Problem 2

For this problem set, we will use
https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fcovid19_1

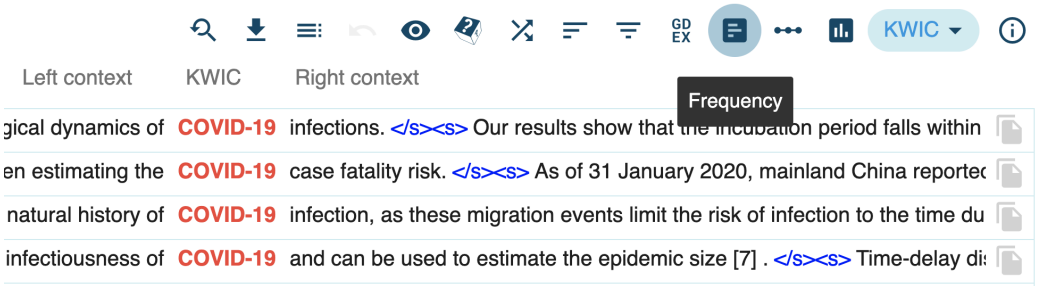
Problem 2.1

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

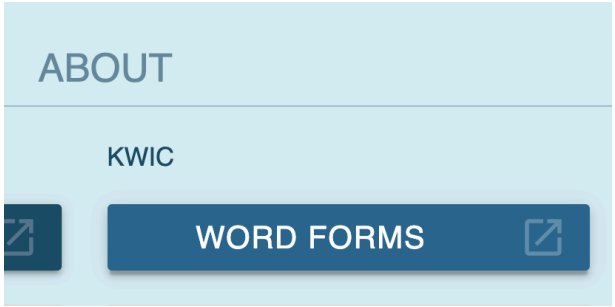
Now write a query to find sentences containing all forms of covid and execute it. Some forms include covid-19, covid19, COVID19, covid-36, covid-54.

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	Relative ?
1	COVID-19	4,062,440	2,263.77
2	COVID	163,675	91.21
3	Covid-19	142,264	79.28
4	COVID19	22,595	12.59
5	covid-19	17,503	9.75

Note: Your list of words and frequencies doesn't have to match exactly, but should be approximately the same. (There is some subjectivity in exactly what strings are considered a form of covid.)

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]-?[0-9]*"]

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

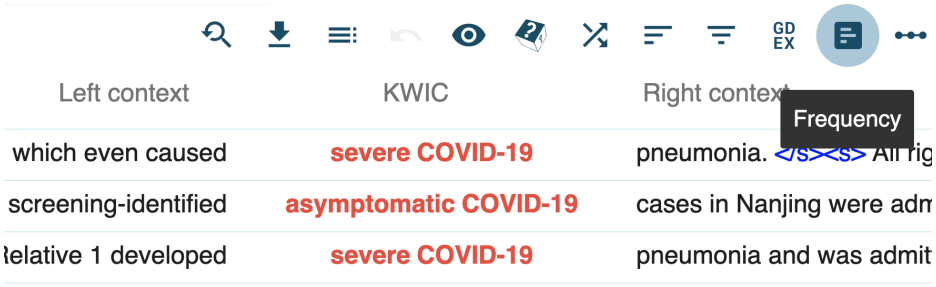
	Word	Frequency	Relative ?
1	COVID-19	4,062,440	2,263.77
2	COVID	163,675	91.21
3	Covid-19	142,264	79.28
4	COVID19	22,595	12.59
5	covid-19	17,503	9.75
6	Covid	13,730	7.65
7	covid	6,349	3.54
8	Covid19	2,396	1.34
9	CoVID-19	2,243	1.25
10	COVID-2019	1,858	1.04
11	CoViD-19	1,743	0.97
12	covid19	1,350	0.75
13	cOVId-19	808	0.45
14	COVID-10	525	0.29
15	COviD-19	522	0.29
16	COVID-9	470	0.26
17	CoVid-19	325	0.18
18	COvid-19	312	0.17
19	coVid-19	285	0.16
20	COViD-19	245	0.14

Problem 2.2

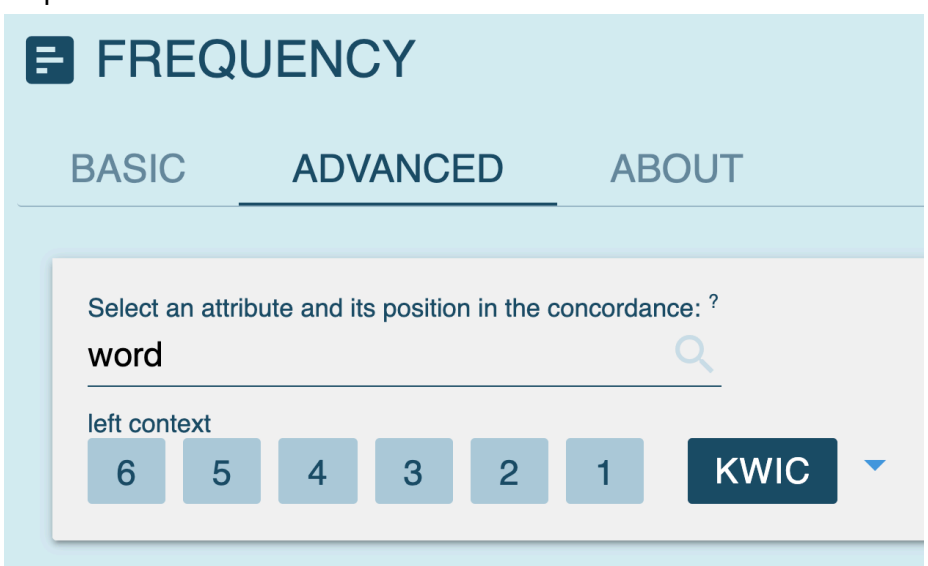
Let's write CQL queries to find interesting words that occur in specific syntactic relations with covid (all forms). We did similar things in class. You will have to use tag and lemma in CQL queries. This tagset could be useful

I will demonstrate how to get the modifiers of covid:

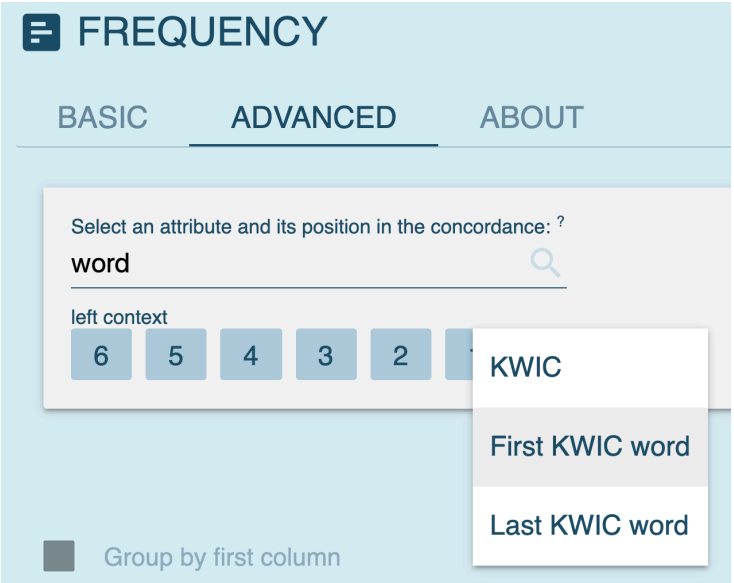
Step 1: First write a CQL query that produces concordance (examples) like this:



Step 2:



Step 3:



Step 4:

	Word	Frequency
1	severe	110,937
2	current	18,238
3	ill	10,926
4	first	10,507

What is the CQL query for modifiers of covid (all forms)?
Answer:
[tag="J.*"][word="[cC][oO][vV][iI][dD]-?[0-9]*"]

Include the snapshot of the 20 most frequent modifiers modifiers (top four are shown above):

(8,845 items, 475,248 total frequency)

	Word	Frequency	Relative ?
1	severe	110,937	61.82
2	current	18,238	10.16
3	ill	10,926	6.09
4	first	10,507	5.85
5	confirmed	10,309	5.74
6	ongoing	9,728	5.42
7	mild	9,252	5.16
8	suspected	9,159	5.10
9	long	9,130	5.09
10	critical	9,046	5.04
11	positive	8,164	4.55
12	acute	8,008	4.46
13	new	6,889	3.84
14	symptomatic	6,771	3.77
15	moderate	5,904	3.29
16	global	5,336	2.97
17	recent	4,752	2.65
18	asymptomatic	4,110	2.29
19	laboratory-confirmed	3,724	2.08
20	potential	3,365	1.88

Note: The modifiers found will depend on what exactly you considered to be a form of covid, and therefore frequency won't necessarily be the exact same numbers as in the example. However, frequencies should be approximately the same (I expect that the top modifier you find is also the word *severe*.)

What is the CQL query of words that are modified by covid (all forms)?

Answer:

[word="[cC][oO][vV][iI][dD]-?[0-9]*"][tag="N.*"]

Include the snapshot of those words

(19,268 items, 2,256,497 total frequency)

	Word	Frequency	Relative ?
1	pandemic	488,655	272.30
2	patients	325,972	181.65
3	infection	142,806	79.58
4	cases	106,538	59.37
5	vaccine	67,565	37.65
6	outbreak	59,394	33.10
7	disease	46,089	25.68
8	vaccination	38,590	21.50
9	vaccines	37,396	20.84
10	pneumonia	36,799	20.51
11	crisis	27,514	15.33
12	epidemic	25,882	14.42
13	symptoms	23,973	13.36
14	infections	20,410	11.37
15	diagnosis	19,844	11.06
16	severity	19,317	10.76
17	lockdown	18,516	10.32
18	mortality	18,505	10.31
19	case	16,502	9.20
20	transmission	15,827	8.82

An example:

	Word	Frequency
1	pandemic	511,065
2	patients	353,157
3	infection	147,198

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 conjuncts/coordinates).

Answer:

[lemma="[cC][oO][vV][iI][dD]-?[0-9]*"][tag="CC"|word=","][tag="N.*"]

Include the snapshot of those words

(12,007 items, 84,818 total frequency)

	Word	Frequency	Relative ?
1	influenza	2,194	1.22
2	SARS-CoV-2	1,855	1.03
3	coronavirus	1,680	0.94
4	patients	1,345	0.75
5	SARS	1,131	0.63
6	COVID-19	933	0.52
7	pneumonia	889	0.50
8	cancer	697	0.39
9	diabetes	688	0.38
10	mortality	649	0.36
11	death	627	0.35
12	non-COVID-19	593	0.33
13	people	572	0.32
14	MIS-C	562	0.31
15	health	522	0.29
16	control	513	0.29
17	ARDS	460	0.26
18	vaccination	401	0.22
19	risk	384	0.21
20	HIV	382	0.21

You are only allowed to access 1,000 items. Get more

Rows per page: 20 1–20 of 1,000 1 / 50

An example:

	Word	Frequency
1	influenza	2,194
2	SARS-CoV-2	1,855
3	coronavirus	1,680

What is the CQL query for verbs that can take covid (all forms) as subject?

Answer:

[lemma="[cC][oO][vV][iI][dD]-?[0-9]*" & tag="N.*"][]{0,2}[tag="V.*" & !lemma="be|have"]

Include the snapshot of verbs that take covid as subject

(12,816 items, 1,058,393 total frequency)

	Word	Frequency	Relative ?
1	compared	23,420	13.05
2	reported	21,238	11.83
3	associated	20,753	11.56
4	caused	19,314	10.76
5	including	16,792	9.36
6	using	16,493	9.19
7	confirmed	13,434	7.49
8	based	12,963	7.22
9	admitted	10,629	5.92
10	found	10,544	5.88
11	increased	10,244	5.71
12	showed	10,194	5.68
13	affected	8,906	4.96
14	related	8,311	4.63
15	included	8,217	4.58
16	led	7,788	4.34
17	remains	7,244	4.04
18	spread	7,009	3.91
19	did	6,928	3.86
20	include	6,928	3.86

You are only allowed to access 1,000 items. Get more

Rows per page: 20 1-20 of 1,000

An example:

	Word	Frequency
1	compared	23,820
2	reported	21,448
3	associated	21,206
4	caused	19,684

What is the CQL query for verbs that can take covid (all forms) as object?

Answer:

[tag="V.*" & !lemma="be|have"][]{0,3}[lemma="[cC][oO][vV][iI][dD]-?[0-9]*" & tag="N.*"]

Include the snapshot of verbs that take COVID as object.

(13,553 items, 1,518,175 total frequency)

	Word	Frequency	Relative ?
1	confirmed	43,239	24.09
2	associated	41,392	23.07
3	related	35,643	19.86
4	hospitalized	31,284	17.43
5	reported	23,724	13.22
6	diagnosed	22,943	12.78
7	infected	19,897	11.09
8	increased	19,601	10.92
9	affected	19,384	10.80
10	regarding	16,460	9.17
11	tested	16,050	8.94
12	used	15,107	8.42
13	caused	15,105	8.42
14	following	13,754	7.66
15	prevent	12,864	7.17
16	treat	11,860	6.61
17	including	11,569	6.45
18	contracting	11,551	6.44
19	observed	10,569	5.89
20	suspected	10,513	5.86

An example:

	Word	Frequency
1	confirmed	45,979
2	associated	38,051
3	hospitalized	35,389

Problem 2.3

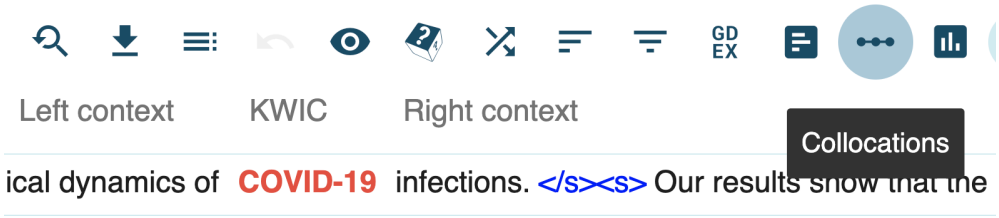
What are the most important words that form collocations with COVID (where covid is the right word)?

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:

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

Sorted by MI

	Word	Cooccurrences ?	Candidates ?	T-score	MI ↓	LogDice
1	U-TOPTM	5	5	2.23	8.66	-4.76
2	30,901	5	5	2.23	8.66	-4.76
3	SGTI-flex	16	16	3.99	8.66	-3.08
4	vACCine	6	6	2.44	8.66	-4.50
5	unittreated	5	5	2.23	8.66	-4.76
6	symptom-predicted	7	7	2.64	8.66	-4.28
7	InfectCheck	5	5	2.23	8.66	-4.76
8	London-YouGov	7	7	2.64	8.66	-4.28
9	AFIASTM	6	6	2.44	8.66	-4.50
10	respiratory-confirmed	6	6	2.44	8.66	-4.50

	Word	Cooccurrences ?	Candidates ?	T-score	MI ↓	LogDice
11	adsorvida	5	5	2.23	8.66	-4.76
12	Terminos	7	7	2.64	8.66	-4.28
13	UKILD-Long	11	11	3.31	8.66	-3.62
14	Di20K	23	23	4.78	8.66	-2.56
15	ActualMed	16	16	3.99	8.66	-3.08
16	SMAtteo	8	8	2.82	8.66	-4.08
17	395,926	5	5	2.23	8.66	-4.76
18	Posthospitalisation	6	6	2.44	8.66	-4.50
19	#Long	5	5	2.23	8.66	-4.76
20	GTPP	9	9	2.99	8.66	-3.91

Problem 3:

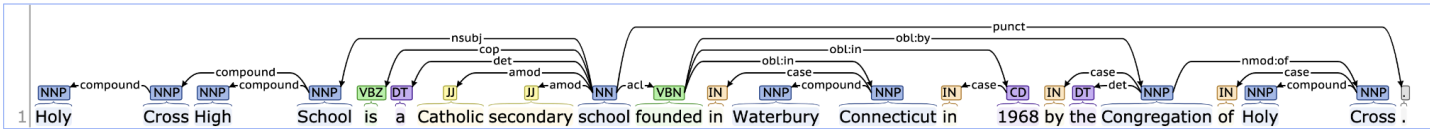
Write SemGrex regular expressions that can detect organizations and their founders. Make use of https://corenlp.run to parse sentences to syntactic graphs and for running SemGrex expressions.

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:

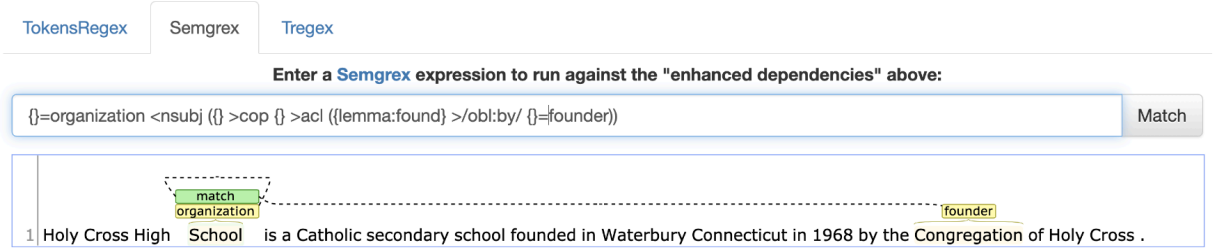
Enhanced++ Dependencies:



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

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

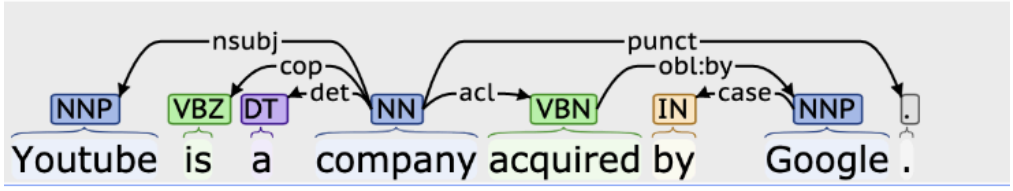
CoreNLP Tools:

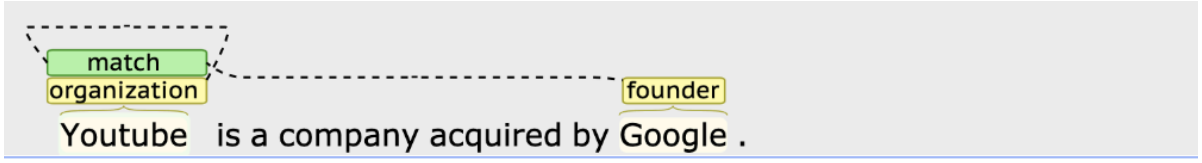


This pattern can be read as the “organization” that is a subject of something, and this something is founded by the founder.

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

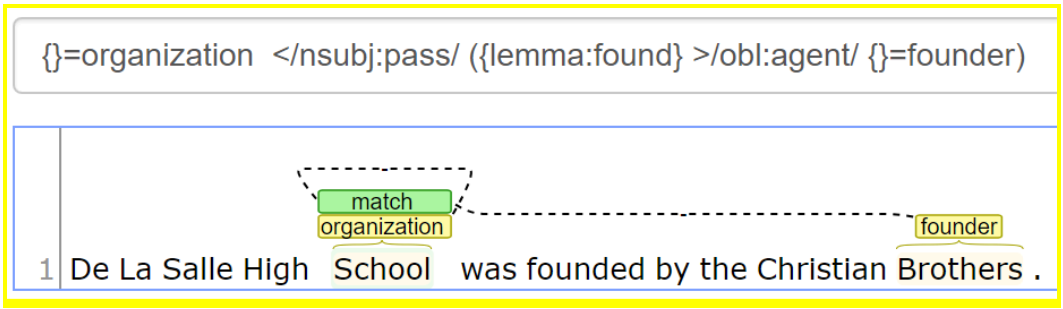
Write the SemGrex patterns for the following sentences that extract the organization name (headword is enough) and its founder (headword is enough).

Include a snapshot of each SemGrex expression that you write (containing Enhanced++ Dependencies, Semgrex expression, and the matchings).

Sentences that can make use of the same expression should be in the same snapshot.

Sentences:

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



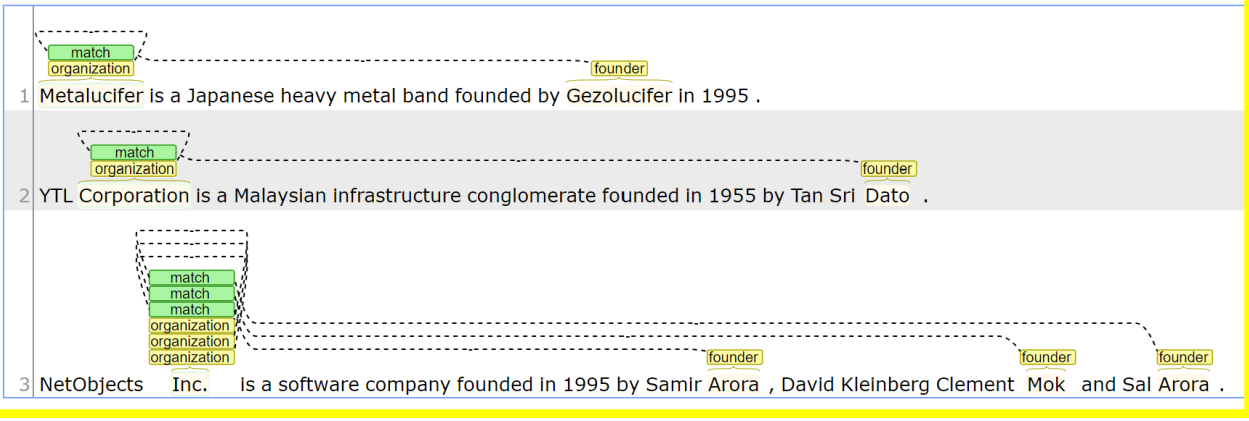
Metalucifer is a Japanese heavy metal band founded by **Gezolucifer** in 1995 .

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 .**

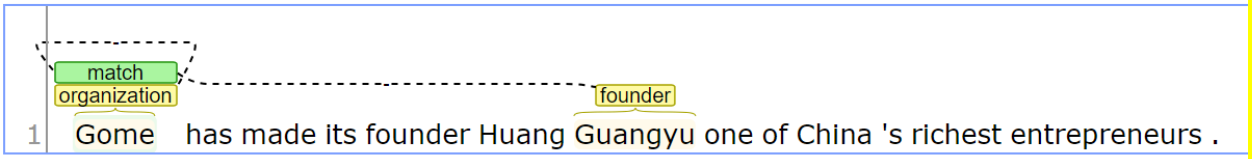
(If there are multiple founders, you have to extract headword corresponding to each founder)

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



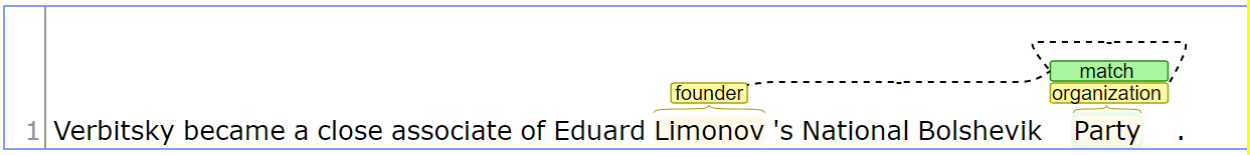
Gome has made its founder **Huang Guangyu** one of China's richest entrepreneurs.

```
{}=organization <nsubj ({} >iobj ({} >dep {pos:/NNP/}=founder))
```



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

```
{}=organization >/nmod:poss/ {}=founder
```



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

```
{}=organization </nmod:poss/ ({} >dep {}=founder)
```