Problem 2

For this problem set, we will use
https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fcovid19

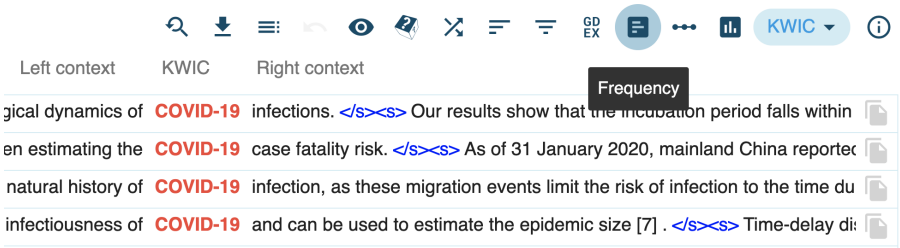
Problem 2.1

Click the above link, and follow this: Dashboard -> Concordance -> Advanced -> CQL.

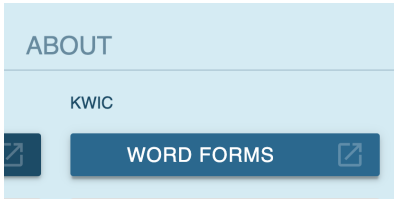
Now write a query to find sentences containing all forms of covid and execute it. Some forms include covid-19, covid19, COVID19, covid-36, covid-54.

Once you get the sentences, click `Frequency -> KWIC > WORD FORMS` to generate the frequency of words. These steps are shown below:

Step 1:



Step 2:



Step 3: The word list looks something like this:

		Word	↓ Frequency	Per million tokens
1	☐	COVID-19	20,773	73.99
2	☐	Covid-19	429	1.53
3	☐	COVID19	169	0.60
4	☐	COVID-2019	157	0.56
5	☐	CoVID-19	32	0.11

What is the CQL query that you used for getting all forms of covid (i.e. the query that is used to generate the above figure)?
Answer: [word = “[cC][oO][vV][iI][dD]-?\d+”]

Include the snapshot of the top 20 words (5 words are shown above)?
Answer:

(31 items, 21,642 total frequency)

	Word	Frequency ↓	Relative ?
1	COVID-19	20,773	73.99
2	Covid-19	429	1.53
3	COVID19	169	0.60
4	COVID-2019	157	0.56
5	CoVID-19	32	0.11
6	covid-19	30	0.11
7	CoViD-19	10	0.04
8	COVID-10	7	0.02
9	COVID-9	7	0.02
10	Covid-2019	4	0.01
11	covid19	3	0.01
12	Covid19	2	< 0.01
13	covid-10	1	< 0.01
14	COVID-138	1	< 0.01
15	Covid-10	1	< 0.01
16	Covid-56	1	< 0.01
17	COVID-173	1	< 0.01
18	COVID-27	1	< 0.01
19	COVID-110	1	< 0.01
20	COVID-2	1	< 0.01

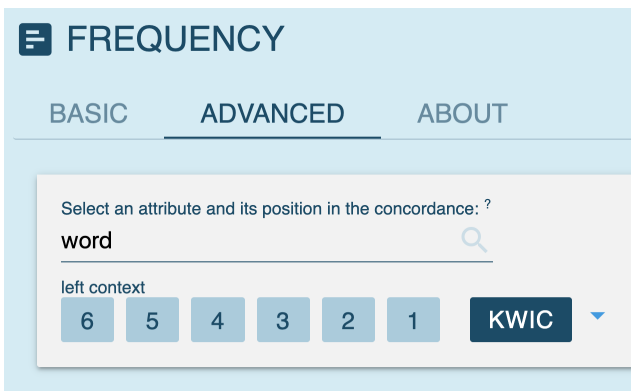
Problem 2.2

Let's write CQL queries to find interesting words that occur in specific syntactic relations with covid (all forms). We did similar things in class. You will have to use tag and lemma in CQL queries. This [tagset]() could be useful

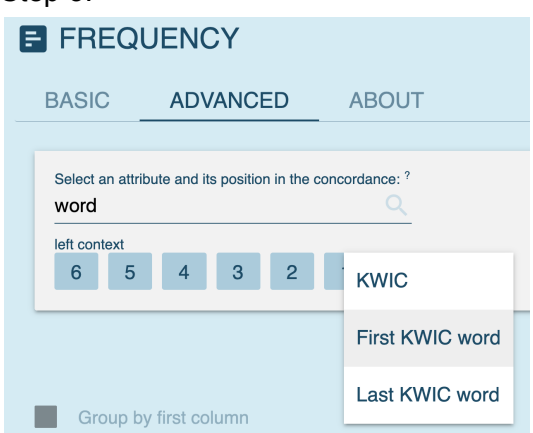
I will demonstrate how to get the modifiers of covid:

Step 1: First write a CQL query that produces concordance (examples) like this:

Step 2:



Step 3:



Step 4:

		Word	↓ Frequency	Per million tokens
1	☐	severe	298	1.06
2	☐	confirmed	115	0.41
3	☐	current	103	0.37

What is the CQL query for modifiers of covid (all forms)?
Answer: [tag = "J.*"][word = "[cC][oO][vV][iI][dD]-?\d+"]

Include the snapshot of modifiers (top three are shown above)

(307 items, 1,712 total frequency)

		Word	Frequency ↓	Relative ?
1	☐	severe	298	1.06
2	☐	confirmed	115	0.41
3	☐	current	103	0.37
4	☐	suspected	81	0.29
5	☐	laboratory-confirmed	69	0.25
6	☐	ongoing	64	0.23
7	☐	new	43	0.15
8	☐	first	42	0.15
9	☐	mild	40	0.14
10	☐	reported	31	0.11
11	☐	critical	27	0.10
12	☐	potential	26	0.09
13	☐	global	24	0.09
14	☐	ill	22	0.08
15	☐	asymptomatic	18	0.06
16	☐	early	17	0.06
17	☐	moderate	16	0.06
18	☐	active	15	0.05
19	☐	novel	14	0.05
20	☐	recent	14	0.05

What is the CQL query of words that are modified by covid (all forms)?
Answer: [word = "[cC][oO][vV][iI][dD]-?\d+"][tag = "N.*"]

Include the snapshot of those words

(549 items, 8,856 total frequency)

	Word	Frequency ↓	Relative ?
1	patients	1,720	6.13
2	cases	954	3.40
3	outbreak	721	2.57
4	infection	696	2.48
5	epidemic	540	1.92
6	pneumonia	496	1.77
7	pandemic	409	1.46
8	resource	396	1.41
9	virus	153	0.54
10	case	147	0.52
11	infections	144	0.51
12	transmission	141	0.50
13	disease	125	0.45
14	patient	104	0.37
15	spread	63	0.22
16	diagnosis	58	0.21
17	outbreaks	55	0.20
18	testing	55	0.20
19	treatment	51	0.18
20	mortality	50	0.18

What is the CQL query for words that occur in right coordination with covid (all forms)
(e.g., in COVID-19 , SARS-2002 , and HCoV-NL63, the words iSARS-2002 and HCoV-NL63 are the right conjucts/coordinates).
Answer:
[word = "[cC][oO][vV][iI][dD]-?\d+" & tag = "N.*"][word = "\," | tag = "N.*"]{0,}[tag = "CC"][tag = "N.*"]

Include the snapshot of those words

(210 items, 426 total frequency)

	Word	Frequency	Relative
1	SARS	31	0.11
2	MERS-COV	14	0.05
3	MERS	13	0.05
4	Treatment	11	0.04
5	SARS-CoV-2	11	0.04
6	control	10	0.04
7	prevention	9	0.03
8	H1N1	9	0.03
9	treatment	8	0.03
10	HAPE	6	0.02
11	death	6	0.02
12	deaths	6	0.02
13	influenza	5	0.02
14	SARS-CoV	5	0.02
15	mortality	5	0.02
16	RoS	4	0.01
17	cancer	4	0.01
18	SARS-2002	4	0.01
19	diagnosis	4	0.01
20	B	4	0.01

What is the CQL query for verbs that can take covid (all forms) as subject?
Answer: [word = "[cC][oO][vV][iI][dD]-?\d+" & tag = "N.*"][]{0,3}[tag = "VV.*"]

Include the snapshot of verbs that take covid as subject

(1,436 items, 8,273 total frequency)

	Word	Frequency	Relative
1	remains	444	1.58
2	reported	228	0.81
3	confirmed	191	0.68
4	caused	179	0.64
5	spread	138	0.49
6	using	133	0.47
7	based	113	0.40
8	including	105	0.37
9	identified	98	0.35
10	found	90	0.32
11	admitted	74	0.26
12	showed	72	0.26
13	compared	69	0.25
14	spreading	69	0.25
15	included	66	0.24
16	include	64	0.23
17	associated	60	0.21
18	occurred	56	0.20
19	according	56	0.20
20	become	51	0.18

What is the CQL query for verbs that can take covid (all forms) as object?
Answer: [tag = "VV.*"][]{0,5}[word = "[cC][oO][vV][iI][dD]-?\d+" & tag = "N.*"]

Include the snapshot of verbs that take COVID as object.

(1,637 items, 11,070 total frequency)

	Word	Frequency ↓	Relative ?
1	confirmed	663	2.36
2	reported	252	0.90
3	diagnosed	242	0.86
4	infected	211	0.75
5	used	130	0.46
6	suspected	128	0.46
7	associated	127	0.45
8	related	107	0.38
9	caused	103	0.37
10	control	99	0.35
11	named	95	0.34
12	hospitalized	94	0.33
13	found	90	0.32
14	including	87	0.31
15	declared	85	0.30
16	prevent	83	0.30
17	treat	83	0.30
18	treating	78	0.28
19	regarding	77	0.27
20	contain	73	0.26

Problem 2.3

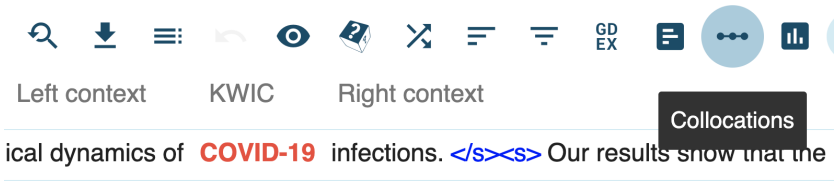
What are the most important words that form collocations with COVID (where covid is the right word)?

You can generate collocations as follows: First get concordance of all forms of covid.

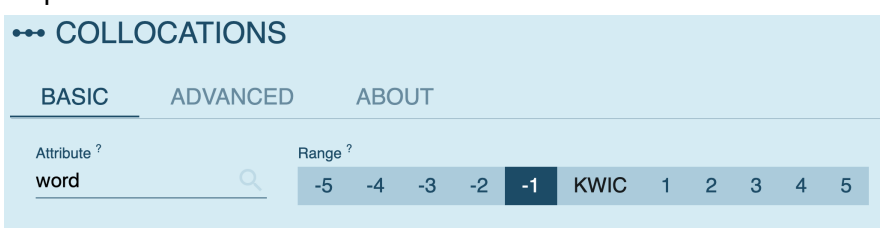
Step 1:

Left context	KWIC	Right
lemiological dynamics of	COVID-19	infect
red when estimating the	COVID-19	case
udy the natural history of	COVID-19	infect

Step 2:



Step 3:



Step 4:

		Word	Cooccurrences ?
1	☐	confirmed	458
2	☐	suspected	133

Show the collocations sorted according to what you think is the best metric (T-Score, MI, LogDice). Indicate the metric you used.

		Word	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓
1	☐	confirmed	458	65,495	21.17	6.50	7.43
2	☐	suspected	133	21,439	11.39	6.33	6.66
3	☐	laboratory-confirmed	75	3,601	8.63	8.08	6.61
4	☐	severe	298	112,078	16.76	5.11	6.19
5	☐	ongoing	64	12,451	7.88	6.06	5.94
6	☐	treat	57	14,546	7.40	5.67	5.69
7	☐	treating	46	9,478	6.67	5.98	5.60
8	☐	current	103	50,596	9.76	4.72	5.55
9	☐	declared	33	4,219	5.69	6.66	5.39
10	☐	towards	44	17,999	6.42	4.99	5.18

		Word	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓
11	☐	hospitalized	40	15,225	6.14	5.09	5.15
12	☐	named	33	9,921	5.61	5.43	5.10
13	☐	about	114	98,409	9.97	3.91	4.96
14	☐	non-severe	21	825	4.57	8.37	4.94
15	☐	with	1,999	2,412,053	40.55	3.43	4.75
16	☐	mild	40	30,282	5.96	4.10	4.66
17	☐	against	151	180,158	11.16	3.44	4.62
18	☐	contracted	17	1,822	4.09	6.92	4.57
19	☐	having	34	27,817	5.46	3.99	4.49
20	☐	of	5,930	8,766,274	68.23	3.13	4.47

I used LogDice.

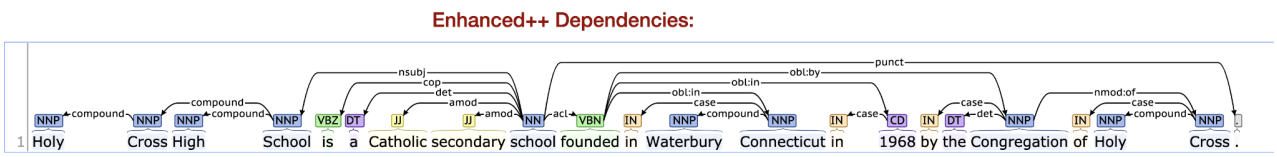
Problem 3:

Write SemGrex regular expressions that can detect organizations and their founders. Make use of https://corenlp.run to parse sentences to syntactic graphs and for running SemGrex expressions.

Here is an example:

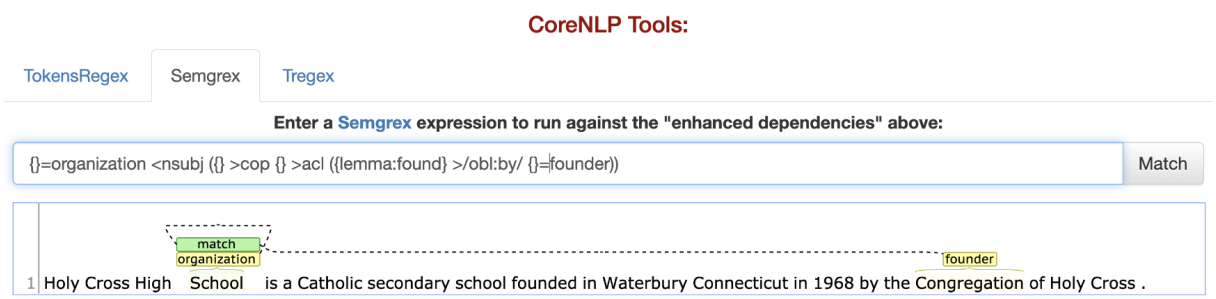
Holy Cross High School is a Catholic secondary school founded in Waterbury Connecticut in 1968 by the **Congregation of Holy Cross .**

The corresponding Enhanced++ Dependencies syntactic graph is as follows:



The below SemGrex pattern extracts the headword of the organization and the headword of the founder.

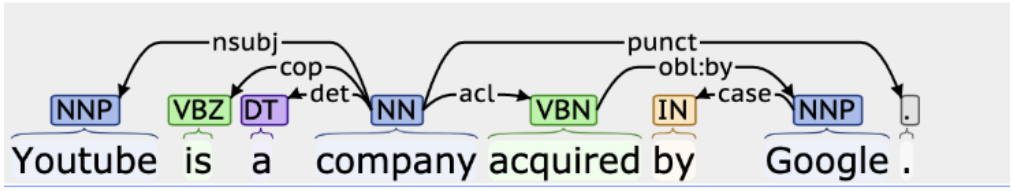
{}=organization <nsubj ({} >acl ({lemma:found} >/obl:by/ {}=founder))

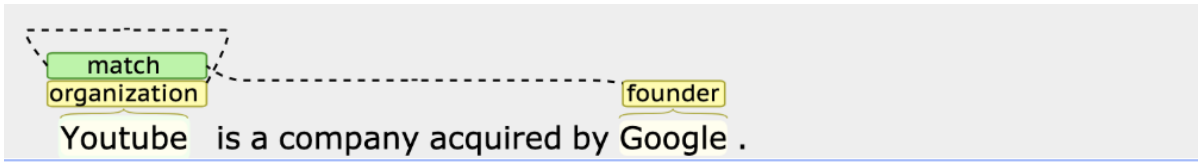


This pattern can be read as the “organization” that is a subject of something, and this something is founded by the founder.

Here it extracts School (i.e., the headword of Holy Cross High School) as the organization and Congregation (i.e., the headword of the Congregation of Holy Cross) as the founder.

Your goal is to write SemGrex expressions that can generalize to multiple sentences but at the same time don't match incorrect sentences. For example, if you don't use {lemma:found} in the above sentence, your pattern will also match a sentence like "Youtube is a company acquired by Google" (see below.)



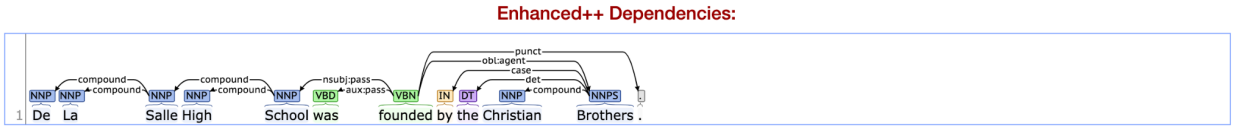


Problem 3.1

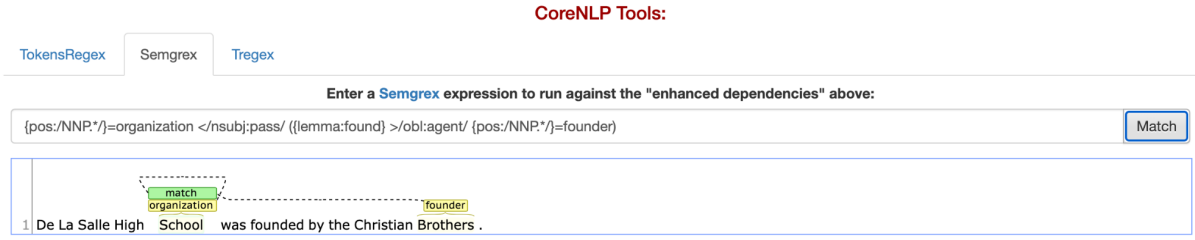
Write the SemGrex patterns for the following sentences that extract the organization name (headword is enough) and its founder (headword is enough). Sentences that can make use of the same expression should be in the same snapshot (containing Enhanced++ Dependencies, Semgrex expression, and the matchings):

De La Salle High School was founded by **the Christian Brothers** .

Enhanced++ Dependencies:

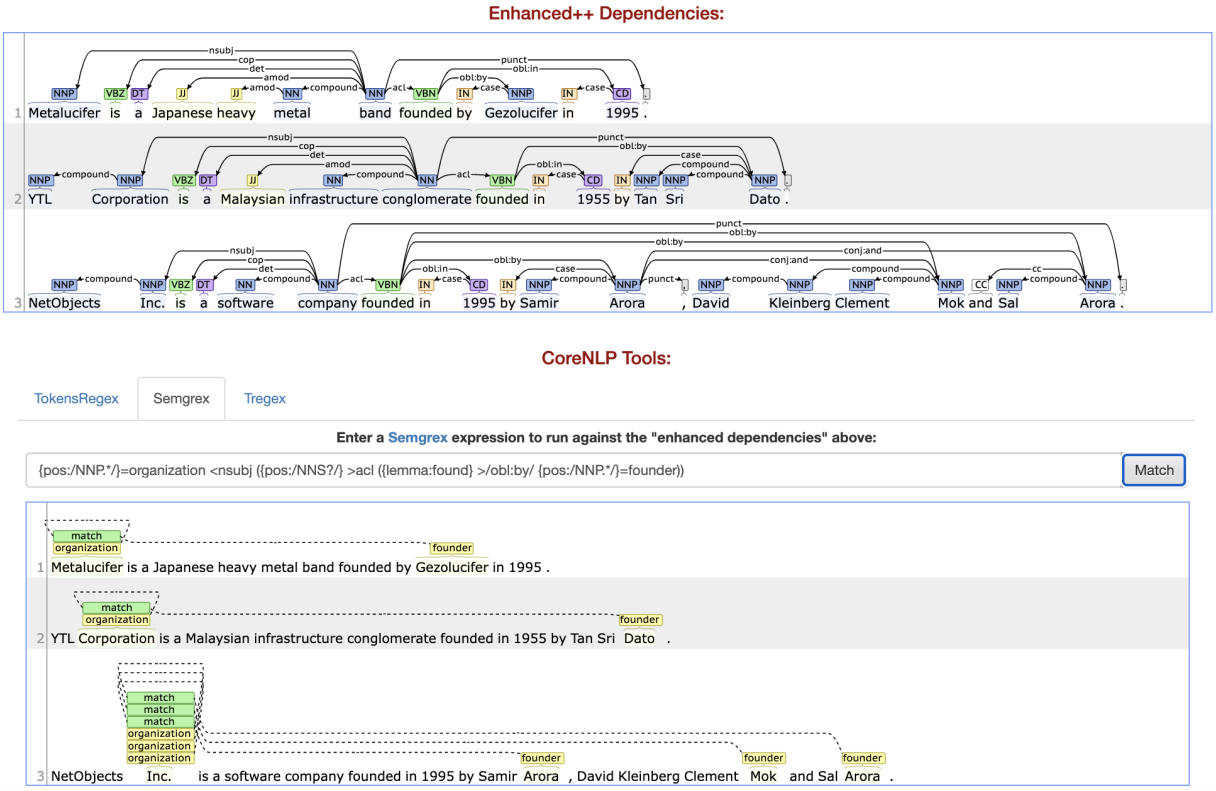


CoreNLP Tools:

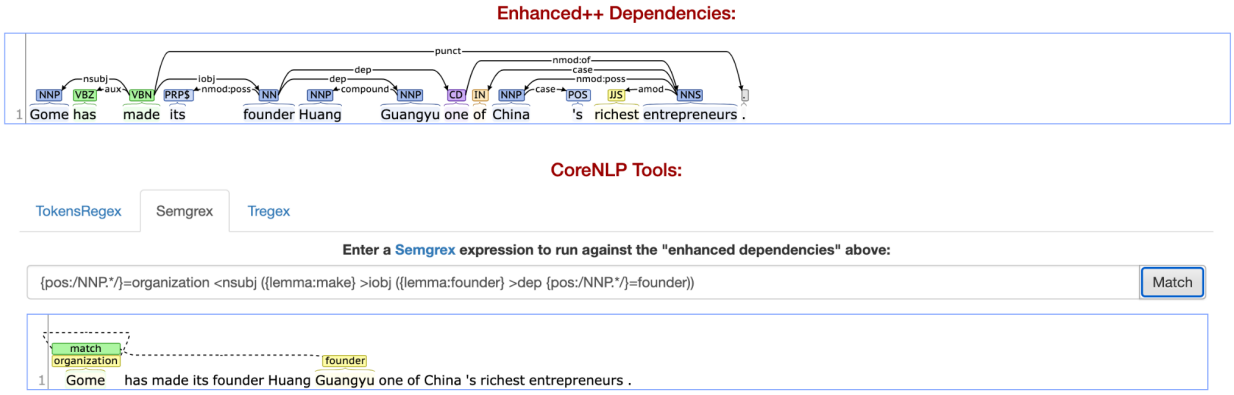


Metalucifer is a Japanese heavy metal band founded by **Gezolucifer** in 1995 .
YTL Corporation is a Malaysian infrastructure conglomerate founded in 1955 by **Tan Sri Dato**.
NetObjects Inc. is a software company founded in 1995 by **Samir Arora, David Kleinberg Clement Mok** and **Sal Arora .**

(If there are multiple founders, you have to extract headword corresponding to each founder)

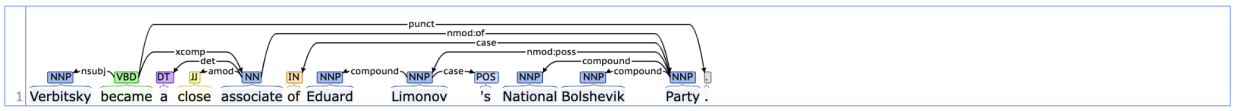


Gome has made its founder **Huang Guangyu** one of China's richest entrepreneurs.

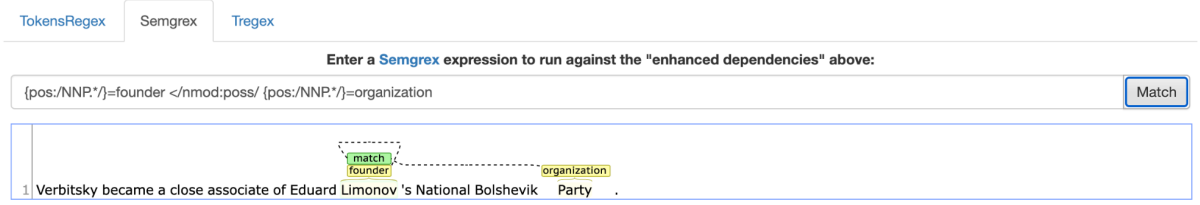


Verbitsky became a close associate of **Eduard Limonov**'s **National Bolshevik Party**.

Enhanced++ Dependencies:

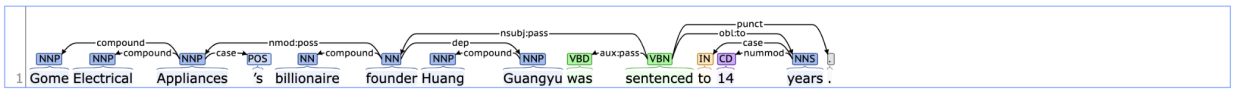


CoreNLP Tools:



Gome Electrical Appliances's billionaire founder **Huang Guangyu** was sentenced to 14 years.

Enhanced++ Dependencies:



CoreNLP Tools: