Bulgarian X-language Parallel Corpus

Svetla Koeva, Ivelina Stoyanova, Rositsa Dekova, Borislav Rizov, Angel Genov
Department of Computational Linguistics, Institute for Bulgarian, Bulgarian Academy of Sciences
52, Shipchenski Prohod Blvd, Building 17, Sofia 1113, Bulgaria
E-mail: svetla@dcl.bas.bg, iva@dcl.bas.bg, rosdek@dcl.bas.bg, boby@dcl.bas.bg, angel@dcl.bas.bg

Abstract

The paper presents the methodology and the outcome of the compilation and the processing of the Bulgarian X-language Parallel Corpus (Bul-X-Cor) which was integrated as part of the Bulgarian National Corpus (BulNC). We focus on building representative parallel corpora which include a diversity of domains and genres, reflect the relations between Bulgarian and other languages and are consistent in terms of compilation methodology, text representation, metadata description and annotation conventions. The approaches implemented in the construction of Bul-X-Cor include using readily available text collections on the web, manual compilation (by means of Internet browsing) and preferably automatic compilation (by means of web crawling – general and focused). Certain levels of annotation applied to Bul-X-Cor are taken as obligatory (sentence segmentation and sentence alignment), while others depend on the availability of tools for a particular language (morpho-syntactic tagging, lemmatisation, syntactic parsing, named entity recognition, word sense disambiguation, etc.) or for a particular task (word and clause alignment). To achieve uniformity of the annotation we have either annotated raw data from scratch or transformed the already existing annotation to follow the conventions accepted for BulNC. Finally, actual uses of the corpora are presented and conclusions are drawn with respect to future work.

Keywords: parallel corpora, corpora construction, annotation

1. Introduction

The paper outlines the results of the compilation and the processing of the Bulgarian X-language Parallel Corpus (Bul-X-Cor)[1] – part of the Bulgarian National Corpus (Koeva et al., 2010). The Bulgarian National Corpus (BulNC) is a publicly available constantly enlarged corpus focused on Bulgarian (currently comprising 469.5 million tokens). It is designed as a uniform framework for texts of different modality (written – spoken), period, and number of languages (either Bulgarian monolingual or parallel where one of the counterparts is Bulgarian). Any X-languages in the corpus are equally treated with respect to the metadata description scheme, preprocessing and annotation, data storage format and access.

The paper presents briefly the structure and the content of the Bulgarian X-language corpus in the context of the previously developed parallel corpora where Bulgarian is one of the languages. The focus is set on building representative parallel corpora which include a diversity of domains and genres, reflect the relations between Bulgarian and other languages and are consistent in terms of compilation methodology, representation of texts, metadata description and annotation conventions. Some levels of annotation applied to Bul-X-Cor cover all languages (sentence segmentation and sentence alignment), while others depend mainly on the availability of tools for a particular language – in any case theannotation is harmonized with the one adopted in the BulNC.

The BulNC is compiled mainly for the purposes of computational research and implementations, and the same is the function of the parallel corpora within. The provided access includes not only query searching through a web interface but a corpus collocation web service as well.

2. An Overview of Parallel Corpora Including Bulgarian

The quality and applicability of parallel corpora can be assessed using a number of criteria including target languages, size, variety of domains, levels of processing and linguistic annotation, availability and terms of access. In recent years several corpora were developed focused particularly on Slavic and Balkan languages.

The parallel corpora including Bulgarian are relatively small and usually domain specific – they contain mostly literary or administrative texts. Examples of literary corpora are: the **Multext-East corpus** (Dimitrova et al., 1998) comprised by versions of George Orwell's novel *1984* in six languages; the **SEE-ERA.net Literary Corpus** (Tufiş et al., 2009) consisting of Jules Verne’s novel *Around the world in 80 days* translated in 16 languages; the extended **RuN-Euro Corpus** (Grønn & Marijanovic, 2010) including a small Bulgarian part of 271,000 tokens and the **ParaSol**, known as the **Regensburg Parallel Corpus** (Waldenfels, 2006) with a Bulgarian part of 2 million tokens.

Examples of administrative corpora which include Bulgarian are the **SEE-ERA.NET Administrative Corpus** (Tufiş et al., 2009) with 1.4 million tokens for Bulgarian; the **EuroParl Corpus** (Koehn, 2005) with 6 million tokens for Bulgarian; the **JRC-Acquis** (Steinberger, 2006) in which the Bulgarian counterpart is the smallest represented with 16.1 million tokens compared to 22+ million for other languages. To the best of our knowledge, the **Bulgarian-Polish-Lithuanian Corpus** (Dimitrova et al., 2009) is the only parallel corpus with a Bulgarian part (about 300,000 tokens) that attempts to combine texts from more than one domain – administrative and fiction.

The **OPUS collection** (Tiedemann, 2009) offers access to

[1] X-language should be understood as “not restricted to a particular number of languages” and should be read as “Ex-language”.

several aligned parallel corpora with a Bulgarian part: **medical documents by EMEA** (14.7 million tokens for Bulgarian), **film subtitles** (276.6 million tokens for Bulgarian) and the **SETimes news corpus**. All of the corpora discussed are tokenised, POS tagged and aligned at the sentence level, and some of them are also lemmatised.

The brief overview of the existing parallel corpora with respect to Bulgarian reveals that the resources are limited in terms of volume, variety and levels of processing. Most of the existing corpora represent generally administrative and literally texts and they are built from the available texts on the Internet, rather than being compiled on a planned strategy for developing a balanced and representative parallel corpus. Moreover, not all corpora are freely accessible and for some of them there is no availability information at all. Last but not least, the existing parallel corpora have different type of annotation exploiting different conventions and standards which impedes to a great extent computational linguistic research. Therefore, our goal is to build a large (by means of size and number of languages), balanced and representative (embracing as much as possible different domains), annotated (accumulating different layers of annotation consistent with standards and adopted annotation conventions), accessible for computational research and implementations X-language parallel corpus, centred around Bulgarian.

3. Compilation of the Bulgarian X-language Parallel Corpus

Three basic approaches are implemented in the compilation of the Bulgarian X-language parallel corpus: (1) using readily available text collections, (2) manual compilation (by means of Internet browsing) and preferably (3) automatic compilation (by means of web crawling).

The freely available collections of parallel texts including Bulgarian such as OPUS are reused. As much metadata as possible are extracted from the texts which are further redesigned into the unified format of the X-language corpus. Manual collection of texts is applied in limited cases for small in number but large in size documents when development of a focused crawler is deemed inefficient. These include mainly fictional texts where parallel translations are usually not on the same website and require manual search and download.

The automatic approach for collecting parallel corpora which we adopt is well known and widely used. Tsvetkov and Wintner (2010) describe a similar way for harvesting parallel texts: the candidate sites are manually detected, and then automatically monitored over time. Thus a Hebrew-English parallel corpus is compiled, containing articles on news, politics, sports, economics, literature, etc. by performing a daily crawl of web sites with dynamic contents. A simple script which cleans downloaded web pages from HTML tags and extracts plain text and metadata (date, domain, source URL, etc.) is applied. In addition to this we aim at keeping all editorial metadata for future use. Further, the authors report on a language independent content-based method for identification of parallel articles, by which two documents E and H are defined as mutual translations, if E contains enough translated terms from H and vice versa. According to their evaluation results they have obtained 100% precision and 86.5% recall (threshold values were chosen to favor precision over recall, since the quality of the corpus was crucial for them).

Another similar approach is taken also by Aziz & Specia (2011) who use GNU wget3 and a URL template to download HTML pages from a magazine website. For a given issue they first download its index, a single page containing links to articles, and then parse those links and download the actual articles. Issues are identified by a sequential number that is consistent across the different versions of the website (original in Portuguese, English and Spanish). Due to lack of obvious correspondence between articles' identifiers, content-based document alignment techniques are then applied.

The automatic compilation of the Bulgarian X-language Parallel Corpus involves the stages outlined below.

3.1 Automatic web mining

As parallel resources involving Bulgarian are limited on the web (with the exception of Bulgarian-English counterparts), efficiency of web crawling was an issue overcome by targeting (either automatically or manually) the appropriate resources. Furthermore, the structure of source webpages is analysed and applied in the crawling design by involving links traversal algorithm or URL templates of parallel texts for each source. There has been research into fully automatic web mining and parallel text collection. Among others STRAND (Resnik, 1998, 1999) is a language-independent system for automatic discovery of parallel translations on the web. At present the automatic web mining is exceeded by the manual one and its potential is yet to be explored. We plan to investigate, develop and test some techniques for automatic web mining based on partial word-to-word alignment and text similarity analysis as well as bilingual resources.

3.2 Manual web mining

The manual web mining ensures the high quality of the results in terms of validity of documents and parallel correspondence. It also improves essentially the efficiency of the crawling process. In most cases the websites containing parallel texts are very large (e.g. http://eur-lex.europa.eu) and general (non-focused) crawling would need to deal with several times more documents. Moreover, only documents which have a translational correspondence in Bulgarian are targeted and collecting Bulgarian texts first ensures that only such documents are included.

3.3 Development of a general web crawler

A general form of the crawler was designed which is then used as a base for developing website specific crawlers. Several crawling algorithms were examined (Paramita et al., 2011). The main technique applied is the Breadth-First algorithm (Pinkerton, 1994). Some improvements of the algorithm were considered but due to the manual web mining and URL seeding they were not deemed necessary at present. The crawler starts at the initial webpage of the respective collection of documents and either recursively harvests the links until the relevant pages containing the documents are reached, or uses URL templates to directly access the relevant pages.

3.4 Development of a focused (website specific) crawler

The general crawler was adapted to the structure of each available source site with parallel texts according to the manual web mining. The focused crawler either directly implements the link harvesting technique or uses a set of URL templates specific for the particular source website. Some corpora are static and require a single run of the crawler while others are dynamic (e.g. news websites, http://setimes.com) and require weekly or monthly crawls.

On the one hand focused crawling with preceding web structure mining employed ensures the high quality of the results and on the other hand considerably reduces the number of visited links and improves efficiency. The focused crawler also ensures the extracted documents are relevant by selecting only texts which have Bulgarian counterparts.

3.5 Extraction of plain text and metadata

Original documents are in HTML format, rarely XML. These contain the text and some metadata which is extracted, processed and further added to the corpus description.

The Bulgarian National Corpus and the X-language parallel corpora are supplied with extensive metadata description compliant with the well established standards (Burnard, 2005; Adolphs & Lin, 2010). Metadata comprise 25 fields providing editorial information – e.g. source, year of publishing, etc., and classificatory information – e.g. style, primary and secondary domains if present and genre. Domain and genre are considered style specific. Although the system of classificatory categories is open to new additions, the aim is to use a limited number of well defined categories and to provide some additional information in a separate field ‘note’ in the description. Table 1 presents the number of languages, domains and genres across categories and styles at present.

Metadata are mostly derived automatically using two main techniques: extracting information from the HTML markup of the original files and by simple keywords-based heuristics. HTML pages usually contain editorial information such as author, title, publishing date, specifically tagged which makes it easily extractable from the source HTML page.

Category and Style	**Number of domains**	**Number of genres**	**Number of languages**
Administrative	11	16	23
Science/Administrative	21	16	19
Massmedia	19	12	9
Fiction	13	25	4
Informal/Fiction	17	1	29
Subtitles	21	14	2

Table 1: Distribution of languages, domains and genres across styles in parallel corpora.

Source webpages contain similar texts and thus focused website specific crawling facilitates the provision of additional classificatory information such as domain and genre which is rarely specified in the documents.

When classificatory information is not directly available some heuristics are applied to determine the domain and genre of the text using lists of domain specific or genre descriptive keywords to match in the title. For example, if the title of a text contains a genre word (e.g. *report*), it is assumed to denote the genre of the document.

3.6 Filtering of collected documents

Much attention is paid to the quality of the multilingual resources as it influences greatly their applicability for various research purposes (Resnik, 1999; Paramita et al., 2011, etc.). Some procedures verifying the quality of the results from the automatic crawling in terms of validity of the documents are implemented. As a result the documents considered not appropriate are removed. These include very short documents (usually notification pages – e.g. about missing documents, or redirection pages), documents in invalid language or encoding (very rare errors) and repeating texts. Some additional checks whether documents are parallel are also possible at this stage although such tests are not performed at the moment. The primary task of the manual web mining was to ensure that only parallel texts are collected and the results proved to be of high quality.

3.7 Structuring of the corpus

The last stage involves fitting the newly collected parallel corpora into the unified framework of BulNC both in terms of text classification (structure) and metadata. Each X-language subcorpus of BulNC is stored in a separate directory which mirrors the structure of the Bulgarian part of BulNC. The metadata of Bulgarian parallel texts contain information about the additional languages in which the text is available.

4. General overview of the Bulgarian X-language corpus

At present there are texts in 33 languages included in the parallel corpus. However, the languages are not equally represented: the largest parallel corpus is the Bulgarian-English (201.7 million tokens and 6,098,610 sentences for Bulgarian and 204.2 tokens and 6,688,511 sentences for English), there are 5 other corpora between 100 and 200 million words, 16 parallel corpora of size in the range 30-52 million tokens, further 7 in the range 1-10 million tokens, and the rest are below 1 million, with the smallest corpora being the Chinese, Japanese and Icelandic with less than 50,000 tokens per language.

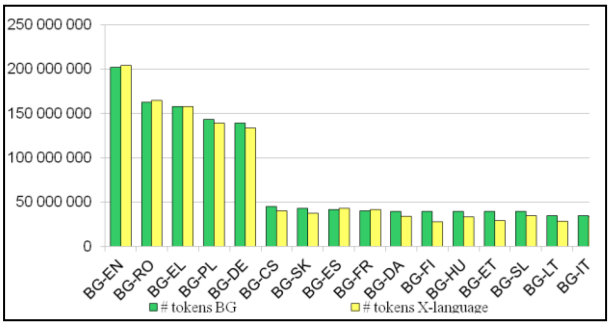


Figure 1 shows the largest parallel corpora.

Fig. 1 Distribution of tokens by pairs of languages in the largest parallel corpora

The data given above indicates the general need to work towards balancing the resources and offering a wider diversity of texts. The imbalance is even more evident for the other language pairs in the corpus. However, the targeted balance is not to be measured merely by the size of the different domains covered (this is not possible as the domains are naturally imbalanced with respect to their usage). Instead, our aim is to cover as many domains, as there are available on the Internet, to the extent similar to their usage. In this respect the structure of the BulNC is flexible to incorporate different categories and domains, while the balance of the parallel corpora will be ensured by variety of styles, genres, and domains in a ratio that reflects both the availability and frequency of usage.

5. Structure of the Bulgarian X-language Parallel Corpus

Resnik and Smith (2003) admit that “even for the top handful of majority languages, the available parallel corpora tend to be unbalanced, representing primarily governmental or newswire-style texts”. Although the web has largely expanded since 2003, the three problems they have reported – too few languages, too little data, difficulty in dissemination – yet remain to be solved. Thus for the moment the main issues concerning representativeness are concentrated over the aim to collect as much as possible parallel to Bulgarian texts in a wide range of languages. At present the Bulgarian X-language Parallel Corpus consists of the following subcorpora.

5.1 Administrative: EU Law Documents in 23 Languages

The subcorpus is collected and compiled automatically from the online repository (http://eur-lex.europa.eu) and contains EU law texts in 23 European languages. The corpus includes all the accessible texts in Bulgarian and their respective counterparts in the remaining languages. The metadata description is created automatically as the editorial information and part of the classificatory information is extracted from the HTML source code of each page and the rest is derived by heuristics. The subcorpus comprises texts created between the years 1958 and 2011 and 79.6% of the texts are administrative acts, published after 2000. The Bulgarian-English part contains 58,364 texts, which total in 2,531,544 sentences and 143,791,703 words in Bulgarian and 2,521,994 sentences and 142,744,872 words in English. Average words per sentence in Bulgarian administrative texts is 56.8 and in English - 56.6.

5.2 Science / Administrative: Healthcare

The subcorpus consists of administrative texts, published by the European Medicines Evaluation Agency (EMEA). It is part of the OPUS collection (Tiedemann, 2009). The corpus was downloaded, the texts were processed into the format of the parallel corpora and the metadata were derived automatically. The resulting corpus includes texts created and published by EMEA in the years between 1978 and 2009. The texts in the subcorpus are grouped in three thematic categories – human medicine (77.5% of all texts), veterinary medicine (13.6%) and general texts (8.9%). The Bulgarian-English part contains 1587 texts, which amount to 580,011 sentences and 12,586,236 words in Bulgarian and 452,808 sentences and 9,735,375 words in English. Average sentence length is 21.7 for Bulgarian and 21.5 for English.

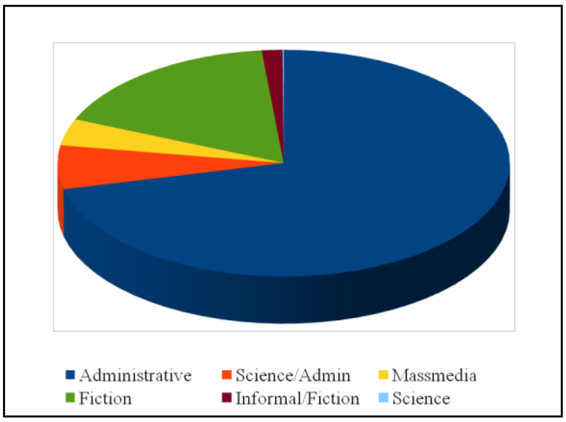


Fig. 2 Domain distribution in the Bulgarian-English Parallel Corpus by number of words

5.3 Massmedia: News in 8 Balkan Languages and English

The subcorpus contains news, as well as some other journalistic texts published since October 2002 on the East Europe information website. It is collected automatically through crawling of the archive of the website. The metadata are also automatically compiled. The corpus is dynamic and is enlarged on a monthly basis by regular crawls of the website. The news are in eight Balkan languages and English. The original language of the texts and the directions of the translation are not indicated. The metadata description of the texts is obtained automatically together with the extraction of the text alone from the HTML file. The subcorpus includes texts from the following genres: news, article, review, and blog. The Bulgarian-English part consists of 35,312 texts, which amount to 7,666,229 words and 304,216 sentences in Bulgarian and 7,965,405 words and 355,418 sentences in English. Average length of sentences is 25.2 words for Bulgarian and 22.4 for English.

5.4 Fiction: Texts in Bulgarian, English, German, and French

The Fiction subcorpus is compiled manually. Many of the Bulgarian texts are taken from BulNC, and others are collected in the process of compilation using various sources – freely available texts on the Internet, scanning, and author's donations. So far, the Fiction subcorpus consists of texts in Bulgarian, English, German, and French but it is also planned to contain documents in other languages. The source language and the direction of translation are not fixed, i.e. the corpus contains translations from English into Bulgarian and translations from a third language into both English and Bulgarian. All texts are classified according to their domain: *Adventure, Biography, Children, Detective, Love, Thriller, General, Horror, Humour, Science Fiction, and Fantasy*. The Bulgarian-English part of the Fiction subcorpus consists of 680 texts, which amount to a total of 39,590,472 words and 2,788,063 sentences for English and 34,553,474 words and 2,262,190 sentences for Bulgarian. Average sentence length is 15.3 for Bulgarian and 14.2 for English.

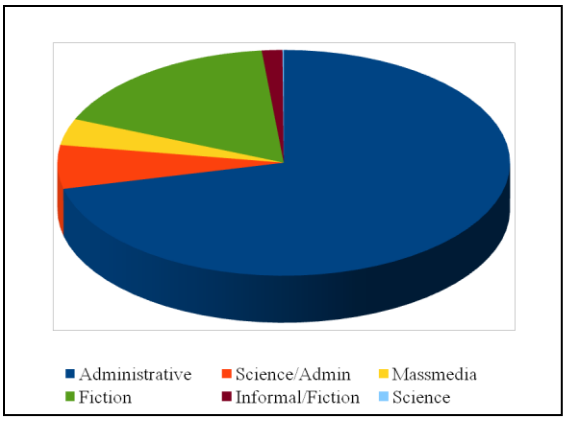


Fig. 3 Domain distribution in the Bulgarian-English Parallel Corpus by number of sentences

5.5 Informal: Subtitles

The subcorpus consists of subtitles of films, documentaries, and animations. It is part of the OPUS collection (Tiedemann, 2009). The texts are downloaded and undergo the same processing as the Healthcare administrative corpus (cf. 4.2) The Bulgarian-English part consists of 497 texts, which amount to 2,970,974 words and 412,635 sentences in Bulgarian and 3,931,784 words and 561,684 sentences in English. Average sentence length is 7.2 words for Bulgarian and 7.0 for English.

5.6 Science

The subcorpus comprises 88 parallel texts of the bulletin of Bulgarian Academy of Sciences since 2004. Texts are in two languages – Bulgarian and English. They were automatically collected from the webpage of the Academy. This is a relatively small corpus containing 8,014 sentences and 174,113 words in Bulgarian and 8,544 sentences and 189,276 words in English with the average sentence length being 21.7 words in Bulgarian and 22.2 in English.

6. Levels of annotation

The Bulgarian-English parallel sub-corpus is supplied with annotation on various levels while the higher level of annotation for other languages has just started. Nevertheless the annotation is uniform and consistent with the standards accepted in the BulNC (Koeva et al, 2010). For the annotation of Bulgarian texts Bulgarian language processing chain is applied. It includes a number of tools (regular expression based sentence splitter and tokeniser, SVM POS tagger, dictionary-based lemmatiser, finite-state chunker, and wordnet senses annotation), designed to work together to ensure interoperability, fast performance and high accuracy.

Apache OpenNLP (http://incubator.apache.org/opennlp/) with pre-trained models is exploited for the English texts annotation - sentence segmentation, tokenisation and POS tagging. OpenNLP is highly flexible and adjustable and could be trained and applied for other languages as well. There are also some pre-trained models for a number of popular languages (German, Spanish, etc.). The OpenNLP annotation conventions for English were transformed in accordance with those adopted in the BulNC. Uniformity of the annotation is achieved in two ways – either by annotation of raw data from scratch or by transforming of already existing annotation. In each case the conventions accepted for the BulNC are followed. Lemmatisation of English texts is performed using RASP – processing system for English (Briscoe et al., 2006).

The identification of the text components (tokens, sentences, paragraphs) is a relatively trivial task usually handled by simple methods such as scripts of regular expressions (Mikheev 2003). The automatic identification of sentence borders is based on regular rules possibly complemented by lexicons of abbreviations after which there must be or there might be a capital letter, a number, etc. in the middle of the sentence. Although the sentence splitters might exploit language specific knowledge for word and sentence graphical structure the general rules are applicable for the alphabetic languages.

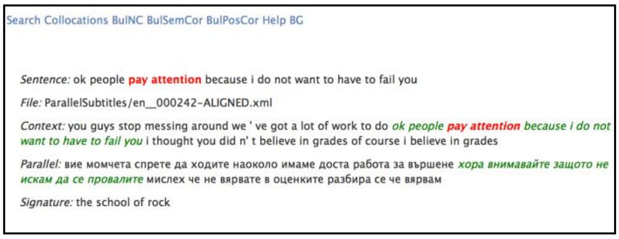
Sentence splitting is performed on Bulgarian and English as the Bulgarian-English corpus has already had several applications for research purposes. The task of sentence splitting of Bulgarian is carried out using a specifically designed tool, part of the Bulgarian language processing chain. For sentence splitting of English texts we use

OpenNLP and a pre-trained model. There are several other OpenNLP sentence splitting models already available – e.g. for Danish, German, and Portuguese, and we plan at training more models on demand.

High quality sentence segmentation is very important for good alignment. The vast majority of the errors in the alignment was related to errors in the sentence segmentation. The sentence alignment relies on HunAlign (Varga et al., 2005) which is based on Gale-Church algorithm using sentence length information. HunAlign can exploit bilingual dictionaries in order to improve results. It uses texts with segmented sentences, dictionaries if available, and outputs a sequence of sentence pairs (bisentences). The produced output contains the alignment along with a number showing the alignment "confidence" i.e. the certainty of the alignment. Main reasons to choose the HunAlign are: it performs well without the assistance of external resources, it works considerably faster comparing to other aligners, i.e. Bilingual Sentence Aligner (Moore, 2002) and shows better results. For example the test results of the HunAlig and Bilingual Sentence Aligner over manually aligned Jules Verne's novel *Around the world in 80 days* (Tufiş et al., 2009) are: precision 96.54, recall 97.48, F1 score 97 for HunAlign versus precision 96.84, recall 89.24, F1 score 92.88. As the F1 score of HunAlign is higher we plan at using it for aligning parallel texts in any pair of languages.

7. Access and use of Bul-X-Cor

The BulNC search system (http://search.dcl.bas.bg) suitable for corpora investigations and lexicographic research is designed to support monolingual and parallel corpora in a uniform way. For a given query the system retrieves matches in all documents irrespectively of the language. Due to the alignment the corresponding sentences in parallel documents are also accessible. The results are paginated and the matches are highlighted. The user can view the detailed information for a given sentence in the result set. That includes the sentence metadata, its context and correspondence in the other languages.



Sentence: ok people pay attention because i do not want to have to fail you

File: ParallelSubtitles/en__000242-ALIGNED.xml

Context: you guys stop messing around we ' ve got a lot of work to do *ok people pay attention because i do not want to have to fail you* i thought you did n' t believe in grades of course i believe in grades

Parallel: вие момчета спрете да ходите наоколо имаме доста работа за вършене *хора внимавайте защото не искам да се провалите* мислех че не вярвате в оценките разбира се че вярвам

Signature: the school of rock

Fig. 4 Result for the ordered query <pay attention>

The designed query language (DQL) is implemented (Tinchev et al., 2007). As compared to the CQL (Christ and Schulze, 1994), DQL supports terms: word – i.e. word; feature – i.e. *{POS=A POS=ADV}, relation – i.e. word/F/, and their combinations – i.e. word/S/{POS=N}. It is not restricted to a predetermined set of relations – at the moment queries for word forms, synonyms, hypernyms, and *similar to* adjectives are allowed, but other wordnet relations can be provided as well. The atomic formulae support both ordered and unordered queries, the later being appropriate for matching adjacent constituents with free word order, e.g. verbal clitics in Slavic languages. The DQL is recursive and all Boolean combinations of formulae are formulae. This for example allows disjunction of ordered queries, i.e. searching for paraphrases, synonymous idioms and so on. The system supports queries with regular expressions as well. The visualisation of aligned sentences in parallel documents and the metadata information for the results are handled uniformly.

The Corpus Collocation web service gives also access to the Bulgarian National Corpus. The Corpus Collocation service employs the free of charge NoSketchEngine, a system for corpora processing that combines Manatee and Bonito (Rychlý, 2007). The Collocation service is a RESTful webservice, supporting complicated queries through http. For example the query http://dcl.bas.bg/collocations/?cmd=collocations&word=cat&cbgrfns=3td returns statistical significance calculated with MI3, T-score, and logDice. The query returns the collocations of a given word in the NoSketchEngine format. The system also supports additional arguments, namely all that are accepted by NoSketchEngine, provided with default values. The main purpose of the Bul-X-Cor is computational research and implementations – obtaining various types of statistical data, building language models, etc. with view to applications in the field of machine translation, extraction of multilingual information, compilation of multilingual lexical resources, etc.

8. Conclusion and future work

The parallel corpora have already been used for particular research projects among which the study of translational asymmetries between Bulgarian and English, or the development and testing of new language independent methods for clause alignment.

Development of more efficient and productive techniques for web crawling and compilation of parallel corpora is one of the main directions of future work. It is also necessary to consider more general but yet successful approaches for focused crawling. We also aim at establishing good practices for evaluation of parallel multilingual resources and ensuring high quality and good coverage of the corpus. As balance and representativeness of corpora are their key features, our work focuses on exploring their qualitative and quantitative characteristics and the ways to improve them. Moreover, we work towards ensuring the high quality and diversity of linguistic annotation at all levels and for different languages.

9. Acknowledgements

The research work described in the paper was partially supported by CESAR (Central and South-East European Resources) – a European project funded under the CIP ICT Policy Support Programme, Grant Agreement 271022; and the Project Multilingual Parallel Corpora – In Aid of Contemporary Language Technologies (MU03), financed by the Program “Young Researchers” of the Bulgarian National Science Fund.

10. References

Adolphs, S., Lin, P. (2010). Corpus Linguistics. In J. Simpson (ed.) *The Routledge Handbook of Applied Linguistics*, Routledge, pp. 597--610.

Aziz, W., Specia, L. (2011). Fully Automatic Compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora. 8th Brazilian Symposium in Information and Human Language Technology (STIL-2011), Cuiaba, Brazil.

Briscoe, E., Carroll, J. and Watson, R. (2006). The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia, pp. 77--80.

Burnard, L. (2005). Metadata for Corpus Work. In M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books.

Christ, O., Schulze, B. M. (1994). *The IMS Corpus Workbench: Corpus Query Processor (CQP) User’s Manual*. University of Stuttgart, Germany.

Dimitrova L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H.J., and Tufis, D. (1998). Multext-East: parallel and comparable corpora and lexicons for six Central and Eastern European languages. In *Proceedings of COLING-ACL 1998*, Montreal, Canada: Morgan Kaufmann Publishers, pp. 315--319.

Dimitrova, L., Koseska, V., Roszko, D., and Roszko, R. (2009). Bulgarian-Polish-Lithuanian Corpus – Current Development. In *Proceedings of the International Workshop “Multilingual resources, technologies and evaluation for Central and Eastern European languages” in conjunction with International Conference RANPL’2009. Borovec, Bulgaria, 17 September 2009*, pp. 1--8.

Grønn, A., Marijanovic, I. (2010). Russian in Contrast: Form, meaning and parallel corpora. In *Oslo Studies in Language (OSLa), 2*(1), pp. 1--24.

Koehn, Ph. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation, In *Proceedings of MT Summit*, pp. 79--86.

Koeva Sv., Blagoeva, D., Kolkovska, S. (2010). Bulgarian National Corpus Project. In *Proceedings of LREC2010*, Valletta, ELRA, pp. 3678--3684.

Mikheev, A. (2003). Text segmentation. In R. Mitkov, (ed.) *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, pp. 376--394.

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Stephen D. Richardson (Ed.) Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (AMTA '02).* London, UK: Springer-Verlag, pp. 135--144.

Paramita, M., Aker, A., Gaizauskas, R., Clough, P., Barker, E., Mastropavlos, N., and Tufis, D. (2011) Report on methods for collection of comparable corpora, ACCURAT - Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation.

Pinkerton, B. (1994). Finding what people want: Experiences with the WebCrawler. In *Proceedings of the First World Wide Web Conference, Geneva, Switzerland, 1994*.

Resnik, Ph. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, In D. Farwell, L. Gerber, and E. Hovy (eds.), *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA '98)*, London, UK: Springer-Verlag, pp. 72--82.

Resnik, Ph. (1999). Mining the Web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 527--534.

Resnik, P., Smith, N.A. (2003). The Web as a parallel corpus. *Computational Linguistics* 29 (3), pp. 349--380.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 65--70.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (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)*, pp. 2142--2147.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.) *Recent Advances in Natural Language Processing* (vol. V), Amsterdam/Philadelphia: John Benjamins, pp. 237--248.

Tinchev, T., Koeva, Sv., Rizov, B., Obreshkov, N. (2007). System for advanced search in corpora. In: *Literature and writing in Internet*, Sofia: St. Kliment Ohridski University Press, pp. 92--111.

Tsvetkov, Y., Wintner, S. (2010). Automatic Acquisition of Parallel Corpora from Websites with Dynamic Content. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC-2010)*, pp 3389--3392.

Tufis, D., Koeva, Sv., Erjavec, T., Gavrilidou, M. and C. Krstev (2009). ID10503 Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In *Scientific results of the SEE-ERA.NET Pilot Joint Call, Vienna*, pp. 37--48.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages In Proceedings of the RANLP 2005, pp. 590-596.

Waldenfels, R. (2006). Compiling a Parallel Corpus of Slavic Languages. Text strategies, Tools and the Question of Lemmatization in Alignment. In B. Brehmer, V. Zhdanova, R. Zimny (eds.) *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, München: Sagner, pp. 123--138.