South-East European Times: A parallel corpus of Balkan languages

Francis M. Tyers, Murat Serdar Alperen

Universitat d'Alacant, Kurupelit,
E-03071 Alacant (Spain), Samsun (Turkey)
ftyers@dlsi.ua.es, msalperen@gmail.com

Abstract

This paper describes the creation of a parallel corpus from a multilingual news website translated into eight languages of the Balkans (Albanian, Bulgarian, Croatian, Greek, Macedonian, Romanian, Serbian, and Turkish) and English. The corpus is then applied to the task of machine translation, creating 72 machine translation systems. The performance of these systems is then evaluated and thought is given to where future work might be focussed.

1. Introduction

The article has a twofold aim, the first is to describe the creation and status of a *free*[1] parallel corpus for the Balkan languages. The second is to describe the use of this corpus to create 72 machine translation (MT) systems between the Balkan languages and evaluate the differing challenges facing MT between these languages. It also presents the first published results for systems between, for example Macedonian and Albanian and gives some thought on where further research might be aimed.

This is a parallel corpus in the vein of EuroParl (Koehn, 2005) or JRC-Acquis (Steinberger et al., 2006), that is designed to be useful for machine translation and other multilingual natural language processing research, not necessarily useful for corpus-linguistic research due to uncertainty of which is the source and which is the target language of the translated sentences.

Aside from the ubiquitous English, the languages contained in this corpus fall into several linguistic groups, Turkic (in the case of Turkish), Slavic (Bulgarian, Croatian, Macedonian and Serbian), Hellenic (Greek), Romance (Romanian) and Albanic (Albanian).

A number of these languages also form what is known as the *Balkan Sprachbund*, or Balkan linguistic area. This is a group of languages which have similar lexical and grammatical features, but as a result of geographical proximity rather than genetic relationship (Lindstedt, 2000).

It has been shown that translating between genetically and typologically related languages is easier than languages with less relation. Homola and Kuboň (2004) for example discuss the relative ease of translating between genetically related Slavic languages, typologically related Baltic languages and English. Part of the purpose of this paper is to see if this holds for the languages of the Balkan Sprachbund.

Although this corpus is described as a corpus of the Balkan languages, it is worth noting that it is not comprehensive. It does not for example include the smaller regional and minority languages of the Balkans, such as Aromanian and Romani, nor does it include Slovenian.

This corpus has also been aligned before (Paskaleva, 2007). However, it was only aligned to English, not between Balkan languages. This paper is motivated by the fact that the corpus in Paskaleva (2007) was not made public, and by producing more aligned text between the Balkan languages, not just with respect to English.

2. Data preparation

The South-East European Times (`http://www.setimes.com`) website is a news site which covers current events in the Balkans in the languages of the Balkans and English. The text content of the site is released as *public domain*, meaning it can be used, modified and redistributed for any purpose without permission. Content has been published starting 2002 and is ongoing.

The website has four main sections: Features, which are mainly longer articles, News Briefs, which are usually shorter articles summarising the news, Articles, which contains news articles somewhat shorter than Features, and Round-up, which is usually a page of extracts of longer articles.

To download all the files we have derived a list of English files using the XENU web-spider.[2]

The unique URL structure of the website made it easy to derive the correspondent language, i.e. `en_GB/.../2009/08/07/feature-02` has the corresponding Turkish version at `tr/.../2009/08/07/feature-02`. Not only could we easily locate the translation , we were also able to apply batch alignments without difficulty.

2.1. Collection

After collating the links, pages were downloaded with `httrack`[3] and stripped of HTML with `funduc`[4]. The encoding of the files (variously in ISO-8859-1, ISO-8859-9 and UTF-8) was normalised to UTF-8.

[1] Here we follow the use of free as defined by the Free Software Foundation; `http://www.gnu.org/philosophy/free-sw.html`, meaning free to *use*, *modify* and *redistribute* for any purpose – including commercial.

[2] `http://home.snafu.de/tilman/xenulink.html`

[3] `http://www.httrack.com/`

[4] `http://www.funduc.com/search_replace.htm`

2.2. Sentence splitting

Sentences were split using the `SentParBreaker` splitter.[5] This splitter unfortunately only accepted input in ISO-8859-1 encoding, so a transliteration scheme was devised for languages which were not written in a script which was representable in this encoding (Bulgarian, Greek and Macedonian). Languages which used a different single-byte encoding (Serbian, Croatian, etc.) were transliterated to ISO-8859-1.

2.3. Sentence alignment

Once split, sentences were aligned pairwise between the languages using *hunalign* (Varga et al., 2005).

A preliminary eye-ball evaluation showed the sentence alignment accuracy to be less than perfect. We performed a more complete evaluation of the whole corpus by selecting from each of the alignments one hundred sentences semi-randomly.[6] These alignments were then checked manually and an accuracy figure calculated for each pair.

The results of this evaluation are presented in table 1. For comparison we applied the above method to the EuroParl corpus (Koehn, 2005) alignments for English to Spanish which received a comparable 93% accuracy. These results are not, however, entirely comparable as the SETimes corpus has a smaller number of sentences and the sentences tend to be of a shorter length.

2.4. Common test set

Unlike other parallel corpora covering many languages, the SETimes corpus does not contain all of the text in all of the languages, some translations on the site were missing in some of the languages. In creating the corpus we attempted to maximise the number of aligned sentences in all language pairs, so sentence pairs were included even if they were not translated into all of the languages.

However, in order to effectively evaluate the machine translation systems produced it was desirable to have a subset of sentences which translated into all of the languages as to make the results comparable.

An example sentence from this test set is given with all translations in figure 1.

For the common test set and training set, 1,000 sentences were extracted and the alignments were manually validated. These 1,000 sentences were split into 400 held out and 600 for testing.

2.5. Corpus statistics

Some simple statistics were calculated over corpus as a whole, and over each of the pairwise alignments. Table 2 gives the number of words in the target language per pair, which is between four million and five million words, excepting pairs with Bulgarian. It is suspected that there was an error in the Bulgarian data that caused this anomaly, possibly due to a sentence in the middle of the data which had no alignment, causing the final sentence extraction script to stop processing.

Table 3 gives the type-token ratio for each of the languages, this presents some kind of measure of their morphological richness, with morphologically rich languages having a larger number of types per token.

Language	Tokens	Types	Ratio
Turkish (`tr`)	34,246,226	139,412	0.40
Croatian (`hr`)	34,968,453	127,756	0.36
Serbian (`sr`)	37,989,711	133,073	0.35
Macedonian (`mk`)	37,623,521	113,393	0.30
Bulgarian (`bg`)	38,419,402	104,669	0.27
Greek (`el`)	41,599,313	105,221	0.25
Albanian (`sq`)	41,741,782	104,322	0.24
Romanian (`ro`)	41,501,934	94,268	0.22
English (`en`)	38,463,808	68,005	0.17

Table 3: Type-token ratio for each of the languages calculated from the raw corpora.

3. Machine translation

For each of the language pairs we trained a phrase-based[7] statistical machine translation system using the Moses toolkit (Koehn et al., 2007). The training process followed the instructions for the baseline system in WMT09, the shared task in the ACL 2009 workshop on statistical machine translation (Callison-Burch et al., 2009) with the following changes: The IRSTLM (Marcello et al., 2008) toolkit was used for the target language model, MERT training was skipped due to time constraints, and text was not recased. The language model trained was a five-gram language model was trained on the target language side of the bilingual aligned text. The total training time for the 72 systems, including time to build the language models and binarise the phrase and reordering tables was around ten days.

4. Evaluation

We evaluated the system with BLEU (Papineni et al., 2002), an automatic metric which attempts to measure translation quality by comparing the source text with one or more pre-translated reference texts. While this has been shown to be problematic when comparing different systems (Callison-Burch et al., 2006; Labaka et al., 2007), we consider it to provide a reasonable measure for comparing the quality of translations output by models trained using the same system for different languages on comparable data.

Table 4 shows the results for all of the systems trained. The scores were calculated using the NIST `mteval-v13a`

[5] `http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector`

[6] The standard program `unsort` was used for this purpose.

[7] The *phrases* in phrase-based statistical machine translation are not syntactic constituents and might be better termed segments or chunks, but here we follow the normal SMT nomenclature.

	bg	el	en	hr	mk	ro	sq	sr	tr
bg	-	97.26	87.14	93.50	98.76	93.15	91.89	89.74	91.54
el	98.63	-	90.27	90.78	97.56	93.50	98.70	92	97.10
en	93.42	92.75	-	100	92.85	98.63	98.59	98.66	98.63
hr	98.55	91.35	95.83	-	92	98.68	96	98.79	100
mk	95.94	97.18	93.84	94.28	-	84.375	87.67	84	97.22
ro	98.48	98.59	100	100	94.20	-	100	98.68	100
sq	92.40	91.54	98.66	97.61	90.41	100	-	91.30	100
sr	89.04	84.81	97.26	100	95.65	96.92	95.94	-	97.84
tr	90.66	90.90	97.46	100	91.54	100	100	98.48	-

Table 1: Percentage of correct alignments out of a semi-randomly selected one hundred for each of the language pairs, with lines containing only formatting excluded.

English:	Ivaylo Markov, 42, was shot dead in his underground parking garage.
Bulgarian:	42-годишният Ивайло Марков бе застрелян в подземния си гараж.
Macedonian:	42-годишниот Ивајло Марков беше убиен во неговата подземна паркинг гаража.
Croatian:	Ivaylo Markov, 42, ubijen je iz vatrenog oružja u svojoj podzemnoj garaži.
Serbian:	Ivajlo Markov, 42, ubijen je iz vatrenog oružja u svojoj podzemnoj garaži.
Greek:	Ο Ιβαίλο Μάρκοφ, 42 ετών, πυροβολήθηκε σε υπόγειο χώρο στάθμευσης.
Romanian:	Ivailo Markov, în vârstă de 42 de ani, a fost ucis prin împușcare în garajul său subteran.
Albanian:	Ivajlo Markov, 42 vjeç u qëllua për vdekje në garazhin e tij nëntokësor.
Turkish:	42 yaşındaki İvaylo Markov yeraltı otoparkında vuruldu.

Figure 1: An example sentence from the manually-validated aligned test set in all nine languages

	bg	el	en	hr	mk	ro	sq	sr	tr
bg	-	3,907,720	4,179,847	3,774,060	3,407,459	4,521,592	4,549,607	4,092,154	4,244,756
el	2,962,995	-	4,920,462	4,454,243	4,822,344	5,360,872	5,385,591	4,857,574	4,463,723
en	2,810,557	5,072,596	-	4,376,267	4,531,170	5,232,684	5,259,457	4,618,166	4,229,978
hr	2,886,604	5,256,393	4,927,971	-	4,712,449	5,302,851	5,336,016	4,775,975	4,322,212
mk	2,193,720	5,005,785	4,498,495	4,085,597	-	4,892,898	4,899,071	4,643,575	4,091,469
ro	2,494,235	5,378,388	4,927,260	4,455,043	4,725,079	-	5,347,961	4,706,146	4,292,371
sq	2,505,685	5,376,278	4,922,668	4,447,139	4,709,619	5,322,065	-	4,715,860	4,296,759
sr	2,738,029	5,029,567	4,666,445	4,335,380	4,793,010	5,062,830	5,098,649	-	4,203,352
tr	3,466,788	5,212,746	4,780,958	4,366,332	4,604,138	5,133,404	5,164,828	4,685,690	-

Table 2: Total number of aligned words per language pair. Number of words words in target language and calculated with the standard `wc` command.

	Target language									
	bg	el	en	hr	mk	ro	sq	sr	tr	Mean from
bg	-	0.1508	**0.2898**	0.1537	0.2501	0.2284	0.2135	0.1560	0.1694	0.2015
el	0.1539	-	**0.4269**	0.2871	0.3979	0.3731	0.3617	0.3051	0.1757	0.3102
en	0.2151	*0.4055*	-	0.3162	***0.4506***	*0.4299*	*0.4464*	0.3477	*0.2090*	0.3526
hr	0.1297	0.3251	0.3952	-	0.4041	0.3478	0.3271	***0.6556***	0.1847	0.3462
mk	*0.2361*	0.3514	**0.4316**	0.3090	-	0.3654	0.3442	0.3376	0.1702	0.3182
ro	0.1735	0.3490	**0.4364**	0.3018	0.3970	-	0.3754	0.3263	0.1894	0.3186
sq	0.1722	0.3760	**0.4908**	0.3011	0.4147	0.4064	-	0.3269	0.1851	0.3341
sr	0.1382	0.3349	0.4016	***0.6524***	0.4118	0.3586	0.3327	-	0.1873	0.3521
tr	0.0566	0.1995	**0.2522**	0.1620	0.2396	0.2215	0.1928	0.1851	-	0.1886
Mean to	0.1235	0.3108	0.3905	0.3104	0.3707	0.3413	0.3242	0.3300	0.1838	-

Table 4: BLEU scores on the test set for all the language pairs. Highest scores translating *to* a language are given in **bold face**, while highest scores translating *from* a language are given in *italics*.

script,[8] and are presented *as output*. Tests for statistical significance have not been made.

It is interesting to note that the scores for Bulgarian are much worse than could be expected, this is probably due to the much lower number of aligned sentences. The lower number of sentences for Bulgarian is probably due to an error in the alignment process, although the alignment validation gave similar alignment quality. This is a subject for further investigation. Other scores are comparable with similar systems.

For an idea of how morphological richness affected translation quality, we calculated the type-token ratio[9] and plotted this against the average BLEU score for translating into the language (see Figure 4.). We consider the type-token ratio a measure of morphological richness, the higher the ratio, the richer the language. For translating into genetically unrelated languages, there is a good correlation between type-token ratio and BLEU score, there is also a good correlation when only considering translation between genetically related languages. For translating from a language, the picture is more clear, and genetic relatedness does not play so much a part, except for the case of Serbian and Croatian where the mean is skewed by the exceptionally high results between these two languages. In fact, considering that for some time, and still to a certain extent these two languages are considered as two written standards of the same language, we calculated the scores for using the Serbian source text as a translation of Croatian and vice-versa. Considering Croatian source text as a Serbian test text gives a score of 0.4114, while the other way around gives a score of 0.4112. This is comparable with the scores of the MT system for *translating* between other languages.

5. Discussion

We have presented, to our knowledge the first pan-Balkan parallel corpus. The corpus is available publically,[10] so that other researchers can reproduce and expand on our results. As the SETimes website continues to publish daily, we intend to continue adding text to the corpus. Targets for the next release will be fixing the Bulgarian data, and rerunning with a Unicode-aware sentence tokeniser.

As can be seen from the results, both translating to and from English gives the best scores for unrelated languages, this could be a result of a number of factors, one is that as English is a very weakly inflected language, the amount of distinct word forms will be lower making translation easier. Also, as noted by Virpioja et al. (2007), in morphologically rich languages words include more information on average, and one mistake in a suffix is enough to mark the whole word as incorrect, although it may not prevent understanding. Another factor is that the texts are probably all translated *from* English and not from each other, which could provide a less literal translation, effectively making any translation not aligned with English more of a paraphrase.

While membership of the *Sprachbund* does not seem to have any relationship with the translation quality between typologically related languages, further work might look at how morphological or syntactic information might be included, for example in factored translation models (Koehn and Hoang, 2007), or even look at creating rule-based systems between these languages, which have been shown to outperform phrase-based SMT between related languages (Tyers and Nordfalk, 2009). However, the lack of free morphological and syntactic analysers for the Balkan languages (with the exception of Romanian and Bulgarian) makes this more difficult.

We hope that the release of this corpus can provide a basis for other multilingual projects for the Balkan languages, one could, for instance, envisage a pan-Balkan aligned tagged corpora or treebanks based on this text.

[8] Available for download from `http://www.itl.nist.gov/iad/mig/tools/`.

[9] The type-token ratio is the ratio between the number of unique tokens in the corpus and the number of tokens in the corpus.

[10] The full corpus, including the translation models trained can be downloaded from `http://elx.dlsi.ua.es/~fran/SETIMES/` and is mirrored at `http://www.statmt.org/setimes/`.

Acknowledgements

This article is inspired by Philipp Koehn's article on EuroParl, any similarities are intentional. We are thankful to Miloš Stolić for his help in validating alignments for the Slavic languages and English. This work has also received the support of the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01.

6. References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. *Proceedings of EACL-2006*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. *EACL 2009 Fourth Workshop on Statistical Machine Translation*, pages 1–28.

Petr Homola and Vladislav Kuboň. 2004. A Translation Model For Languages of Acceding Countries. *Proceedings of the Conference of the European Association of Machine Translation 2004*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *ACL 2007, demonstration session*.

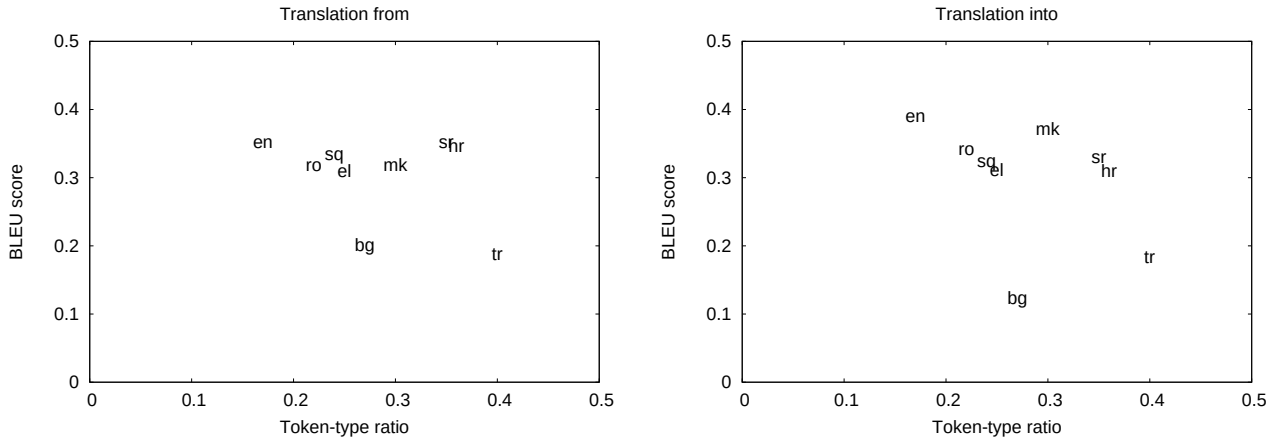

Figure 2: Plots showing token-type ratio versus mean BLEU score for translating into and from a given language. As a trend the token-type ratio increases, the BLEU score decreases. In both cases, Bulgarian is an outlier due to bad training data, or simply less training data, and Macedonian, Serbian and Croatian are outliers due to high translation quality between very closely-related languages regardless of token-type ratio.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005*.

Gorka Labaka, Nicholas Stroppa, Andy Way, and Kepa Sarasola. 2007. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal*.

Jouko Lindstedt. 2000. Linguistic Balkanization: Contact-induced change by mutual reinforcement. *Languages in Contact*, 28:231–246.

Federico Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: An open source toolkit for handling large scale language models. *Proceedings of Interspeech 2008*, pages 1618–1621.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Elena Paskaleva. 2007. Balkan South-East Corpora Aligned to English. *Proceedings of the Workshop on Common Natural Language Processing Paradigm for Balkan Languages, RANLP 2007*, pages 35–42.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy, June.

Francis M. Tyers and Jacob Nordfalk. 2009. Shallow-transfer rule-based machine translation for Swedish to Danish. *Proceedings of FREERBMT2009, the First Workshop on Free/Open-Source Rule-based Machine Translation*, pages 28–33.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. *Proceedings of RANLP 2005*, pages 590–596.

Jaakko J. Virpioja, Mathias Creutz Väyrynen, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, September.