



# A Comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian

Mihael Arcan

Insight Center for Data Analytics, NUI Galway, Ireland

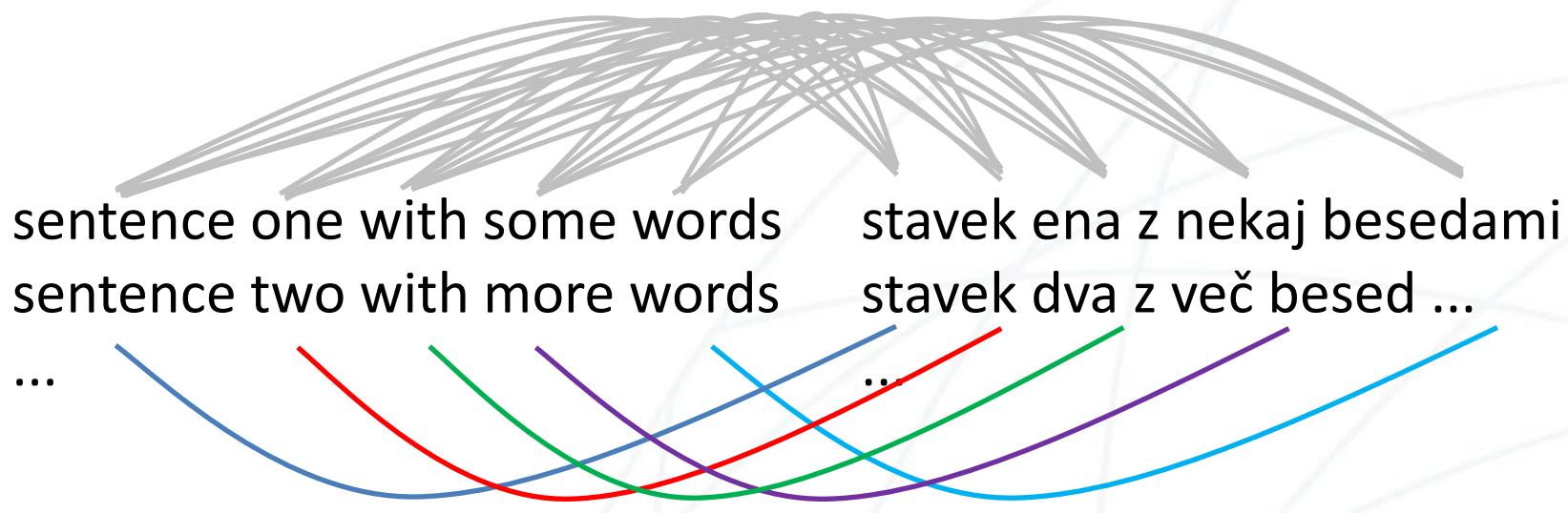
# A Comparison of Statistical ...

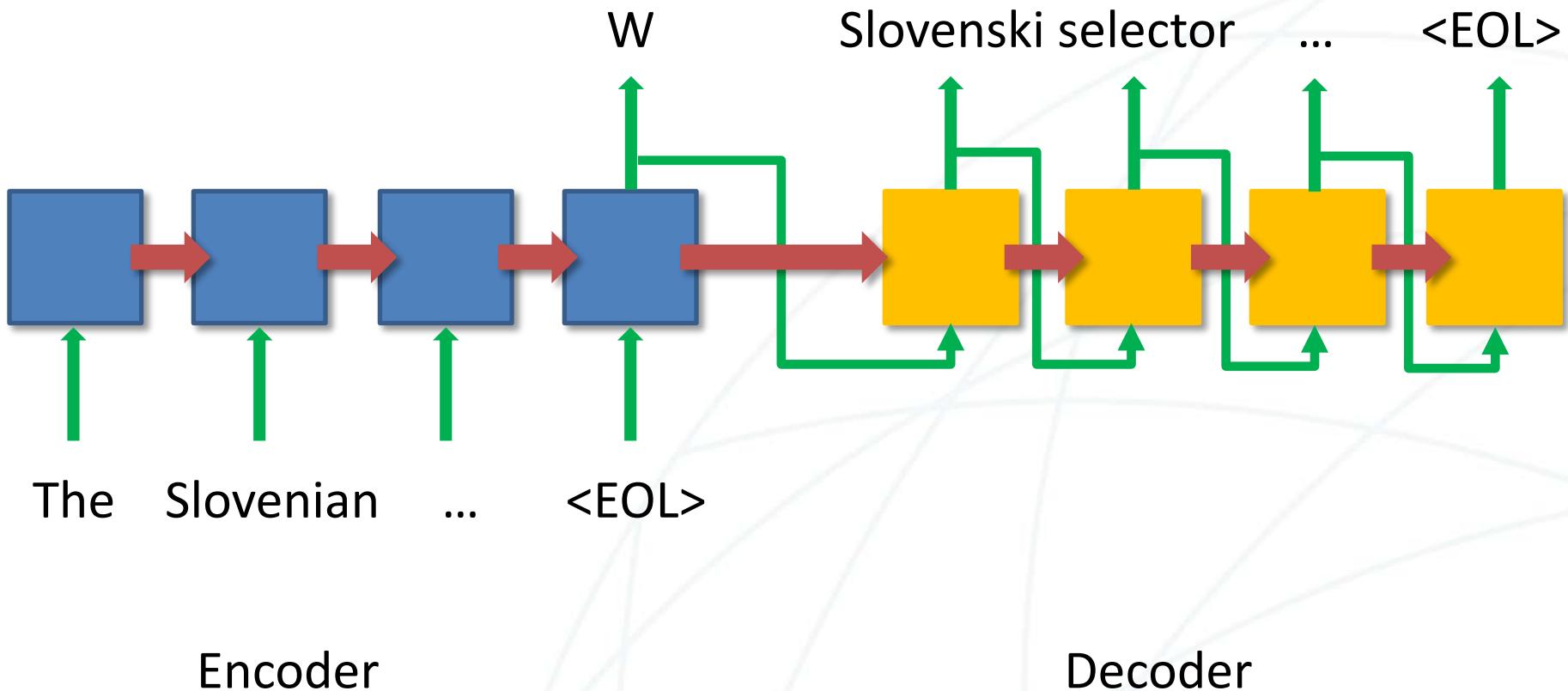
sentence one with some words  
sentence two with more words

...

stavek ena z nekaj besedami  
stavek dva z več besed ...

# A Comparison of Statistical ...





# Word Embeddings



slovenian 0.157417 0.013155 -0.168669 0.258004 -0.261520 -0.004266 0.064963 -0.095019 0.110000  
0.139827 -0.009515 0.102139 -0.014292 0.064913 0.116374 -0.248112 -0.179980 0.143360 -0.058822 -  
0.081025 -0.204884 0.024653 -0.008157 -0.208811 -0.351810 -0.020040 -0.191807 -0.004969 0.148161  
-0.124785 0.073630 -0.063658 -0.325918 0.170596 0.123046 0.106464 -0.141023 0.037222 0.095813  
0.349635 0.166654 -0.160324 -0.036104 0.016911 0.002022 0.026284 -0.129991 -0.061656 0.075007 -  
0.200100 -0.026334 -0.169233 0.033682 0.031311 0.034157 -0.056466 -0.010009 0.373934 -0.461444  
0.137428 -0.080984 -0.214958 0.133438 -0.206060 -0.044642 -0.020855 0.001350 -0.032700 0.054392  
-0.138733 0.271986 0.074465 -0.383632 -0.060489 0.170901 -0.007525 -0.213951 -0.215574 -0.237169  
0.047382 0.059949 -0.167393 -0.086040 0.037600 -0.064871 0.099520 -0.137147 0.170327 -0.044044  
0.238244 0.082285 -0.030852 0.061542 -0.099835 -0.154609 -0.162004 0.061578 -0.172150 -0.163639  
-0.012122 -0.085171 0.197971 0.483274 0.020190 0.203729 0.267945 -0.159703 -0.352687 0.009871  
0.032857 0.076368 0.090400 0.143115 0.161378 -0.203485 -0.020512 -0.139726 0.174444 -0.035236  
0.069389 0.157347 -0.254437 0.016946 -0.190948 -0.186738 0.060985 0.376371 -0.084888 -0.330832 -  
0.032036 0.072837 -0.381461 0.088491 -0.315386 0.119776 -0.174684 0.337164 0.031623 0.036908  
0.297419 -0.208412 0.316575 0.036308 0.119306 -0.025103 0.110816 -0.072111 -0.197073 0.154872  
0.285448 0.063191 -0.018339 -0.117289 -0.089172 0.033924 -0.297626 -0.094350 -0.106417 -0.283593  
-0.126537 0.131320 -0.087875 -0.044319 -0.143214 -0.086091 0.147566 0.043141 -0.121772 -0.123281  
-0.044752 0.036134 -0.273063 -0.046211 0.014122 -0.001258 0.216371 0.027509 -0.117638 0.249159  
0.137707 0.087296 0.059431 0.029131 -0.079217 0.007073 0.006122 -0.015539 0.112570 0.250844 -  
0.012353 -0.162030 -0.107101 0.072732 -0.159938 0.064535 0.069997 0.155462 -0.015620 0.390249  
0.096999

# ... Machine Translation for Slovene, Serbian and Croatian

## META-NET Language White Paper Series (2012)

### Machine Translation

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	<ul style="list-style-type: none"><li>▪ English</li></ul>	<ul style="list-style-type: none"><li>▪ French</li><li>▪ Spanish</li></ul>	<ul style="list-style-type: none"><li>▪ Catalan</li><li>▪ Dutch</li><li>▪ German</li><li>▪ Hungarian</li><li>▪ Italian</li><li>▪ Polish</li><li>▪ Romanian</li></ul>	<ul style="list-style-type: none"><li>▪ Basque</li><li>▪ Bulgarian</li><li>▪ Croatian</li><li>▪ Czech</li><li>▪ Danish</li><li>▪ Estonian</li><li>▪ Finnish</li><li>▪ Galician</li><li>▪ Greek</li><li>▪ Icelandic</li><li>▪ Irish</li><li>▪ Latvian</li><li>▪ Lithuanian</li><li>▪ Maltese</li><li>▪ Norwegian (Bokmål, Nynorsk)</li><li>▪ Portuguese</li><li>▪ Serbian</li><li>▪ Slovak</li><li>▪ Slovene</li><li>▪ Swedish</li><li>▪ Welsh</li></ul>

# Relevant Sentence Selection



each hard capsule contains 182.2 mg of anhydrous lactose .

each hard capsule contains 182.2 mg of anhydrous lactose

infections at the injection site ( including abscess and cellulitis ) occurred in 1.5 % of patients .

# Relevant Sentence Selection



each hard capsule contains 182.2 mg of anhydrous lactose



each hard capsule contains 182.2 mg of anhydrous lactose



infections at the injection site ( including abscess and cellulitis ) occurred in 1.5 % of patients .



# Byte Pair Encoding (BPE)



Sentence	Tokenised Sentence	BPE Segmented Sentence
procedures for proper handling and disposal of antineoplastic medicinal products should be used.	procedures for proper handling and disposal of antineoplastic medicinal products should be used .	procedures for proper handling and disposal of ant■ ine■ op■ lastic medicinal products should be used ■ .
many thanks to mrs van lancker, mr berman, mr lambsdorff, mr hutchinson, mrs scheele, mrs doyle, mrs weber, mr varvitsiotis, mrs hassi and mrs gomes.	many thanks to mrs van lancker , mr berman , mr lambsdorff , mr hutchinson , mrs scheele , mrs doyle , mrs weber , mr varvitsiotis , mrs hassi and mrs gomes .	many thanks to mrs van lanc ■ ker ■ , mr berman ■ , mr lamb■ s■ dor■ ff ■ , mr hut■ chin■ son ■ , mrs sche■ ele ■ , mrs doyle ■ , mrs weber ■ , mr var■ vit■ si■ otis ■ , mrs hass■ i and mrs g■ omes ■ .
kdo vam daje pravico, da invalidom odrekate neomejen dostop do izobraževanja, ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah?	kdo vam daje pravico , da invalidom odrekate neomejen dostop do izobraževanja , ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah ?	kdo vam daje pravico ■ , da invali■ dom odre■ kate ne■ omejen dostop do izo■ braževanja ■ , ali da stare■ jšim ljudem odre■ kate enako obravnavo pri zavarov■ anjih in finančnih storitvah ■ ?

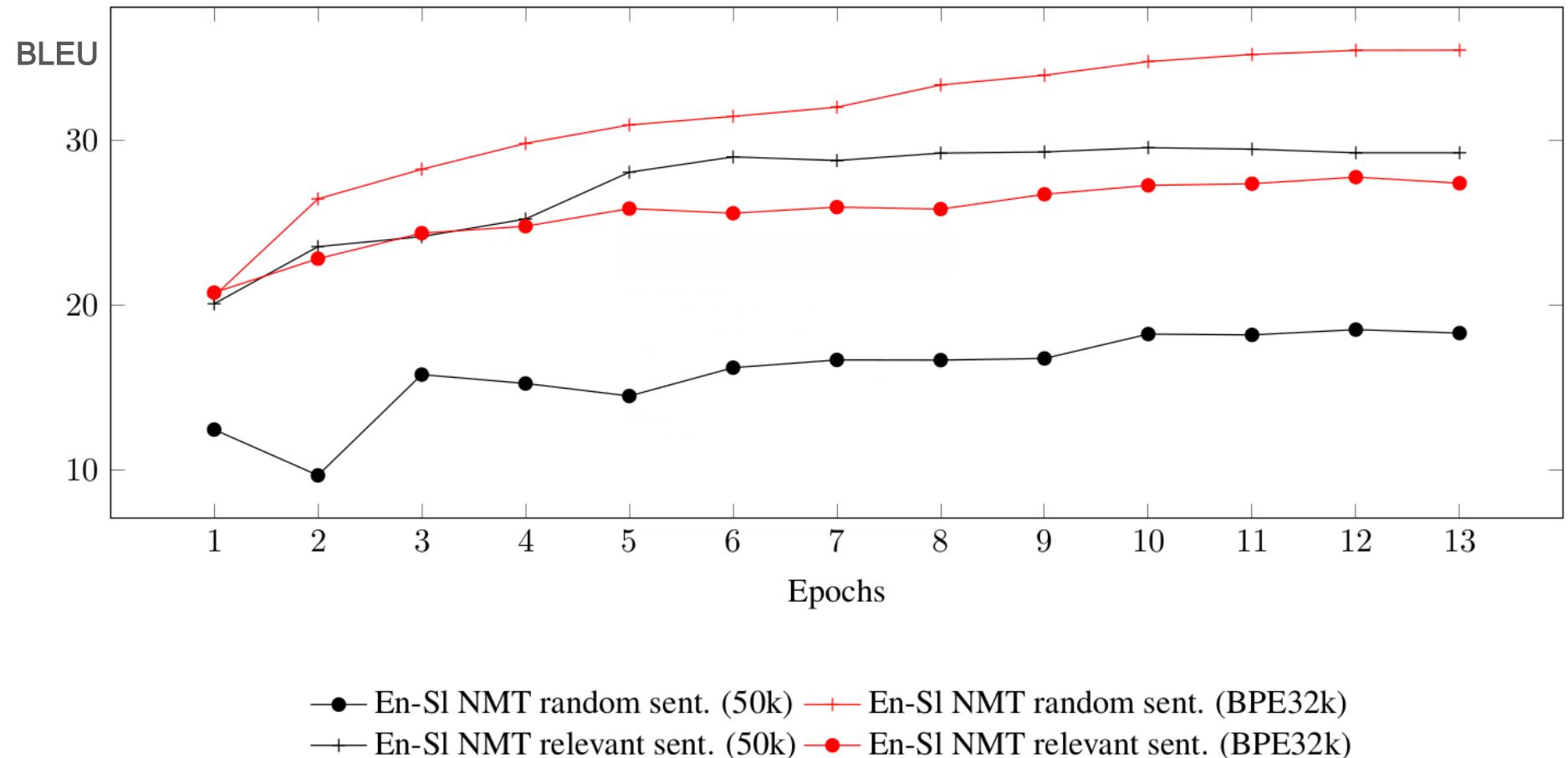
# Byte Pair Encoding (BPE)

Sentence	Tokenised Sentence	BPE Segmented Sentence
procedures for proper handling and disposal of antineoplastic medicinal products should be used.	procedures for proper handling and disposal of antineoplastic medicinal products should be used .	procedures for proper handling and disposal of ant ■ ine ■ op ■ lastic medicinal products should be used ■ .
many thanks to mrs van lancker, mr berman, mr lambsdorff, mr hutchinson, mrs scheele, mrs doyle, mrs weber, mr varvitsiotis, mrs hassi and mrs gomes.	many thanks to mrs van lancker , mr berman , mr lambsdorff , mr hutchinson , mrs scheele , mrs doyle , mrs weber , mr varvitsiotis , mrs hassi and mrs gomes .	many thanks to mrs van lanc ■ ker ■ , mr berman ■ , mr lamb ■ s ■ dor ■ ff ■ , mr hut ■ chin ■ son ■ , mrs sche ■ ele ■ , mrs doyle ■ , mrs weber ■ , mr var ■ vit ■ si ■ otis ■ , mrs hass ■ i and mrs g ■ omes ■ .
kdo vam daje pravico, da invalidom odrekate neomejen dostop do izobraževanja, ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah?	kdo vam daje pravico , da invalidom odrekate neomejen dostop do izobraževanja , ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah ?	kdo vam daje pravico ■ , da invali ■ dom odre ■ kate ne ■ omejen dostop do izo ■ braževanja ■ , ali da stare ■ jšim ljudem odre ■ kate enako obravnavo pri zavarov ■ anjih in finančnih storitvah ■ ?

# Byte Pair Encoding (BPE)

Sentence	Tokenised Sentence	BPE Segmented Sentence
procedures for proper handling and disposal of antineoplastic medicinal products should be used.	procedures for proper handling and disposal of antineoplastic medicinal products should be used .	procedures for proper handling and disposal of ant ■ ine ■ op ■ lastic medicinal products should be used ■ .
many thanks to mrs van lancker, mr berman, mr lambsdorff, mr hutchinson, mrs scheele, mrs doyle, mrs weber, mr varvitsiotis, mrs hassi and mrs gomes.	many thanks to mrs van lancker , mr berman , mr lambsdorff , mr hutchinson , mrs scheele , mrs doyle , mrs weber , mr varvitsiotis , mrs hassi and mrs gomes .	many thanks to mrs van lanc ■ ker ■ , mr berman ■ , mr lamb ■ s ■ dor ■ ff ■ , mr hut ■ chin ■ son ■ , mrs sche ■ ele ■ , mrs doyle ■ , mrs weber ■ , mr var ■ vit ■ si ■ otis ■ , mrs hass ■ i and mrs g ■ omes ■ .
kdo vam daje pravico, da invalidom odrekate neomejen dostop do izobraževanja, ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah?	kdo vam daje pravico , da invalidom odrekate neomejen dostop do izobraževanja , ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah ?	kdo vam daje pravico ■ , da invali ■ dom odre ■ kate ne ■ omejen dostop do izo ■ braževanja ■ , ali da stare ■ jšim ljudem odre ■ kate enako obravnavo pri zavarov ■ anjih in finančnih storitvah ■ ?

# Initial Experiments



# Experimental Setup



## SMT

Moses (phrase based alignment)  
Giza++ (word alignment)  
KenLM (language model)

## NMT

OpenNMT  
(500/600/1000/  
dimensions, 2/4  
layers, bidirectional  
LSTM)

## Evaluation

BLEU (Papineni et al., 2002)  
Meteor (Denkowski and Lavie, 2014)  
chrF (Popović, 2015)

# Datasets (Related sentences)



	L1 Lang.	-	L2 Lang.	L1 Language		L2 Language	
				Sentences	Tokens	Types	Tokens
Training Dataset	English	-	Slovene	2,299,805	37,849,280	631,114	33,379,920
	English	-	Croatian	2,464,895	34,880,415	626,353	29,744,620
	English	-	Serbian	2,152,740	27,465,108	658,660	23,536,540
	Croatian	-	Serbian	2,177,242	22,717,151	1,200,577	22,790,166
	Slovene	-	Croatian	2,004,229	17,759,047	464,392	18,150,733
	Slovene	-	Serbian	2,131,301	20,515,466	769,019	21,257,242
Development Dataset	English	-	Slovene	2,017	38,280	32,918	10,092
	English	-	Croatian	2,114	45,605	40,536	8,264
	English	-	Serbian	2,092	37,757	35,419	10,373
	Croatian	-	Serbian	2,000	14,716	14,774	4,601
	Slovene	-	Croatian	2,000	14,339	14,447	3,924
	Slovene	-	Serbian	2,000	12,985	13,575	3,890
Evaluation Dataset	English	-	Slovene	2,015	44,559	39,561	7,414
	English	-	Croatian	2,113	45,768	40,462	8,218
	English	-	Serbian	2,036	40,349	37,346	6,833
	Croatian	-	Serbian	2,000	12,805	13,043	3,909
	Slovene	-	Croatian	2,000	13,799	14,187	3,794
	Slovene	-	Serbian	2,000	13,090	13,606	3,900

# Evaluation (Related sentences)



	SMT			NMT		
	BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	37.35	29.92	60.74	27.41	24.83	53.02
Slovene → English	46.02	37.42	64.61	27.80	29.28	52.76
English → Croatian	32.44	28.03	57.46	20.94	21.69	49.73
Croatian → English	37.49	35.26	60.66	23.73	26.67	50.41
English → Serbian	31.30	27.33	55.28	17.44	20.24	45.89
Serbian → English	32.49	34.18	57.82	20.64	25.00	48.02
Slovene → Serbian	19.37	22.09	41.52	19.95	21.35	40.99
Serbian → Slovene	21.51	22.82	43.46	21.68	22.30	43.10
Slovene → Croatian	21.29	22.93	43.97	19.54	21.01	40.60
Croatian → Slovene	25.94	25.39	47.60	24.77	24.12	45.71
Serbian → Croatian	68.96	48.24	79.96	61.06	42.76	76.85
Croatian → Serbian	68.15	46.70	78.14	64.25	43.84	76.83

# Evaluation (Related sentences)



	SMT			NMT		
	BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	37.35	29.92	60.74	27.41	24.83	53.02
Slovene → English	46.02	37.42	64.61	27.80	29.28	52.76
English → Croatian	32.44	28.03	57.46	20.94	21.69	49.73
Croatian → English	37.49	35.26	60.66	23.73	26.67	50.41
English → Serbian	31.30	27.33	55.28	17.44	20.24	45.89
Serbian → English	32.49	34.18	57.82	20.64	25.00	48.02
Slovene → Serbian	19.37	22.09	41.52	19.95	21.35	40.99
Serbian → Slovene	21.51	22.82	43.46	21.68	22.30	43.10
Slovene → Croatian	21.29	22.93	43.97	19.54	21.01	40.60
Croatian → Slovene	25.94	25.39	47.60	24.77	24.12	45.71
Serbian → Croatian	68.96	48.24	79.96	61.06	42.76	76.85
Croatian → Serbian	68.15	46.70	78.14	64.25	43.84	76.83

# Evaluation (Related sentences)

	SMT			NMT		
	BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	37.35	29.92	60.74	27.41	24.83	53.02
Slovene → English	46.02	37.42	64.61	27.80	29.28	52.76
English → Croatian	32.44	28.03	57.46	20.94	21.69	49.73
Croatian → English	37.49	35.26	60.66	23.73	26.67	50.41
English → Serbian	31.30	27.33	55.28	17.44	20.24	45.89
Serbian → English	32.49	34.18	57.82	20.64	25.00	48.02
Slovene → Serbian	19.37	22.09	41.52	19.95	21.35	40.99
Serbian → Slovene	21.51	22.82	43.46	21.68	22.30	43.10
Slovene → Croatian	21.29	22.93	43.97	19.54	21.01	40.60
Croatian → Slovene	25.94	25.39	47.60	24.77	24.12	45.71
Serbian → Croatian	68.96	48.24	79.96	61.06	42.76	76.85
Croatian → Serbian	68.15	46.70	78.14	64.25	43.84	76.83

# Experiments on Compressed Data



	2M		
	Sentences	Tokens	Types
random (English)	2,000,000	25,496,017	243,512
random (Slovene)	2,000,000	21,175,388	415,965
compressed (English)	2,299,805	37,849,280	631,114
compressed (Slovene)	2,299,805	33,379,920	587,018

	5M		
	Sentences	Tokens	Types
random (English)	4,980,012	61,802,915	386,260
random (Slovene)	4,980,012	51,219,305	640,471
compressed (English)	5,003,508	85,675,760	618,112
compressed (Slovene)	5,003,508	72,868,553	964,763

# Experiments on Compressed Data

	2M		
	Sentences	Tokens	Types
random (English)	2,000,000	25,496,017	243,512
random (Slovene)	2,000,000	21,175,388	415,965
compressed (English)	2,299,805	37,849,280	631,114
compressed (Slovene)	2,299,805	33,379,920	587,018

	5M		
	Sentences	Tokens	Types
random (English)	4,980,012	61,802,915	386,260
random (Slovene)	4,980,012	51,219,305	640,471
compressed (English)	5,003,508	85,675,760	618,112
compressed (Slovene)	5,003,508	72,868,553	964,763

# Experiments on Compressed Data

	2M		
	Sentences	Tokens	Types
random (English)	2,000,000	25,496,017	243,512
random (Slovene)	2,000,000	21,175,388	415,965
compressed (English)	2,299,805	37,849,280	631,114
compressed (Slovene)	2,299,805	33,379,920	587,018
	5M		
	Sentences	Tokens	Types
random (English)	4,980,012	61,802,915	386,260
random (Slovene)	4,980,012	51,219,305	640,471
compressed (English)	5,003,508	85,675,760	618,112
compressed (Slovene)	5,003,508	72,868,553	964,763

# Experiments on Compressed Data

	2M		
	Sentences	Tokens	Types
random (English)	2,000,000	25,496,017	243,512
random (Slovene)	2,000,000	21,175,388	415,965
compressed (English)	2,299,805	37,849,280	631,114
compressed (Slovene)	2,299,805	33,379,920	587,018

	5M		
	Sentences	Tokens	Types
random (English)	4,980,012	61,802,915	386,260
random (Slovene)	4,980,012	51,219,305	640,471
compressed (English)	5,003,508	85,675,760	618,112
compressed (Slovene)	5,003,508	72,868,553	964,763



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51
Slovene → English		BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → English	random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19
5M				5M++			
English → Slovene		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	36.60	29.94	60.33	38.01	30.75	61.54
English → Slovene	compressed	37.56	30.57	61.75	38.75	31.19	62.58
Slovene → English		BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → English	random	36.72	33.90	59.58	38.02	34.08	59.70
Slovene → English	compressed	37.71	34.38	60.09	40.30	35.61	62.16



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51
Slovene → English		BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → English	random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19
5M				5M++			
English → Slovene		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	36.60	29.94	60.33	38.01	30.75	61.54
English → Slovene	compressed	37.56	30.57	61.75	38.75	31.19	62.58
Slovene → English		BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → English	random	36.72	33.90	59.58	38.02	34.08	59.70
Slovene → English	compressed	37.71	34.38	60.09	40.30	35.61	62.16



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
		38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	random	27.80	29.28	52.76	28.76	29.53	53.19
		5M			5M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	36.60	29.94	60.33	38.01	30.75	61.54
English → Slovene	compressed	37.56	30.57	61.75	38.75	31.19	62.58
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
		36.72	33.90	59.58	38.02	34.08	59.70
Slovene → English	random	37.71	34.38	60.09	40.30	35.61	62.16



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
		random	38.77	35.29	61.15	36.86	33.90
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19
		5M			5M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	36.60	29.94	60.33	38.01	30.75	61.54
English → Slovene	compressed	37.56	30.57	61.75	38.75	31.19	62.58
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
		random	36.72	33.90	59.58	38.02	34.08
Slovene → English	compressed	37.71	34.38	60.09	40.30	35.61	62.16



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
		38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19
		5M			5M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	36.60	29.94	60.33	38.01	30.75	61.54
English → Slovene	compressed	37.56	30.57	61.75	38.75	31.19	62.58
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
		36.72	33.90	59.58	38.02	34.08	59.70
Slovene → English	random	37.71	34.38	60.09	40.30	35.61	62.16
Slovene → English	compressed						



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52	
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51	
		Slovene → English			Slovene → English			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19	
		5M			5M++			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	36.60	29.94	60.33	38.01	30.75	61.54
		compressed	37.56	30.57	61.75	38.75	31.19	62.58
Slovene → English	random	36.72	33.90	59.58	38.02	34.08	59.70	
Slovene → English	compressed	37.71	34.38	60.09	40.30	35.61	62.16	



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52	
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51	
		Slovene → English			Slovene → English			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19	
		5M			5M++			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	36.60	29.94	60.33	38.01	30.75	61.54
		compressed	37.56	30.57	61.75	38.75	31.19	62.58
Slovene → English	random	36.72	33.90	59.58	38.02	34.08	59.70	
Slovene → English	compressed	37.71	34.38	60.09	40.30	35.61	62.16	



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52	
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51	
		Slovene → English			Slovene → English			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19	
		5M			5M++			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	36.60	29.94	60.33	38.01	30.75	61.54
		compressed	37.56	30.57	61.75	38.75	31.19	62.58
		English → Slovene			Slovene → English			
		BLEU	Meteor	chrF	BLEU	Meteor	chrF	
		random	36.72	33.90	59.58	38.02	34.08	59.70
English → Slovene	compressed	37.71	34.38	60.09	40.30	35.61	62.16	



# Evaluation on Compressed Data and different architecture

Insight

		2M			2M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	random	35.49	29.47	60.52	35.26	29.27	59.52
English → Slovene	compressed	27.41	24.83	53.02	29.30	25.75	54.51
		5M			5M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → English	random	38.77	35.29	61.15	36.86	33.90	59.31
Slovene → English	compressed	27.80	29.28	52.76	28.76	29.53	53.19
English → Slovene	random	36.60	29.94	60.33	38.01	30.75	61.54
English → Slovene	compressed	37.56	30.57	61.75	38.75	31.19	62.58
		5M			5M++		
		BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → English	random	36.72	33.90	59.58	38.02	34.08	59.70
Slovene → English	compressed	37.71	34.38	60.09	40.30	35.61	62.16

## Comparison between different SMT and NMT methods

- better understanding and evaluation
- NMT for an under-resourced scenario
- multiway models

**ASISTENT** English-Slavic Translation System

Home About Team Docker Other Projects

Add text to be translated and specify the translation direction:

[English Example](#) [Slovene Example](#) [Croatian Example](#) [Serbian Example](#)

Translate

This service is brought to you by  
<http://nuig.insight-centre.org/unlp/>

ASISTENT (or `assistant@en`) was developed to translate text between English and the morphological complex south Slavic languages: Slovene, Serbian and Croatian.

Select translation option/approach:

phrase based  hierarchical  neural  
 direct translation  pivot translation

Please choose the translation direction:

English → Slovene  Slovene → English  
 English → Croatian  Croatian → English  
 English → Serbian  Serbian → English  
 Slovene → Croatian  Croatian → Slovene  
 Serbian → Slovene  Slovene → Serbian  
 Croatian → Serbian  Serbian → Croatian

Insight NUI Galway SFI

This resource has been funded by Grant No. SFI/12/RC/2289 - INSIGHT - National University of Ireland, Galway

<http://server1.nlp.insight-centre.org/asistent/>

# A Comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian

Mihael Arcan

[mihael.arcan@insight-centre.org](mailto:mihael.arcan@insight-centre.org)

Thank You

# Comparison of different RNNs



	English-Irish		English-Basque	
	Word	BPE	Word	BPE
RNN	48.78	49.39	38.07	26.57
bidirectional RNN	50.99	51.6	40.43	27.88
Deep bidirectional RNN	51.42	52.44	41.37	28.21
Pyramidal Deep bidirectional RNN	49.78	50.78	40.42	27.76
GNMT (transformer)	50.15	50.87	39.27	27.22

# Evaluation

		BLEU	Meteor	chrF		
Slovene-Serbian	5.30	11.48	25.61			
Serbian-Slovene	5.24	11.97	26.61			
Slovene-Croatian	4.80	11.51	26.34			
Croatian-Slovene	4.76	11.98	27.80			
Serbian-Croatian	66.78	46.70	78.53			
Croatian-Serbian	67.20	46.08	77.85			
	SMT	NMT				
	BLEU	Meteor	chrF	BLEU	Meteor	chrF
Slovene → Serbian	19.37	22.09	41.52	19.95	21.35	40.99
Serbian → Slovene	21.51	22.82	43.46	21.68	22.30	43.10
Slovene → Croatian	21.29	22.93	43.97	19.54	21.01	40.60
Croatian → Slovene	25.94	25.39	47.60	24.77	24.12	45.71
Serbian → Croatian	68.96	48.24	79.96	61.06	42.76	76.85
Croatian → Serbian	68.15	46.70	78.14	64.25	43.84	76.83

# Multiway Models experiments



	500k				2M			
	2/500/500		4/600/1000		2/500/500		4/600/1000	
	Word	BPE	Word	BPE	Word	BPE	Word	BPE
En-SI	6.37	3.54	7.28	5.8	11.11	6.63	11.77	8.14
SI-En	10.29	3.18	8.65	5.04	13.77	7.88	13.77	7.01
En-Sr	7.56	4.2	6.2	3.25	10.22	3.83	9.15	6.83
Sr-En	10.26	7.67	7.23	5.98	14.32	11.21	13.9	11.87
En-Hr	6.66	3.97	5.75	3.99	9.67	2.4	9.09	6.76
Hr-En	8.55	7.21	9.18	6.37	13.18	9.78	13.2	8.96
Sl-Sr	15.66	7.08	14.58	7.14	10.99	18.25	9.14	18.35
Sr-Sl	18.22	13.8	17.56	11.34	21.38	17.22	21.02	15.96
Sl-Hr	16.98	10	16.33	10.14	19.25	12.1	19.43	13.14
Hr-Sl	18.96	12.84	18.37	12.74	21.99	16.92	21.26	13.62
Sr-Hr	56.28	44.76	53.6	52.61	58.71	53.75	58.56	54.25
Hr-Sr	54.13	43.24	52.11	41.3	56.26	52.23	55.95	51.34

# All available data vs 2M parallel data



	SMT		
	BLEU	Meteor	chrF
English → Slovene	37.35	29.92	60.74
Slovene → English	46.02	37.42	64.61
English → Croatian	32.44	28.03	57.46
Croatian → English	37.49	35.26	60.66
English → Serbian	31.30	27.33	55.28
Serbian → English	32.49	34.18	57.82
Slovene → Serbian	19.37	22.09	41.52
Serbian → Slovene	21.51	22.82	43.46
Slovene → Croatian	21.29	22.93	43.97
Croatian → Slovene	25.94	25.39	47.60
Serbian → Croatian	68.96	48.24	79.96
Croatian → Serbian	68.15	46.70	78.14

English → Slovene			Slovene → English			
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	34.46	32.90	61.75	44.41	40.59	62.94
Asistent	<b>49.82</b>	<b>36.36</b>	<b>69.26</b>	<b>64.14</b>	<b>45.77</b>	<b>76.71</b>
English → Croatian			Croatian → English			
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	29.27	26.87	57.19	46.08	39.35	67.61
Asistent	<b>42.15</b>	<b>33.02</b>	<b>64.95</b>	<b>48.07</b>	<b>40.05</b>	<b>68.18</b>
English → Serbian			Serbian → English			
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	27.48	26.33	55.80	<b>46.05</b>	<b>39.35</b>	<b>67.53</b>
Asistent	<b>42.47</b>	<b>32.63</b>	<b>63.59</b>	42.35	39.11	64.96
Slovene → Serbian			Serbian → Slovene			
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	12.27	16.83	34.27	14.05	18.32	37.16
Asistent	<b>23.46</b>	<b>23.23</b>	<b>43.23</b>	<b>29.23</b>	<b>25.33</b>	<b>47.42</b>
Croatian → Serbian			Serbian → Croatian			
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	59.88	41.72	73.84	64.90	46.11	78.65
Asistent	<b>67.39</b>	<b>46.34</b>	<b>77.98</b>	<b>70.09</b>	<b>48.97</b>	<b>80.89</b>
Slovene → Croatian			Croatian → Slovene			
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	13.47	17.81	37.36	16.07	19.60	39.54
Asistent	<b>34.63</b>	<b>27.02</b>	<b>50.98</b>	<b>38.64</b>	<b>29.78</b>	<b>54.98</b>