STSM report: MWEs in corpus query systems
COST-STSM-IC1207-33762

Participant: Natalia Klyueva
Institute of formal and Applied Linguistics
Charles University in Prague
E-mail: kljueva@ufal.mff.cuni.cz

Host: Shuly Wintner
Department of Computer Science
University of Haifa
E-mail: shuly@cs.haifa.ac.il

During my three week stay at University of Haifa, I collaborated with Shuly Wintner and his colleagues in several directions. First, I got acquainted with the HPSG grammar developed in Haifa, and how multiword expressions are represented within this formalisms. Secondly, we explored the corpus query system in which one can query for different types of MWEs. Unfortunately, the latter system was not properly maintained. Therefore, we concentrated on the main purpose of the visit - exploitation of query system SketchEngine in the task of identifying MWEs in annotated corpora.

1 Description of the work carried out during the STSM

1.1 MWE identification within Sketch Engine

The main goal was to create the set of rules (queries) over the syntactically parsed corpus and to extract the most frequent mwes. The work was inspired and can be viewed as continuation of another STSM - by Noam Ordan[1]. While in the latter research corpora prepared by SketchEngine annotated only on the morphological layer (enTenTen and heTenTen) were exploited, in the present project the processing up to syntactic level and compilation of the corpora was done.

In the first phase, I parsed both sides of parallel Hebrew-English corpus Haaretz with the UDpipe[2]. Then, I separately compiled the two parsed corpora within Sketch Engine, with the attributes as they are defined in the conllu format. Though

[1] http://typo.uni-konstanz.de/parseme/images/STSM-reports/Ordan_report_STSM.pdf

[2] https://ufal.mff.cuni.cz/udpipe

initially we planned to exploit the parallel data, it came out that the original segmentation from Haaretz does not fit the segmentation done by the parser, which made problems especially for m:n aligned sentences. Parallelising the corpus was not planned in the initial proposal, so we left it for the future cooperation and concentrated on MWE identification.

1.2 Rules for Sketch Engine (SkE)

During Noam's mission, verbal MWEs in the corpus were identified using the rules written in a form of regular expressions.

I improved the rules made over Hebrew corpus and adjusted them so to make use of syntactic attributes. This approach allows to use simple queries without the need of word sketches, because the query can involve such syntactic relations as root (predicate) or dobj (direct object).

Following is the query to extract verb-noun collocations:

```
[deprel="root"][deprel="DET"]{0,1}[deprel="dobj" & pos="NOUN"]
```

As for the noun-noun collocations, the query for the morphological tag without exploiting syntactic queries was enough:

```
[penn_tag="NN"][penn_tag="NN"]
```

Interesting to note, that almost all of the noun-noun collocations from the list had dependency relation "compound".
In the Appendix section, I included the frequency lists for the queries above for English. We also had the idea to compare the candidates from the lists with those generated by the mwetoolkit.

2 Dependency relation 'mwe' from Universal Dependencies

Another part of the project concerned extracting MWEs in parsed parallel corpus based on the dependency relation attribute. The dependency relation (deprel) 'mwe' in the Universal Dependencies (UD) [3] is related mostly to functional mwes, e.g. multiword prepositions. Again I exploited the parallel corpus Haaretz parsed with UDpipe, and concatenated the nodes marked with the relation **mwe** into one, e.g. as well as -> as+well+as. The idea was then to run GIZA++ and to exploit the cases of differences, e.g. when in English the mwe is a concatenated token whereas in Hebrew not, and the other way around. This would also reveal the cases of differences in annotation in UD, which can be applied to other treebanks as well. However, the experiment was not finished, because of the problem with parallelising sentences after parsing that was described above. We plan to make the parsed corpus parallel in the nearest future.

[3] http://universaldependencies.org/u/dep/mwe.html

3 Description of the main results obtained

The main results are the scripts to process the corpora and the queries to generate candidate lists of MWEs with the corpus query system. Some scripts and description of the project can be found here: `https://github.com/natalink/udpipe_mwe`.

4 Confirmation by the host institution of the successful execution of the STSM

Natalia Klyueva visited my lab in Haifa in July. I confirm that the visit was completed to my satisfaction. Not all our goals were successfully met, but I am looking forward to continuing my collaboration with Natalia in the near future, and I am positive that more results are imminent. We will update you on any joint publications stemming from this work, of course.

	lemma	Frequency
P | N	take place	22
P | N	take part	8
P | N	have something	8
P | N	have nothing	8
P | N	have trouble	7
P | N	do everything	7
P | N	want peace	6
P | N	make peace	6
P | N	wag war	4
P | N	have time	4
P | N	do nothing	4
P | N	arouse concern	4
P | N	hold talk	3
P | N	have difficulty	3
P | N	give rise	3
P | N	transfer service	2
P | N	take shape	2
P | N	take responsibility	2
P | N	take office	2
P | N	take control	2
P | N	take care	2
P | N	take advantage	2
P | N	stop shoo	2
P | N	stop free	2
P | N	stop fighting	2
P | N	stop construction	2
P | N	set precondition	2
P | N	seek peace	2
P | N	see thing	2
P | N	see photograph	2
P | N	save money	2
P | N	remain mum	2
P | N	read book	2
P | N	raise tax	2
P | N	play poker	2
P | N	play hardball	2
P | N	need space	2
P | N	make mistake	2
P | N	make effort	2
P | N	make do	2
P | N	make decision	2
P | N	know nothing	2
P | N	know everything	2
P | N	impose sanction	2
P | N	hold water	2
P | N	have value	2
P | N	have plenty	2

Figure 1: Verb-noun frequency list

	word	Frequency
P | N	peace process	120
P | N	defense minister	96
P | N	peace agreement	70
P | N	state solution	65
P | N	finance minister	50
P | N	settlement construction	45
P | N	road map	43
P | N	construction freeze	42
P | N	justice minister	36
P | N	settlement freeze	34
P | N	separation fence	28
P | N	status quo	27
P | N	attorney general	27
P | N	defense establishment	25
P | N	peace initiative	24
P | N	education system	23
P | N	defense budget	22
P | N	peace plan	21
P | N	Goldstone report	21
P | N	wing government	19
P | N	status agreement	18
P | N	unity government	17
P | N	state prosecutor	17
P | N	security service	17
P | N	law enforcement	17
P | N	peace treaty	16
P | N	peace camp	16
P | N	education minister	16
P | N	air force	16
P | N	negotiating table	14
P | N	labor federation	13
P | N	election campaign	13
P | N	retirement age	12
P | N	prisoner exchange	12
P | N	turpitude law	11
P | N	state commission	11
P | N	budget deficit	11
P | N	state education	10
P | N	security situation	10
P | N	refugee problem	10
P | N	opposition leader	10
P | N	justice system	10
P | N	home front	10
P | N	coalition agreement	10
P | N	state comptroller	9
P | N	price tag	9
P | N	income tax	9

Figure 2: Noun-noun frequency list