Extraction of potential multi-word expressions in a parallel corpus of their translations from a multilingual corpus

WG 2: Parsing Techniques for MWEs

Katerina Zdravkova, University Sts Cyril and Methodius, Skopje
Aleksandar Petrovski, International Slavic University, Sveti Nikole

Basic setting

Orwell's 1984 parallel corpora
Resources of 16 languages belonging to Multext-East V.4. (courtesy of Institute Jozef Stefan, Ljubljana)

Orwl-mk.txt

Orwl-mk.xml

Main module

```
import sys
sys.path[:0] = ["../"]

from text.output import writeText
from text.output import writePhrasesXML
from text.output import writePhrasesCount
from text.input import readOrwellMkXML
from text.core import findPhrases
from text.core import trimCountLower
from text.core import trimSubPhrases

text =  readOrwellMkXML('Omk.xml')
phrases = findPhrases(text)
phrases = trimCountLower(phrases, 2)
phrases = trimSubPhrases(phrases)

writeText(text, 'text.txt')
writePhrasesCount(phrases, 'phrases_count.txt')
writePhrasesXML(phrases, 'phrases.xml')
```

The core part

```
from collections import defaultdict

def findPhrases(sentences):

    phrases = defaultdict(list)

    for sentence in sentences:
        words = sentence.text

        for i in xrange(2, len(words)):

            for x in xrange(0,len(words)-i):
                s=words[x]
                for y in xrange(x+1, x+i+1):
                    if(y<len(words)):
                        s+=" "+words[y]

                phrases[s].append(sentence)

    return phrases

def trimCountLower(phrases, count):

    keys = phrases.keys()

    for phrase in keys:
        if len(phrases[phrase]) < count:
            del phrases[phrase]

    return phrases
```

Separation of all the repeated blocks of words

15463 blocks of words (potential MWEs)
They appeared in total 53246 times
Average length: 3.44 words / block

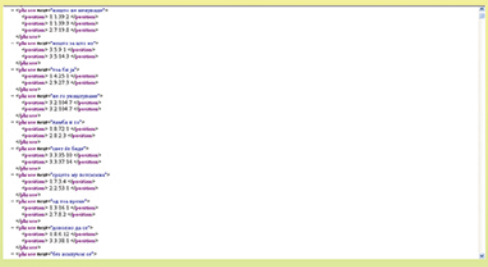
Generated smaller blocks

length of the block	frequency	generated smaller blocks
39 words	2	741
24 words	2	276
20 words	2	190
17 words	2	136
12 words	2	66
...	...	...

Removing of the blocks generated by longer blocks

4690 blocks of words (potential MWEs)
They appeared in total 7621 times
Average length: 2.07 words / block

Part of extracted MWEs



Frequency of extracted MWEs

Filtering of nominal phrases

Adj N (безнадежна љубов - beznadezhna ljubov = hopeless love)
Adj Adj Noun (друго човечко суштество - drugo chovechko sushtestvo = another human being)
Adj N N (мал број луѓе - mal broj lugje = very few people)
N N (безумие безумие - bezumie bezumie = folly folly)
Adj N Prep N (будното око на полицијата - budnoto oko na policijata = under the eyes of the police)
Adj N Adv (неколку минути подоцна - nekolku minuti podocna = after only a few minutes' delay)
Adj N Prep Adj N (дневната заповед на големиот брат - dnevnata zapoved na golemiot brat = big brother's order for the day)
Adv N Adv N (повеќе храна повеќе облека - povekje hrana povekje obleka = more food more clothes)

Blocks that contain a verb