Sentence-Alignment and Application of Russian-German Multi-Target Parallel Corpora for Linguistic Analysis and Literary Studies

DESISLAVA ZHEKOVA, ROBERT ZANGENFEIND, ALENA MIKHAYLOVA, TETIANA NIKOLAIENKO
Ludwig-Maximilians-Universität München

Abstract
This paper presents the application of multi-target parallel corpora consisting of a single source text and multiple target translations of it for linguistic analysis. We discuss the alignment, interactive search and visualization of this type of data within a specific tool called ALuDo (Alignment with Lucene for Dostoyevsky). This is a Java implementation that uses local grammars, ontological information, bilingual dictionaries and statistical approaches for alignment and search. The data set in use is the Russian novel *Crime and Punishment* by Fyodor Dostoyevsky and three German translations of it. With this bilingual corpus quite a number of investigations in the field of linguistics and of literary studies are possible. Additionally, we release part of the resulting parallel corpus. **Keywords:** interactive alignment; rule-based alignment; statistical alignment; coreference resolution; paraphrase identification.

Resumo
Este artigo apresenta a aplicação de *corpora* multialvo paralelos – compostos por um único texto-fonte e múltiplas traduções-alvo desse texto – para análise linguística. Discute-se o alinhamento, busca interativa e visualização deste tipo de dados usando uma ferramenta específica chamada ALuDo (Alinhamento com Lucene para Dostoievski). Trata-se de uma aplicação Java que utiliza gramáticas locais, informação ontológica, dicionários bilingues e abordagens estatísticas para alinhamento e pesquisa. O conjunto de dados utilizado é constituído pelo romance russo *Crime e Castigo* de Fiodor Dostoievski e três traduções do romance em alemão. Com este corpus bilingue é possível levar a cabo investigação significativa no campo da linguística e dos estudos literários. Adicionalmente, publicamos parte do *corpus* paralelo resultante. **Palavras-chave:** alinhamento interativo; alinhamento baseado em regras; alinhamento estatístico; resolução de correferência; identificação de paráfrase.

1. Introduction

In the last decade, parallel corpora have become an important resource for a wide range of Natural Language Processing (NLP) tasks. As Wetzel and Bond (2012) note, large and qualitative parallel corpora are a vital factor for achieving good translations produced by Statistical Machine Translation (SMT) systems. Lefever *et al.* (2011) use parallel corpora to derive

MATLIT 4.1 (2016): 45-61. ISSN 2182-8830
http://dx.doi.org/10.14195/2182-8830_4-1_3

word senses from word alignments for their multilingual classification-based approach to Word Sense Disambiguation. For this application, however, it is sufficient to make use of general bilingual corpora – collections of source texts and their translations. Another combination of data are the so-called monolingual parallel corpora which can be achieved by aligning multiple translations of a given text in one language (cf. examples (1) and (2)).

(1) Emma burst into tears and he tried to comfort her, saying things to make her smile.
(2) Emma cried, and he tried to console her, adorning his words with puns.

A combination of general bilingual and monolingual parallel corpora results in the data that we are interested in our work – bilingual multi-target parallel corpora (a given collection of texts and more than one (thus multi-target) aligned translations of it, shown in the lower part of figure 1).

Yet, bilingual multi-target parallel corpora are significantly scarce, if at all existent. Thus, one major contribution of our work is the release of a corpus for the Russian-German language pair, for which no such freely available data is accessible. This type of data often results in smaller data sets than traditional bilingual corpora and thus its use has been considerably neglected with respect to most NLP tasks so far. However, a number of areas can highly profit from the richness of information in it, which we also aim to underline within this discussion. Since some fields of study require more than simply aligned data (e.g. Contrastive Language Studies and Literary Studies), we include the released corpus into the interactive alignment tool ALuDo (also made freely available, see section 4) for visualizing concrete areas of interest for linguistic and literary analysis (cf. figure 1).

The rest of this paper is structured in the following way: In section 2, a motivation for the use of bilingual multi-target parallel corpora for linguistic and literary analysis is given. Then, we provide an overview of previous approaches to alignment of parallel corpora (section 3) after which we delineate the system setup we make use of to statistically align the data set of a Russian-German novel and its three translations (section 4). In section 5, we discuss the resulting alignation and in section 6 an appropriate conclusion is provided.

Figure 1. Interactive alignment of bilingual multi-target parallel corpora.

2. Application of Parallel Corpora with Multiple Target Texts

Parallel corpora with multiple target texts are often avoided by the research community due to the scarcity of such data. However, this type of corpora can be highly useful for a number of linguistic applications. This section is devoted to three such areas and aims to underline the importance of such corpora to the enhancement of these tasks.

2.1. Coreference Resolution

Coreference Resolution (CR) aims to identify all phrases in a text that refer to the same discourse entity in it and group them in equivalence classes (also called entities or coreference chains). In the last decade, multilinguality has become an important factor for CR (Pradhan *et al.*, 2012; Zhekova, 2013). Thus, the interest for parallel corpora for CR has also drastically increased (Kobdani *et al.*, 2011). However, all these approaches made use of bilingual data mostly with respect to projecting coreferential chains from one language to another. Aligned bilingual multi-target corpora can be used to extend this purpose. In fact, they have not yet been made use of for CR at all, although, as we will show, they can be highly useful for the enhancement of discourse entity representations.

Most of the state-of-the-art statistical approaches to CR use the mention-pair model (Rahman and Ng, 2009) to represent coreferential relations (Pradhan *et al.*, 2012). World knowledge is vital for these approaches, since for the last decades the feature sets in use, representing information about the mention pair, have been mainly covering features in various groups (e.g. lexical, grammatical, positional, semantic, etc.), which largely excluded world knowledge.

In fact, recently released large CR datasets, which aimed at providing exact world knowledge, such as the Wikipedia Links Corpus (WLC) (Singh *et al.,* 2011), also fail in collecting extensive world knowledge. This is the case, because WLC was collected in such a way that mentions with large string edit distances are discarded in order to avoid noise in the data. However, such a case is for example the pair President – Barack Obama. Excluding such mentions results in trivial pairs with large string overlaps, such as President Obama – Barack Obama or Obama – Barack Obama. However, lexical features used in the majority of the state-of-the-art CR systems, already monitor, exactly, string overlap between the mentions. The pair President – Barack Obama is thus a lot more important to these systems because it is a pair of mentions with low string overlap.

Our hypothesis is that the aligned monolingual texts can provide a better solution for the acquisition of diverse entity representations and all their corresponding mentions than what was achieved by Singh *et al.* (2011). These corpora provide higher diversity of paraphrases and world knowledge, which are not necessarily based on small string edit distance without introducing noise in the resulting chains. This is achieved when the representation of the aligned mentions are merged together – e.g. instead of using only Jack Nicholson as a mention (shown in sentence (3)) all equivalent phrases to it (e.g. the actor, the producer, etc.) from potentially aligned texts can be used as well (e.g. examples (4) through (7)).

(3) **Jack Nicholson** is becoming quite old.
(4) **The actor** is becoming quite old.
(5) **The producer** is becoming quite old.
(6) **The screenwriter** is becoming quite old.
(7) **The director** is becoming quite old.

2.2. Rule-Based Paraphrase Identification and Generation

Other works (Barzilay and Lee, 2003; Barzilay and McKeown, 2001; Pang *et al.*, 2003; Ibrahim et al., 2003) have examined the use of monolingual parallel corpora for paraphrase extraction based on statistical methods. For this purpose, parallel corpora already exist, e.g. the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013), or the Microsoft Research Paraphrase Corpus (MSRP) (Dolan and Brockett, 2005), but these are mainly in English and contain a lot of examples that are either trivial (e.g. material–materials, microdata–microdata), pure synonyms (e.g. president–chairperson, staff–personell) or rather too vague paraphrases conveying indeed different information (cf. PPDB).

Barzilay and Lee (2003) not only identified paraphrases for the construction of paraphrase corpora, but also tried to generate new paraphrased sentences,

which, however, are partly wrong. As a consequence, these statistically-based corpora contain valuable examples, on the one side, but also numerous expressions that are not easily described as paraphrases by formal means, which is important for the purpose of rule-based paraphrase generation.

Previous examinations concerning rule-based identification and generation of paraphrases (Zangenfeind, 2010) show that an important part of rules for paraphrasing natural language texts can be described by means of lexical functions (LFs) (Mel'chuk, 1996). LFs are a part of a very elaborated paraphrasing system in the Meaning-Text-Theory (MTT) (Mel'chuk, 1974; Mel'chuk *et al.*, 1992; Apresjan, 1974; Apresjan and Cinman, 2002).

As shown in Zangenfeind (2009; 2010), in a corpus consisting of the opening part of Tolstoj's novel *Anna Karenina* and its 22 German translations, almost 80% of all paraphrases of predicates which are found in the translations can be described with the help of LFs. These results are very encouraging to make use of LFs and the paraphrasing system of MTT for the alignment of parallel corpora in form of different translations of the same novel and to acquire data for the generation of new paraphrases. By means of LFs, about one hundred paraphrase rules are postulated in different publications of MTT (cf. Zangenfeind 2010).

It is true that for the pure alignment of two texts it is not essential to recognize, for example, which kind of support verb has been used in the paraphrase – it should be sufficient to recognize for instance that a full verb (e.g. [she] helped [us]) has a corresponding noun plus any of the according support verbs ([she] gave [us] help or [we] got help [from her]) in another translation or a corresponding adjective plus support verb or copula ([she] has been helpful [to us]). However, if new paraphrases are to be generated it is absolutely essential to know which kind of support verb to use (which should be described by LFs), because the connection of actants vitally depends on that fact.

The multi-target corpora can be highly helpful for the annotation of LFs to all according lexemes in the data, which can be used for a dictionary of German, similar to the Russian and English dictionaries of the machine translation system ETAP-3 (Apresjan *et al.*, 2003). With the help of LFs, paraphrases in different translations can be automatically identified and further generated.

2.3. Contrastive Language Studies and Literary Studies

Freely available bilingual multi-target parallel corpora for the pair Russian-German, such as the one we release with this work, can help answer a number of linguistic questions. Concerning syntactic constructions of Russian and German it is interesting, for example, to know how they alter when translated. This is a question we are specifically interested in, because with the help of this information (machine) translation rules from Russian to German can be

further developed (Zangenfeind, 2011; Zangenfeind, 2012). When one sentence is segmented into two (or more): How are semantic references realized? Concerning the field of discourse, continuity of topic and anaphoric reference (see section 2.1) can also be researched with the help of such data.

In the field of morphology: How are Russian tenses and verbal aspects translated? This still is an intricate problem for machine translation (Sonnenhauser and Zangenfeind, 2013). How are numbers of nouns translated? By which means is definiteness expressed in Russian (word order, aspect), i.e. which means in Russian corresponds to articles in German translations? This includes the examination of the presumed functional equivalence of article and aspect. How is word formation (compounds etc.) done in the two languages? Idiomatic expressions (phrasemes) are interesting for investigation (Biber *et al.*, 2002). Collocations, especially support verb constructions are very important for the description of paraphrases (Zangenfeind, 2009; Zangenfeind, 2010). Idioms, in the narrow sense, are also part of the focus.

Moreover, there are a lot of prospects for using the described parallel corpus for further investigation in regard to literary studies. The main goal of literary studies is the systematic analysis of literature and its adherence to poetic norms and values. To achieve this goal, scholars need to analyze literary texts, which has been realized as a computer-aided activity for the past few decades. However, literary scholars can identify and analyze in an automated way only phenomena that are realized as surface features of text, which makes them easily searchable and extractable (e.g. repetitions, concordances, general patterns in text, etc.). The aligned multi-target corpora that we provide enables literary analysis in parallel allowing for the simultaneous observation of phenomena across the different translations. Moreover, phenomena that are not realized as surface features of text (and are thus not automatically identifiable) would require manual observation of the data. For the latter, an appropriate visualization of the multiple alignments is needed, which we also provide via ALuDo, presented in (Zhekova *et al.*, 2014), which is a tool for interactively aligning search results within both bilingual and monolingual comparable texts (see section 4.1).

Some aspects that could be brought into the focus of the comparison of different translations and the original text are: Concerning stylistics, according to temporal classification of different translations, interesting issues specifically related to Dostoyevsky might be the following questions: How is Dostoyevsky's style of "spoken language" translated into German? How is Dostoyevsky's polyphonic language ("open dialogue" of the narrator with his characters) translated? Do translators try to "improve" the original text? To what extent are the different translations relevant for having established Dostoyevsky's oeuvre as world literature? How are diminutives translated? How are particles translated? In what way is there a guidance of the recipients (depending on the target audience)? Is there a correlation between the success of a certain translation and the language/style used in this translation? A further point of

interest is the question whether there is more orientation to the original language or to the target language with respect to the different translations.

3. Related Work

Most of the contemporary parallel corpora are mainly translations of foreign novels into English. For example, Barzilay and McKeown (2001) collect altogether 11 different English translations of five different books, which they used for the development of a supervised learning algorithm for the identification of paraphrases in English. Nine of the eleven translations were released and are currently available upon request. Another source for monolingual parallel corpora is, for example, the data collected for machine translation evaluation competitions, such as Bleu (Papineni *et al.,* 2002), where a number of different target texts in the form of translations are made use of.

Unfortunately, there are not many data sets that include the language pair Russian-German. An example is ParaSol (Waldenfels, 2006). In general, ParaSol is a parallel corpus of belletristic texts that includes mainly a collection of Slavic languages. In this data, the pair Russian-German is included, which is of interest to us (since as we noted in section 1 for this pair there are no freely available multi-target corpora). However, the corpus can only be viewed through its web interface, while the actual data is not available for download. Another similar collection that is also available only through a web interface and that includes the pair Russian-German is RusCorpora.[1] Furthermore, the Austrian Academy Corpus[2] (AAC) (Biber *et al.,* 2002) consists of the Russian text of Dostoyevsky's novel *The Idiot* and three German translations of it. This corpus has been used for contrastive investigations in the areas of collocations, cultural-specific lexis, forms of address, etc. (Dobrovol'skij, 2014). Unfortunately, once more, this corpus is not freely available. In fact, to the best of our knowledge, for this language pair, there are no large-scale freely-available aligned corpora with respect to both types: bilingual or bilingual multi-target.

Altogether, multi-target parallel corpora are a highly scarce resource especially with respect to literary texts. One of the freely-available resources for online search, CorTrad corpus[3] (Tagnin *et al.,* 2009), includes literary texts and multiple translations of them, but only for the language pair English-Portuguese. Additionally, the underlying texts can only be browsed with the online web interface available in Portuguese and cannot be downloaded for further use. The Russian language is interesting on its own most specifically in the form of literary texts from the 19th century. Biber *et al.* (2002) note e.g, that the use of collocations in Dostoyevsky's work differs significantly from

[1] http://www.ruscorpora.ru/en
[2] http://www.aac.ac.at
[3] http://www.linguateca.pt/CorTrad/

modern Russian literature. The multi-target data will enable an investigation not only within Russian texts, but also into how the lexical structure and contextual usage of such constructions is translated into a number of target texts of a different language. An interesting question for literary studies, for example, would be a comparison between a diachronic and a synchronic view of different styles in translations of such constructions. In order to make this possible, we need to select translations from significantly different dates, which are available for German – the translations by Alexander Eliasberg (Dostoyevsky, 1924), Hermann Röhl (Dostoyevsky, 1956) and Swetlana Geier (Dostoyevsky, 2012).

4. Sentence Alignment, Interactive Search and Visualization

Creating a bilingual multi-target parallel corpus for Russian-German is not an easy task, since the lack of parallel data for this pair excludes the use of superviszed approaches. One very powerful framework – the IMS Open Corpus Workbench[4] (CWB) presented in (Evert and Hardie, 2011) also provides support for sentence- or chunk-level alignment across two or more corpora. However, our intention is also to include word-alignment in further stages of development, which is not available in CWB. Moreover, we are not simply aiming at sentence alignment for any parallel corpora, but rather we intend to improve on the sentence alignment in AluDo, which provides us with graphical visualization and interactive search capabilities for the multi-target corpora of out interest, which are also not available in CWB. Thus, in section 4.1, we introduce a rule-based alignment procedure and in section 4.2 we discuss how it can be improved via an unsupervised statistical approach.

4.1. ALuDo

While coreference resolution and paraphrase identification/generation can well work with the underlying data itself, Contrastive Language Studies as well as other analysis-oriented applications would require an assistive environment that would enable an easy and targeted exploration of the texts. Thus, we make use of ALuDo (Alignment with Lucene for Dostoyevsky), presented in (Zhekova *et al.*, 2014), which is a tool for interactively aligning search results within both bilingual and monolingual comparable texts. ALuDo is a Java implementation, which could be easily applied for any language pair. The tool was also made accessible as a freely available web interface.[5] During the development phase, data for the target language pair Russian-German was selected, consisting of the Russian novel *Crime and*

[4] http://cwb.sourceforge.net
[5] http://dostojewski.cis.uni-muenchen.de Initial registration is required, but can be obtained by sending a mail to the first author.

Punishment (Dostoyevsky, 1866) by Fyodor Dostoyevsky and three of its translations in German by Alexander Eliasberg (Dostoyevsky, 1924), Hermann Röhl (Dostoyevsky, 1956) and Swetlana Geier (Dostoyevsky, 2012). Both former translations (Dostoyevsky, 1924; Dostoyevsky, 1956) are freely available online, while the latter translation by Swetlana Geier (Dostoyevsky, 2012) was and, unfortunately, still is not freely available as a digital copy. As Zhekova *et al.* (2014) note, it was provided for research purposes under specific restrictions by Fischer Verlag.[6]

As discussed in (Zhekova *et al.*, 2014), ALuDo was initially designed as a rule-based interactive aligner that made use of a query term in order to align the source text with the three target translations and display only the relevant aligned snippets. An example of the web interface and resulting interactive alignment is visualized in figure 1. Overall, there are two independent alignment modules in ALuDo: alignment via local grammars and alignment via ontological information and bilingual dictionaries.

Alignment via local grammars: Local grammars are built upon the finite-state formalism and are used to specify syntactic constraints for specific phenomena observed in the text. They were previously used for the representation of linguistic expressions (Maurel, 1989; Gross, 1997), however, specifically for alignment of parallel corpora this formalism was also applied in (Brown *et al.*, 1991; Kay and Röscheisen, 1993). In the latter works, proper names were indicated as anchors for the alignment. The same principle was also considered in ALuDo. Local rules represent, indeed, a simplistic rule-based baseline alignment approach using proper names as anchors that can be easily applied in a language independent manner. However, for every new text it requires time-consuming manual development of rules, which can become inefficient for larger collections of texts. Thus, the next paragraph delineates an extension of this approach that ALuDo uses simultaneously with local grammars.

Alignment via ontological information and bilingual dictionaries: Not only do local grammars (using proper names as anchors) need to be developed for every new text, but as well the alignment that they provide is not optimal. Proper names do not occur in every sentence in the text and not all linguistically interesting slots, which are potentially targeted by the interactive search, contain them. Additionally, anaphoric use of pronouns as referents to proper names increases the inefficiency of this approach even more. Thus, bilingual dictionaries and ontological information was also integrated into ALuDo. With these additional resources the alignment procedure in ALuDo supports interactive search queries between the two languages based not only on proper names, but rather based on any word found in the bilingual dictionary.

Ontological information can then be used to enhance the base search enabling the identification of words that stand in particular linguistic relation

[6] http://www.fischerverlage.de

with the query term. Ontologies, such as WordNet[7] (Miller, 1995; Fellbaum, 1998), have been made use of in a number of NLP tasks in the last decades. However, one of the biggest problems of ontologies is their applicability and coverage across languages. In order to avoid this hurdle, ALuDo uses only one ontology independently of the language pair at hand and integrates the bilingual dictionaries as a bridge between the texts and the ontology. In this case, the German version of WordNet, GermaNet[8] (Hamp and Feldweg, 1997), is used. Russian-German bilingual dictionaries are extracted from a web source[9] and merged together to bridge the gap between the language pair as well as to provide a proper connection to the ontology. Yet, neither the search terms nor the texts are restricted to the sole use of lemmas. Hence, a lemmatization module was also integrated in ALuDo that uses CISLEX (Guenthner and Maier, 1994; Langer *et al.*, 1996), a morphologically-rich dictionary. The disambiguation of polysemous entries was achieved by the most frequent sense baseline approach.

For a given search term, the actual alignment of results is achieved by transforming the search term into two sets of words, one for each language in the pair, containing all morphological derivations and semantically-related words of the query term. Then, sentences that contain these are identified and finally an overlap score between them is calculated. Sentences with a low overlap score are discarded (unless there is no other candidate sentence), while the ones with the highest overlap score and also situated positionally close are aligned.

The approach used in ALuDo is simple and easily implemented for any other language pair, but the alignment achieved through the overlap score on the bag-of-words principle does not always provide an optimal result for the sentences. This can be seen in figure 2, where the third result in the translation by Eliasberg is also positionally very close to the others (same part, same chapter, etc.) and contains the searched word (Sonne), but does not represent a correct alignment.

In order to avoid such cases, we consider the integration of another alignment module based on statistical methods. Aligning the texts in a preprocessing step allows for the exclusion of sentences, such as the example from figure 2. Additionally, ALuDo identifies sentences as potentially relevant only if the query term has been translated directly or replaced by a semantically related word (e.g. hyponym, meronym, etc.). However, translations are known to be frequently unrestricted and thus, there is no guarantee that this term will be present. So far, ALuDo discards such sentences. Yet, for Contrastive Language Studies, such sentences are also very important. Hence, applying statistical alignment would also allow for the preservation of these parts.

[7] http://wordnet.princeton.edu/wordnet
[8] http://www.sfs.uni-tuebingen.de/lsd
[9] http://www.dicts.info/uddl.php

Figure 2. An example of wrongly aligned sentences (result 3) in the translation by Eliasberg. The difference in visualization in comparison to figure 1 is due to the use of an older visualization standard needed to better demonstrate the issue.

4.2. Gargantua

Supervised statistical approaches to alignment are often used in various NLP applications, such as SMT. However, these need a large number of previously-aligned texts for training, which is a great hurdle for language pairs, such as Russian-German. Moreover, as Braune and Fraser (2010) note, a large number of them are also not completely language independent and not flexible to other language pairs (Chen, 1993; Fattah *et al.*, 2007). Thus, supervised alignment cannot be easily applied to this data and we turn back to unsupervised approaches.

A number of works have previously tackled unsupervised alignment (Brown *et al.*, 1991; Gale and Church, 1993; Moore, 2002; Deng *et al.*, 2006), but the approach proposed by Braune and Fraser (2010) is efficient and achieves high performance for both symmetrical and asymmetrical parallel corpora. The authors also release the aligner, called Gargantua,[10] which makes it easily available and applicable to this task.

Gargantua is a language-independent aligner that aligns two texts (a source and a target) at a time. Using Gargantua, we aligned each of the translations of Dostoyevsky's novel with the original Russian text resulting in three aligned pairs [(Dostoyevsky, 1866) – (Dostoyevsky, 2012); (Dostoyevsky, 1866) – (Dostoyevsky, 1956); (Dostoyevsky, 1866) – (Dostoyevsky, 1924)]. Part of the resulting alignments is also made freely available.[11] The translation from Swetlana Geier was unfortunately not released for download by Fischer Verlag and thus we could not include it in the package. Yet, this translation can be used via the interactive web aligner ALuDo where only the aligned parts are displayed.

[10] http://sourceforge.net/projects/gargantua
[11] http://www.cis.uni-muenchen.de/desi/rgdata/data.zip

5. Multiple Target Text Alignment

Aligning each translation with the original novel results in a regrouping and pairing up of the sentences that fit the so called 1-to-0/0-to-1 and 1-to-many/many-to-1 sentence correspondences between the novel and a particular translation.

In order to be able to easily merge all three translations within ALuDo, a post-processing step was also integrated in which the output from Gargantua (a single document containing all pairs of aligned sentences for all three translations) was transformed to three separate data sets of the same length. Each document represents one of the three pairs and each line in a document is positionally located as the corresponding aligned line from the novel in another document. Empty lines are used to account for mismatches caused by 1-to-0/0-to-1 relations (visualized in table 1).

Geier (2010)	**Röhl (1956)**	**Eliasberg (1924)**
За колечко … Den Ring habe …	За колечко … Auf den Ring …	За колечко … Für den Ring …
– Рубля-то … »Geben Sie vier …«		– Рубля-то … »Geben Sie mir …«
Я скоро … Ich werde …	Я скоро … Ich bekomme …	Я скоро … Ich bekomme …

Table 1. An example of 1-to-0/0-to-1 relations across the alignment of the three translations.

Geier (2010)	**Röhl (1956)**	**Eliasberg (1924)**
Оно лучше. Пусть побьет, душу отведет... оно лучше... А вот и дом. Козеля дом.	Оно лучше. Пусть побьет, душу отведет... оно лучше... А вот и дом. Козеля дом.	Оно лучше. Пусть побьет, душу отведет... оно лучше... А вот и дом. Козеля дом. Слесаря, немца, богатого... веди!
Es ist besser so ... Da ist das Haus, Haus Kosel. Der Schlosser Kosel, ein reicher Deutscher ... Führe mich	Es ist besser so. Mag sie mich schlagen, das macht ihr das Herz leichter ... Es ist besser so ... Aber da ist das Haus, das Koselsche Haus. Herr Kosel ist Schlosser, ein reicher Deutscher ... Kommen Sie mit!	Es ist besser so. Soll sie mich nur schlagen und ihrem Herzen Luft machen ... es ist besser ... Da ist schon das Haus. Das Koselsche Haus. Kosel ist Schlosser. Ein reicher Deutscher ... Führe mich!

Table 2. An example of partially incorrect many-to-many correspondences merged on a single line.

As can be seen in table 2 the post-processing step does not always produce fully correct output: for the translation of Geier only the second part of the Russian original text that is assumed to correspond to the German translation is really equivalent with the first part of the translation ("ono lučše ... A vot i dom. Kozelja dom." – "Es ist besser so ... Da ist das Haus, Haus Kosel."). Neither the first part of the erroneously aligned Russian paragraph ("Ono lučše ... Pust' pob'et, dušu otvedet ...") nor the second part of the erroneously-aligned German paragraph ("Der Schlosser Kosel, ein reicher Deutscher ... Führe mich") have any equivalent part in the corresponding paragraph in the other language respectively. For the translation of Röhl the second part of the German translation ("Herr Kosel ist Schlosser, ein reicher Deutscher ... Kommen Sie mit!") has no equivalent part in the corresponding paragraph of the Russian part. Only the alignment for the translation of Eliasberg is correct.

Dostojewskij (1866)	Geier (2010)	Röhl (1956)	Eliasberg (1924)
Небольшая комната, в которую прошел молодой человек, с желтыми обоями, геранями и кисейными занавесками на окнах, была в эту минуту ярко освещена заходящим солнцем.	Das mittelgroße Zimmer, das der junge Mann nun betrat, mit gelben Tapeten, Geranientöpfen und Musselin-Gardinen, war in diesem Augenblick von dem grellen Licht der untergehenden Sonne erfüllt.	Das kleine Zimmer, in welches der junge Mann eintrat, war gelb tapeziert; an den Fenstern hingen Musselingardinen; auf den Fensterbrettern standen Geranientöpfe; in diesem Augenblick war das Zimmer von der untergehenden Sonne hell erleuchtet.	Das kleine Zimmer, mit den gelben Tapeten, Geranien und Mullvorhängen an den Fenstern, in das der junge Mann kam, war in diesem Augenblick grell von der untergehenden Sonne erleuchtet.
«И тогда, стало быть, так же будет солнце светить!..» — как бы невзначай мелькнуло в уме Раскольникова, и быстрым взглядом окинул он всё в комнате, чтобы по возможности изучить и запомнить расположение.	"Auch dann wird also die Sonne so leuchten", fuhr es Raskolnikow unversehens durch den Kopf, und er sah sich alles im Zimmer rasch an, um die Lokalität so gut wie möglich kennenzulernen und sich einzuprägen.	›Die Sonne wird also auch dann so scheinen!‹ dachte Raskolnikow unwillkürlich und ließ einen schnellen Blick über das ganze Zimmer gleiten, um die Lage und Einrichtung möglichst kennenzulernen und sich einzuprägen.	– Also wird die Sonne auch dann ebenso leuchten! – ging es Raskolnikow unwillkürlich durch den Kopf, und er überflog mit einem schnellen Blick das ganze Zimmer, um alles zu studieren und sich nach Möglichkeit zu merken.
Пусть, пусть даже нет никаких сомнений во всех этих расчетах, будь это всё, что решено в этот месяц, ясно как день, справедливо как арифметика.	Und wenn, und wenn auch alle Berechnungen sogar über jeden Zweifel erhaben und wenn auch alles, was ich in diesem Monat beschlossen habe, sonnenklar und so richtig wie die Arithmetik sein sollte.	Und wenn auch in all diesen Berechnungen kein einziger zweifelhafter Punkt ist; und wenn auch alles, was ich mir in diesem Monate zurechtgelegt habe, klar wie der Tag und richtig wie das Einmaleins ist.	Wenn in meinen Berechnungen auch gar keine Fehler enthalten sind, wenn auch alles, was ich mir in den letzten vier Wochen zurechtgelegt habe, so klar wie die Sonne, so logisch wie die Mathematik ist ...

Figure 3. The correct alignment for the query presented in figure 2 achieved with the help of Gargantua.

The final integration of the statistically aligned corpora into ALuDo is demonstrated in figure 3 where, irrespective of the above-described problems, the wrongly-aligned third result for the same search, as in figure 2, is corrected and the equivalent parts in all texts are properly aligned.

6. Conclusion and Future Work

We have discussed the application of an unsupervised alignment approach of bilingual multi-target parallel corpora for error correction in the interactive aligner ALuDo. We have also argued that this type of corpora has not been sufficiently explored for linguistic analysis and have tried to show it can be useful for various linguistic tasks: Contrastive Language Studies and Literary

Studies, Coreference Resolution and Paraphrase Identification and Generation. We have also made the aligned data freely available, since such a corpus for Russian-German had not yet been released. We believe that corpora of this type can be applied to a large number of crosslingual investigations and thus strongly encourage the creation and release of other similar data sets.

In the future, we plan to perform task-specific explorations and integrate this data into, for example, a state-of-the-art Coreference Resolution system in order to be able to objectively evaluate the usefulness of this type of data for this task. Additionally, since for both Paraphrase Identification and Generation as well as Coreference Resolution word aligned data is needed, we also plan to word-align the dataset and make it freely available to the research community.

References

APRESJAN, Jurij D. (1974). *Leksicheskaja semantika.* Moskva: Nauka.

APRESJAN, Jurij D., and Leonid L. Cinman (2002). "Formal'naja model' perifrazirovanija predlozhenij dlja sistem pererabotki tekstov na estestvennyh jazykah." *Russkij jazyk v nauchnom osveshhenii*. 4.2: 102-146.

APRESJAN, Jurij D. *et. al.* (2003). "ETAP-3 Linguistic Processor: a Full-fledged NLP Implementation of the Meaning <=> Text Theory." *Conference Proceedings of MTT 2003*. Paris: 279-288.

BARZILAY, Regina, and Lillian Lee (2003). "Learning to Paraphrase: An unsupervised Approach using Multiple-sequence Alignment." *Proceedings of the 2003 Conference of the North American Chapter of the ACL-HLT*. Stroudsburg, PA, USA. ACL: 16-23.

BARZILAY, Regina, and Kathleen R. McKeown (2001). "Extracting Paraphrases from a Parallel Corpus." *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*: 50–57.

BIBER, Hanno, Evelyn Breiteneder, and Dmitrij Dobrovol'skij (2002). "Corpus-based Study of Collocations in the AAC." *Proceedings of the Tenth EURALEX International Congress*, Vol. 1. Eds. Anna Braasch and Claus Povlsen. Center for Sprogtknologi, Kopenhagen. 85–95.

BRAUNE, Fabienne, and Alexander Fraser (2010). "Improved unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora." *Proceedings of the 23nd International Conference on CL (Coling 2010)*. Beijing, China, August. 81-89.

BROWN, Peter F., Jennifer C. Lai, and Robert L. Mercer (1991). "Aligning Sentences in Parallel Corpora." *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL '91*. Stroudsburg, PA, USA. ACL. 169–176.

CHEN, Stanley F. (1993). "Aligning Sentences in Bilingual Corpora using Lexical Information." *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Ed. Lenhart K. Schubert. ACL. 9-16.

DENG, Yonggang, Shankar Kumar, and William Byrne (2006). "Segmentation and Alignment of Parallel Text for Statistical Machine Translation." *NLE* 12.4: 1-26.

DOBROVOL'SKIJ, Dmitrij (2014). "Russkie obrashchenija v parallel'nyh korpusah." *Die Welt der Slaven* 59.1: 1-21.

DOLAN, William B., and C. Brockett (2005). "Automatically Constructing a Corpus of Sentential Paraphrases." *Proceedings of IWP*. 9-16.

DOSTOYEVSKY, Fyodor Mihailovich (1866). *Prestuplenie i nakazanie*. Moskva: Editora.

———— (1924). *Verbrechen und Strafe*. Tr. Alexander Eliasberg. Potsdam: Gustav Kiepenheuer.

———— (1956). *Schuld und Sühne*. Tr. Hermann Röhl. Berlin: Aufbau Verlag.

———— (2012). *Verbrechen und Strafe*. Tr. Swetlana Geier. Frankfurt am Main: Fischer Taschenbuch Verlag.

EVERT, Stefan, and Andrew Hardie (2011). "Twenty-first century corpus workbench: Updating a query architecture for the new millennium." *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.

FATTAH, Mohamed Abdel, *et. al.* (2007). "Sentence Alignment using P-NNT and GMM." *Computer Speech & Language* 21.4: 594-608.

FELLBAUM, Christiane, ed. (1998). *WordNet: an electronic lexical database*. Cambridge: MIT Press.

GALE, William A., and Kenneth W. Church (1993). "A Program for Aligning Sentences in Bilingual Corpora." *Computational Linguistics* 19.1: 75-102.

GANITKEVITCH, Juri, Benjamin Van Durme, and Chris Callison-Burch (2013). "PPDB: The Paraphrase Database." *HLT-NAACL*. ACL. 758–764.

GROSS, Maurice (1997). "The Construction of Local Grammars." *Finite-State Language Processing*. Eds. E. Roche & Y. Schabès. Cambridge: MIT Press.329–354.

GUENTHNER, Franz, and Petra Maier (1994). *Das CISLEX Wörterbuchsystem*. CIS.

HAMP, Birgit, and Helmut Feldweg (1997). "GermaNet – a Lexical-semantic Net for German." *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain. 9-15.

IBRAHIM, Ali, Boris Katz, and Jimmy Lin (2003). "Extracting Structural Paraphrases from Aligned Monolingual Corpora." *Proceedings of the Second International Workshop on Paraphrasing*. Vol.16 *PARAPHRASE '03*. Stroudsburg, PA, USA. ACL. 57-64.

KAY, Martin, and Martin Röscheisen (1993). "Text-translation Alignment." *Computational Linguistics* 19.1: 121-142.

KOBDANI, Hamidreza et. al. (2011). "Bootstrapping Coreference Resolution using Word Associations." *Proceedings of the 49th Annual Meeting of the*

ACL: Human Language Technologies. Vol. 1 *HLT '11*. Stroudsburg, PA, USA. ACL. 783-792.

LANGER, Stefan, Petra Maier, and J. Oesterle (1996). *CISLEX – an Electronic Dictionary for German: Its Structure and a Lexicographic Application. CIS-Bericht*. CIS.

LEFEVER, Els, Véronique Hoste, and Martine De Cock (2011). "ParaSense or how to use Parallel Corpora for Word Sense Disambiguation." *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*. Portland, Oregon, USA, June. ACL. 317-322.

MAUREL, Denis (1989). *Reconnaissance de séquences de mots par automate: Cas des adverbes de date du français 1 microfiche*. Ph.D. thesis, Université Paris 7. Grenoble. Th.: informatique fondamentale.

MEL'CHUK, Igor' A. (1974). *Opyt teorii lingvisticheskih modelej "Smysl <=> Tekst"*. Moskva: Editora.

———— et. al. (1992). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques*. Montréal : Les Presses de l'Université de Montréal.

———— (1996). "Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon." *Lexical Functions in Lexicography and Natural Language Processing*. Ed. Leo Wanner. Amsterdam/Philadelphia: John Benjamins. 37-102.

MILLER, George A. (1995). "WordNet: a Lexical Database for English." *Commun. ACM* 38.11: 39-41.

MOORE, Robert C. (2002). "Fast and accurate Sentence Alignment of Bilingual Corpora." *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. 135-144.

PANG, Bo, Kevin Knight, and Daniel Marcu. (2003). "Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and generating new Sentences." *Proceedings of the 2003 Conference of the North American Chapter of the ACL on Human Language Technology, NAACL '03* Vol. 1. Stroudsburg, PA, USA. ACL. 102-109.

PAPINENI, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2002). "Bleu: A Method for automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, *ACL '02*. Stroudsburg, PA, USA. ACL. 311-318.

PRADHAN, Sameer et. al.. (2012). "CoNLL – 2012 shared task: Modeling multilingual unrestricted Coreference in Ontonotes." *Joint Conference on EMNLP and CoNLL – Shared Task*. Jeju Island, Korea, July. ACL. 1-40.

RAHMAN, Altaf, and Vincent Ng. (2009). "Supervised Models for Coreference Resolution." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP '09* Vol. 2. Stroudsburg, PA, USA. ACL. 968-977.

SINGH, Sameer et. al.. (2011). "Large-scale cross-document Coreference using distributed Inference and hierarchical Models." *Proceedings of the*

49th Annual Meeting of the ACL: Human Language Technologies. Portland, Oregon, USA, June. ACL. 793-803.

SONNENHAUSER, Barbara, and Robert Zangenfeind (2013). "Towards Machine Translation of Russian Aspect." *Proceedings of the 6th International Conference on Meaning-Text Theory*. Eds. Valentina Apresjan, Boris Iomdin, and E. Ageeva. Prague. 192-201.

WALDENFELS, Ruprecht von (2006). "Compiling a Parallel Corpus of Slavic Languages. Text Strategies, Tools and the Question of Lemmatization in Alignment." *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9: 123-138.

WETZEL, Dominikus, and Francis Bond (2012). "Enriching Parallel Corpora for Statistical Machine Translation with semantic Negation Rephrasing." *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Jeju, Republic of Korea, July. ACL. 20-29.

ZANGENFEIND, Robert (2009). "Types of Paraphrase Rules in Practice. German Paraphrases of a Russian Text." *Meaning – Text Theory* 2009. Eds. David Beck, Kim Gerdes, Jasmina Milićević, and Alain Polguère. Montréal. 389-398.

———— (2010). *Grammatik der Paraphrase* (= Linguistic Resources for Natural Language Processing, 4). München: Lincom Europa.

———— (2011). "Transfer of Russian Actantial Syntactic Relations into German." *Meaning – Text Theory* 2011. Eds. Igor Boguslavsky, and Leo Wanner. Barcelona. 306-31.

———— (2012). "Towards a System of Syntactic Dependencies of German." *Komp'juternaja lingvistika i intellektual'nye tehnologii – Computational Linguistics and Intellectual Technologies* 11.18. Ed. Kibrik, A.E., RGGU. Moscow: 706-715.

ZHEKOVA, Desislava (2013). *Towards Multilingual Coreference Resolution.* Ph.D. thesis, University of Bremen.

ZHEKOVA, Desislava, *et al.* (2014). "Alignment of Multiple Translations for Linguistic Analysis." *Proceedings of The 3rd Annual International Conference on Language, Literature and Linguistics (L3)*.