PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 94 SEPTEMBER 2010 67–76

CorporAl: a Method and Tool for Handling Overlapping Parallel Corpora

Mark Fishel, Heiki-Jaan Kaalep

Institute of Computer Science, University of Tartu

Abstract

This work introduces a method and tool for handling overlapping parallel corpora – i.e. corpora that are based on the same source material. The method is insensitive to minor changes in the text, different segmentation levels of the corpora and omitted material from either corpora. The aim is to detect matching sentence pairs and either produce combinations of the overlapping corpora or compare them and assess their quality in comparison to each other. The introduced tool enables the user to define the desired behavior when combining corpora pairs, resulting in pure comparison, maximum-size or maximum-quality versions of the combinations. We test the tool on two cases of overlapping parallel corpora and five language pairs. We also evaluate the impact of using the method on two translation systems – a phrase-based and a parsing-based one.

1. Introduction

The target of this research is parallel corpora that are based on partially or fully overlapping sources of the same language pair – overlapping parallel corpora. Such corpora can exist, for instance, when the same source documents are independently used to create corpora at different times or different institutions.

Processing such corpora can be quite problematic. Simply concatenating them is not a valid solution, since the data distribution of the combined corpus will be skewed. At the same time using the standard `diff` utility is not guaranteed to elegantly solve the problem of detecting the repeated and unique samples. Typically the texts have differences in representation, or some typing or aligning errors fixed or introduced in one of the corpora. In addition sentence pairs might be segmented differently in the two corpora or be omitted from one of them.

Corresponding author: fishel@ut.ee

Cite as: Mark Fishel, Heiki-Jaan Kaalep. CorporAl: a Method and Tool for Handling Overlapping Parallel Corpora. The Prague Bulletin of Mathematical Linguistics No. 94, 2010, pp. 67–76. doi: 10.2478/v10108-010-0021-7.

On the other hand, if those difficulties could be overcome, the overlap could be exploited to many advantages. By comparing the two corpora the level of segmentation of both can be increased, the potential alignment error spots can be found and the size of both can be increased on the account of omitted sentence pairs from one or the other corpus. Finally, if it can be assumed that one of the corpora is much more accurate, the other corpus can be proofed against it to evaluate or improve its quality.

Here we present a method that can be used to do all of the tasks mentioned above, together with its implementation. We apply the method to two cases of overlapping parallel corpora and evaluate its influence on the scores of statistical translation systems, trained on the resulting corpora.

2. Overlapping Parallel Corpora

Let us first look at some examples of overlapping parallel corpora.

(Kaalep and Veskis, 2007) compare the JRC-Acquis corpus (version 2.2) (Steinberger et al., 2006) and the corpus of the University of Tartu[1]. The latter also includes Estonian laws with their English translations, in addition to the EU legislation. To our knowledge (Kaalep and Veskis, 2007) is the only work addressing the issue of overlapping parallel corpora.

Another example is the JRC-Acquis corpus itself, since it provides two alternative alignments for every language pair it includes – done with Vanilla[2] and HunAlign (Varga et al., 2005). This means that, although the text might be exactly the same, the level of segmentation can be different in the two versions. In addition, it is common practice for aligners to exclude sentence pairs in which they are not confident enough.

In the experimental part of this work we focus on the two presented cases; however there are other examples as well. The Hunglish corpus (Varga et al., 2005) includes EU legislation, obtained from the same sources as the JRC-Acquis. One part of the CzEng corpus (Bojar and Žabokrtský, 2009) also consists of EU legislation, whereas the source documents were taken directly from JRC-Acquis, but the text processing and alignment was done all over. Also a whole domain of corpora is a potential source for multiple versions of the same text – movie subtitles.

3. Method Description

Let us start with an example of two parallel corpora containing an overlap (figure 1). The third sentence pair of corpus B is omitted from corpus A and the third sentence pair of corpus A is segmented into two sentence pairs in corpus B. Also there are slight differences in punctuation between the two corpora.

[1] http://www.cl.ut.ee/korpused/paralleel/?lang=en

[2] http://nl.ijs.si/telri/Vanilla/

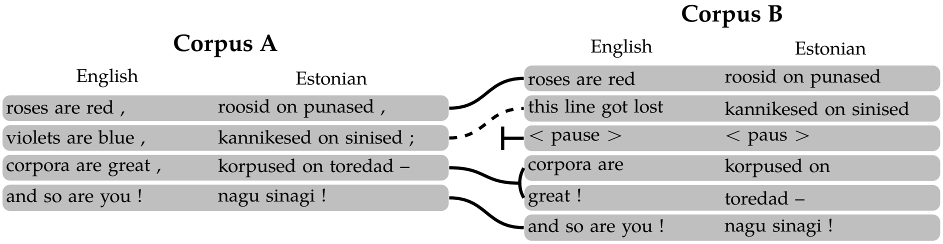


Figure 1. An example of overlapping parallel corpora with the correspondence of the two corpora shown. Second sentence pair of corpus B is an erroneous alignment.

Knowing both English and Estonian, it is easy to see that the English sentence from second sentence pair in corpus B got distorted, which makes the pair an erroneous alignment. Without knowing either of the languages, it can still be detected that one of the second sentence pairs in both corpora is probably erroneous – since the Estonian parts are practically the same, while the English parts are nothing like each other. Very simply put, this language-wise comparison is the basis of the method that we are about to introduce.

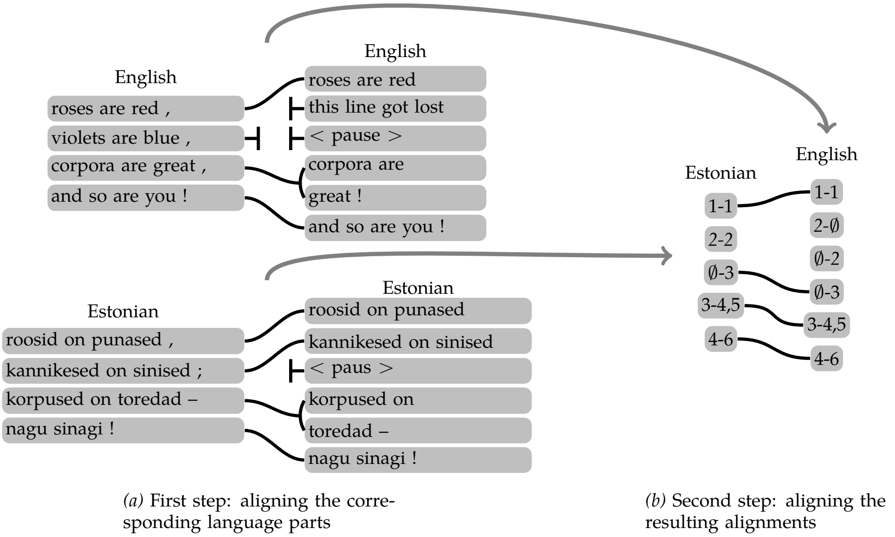
The method involves two steps. The first step consists of aligning the corresponding language parts to each other; see figure 2 (a) for an illustration. In the second step the resulting language alignments are themselves aligned to each other. Here the aim is to find the matching and mismatching alignment chunks. This way whenever in one language two sentences match while in the other language the corresponding sentences do not, this will be detected as an alignment error. See figure 2 (b) for an illustration of the second step; notice the resemblance between the resulting alignment and the correspondence of the parallel corpora in the example on figure 1.

In the following subsections we will describe in detail the two steps of the algorithm, as well as sentence approximate matching.

3.1. Aligning the Corresponding Language Parts

The first step is in essence very similar to the original task of bilingual sentence alignment itself. However, whereas the latter means comparing different languages and therefore requires, for instance, probabilistic solutions, in this case the task is much simpler, since both parts are in the same language and it suffices to compare the sentences using simple text processing. The only problem is that instead of strict comparison of the sentences, here approximate comparison is required due to possible slight differences in different corpora.

The aligning task is therefore analogical to the longest common subsequence problem, where corpora units (i.e. sentences or paragraphs) are matched to each other.



(a) First step: aligning the corresponding language parts

(b) Second step: aligning the resulting alignments

Figure 2. The two steps of processing the overlapping parallel corpora from the example on figure 1.1. ∅ stands for an empty counterpart (in zero-to-one alignments).

Here the alignment of the two texts is computed using generalized edit distance. The cost of substituting a unit for another equals the similarity between them (obtained with approximate sentence matching, explained in the next subsection). In addition all N-to-M pairs are also considered (up to a predefined limit). This enables matching aligned units even if the segmentation level is very different in the two corpora.

3.2. Approximate Sentence Matching

(Kaalep and Veskis, 2007) use Levenshtein distance with 1% of the average of the two sentence's length as a threshold. Other string similarity metrics applied to written text include several from the edit distance family (the Needleman-Wunsch metric, the Smith-Waterman metric, etc), the Jaro metric and others.

Here we use the method of (Kaalep and Veskis, 2007), extended to generalized edit distance. For instance the weight of replacing/inserting digits is extremely high, so that e.g. sentences "article 3" and "article 5" will not be considered to match with no matter what edit distance percentage threshold. On the other hand operations on empty symbols (spaces, tabs) and punctuation have low weights. This allows to set the percentage threshold higher without adding obvious matching errors.

3.3. Aligning the Alignments

As soon as the language part alignments are obtained, their correspondence to each other is to be determined. Although different language parts are to be compared here, only the alignments between unit numbers are compared, which again enables using direct comparison. In this case it is accomplished with the Levenshtein distance of the alignment cells.

It is important to note that a mismatch between two alignments does not indicate, which of the corpora has an erroneous alignment; instead, it shows a potential spot, where at least one of the corpora has an error. If one of the corpora is known to be accurately aligned, the errors of the other corpus can be corrected automatically this way. Otherwise the spots can be manually post-processed and the errors in the appropriate corpus – corrected.

On the other hand a match between alignments also merely indicates that the two corpora have matching alignments. This can occur both in case of correct alignments and coinciding erroneous alignments, though the latter is less likely (depending on the used alignment method).

3.4. Implementation – the CorporAl Tool

The CorporAl open-source project is available from Sourceforge[3]. The implementation is done as a PERL script and thus can be run on any platform with a PERL interpreter; the interface of the tool is command-line-based.

The tool name is meant to reflect the core idea of the method – "aligning" the corpora to each other. Using the alignment between the corpora the tool generates a new combined corpus. The exact behaviour can be controlled with input parameters: whether to include or exclude sentences from the unique and the matching parts, whether to skip mismatched sentence pairs or define one of the corpora as the more trustworthy one and include sentences from it. If sentence pairs match, the side with a higer level of segmentation is automatically included. Also it is possible to just output the alignment of the corpora to be used for further processing.

The main direction of further development of CorporAl is extending it to support monolingual corpora with annotation, in addition to parallel corpora. If the two overlapping corpora are augmented with the same annotation then both the text and the annotation can be compared, just like the two language parts of parallel corpora.

Alternatively, if the annotations differ, only the text can be matched and not the annotation. As a result the tool would allow to produce a text corpus with both annotations, regardless of differences of the texts. Also it could be applied to parallel corpora where all languages are aligned to one, like Europarl (Koehn, 2005), to produce a corpus of any two languages without re-applying the aligners. This just requires making the alignment of the annotation or second part of a parallel corpus optional.

[3] http://corporal.sf.net

4. Experiments

Our final aim was to test the presented method in practice. We focused the experiments on two cases of overlapping parallel corpora, described in section 2: first, the corpus of the University of Tartu (UT) and the English-Estonian (en-et) part of JRC-Acquis version 2.2 (JRC2) and second, the HunAlign and Vanilla versions of the English-Estonian (en-et), English-Latvian (en-lv), Estonian-Latvian (et-lv) and German-English (de-en) parts of JRC-Acquis version 3 (JRC3). First we present the results of processing the corpora and then go on to testing the effect of our method on statistical translation systems.

4.1. Processing Overlapping Parallel Corpora

We first grouped the documents in all corpora by their CELEX codes, which resulted in three groups: documents unique to one of the corpora and the ones present in both corpora in a pair. Then the common parts of the corpora were processed with the CorporAl tool. We generated two different versions of the combination: one (called max-size) prioritized the resulting corpus size and the other one (called max-accuracy) prioritized the resulting accuracy – the latter thus included only the matching sentence pairs, present in both corpora.

The sizes of the documents and the resulting corpora parts are presented in Table 1 and the frequencies of the types of sentence pair matches – in Table 2.

Looking at the match type frequencies it can be seen that the many-to-one matches constitute just a small percent of all the matches (below 1% on both sides). Thus, contrary to our initial assumption, the levels of segmentation of the UT and JRC2 corpora overlapping parts are practically the same. The same goes for the JRC3 pairs, where the total percentage of many-to-many alignments is even lower.

An interesting observation about the JRC3 pairs is the difference between the documents included only in the Vanilla or HunAlign versions. It can be seen in Table 1 that while the HunAlign versions of all the four pairs include only three to five documents that are not included in the Vanilla versions, the total numbers of words and sentence pairs in these documents are much higher than their counterparts in the Vanilla versions. In addition the total sizes of the common parts of the HunAlign versions are bigger than the same document sets of Vanilla versions. These two facts might indicate that in the HunAlign version documents and sentences were more confidently included into the corpus than in the Vanilla versions.

As a result of similarity of the JRC3 pairs the max-size combinations are practically of the same size as the bigger HunAlign common parts (with only 100-150 extra sentence pairs). The max-size combination of UT and JRC2 is visibly bigger than both corpora. On the other hand the max-accuracy combinations are slightly smaller than the source corpora in all five cases, which is caused by the portion of mismatching and omitted sentence pairs.

		#docs	**#snt pairs**	**#lang-1 words**	**#lang-2 words**
UT+JRC2		**UT/JRC**	**UT/JRC (·10^3)**	**UT/JRC (·10^6)**	**UT/JRC (·10^6)**
en-et	Unique	2048/5807	134.7/205.0	3.12/4.86	2.17/3.25
	Common	2009	93.2/68.2	1.9/1.7	1.3/1.1
	Max-size	2009	98946	2.03	1.36
	Max-acc	2009	56234	1.35	0.88
		#docs	**#snt pairs**	**#lang-1 words**	**#lang-2 words**
JRC3		**Hun/Van**	**Hun/Van (·10^3)**	**Hun/Van (·10^6)**	**Hun/Van (·10^6)**
en-et	Unique	5/173	63.5/8.4	0.80/0.28	0.73/0.22
	Common	23181	1247.3/1183.9	31.26/31.12	22.49/22.29
	Max-size	23181	1247.4	31.26	22.49
	Max-acc	22512	1084.5	18.27	20.00
en-lv	Unique	4/183	63.5/9.1	0.80/0.26	0.75/0.30
	Common	22560	1235.2/1175.8	30.84/30.77	25.34/25.10
	Max-size	22560	1235.3	30.84	25.34
	Max-acc	21975	1080.1	28.22	22.43
et-lv	Unique	3/54	63.5/3.4	0.73/0.06	0.75/0.14
	Common	22681	1293.7/1272.0	22.31/22.29	25.51/25.41
	Max-size	22681	1293.7	22.31	25.51
	Max-acc	22588	1242.3	21.67	24.44
de-en	Unique	4/83	66.1/3.7	0.84/0.11	0.80/0.08
	Common	23331	1272.7/1236.0	29.54/29.44	32.00/31.97
	Max-size	23331	1272.8	29.54	32.00
	Max-acc	22805	1189.9	27.98	30.70

Table 1. Results of processing the corpora: number and sizes of the documents in the common parts, documents present in just one corpus and the resulting max-size and max-accuracy combinations

	UT+JRC2	JRC3 en-et	JRC3 en-lv	JRC3 et-lv	JRC3 de-en
	UT/JRC	Hun/Van	Hun/Van	Hun/Van	Hun/Van
∅	7.1%/9.8%	12.2%/8.4%	11.7%/8.1%	3.9%/2.4%	6.1%/3.9%
0-1	0.0%/8.2%	0.0%/0.0%	0.0%/0.0%	0.0%/0.0%	0.0%/0.0%
1-0	32.5%/0.0%	0.7%/0.0%	0.7%/0.0%	0.0%/0.0%	0.3%/0.0%
1-1	59.3%/81.0%	86.8%/91.4%	87.3%/91.7%	95.8%/97.4%	93.0%/95.8%
N-M	1.0%/0.9%	0.1%/0.1%	0.1%/0.1%	0.2%/0.1%	0.4%/0.2%

Table 2. Frequency of the match types between sentence pairs of the corpora pairs; given as proportion of sentences per match type and corpus.

4.2. Influence on Machine Translation

Whenever it is known that two corpora overlap, concatenating them is an erroneous solution. As a result of straightforward concatenation the sentence pairs present in both parts of the overlap will be overrepresented since their relative frequency will increase in comparison to the sentence pairs outside the overlap or the ones that are present in only one corpus. The correct baseline method of combining overlapping corpora is taking the non-overlapping parts of both corpora and the overlapping part from just one of them. In our case instead of giving preference to either part of UT+JRC2 or JRC3 pairs we used both versions of the baseline, comparing them to the max-size and max-accuracy combinations of CorporAl.

Development and test sets were separated from the rest of the material, prior to processing the common parts. The size of both the dev and test sets was 2500 sentence pairs for all translation directions.

We evaluated the influence of the different corpora versions on two statistical translation systems: the first one is a phrase-based system, implemented in the Moses toolkit (Koehn et al., 2007) and the second one – hierarchical phrase-based, implemented in the Joshua toolkit (Li et al., 2009). Word alignment and language modeling for both systems were done with GIZA++ (Och and Ney, 2003) and SRILM (Stolcke, 2002). We used the BLEU (Papieni et al., 2001) and NIST (NIST, 2002) scores to compare the translation hypotheses.

The resulting scores of all the translation systems are presented in Table 3. In case of the UT+JRC2 pairs a clear pattern is visible: although in some cases the JRC-based results are better than the UT-based results, in general the max-accuracy, UT-based and JRC-based results are very similar and the max-size results noticeably exceed all three. The JRC3 pairs on the other hand do not exhibit any clear pattern. The scales of the differences suggest that there is no significant difference between all four systems in most cases.

Both of these opposite conclusions for UT+JRC2 and JRC3 experiments can be explained by the UT and JRC2 corpora being much more heterogenous than all the JRC3 pairs, as showed by the results of processing them, as well as by the UT+JRC2 max-size combinations being considerably bigger than the other parts and the JRC3 combinations being of the same size.

At the same time the max-accuracy results are roughly the same as the baselines in the UT+JRC2 case. Similarly, although (Kaalep and Veskis, 2007) showed Vanilla alignments to be of worse quality than HunAlign ones, there is no significant difference between the baselines of all the JRC3 pairs. This can be attributed to the frequency-based re-estimation of parameters in statistical machine translation, which results in automatic discarding of noise in the data (such as errors in sentece or word alignments) and thus also in lower sensitivity to alignment quality.

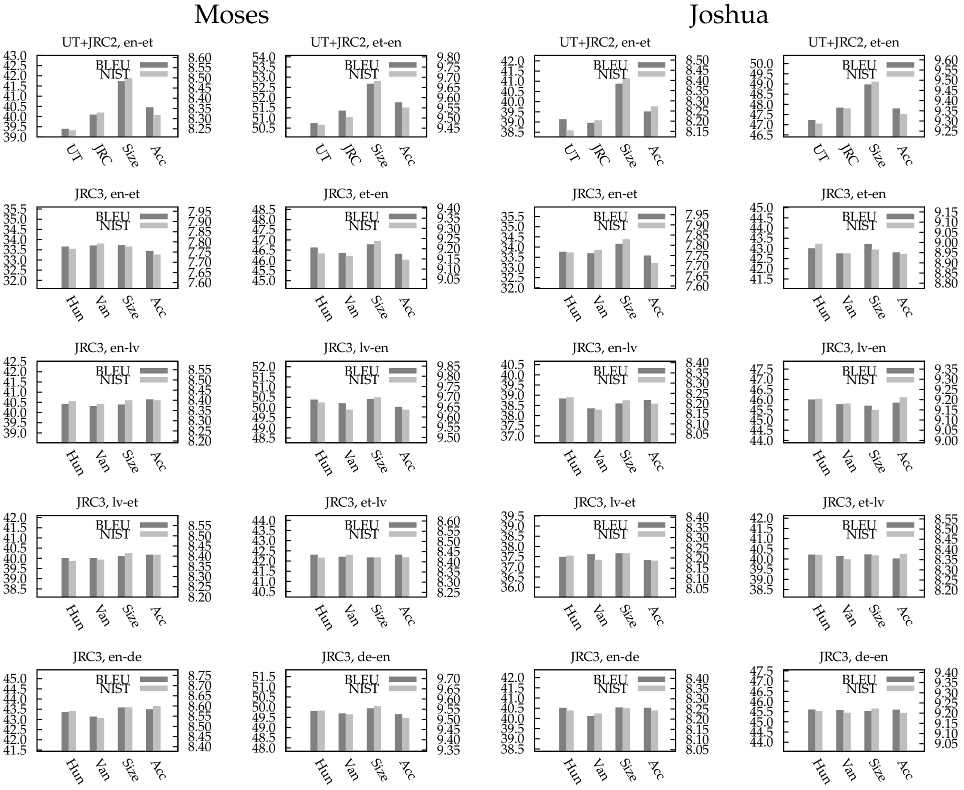


Table 3. Results of the machine translation experiments. The BLEU scale is on the left, and the NIST scale – on the right.

5. Conclusions

In this paper we have introduced a method for handling parallel corpora that are based on the same source material – i.e. overlapping parallel corpora. The method can detect matching and mismatching sentence pairs and omitted sentences. It can cope with minor differences in the text, such as typing errors and different notations. Also it can detect matches between several sentence pairs.

We described the CorporAl tool, which supports flexible combination of overlapping corpora and analysis of their similarities and differences.

The method was tested on two pairs of overlapping parallel corpora: the JRC-Acquis (version 2.2) with the corpus of the University of Tartu and the Vanilla and HunAlign-based versions of the JRC-Acquis (version 3.0); in the second case we included four language pairs. Processing the first pair resulted a bigger joint corpus while in case of the other four language pairs the size practically did not increase. Machine translation results showed dependence on the size and heterogeneity of the initial corpora and low sensitivity to alignment quality.

Bibliography

Bojar, Ondřej and Zdeněk Žabokrtský. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92, 2009.

Kaalep, Heiki-Jaan and Kaarel Veskis. Comparing parallel corpora and evaluating their quality. In *Proceedings of MT Summit XI*, pages 275–279, Copenhagen, Denmark, 2007.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand, 2005.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL'07*, pages 177–180, Prague, Czech Republic, 2007.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, 2009.

NIST. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Technical report, NIST, 2002.

Och, Franz J. and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Papieni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL'01*, pages 311–318, Philadelphia, PA, USA, 2001.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC'06*, pages 2142–2147, Genoa, Italy, 2006.

Stolcke, Andreas. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP'02*, volume 2, pages 901–904, Denver, Colorado, USA, 2002.

Varga, Daniel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of RANLP'05*, pages 590–596, Borovets, Bulgaria, 2005.

Address for correspondence:
Mark Fishel
fishel@ut.ee
University of Tartu, J. Liivi 2, 50409 Tartu, Estonia