces in the Pacific". Cutting
A37 1280 to the Smith-Colmer forces: as chairman of the Appropriations
A37 1610 bipartisan force- from Minnesota's Democratic Hubert Humphrey
A41 0060 grab a colony while the grabbing was still good. By force, he took under
A41 1330 assistance of the United Nations, the national security forces
A41 1720 to February 1961, the U&N& forces in the Congo took little action.
A41 1830 **h the use of force, if necessary, in the last resort". Although
A42 0130 their forces from the Mediterranean base. To continue their important
A42 1520 the armed forces would be increased by only 160,000. The "hold-back",
[maro3214@verbs2 brown]$
```

Egrep - "Obligate"

As we can see here we got on result from the first text


```
[maro3214@verbs2 brown]$ egrep 'obligate' BROWN1_A.TXT
A25 1070 do not feel obligated to make the people adequately aware of this danger;
[maro3214@verbs2 brown]$
```

NLTK - Downloading Books

import nltk

gb = nltk.download('gutenberg')

nltk.download('genesis')

nltk.download('inaugural')

nltk.download('nps_chat')

nltk.download('webtext')

nltk.download('treebank')

from nltk.book import *

nltk.download('brown')

```
[3]: from nltk.book import *

*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

NLTK

convince

text1.concordance("convince")

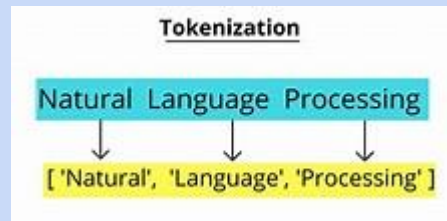text2.concordance("convince")

force

text1.concordance("force")

text2.concordance("force")

obligate

text1.concordance("oblige")

text2.concordance("oblige")

```
df_dic = {"text":[], "convince":[], "force":[], "obligate":[]}
temp_list = []
found = 0
convince = 0
obligate = 0
for i in range(40000):
    if(text1[:40000][i] != "."):
        temp_list.append(text1[:40000][i])
        if(text1[:40000][i] == "convince"):
            convince = 1
        elif(text1[:40000][i] == "force"):
            force = 2
        elif(text1[:40000][i] == "obligate"):
            obligate = 3
    else:
        df_dic["text"].append(temp_list)
        if(convince == 1):
            df_dic["convince"].append(1)
        else:
            df_dic["convince"].append(0)
        if(force == 2):
            df_dic["force"].append(1)
        else:
            df_dic["force"].append(0)
        if(obligate == 3):
            df_dic["obligate"].append(1)
        else:
            df_dic["obligate"].append(0)

        temp_list = []
# text1[:198]
# text1[:20][17] == "to"
df = pd.DataFrame(df_dic)
# df.tail(100)
# df
df.loc[df["force"] == 1]
```


Tokenization

Natural Language Processing

↓ ↓ ↓

['Natural', 'Language', 'Processing']

elf the comfort of endeavouring to convince Lucy that her heart was unwounded
tance they chose ; and he tried to convince me , as thoroughly as he was convi
." Elinor tried very seriously to convince him that there was no likelihood o
uld give no information that would convince Marianne ; and painful as the cons
was so , by that which only could convince her , a better knowledge of mankin
him hopes that his eloquence would convince her in TIME , another visit , anot
Displaying 12 of 12 matches:
in a man of greatly superior natural force , with a globular brain and a ponder
e caught me as I was again trying to force open the door . " I don ' t allow it
ary usages ; and not only that , but force himself to evince all his well known
 such a yawning commander , by sheer force of contrast , acted like a charm upo
whom these spikes fall with impaling force . But far more terrible is it to beh
umped the German ' s aside with such force , that both Derick and his baffled h
ew round the loggerheads with such a force as to gouge deep grooves in them ; w
n the tail the confluent measureless force of the whole whale seems concentrate
 vanished ; then , with the tapering force of his parting momentum , we glided
 a modified form , is to this day in force in England ; and as it offers in var
he above - mentioned law is still in force , I proceed to lay before you a circ
s great buoyancy , rising with great force , the coffin life - buoy shot length
Displaying 11 of 11 matches:
 , it lasted no longer than till the force of her arguments and the brightness
it eligible . I would not attempt to force the confidence of any one ; of a chi
eresting , must be before her , must force her attention , and engross her memo
them immediately , which was in full force at the end of every week , they were
 , as soon as she could speak , " and force him to come to me . Tell him I must
etirement of Barton , where it might force him before her while paying that vis
ith great humility , did not see the force of her objection . " They had alread
nduct and her own . She felt all the force of that comparison ; but not as her
outh , when her visitor entered , to force her upon this greatest exertion of a
r two trifled with or denied , would force itself by increasing ailments on the
t , and a mistaken confidence in the force of his engagement . " I was simple e
Displaying 2 of 2 matches:
s . " Elijah ," said I , " you will oblige my friend and me by withdrawing . W
 this dirty scrape ." " Anything to oblige ye , my sweet and pleasant fellow ,
Displaying 2 of 2 matches:

Beginner's Guide to Building Custom NLP Models with NLTK

NLTK - Frequency Count

Code: text1.count("convince") / len(text1) * 100
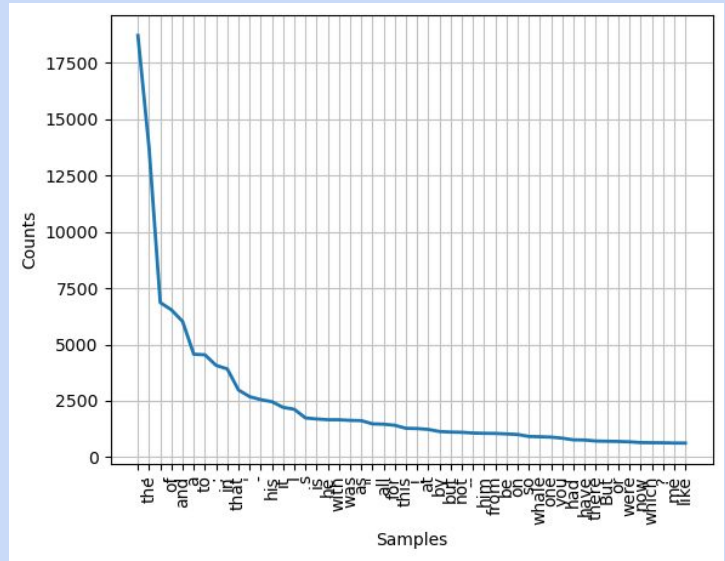
Output: 0.00038340765051625837

Code: text1.count("force") / len(text1) * 100

Output: 0.004600891806195101

Code: text1.count("oblige") / len(text1) * 100

Output: 0.0007668153010325167

Support Vector Machine: Word Classification based on POS tag

```
Displaying 1 of 1 matches:
 ; therefore , I am all anxiety to convince ye , ye landsmen , of the injustic
None
Displaying 2 of 2 matches:
s . " Elijah ," said I , " you will oblige my friend and me by withdrawing . W
  this dirty scrape ." " Anything to oblige ye , my sweet and pleasant fellow ,
None
Displaying 12 of 12 matches:
in a man of greatly superior natural force , with a globular brain and a ponder
e caught me as I was again trying to force open the door . " I don ' t allow it
ary usages ; and not only that , but force himself to evince all his well known
  such a yawning commander , by sheer force of contrast , acted like a charm upo
whom these spikes fall with impaling force . But far more terrible is it to beh
umped the German ' s aside with such force , that both Derick and his baffled h
ew round the loggerheads with such a force as to gouge deep grooves in them ; w
n the tail the confluent measureless force of the whole whale seems concentrate
  vanished ; then , with the tapering force of his parting momentum , we glided
  a modified form , is to this day in force in England ; and as it offers in var
he above – mentioned law is still in force , I proceed to lay before you a circ
s great buoyancy , rising with great force , the coffin life – buoy shot length
None
```

Concordances in nltk.book texts

Support Vector Machine: Word Classification based on POS tag

	target (y)	n-1 TAG	n+1 TAG	n-2 TAG	n+2 TAG	nTAG
1	0	PART	PRON	N	PRON	V
2	1	AUX	DET	PRON	N	V
3	1	PART	PRON	N	DET	V
4	2	ADJ	PREP	ADV	DET	N
5	2	PART	V	V	DET	V
6	2	CONJ	PRON	CONJ	PREP	V
7	2	ADJ	PREP	PREP	N	N
8	2	ADJ	CONJ	PREP	ADV	N
9	2	ADV	CONJ	PREP	CONJ	N
10	2	DET	CONJ	ADV	PREP	N
11	2	ADJ	PREP	ADJ	DET	N
12	2	ADJ	PREP	DET	DET	N
13	2	PREP	PREP	N	N	N
14	2	PREP	PRON	ADJ	V	N
15	2	ADJ	DET	PREP	N	N
16	0	PART	PRON	N	CONJ	V
17	0	PART	PRON	CONJ	ADV	V
18	0	PART	N	V	PREP	V
19	0	PART	N	V	CONJ	V
20	0	PART	PRON	V	ADV	V
21	0	PART	PRON	ADV	CONJ	V
22	0	AUX	N	CONJ	CONJ	V
23	0	AUX	PRON	ADV	DET	V
24	0	AUX	PRON	N	PREP	V
25	1	AUX	PRON	CONJ	PART	V
26	1	PART	PRON	V	CONJ	V
27	2	DET	PREP	PREP	PRON	N

Raw data (categorical)

Support Vector Machine: Word Classification based on POS tag
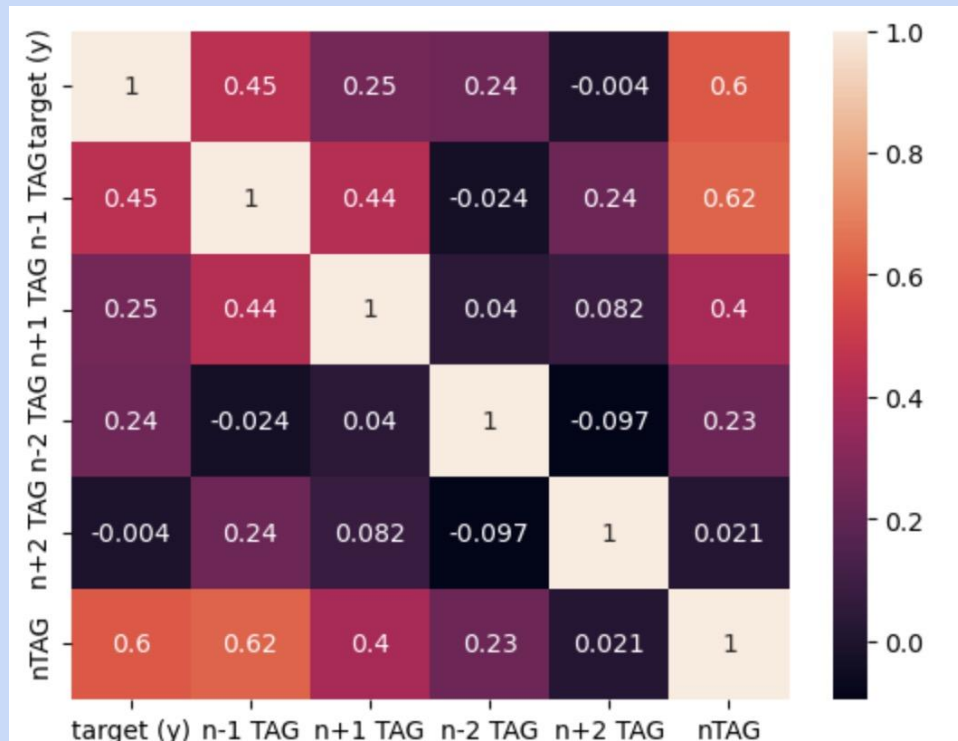
Correlation heatmap

Support Vector Machine: Word Classification based on POS tag

```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         3
           1       0.00      0.00      0.00         2
           2       0.76      0.89      0.82        18

    accuracy                           0.70        23
   macro avg       0.25      0.30      0.27        23
weighted avg       0.60      0.70      0.64        23

Confusion Matrix:
[[ 0  0  3]
 [ 0  0  2]
 [ 2  0 16]]
Test Accuracy:   0.6956521739130435
```

SVM and results
(very bad)

Support Vector Machine: Sentence TF-IDF in Brown Corpus

```
Accuracy: 0.96

Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       108
           1       1.00      0.91      0.95        11
           2       0.00      0.00      0.00         4

    accuracy                           0.96       123
   macro avg       0.65      0.64      0.64       123
weighted avg       0.93      0.96      0.94       123

Confusion Matrix:
[[108   0   0]
 [  1  10   0]
 [  4   0   0]]
```

Original three words

Support Vector Machine: Sentence TF-IDF in Brown Corpus

```
Accuracy: 0.88

Classification Report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.93       108
           1       1.00      0.91      0.95        11
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00         5
           4       0.00      0.00      0.00         1
           5       1.00      0.17      0.29         6
           6       0.00      0.00      0.00         4

    accuracy                           0.88       136
   macro avg       0.41      0.30      0.31       136
weighted avg       0.81      0.88      0.83       136

Confusion Matrix:
[[108   0   0   0   0   0   0]
 [  1  10   0   0   0   0   0]
 [  1   0   0   0   0   0   0]
 [  5   0   0   0   0   0   0]
 [  1   0   0   0   0   0   0]
 [  5   0   0   0   0   1   0]
 [  4   0   0   0   0   0   0]]
```

All words in list A

In Continuation…

Applying this to other languages

Refine the simple models that we have

Experimenting with other models

- RandomForestClassifier
- GradientBoostingClassifier

Training and using a larger Corpus or multiple Corpuses.