# Evaluation of the system for extraction of multi-word expressions and prediction of their translations

## Katerina Zdravkova

University Sts Cyril and Methodius, Skopje katerina.zdravkova@finki.ukim.mk

## Aleksandar Petrovski

International Slavic University, Sveti Nikole a.petrovski.sise@gmail.com

## Four complementary phases:

- extraction of all candidate groups of words that appear in each language at least twice,
- syntactical filtering of obtained candidates, using a predefined set of eligible syntactic expressions,
- prediction of potential translation equivalents from corresponding pairs of aligned sentences
- cross-evaluation of candidate translations, interchanging the source and the target language.

#### Extraction

3500 candidate MWEs, including many useless:

тоа би ја / toa bi ја, instead of тоа би ја усреќило / toa bi ја usrekjilo = that will make her happy рече тој со / reche toj so, instead of рече тој со недоверба / reche toj so nedoverba = he said with mistrust;

# Syntactical filtering

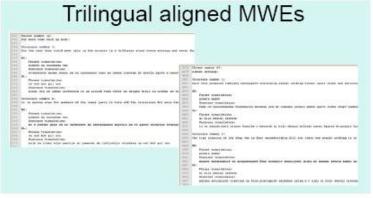
- Less than 500 phrases, sometimes inflections of the same phrase:

атомската бомба / atomskata bomb = the atomic bomb, атомски бомби / atomski bombi = atomic bombs обичен човек / obichen chovek = an ordinary man, обичните луѓе / obichnite lugje = the ordinary men or the ordinary people, шаховска табла / shahovska tabla = a chess board, шаховската табла / shahovskata tabla = the chess board

#### Translation with cross evaluation

- Candidate MWEs obtained without syntactical filtering
  968 English candidate MWEs and their translations
- Parallel aligned MWEs
  Initially: English to Macedonian

Extension: English to Macedonian and Slovene



# Aligned candidate multiword expressions

mannyora	CAPICOSIONS
*** The state of the retain of the state of	The second control of the control of

# Consistency and completeness of MWEs

$$HHI = \sum_{i=1}^{n} S_i^2 \qquad RC = \frac{HHI}{|MWE|}$$

$$DG = \frac{length(generated\ MWE)}{length(complete\ MWE)}$$

$$TC = \frac{1}{m} \sum_{j=1}^{m} S_j^2 DG_j^2$$

## Conclusions and further work

- Incorporation of MWE lexical entries
- Extension with semi-fixed and flexible MWEs
- Study the lexical cohesion, and extend the

document-level translation to a larger collection

- Integration of the model into a hierarchical phrase-based SMT system
- Extraction of the common knowledge about MWEs out of a continuous context and its incorporation into a translation system capable to competently deal with them

# Crucial problems

- Manual translation was either inconsistent or had "an artistic freedom"
- Inflectional paradigms, which are richer in the Slavic languages can influence the translation
- The context in which the same target MWE can influence its translation

Language	English	Mecedonian	Storene
Multiword expression	the seconds were ticking by	опсунурате петрана (отнукува/вк)	setundo se titrakali mino
Mac 1: октунда	пе монувая сторужаўю	14	-
Mac 2: cerryrus	пе минуван београриода	2564	
Multiword expression	atmost on a level with	ренси на исто (се)	no translation
Mic 1 _ pires	м на ното живо со		
Mac 2: payer	и на исто ранинале со		
Slov 1: skors	na navni z		
Slov 2: slore	y lati vistoriz	-0.5 t +0 v	7 677
Multiword expression	the first thing	(урва рабитацито норя) да ја офотици	peva sover kilje reoraš
Eng 1: the first	fring for you to understan	d	
Eng 2: the South	tring you must replace		

Language	English	Macedonian	Slovene
Multiword expression	smelt of her hair	(мирисот) на нејзината коса	vonj njenih ias
Mac 1: (npvj	атниот) вырыс на нејзы	натакоов	
Мас 2: мириос	ут на нејзината госа	1/1	VH S
Multiword expression	ideologically neutral	идеопоции неутрален	ideolotko nevtralen/na
Slav 1: (nobs	ena beseda ni bila) id	eološko nevtralna	
Slov 2: (pred	met govora ni bil) idedk	ško nestralen	A 2000 -01 3
Multiword expression	against us	протие нас-	po robu (proti nam)
Stov 1: in	kdar ne) postavi po robi		

Language	English	Macedonian	Slovene
Multiword expression	for more than half an hour	(за) повеќе од половина час	za več kot pol ure
	ever for more than half a prais notesia og nonces		
	off the telescreen for m драмг исклучан телекр	ore than half an hour deat notes og noncess s	se .
Multiword expression	definitive edition	дефинитивното издание	no translation (dokončna izdaja)
	leventh edition is the j de sta totaja je) dokončna	finitive addisn	
	ere producing a) definiti avljaš smo) končno izdaj		
Multiword expression	(in) the rewrite squad	во одделот за препишување	no translation (prepisovalna ekipa)
Eng 1: (down	apuserre pafora so opp i to final touching-up by) o končne obdelave) prep	the remite squad	
Slov 2: (nikd	rau we particine so ogga- ar nisem bita v) prepisov over in) the rewite squar	alniekipi	