Bulgarian-English Sentence- and Clause-Aligned Corpus

Svetla Koeva, Borislav Rizov, Ekaterina Tarpomanova,
Tsvetana Dimitrova, Rositsa Dekova, Ivelina Stoyanova,
Svetlozara Leseva, Hristina Kukova, Angel Genov

Department of Computational Linguistics,
Institute for Bulgarian Language, BAS
52 Shipchenski Prohod Blvd., 1113 Sofia, Bulgaria
E-mail: {svetla, boby, katja, cvetana}@dcl.bas.bg,
{rosdek, iva, zarka, hristina, angel}@dcl.bas.bg

Abstract

The paper presents the partially automatically annotated and fully manually validated Bulgarian-English Sentence- and Clause-Aligned Corpus. The discussion covers the motivation behind the corpus development, the structure and content of the corpus, illustrated by statistical data, the segmentation and alignment strategy and the tools used in the corpus processing. The paper sketches the principles of clause annotation adopted in the creation of the corpus and addresses some issues related to interlingual asymmetry. The paper concludes with an outline of some applications of the corpus in the field of computational linguistics.

1 Introduction and motivation

Although parallel texts can be aligned at various levels (word, phrase, clause, sentence), clause alignment has proved to have advantages over sentence and word alignment in certain NLP tasks. Due to the fact that many of the challenges encountered in parallel text processing are related to (i) sentence length and complexity, (ii) the number of clauses in a sentence and (iii) their relative order, clause segmentation and alignment can significantly help in handling them. This observation is based on the linguistic fact that differences in word order and phrase structure across languages are better captured and formalised at clause level rather than at sentence level. As a result, monolingual and parallel text processing at clause level facilitates the automatic linguistic analysis, parsing, translation, and other NLP tasks.

Consequently, this strand of research has incited growing interest with regard to machine translation (MT). Clause-aligned corpora have been successfully em-

ployed in the training of models based on clause-to-clause translation and clause reordering in Statistical Machine Translation (SMT) – see [1] for syntax-based German-to-English SMT; [9] for English-to-Japanese phrase-based SMT; [2] for Japanese-to-English SMT; [8] for English-Hindi SMT, among others. Clause alignment has also been suggested for translation equivalent extraction within the example-based machine translation framework [7].

The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC) was created as a training and evaluation data set for automatic clause alignment in the task of exploring the effect of clause reordering on the performance of SMT [6].

The paper is organised as follows. Section 2 describes the structure, content and format of the BulEnAC and the annotation tool. Section 3 summarises the approach to sentence identification and alignment. Section 4 outlines the approach to clause splitting and alignment followed by a discussion on the principles of clause annotation. Section 5 addresses the possible applications of the corpus.

2 Structure of the BulEnAC

2.1 Basic structure

The BulEnAC is an excerpt from the Bulgarian-English Parallel Corpus – a part of the Bulgarian National Corpus (BulNC) of approximately 280.8 million tokens and 8.2 million sentences for Bulgarian and 283.1 million tokens and 8.9 million sentences for English. The Bulgarian-English Parallel Corpus has been processed at several levels: tokenisation, sentence splitting, lemmatisation. The processing has been performed using the Bulgarian language processing chain [5] for the Bulgarian part and Apache OpenNLP[1] with pre-trained modules for the English part[2].

The BulEnAC consists of 366,865 tokens altogether. The Bulgarian texts comprise 176,397 tokens in 14,667 sentences, with average sentence length 12.02 words. The English part totals 190,468 tokens and 15,718 sentences (12.11 words per sentence). The number of clauses in a sentence averages 1.67 for Bulgarian compared with 1.85 clauses per sentence for English.

The text samples are distributed in five broad categories, called 'styles'. A style is a general complex text category that combines the notions of register, mode, and discourse and describes the intrinsic characteristics of texts in relation to the external, sociolinguistic factors, such as the function of the communication act.

Clause-aligned corpora typically contain a limited number of sentences and cover a particular style, domain or genre[3], such as biomedical texts [3], legal texts [4], etc.

[1] `http://opennlp.apache.org/`

[2] The OpenNLP implementations used in the development of the BulEnAC were made by Ivelina Stoyanova.

[3] The further subdivision of the styles includes categorisation into domains (e.g., Administrative: Economy, Law, etc.) and genres (e.g., Fiction: novel, poem, etc.).

The goal in creating the corpus was to cover diverse styles so as to be able to make judgments on the performance of the alignment methods across different text types. As a result, the corpus consists of the following categories: Administrative texts (20.5%), Fiction (21.35%), Journalistic texts (37.13%), Science (11.16%) and Informal/Fiction (9.84%). Figure 1 shows a comparison of the average sentence length across styles for the two languages.

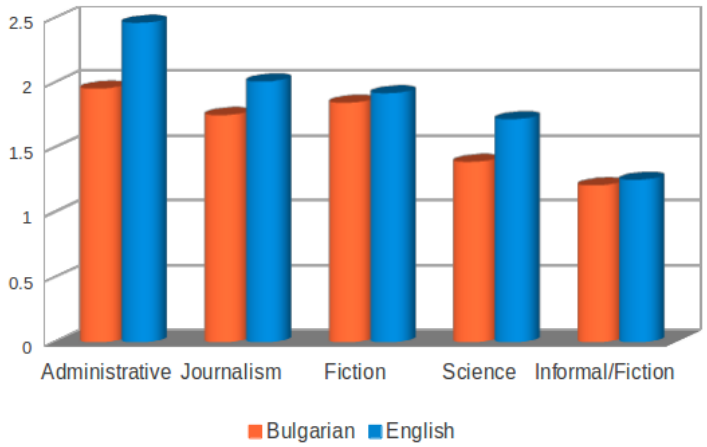

Figure 1: Average length of Bulgarian and English sentences (in terms of number of clauses) across the different styles.

2.2 Format of the Corpus

The files of the corpus are stored in a flat XML format. The words in the text are represented as a sequence of XML elements of the type `word`. Each `word` element is defined by a set of attributes that correspond to different annotation levels:

1. Lexical level (lemmatisation) – the attributes `w` and `l` denote the word form and the lemma, respectively.

2. Syntactic (sentence level) – the combination of two attributes, `e=True` and `sen=senID`, denotes the end of each sentence and the corresponding id of the sentence in the corpus.

3. Syntactic (clause level) – the attribute `cl` corresponds to the id of the clause in which the word occurs.

4. Syntactic (applied only to conjunctions) – the attribute `cl2` is used for conjunctions and other words and phrases that connect two clauses[4], and denotes the id of the clause to which the current clause is connected. The attribute `m`

[4]For brevity and simplicity such words and phrases are also termed 'conjunctions'.

defines the type of the relation between the two clauses `cl` and `cl2` (coordination or subordination), the direction of the relation (in the case of subordination) and the position of the conjunction with respect to the clauses. The inter-clausal relations are discussed in more detail in Section 4.2.

5. Alignment – the attributes `sen_al` and `cl_al` define sentence and clause alignment, respectively. Corresponding sentences/clauses in the two parallel texts are assigned the same id.

Example (1) shows the basic format of the corpus files.

Example 1 *The EU says Romania needs reforms.*

```
<word cl="864" cl_al="6c8f" l="the" w="The"/>
<word cl="864" l="eu" w="EU"/>
<word cl="864" l="say" w="says"/>
<word cl="865" cl2="864" cl_al="19f" l="PUNCT" m="N_S" w="===="/>
<word cl="865" l="Romania" w="Romania"/>
<word cl="865" l="need" w="needs"/>
<word cl="865" e="True" l="reform" sen="bc90" w="reforms.}"/>
```

Empty words (`w="===="`) are artificial elements introduced at the beginning of a new clause when the conjunction is not explicit or the clauses are connected by means of a punctuation mark. For simplicity of annotation punctuation marks are not identified as independent tokens but are attached to the preceding token.

The flat XML format is more suitable for the representation of discontinuous clauses than a hierarchical one; at the same time it is powerful enough to represent the annotation and to encode the syntactic hierarchy between pairs of clauses through the clause relation type.

2.3 The Annotation Tool

The manual sentence and clause alignment, as well as the verification and post-editing of the automatically performed alignment were carried out with a specially designed tool – ClauseChooser[5]. It supports two kinds of operating modes: a monolingual one intended for manual editing and annotation of each part of the parallel corpus, and a multilingual one that allows annotators to align the parallel units.

The monolingual mode includes: (i) sentence splitting; (ii) clause splitting; (iii) correction of wrong splitting (merging of split sentences/clauses); (iv) annotation of conjunctions; and (v) identification of the type of relation between pairs of connected clauses. Figure 2 shows the monolingual mode of ClauseChooser used for sentence and clause segmentation and annotation of clause relations. After having been segmented in the bottom left pane, the clauses are listed to the right. The type

[5]ClauseChooser was developed at the Department of Computational Linguistics by Borislav Rizov.

of relation for each pair of syntactically linked clauses is selected with the grey buttons N_N, N_S, etc.

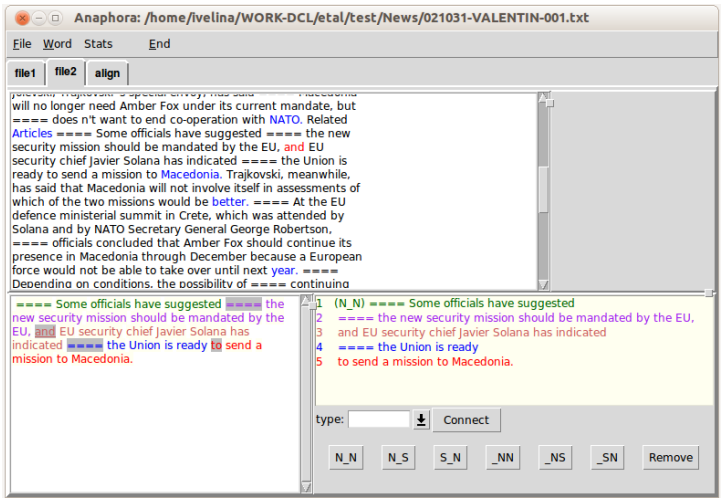

Figure 2: View of the monolingual mode of ClauseChooser

The multilingual mode uses the output of the monolingual sentence and clause splitting and supports: (i) manual sentence alignment; (ii) manual clause alignment.

3 Sentence segmentation and alignment

Both the Bulgarian and the English parts of the corpus were automatically sentence-split and sentence-aligned. The sentence segmentation of the Bulgarian part was performed with the BG Sentence Splitter. The tool identifies the sentence boundaries in a raw Bulgarian text using regular rules and a lexicon [5]. The English part was sentence-split using an implementation of an OpenNLP[6] pre-trained model. Sentence alignment was carried out automatically using HunAlign[7], and manually verified by experts.

The dominant sentence alignment pattern is 1:1 that stands for one-to-one correspondences in the two languages. The 0:1 and 1:0 alignments designate that a sentence in one of the languages is either not translated, or is merged with another sentence. Table 1 shows the distribution of the sentences in the corpus across alignment types. The category 'other' covers models with low frequency, such as 1:3, 3:1, 2:2, etc.

[6] http://opennlp.apache.org/

[7] http://mokk.bme.hu/resources/hunalign/

BG:EN alignment	frequency	in % of all
0:1	1187	7.60
1:0	225	1.44
1:1	13697	87.74
1:2	264	1.69
2:1	187	1.20
other	15	0.33

Table 1: Sentence alignment categories

4 Clause segmentation and alignment

A pre-trained OpenNLP parser[8] was used to determine the clause boundaries in the English part, followed by manual expert post-editing. The Bulgarian sentences were split into clauses manually. Clause segmentation is a language-dependent task that should be performed in compliance with the specific syntactic rules and the established grammar tradition and annotation practices for the respective languages. This approach ensures the authenticity of the annotation decisions and helps in outlining actual language-specific issues of multilingual alignment.

4.1 Clause alignment

After clause segmentation took place, the parallel clauses in the English and the Bulgarian texts were manually aligned. Alignment was performed only between clauses located within pairs of corresponding sentences.

The prevalent alignment pattern for clauses is also 1:1. However, due to some distinct syntactic properties of the languages involved, the different lexical choices, 'information packaging' patterns, etc., various asymmetries arise. The non-straightforward alignments have proved to be considerably more pronounced at clause than at sentence level as reflected in the higher frequency of clause alignment patterns of the type 1:0, 1:N and N:M (N, M>1), and the greater number of patterns that are represented by a considerable number of instances (Table 2).

1:0 and 0:1 alignments are found where a clause in one language does not have a correspondence in the other. For instance, in Example (2) the clause *he said* (2a) is not translated to Bulgarian (2b)[9].

Example 2

(a) [*La Guardia, step on it!*], [***he said.***]

[8] http://opennlp.apache.org/

[9] The Bulgarian examples are transliterated and glossed. We adopted word-by-word glossing with the following abbreviations (cf. Leipzig Glossing Rules, http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf): N – noun; ADJ – adjective; ADV – adverb; PTCP – participle; PST – past; PRS – present; SG – singular; PL – plural; ACC – accusative; COMP – comparative; DEF – definite.

(b) [*La Guardia, po-barzo!*]
[*La Guardia, quick-ADV;COMP!*]

1:N, N:1 patterns (N>1) stand for alignments where a given clause corresponds to a complex of (two or more) clauses. A systemic asymmetry is represented by the participial *-ing* and *-ed* clauses in English – clause 2 in (3a), and their Bulgarian counterparts. Bulgarian lacks non-finite clauses, therefore syntactic units that are headed by non-finite verbs are treated as participial constructions (the bold face part of the sentence in (3b)). In Example (3), the different clause structure of the English and the Bulgarian sentences leads to 2:1 alignment.

Example 3

*(a) [1 The Ministry announced a redistribution of financing,] [2 ==== **shifting funds to private sector projects.**]*

(b) [*1 Ministerstvoto obyavi prerazpredelenie na*
Ministry-the;DEF announce-PST;SG redistribution-N;SG of

finansiraneto, ***prehvarlyayki fondovete kam***
financing-the;N;DEF, shifting-PTCP fund-the;PL;DEF to

proekti v chastniya sektor.]
project-PL in private-the;DEF sector.]

Another frequent pattern is illustrated in Example (4). The two subordinate clauses marked in the sentence as clauses 2 and 3 in (4a), are translated as prepositional phrases PP_2 and PP_3, respectively[10]. As a result, the Bulgarian translation of the 3-clause English sentence consists of a single clause (4b); hence the alignment pattern is 3:1.

Example 4

*(a) [1 This Regulation does not go beyond] [2 **what is necessary**] [3 **to achieve those objectives.**]*

(b) Nastoyasthiyat reglament ne otiva po-dalech
Present-the;DEF regulation not go-PST;SG beyond-ADV;COMP

(PP_2 *ot neobhodimoto (*PP_3 *za postigane na*
from necessary-the;DEF for achievement-N of

tezi tseli)).
this-PL objective-PL.

Alignments of the type N:M (N,M>1) represent complex-to-complex correspondence and are relatively rare (0.84% of the clauses, Table 2). Example (5) illustrates an alignment pattern of the type 3:2. The English matrix clause 1 in (5a)

[10]The phrase labels are given for expository purposes. The clause-aligned corpus does not include annotation of phrasal categories.

is translated into Bulgarian (5b) by means of clause 1 and the part of clause 2 in boldface. The object of the English clause 1 *measures* (BG: merki) is the subject of the Bulgarian subordinate clause 2 *da badat vzeti merki...* (EN: for measures to be taken...) that roughly corresponds to the prepositional phrase in the English counterpart *for measures*. On the other hand, the subordinate clauses 2 and 3 in the English sentence are rendered as the prepositional phrase PP in Bulgarian (5b).

Example 5

*(a) [1 He urged **for measures**] [2 to help displaced persons] [3 return to their homes.]*

(b) [*1 Toy nastoya*] [*2* ***da badat vzeti***
He insist-PST;SG to be-PRS;PL take-PTCP;PL

merki *(PP za podpomagane na zavrashtaneto*
measure-PL for help-N of returning-the;N;DEF

na prinuditelno izselenite po tehnite domove).]
of forcefully displaced-PTCP;DEF;PL to their home-PL.

The distribution of the alignment pairs is given in Table 2.

BG:EN alignment	frequency	in % of all
0:1	1745	7.05
1:0	482	1.95
1:1	18997	76.80
1:2	2256	9.12
1:3	239	1.33
1:4	99	0.40
2:1	621	2.51
2:2	87	0.32
other	128	0.52

Table 2: Clause alignment categories.

Non-straightforward alignment patterns account for considerable number of 0:1 (7.05%) and 1:2 (9.12%) clause alignments in Bulgarian-English, with the reverse types amounting to just 1.95% (1:0) and 2.51% (2:1), respectively. These results suggest that a stronger tendency exists for 1:N (N>1) correspondences for Bulgarian-to-English than for English-to-Bulgarian. Some of the factors for this trend include the different segmentation into clauses as in the case of participial constructions versus participial clauses, and the rendition of prepositional phrases as clauses or vice versa.

4.2 Annotation of clause relations

The BulEnAC is supplied with partial syntactic annotation that includes:

(i) delimiting the sentence and clause boundaries;

(ii) identifying the type of relation (subordination or coordination) between the clauses in a sentence;

(iii) identifying the linguistic markers that introduce clauses – conjunctions, adverbs, pronouns, punctuation marks, etc.

A clause relation is defined between a pair of clauses. We were interested in the type of relation between the clauses, the ordering of clauses that stand in a given relation, the position of the conjunction, and language-specific clause-to-clause ordering constraints. With respect to the relation each clause in the pair is identified as either *main* or *subordinate* with at least one being *main*. In this paper the term *main* is used in a broader sense that encompasses both the meaning of an independent clause and that of a superordinate clause. Thus, *main* (N) denotes either a clause with equal status as the other member of the pair or one that is superordinate to it. *Subordinate* (S) status is assigned to a clause that is syntactically subordinate to the other member of the pair.

The status of the clauses is defined with respect to a particular clause relation and is therefore relative. Consequently, the relationship between a pair of coordinated independent or coordinated subordinate clauses is both N_N, cf. Example (6) for independent and Example (7) for dependent clauses. In the case of coordinated subordinate clauses, the dependent status of the pair is denoted by the relation N_S established between their superordinate and the first of the subordinate clauses (7b).

Example 6

*(a) [*N1 *I usually forget things,] [*N2 ***but***$_{\text{N1\_N2}}$ *I remembered it!]*

*(b) [*N1 *He asked her] [*S ***if***$_{\text{N1\_S}}$ *he could pick her up on the morning of the experiment] [*N2 ***and***$_{\text{N1\_N2}}$ *she agreed gratefully.]*

Example 7

(a) [1 Dutch police authorities said] [2 ***they were illegal immigrants****] [3* ***and would be deported.****]*

(b) [1 N *Dutch police authorities said] [2* S **====**$_{\text{N\_S}}$ *they were illegal immigrants]*

(c) [2 N1 *they were illegal immigrants] [3* N2 ***and***$_{\text{N1\_N2}}$ *would be deported.]*

A syntactically subordinate clause that is superordinate to another clause has the status *main* with respect to it. For instance, in (8a) clause 2 is subordinate to the matrix clause – clause 1 (8b), and a main clause with respect to clause 3 (8c):

Example 8

(a) [1 This Regulation does not go beyond] [2 ***what is necessary****] [3* ***to achieve those objectives.****]*

(b) *[1 N This Regulation does not go beyond] [2 S* ***what***$_{\texttt{N\_S}}$ *is necessary]*

(c) *[2 N ...what is necessary] [3 S* ***to***$_{\texttt{N\_S}}$ *achieve those objectives.]*

In the languages under consideration the following three clause ordering models cover almost all the cases: N_N, N_S and _SN.

4.3 More on translational asymmetries

Translational asymmetries stem also from different information distribution, lexical and grammatical choices, reordering of the clauses with respect to each other and (cross-clause boundary) reordering of constituents. In this section, we point out two types of asymmetry concerning the internal structure of clauses and their relative order within the sentence.

A frequent pattern found in the corpus is the selection of verbs with different types of complements motivated by grammatical structure, lexical choice or other factors. In the aligned sentences in Example (9) the choice of the Bulgarian verb *nastoyavam* (insist) as the translation equivalent of the English object-control verb *urge* predetermines the difference in the structure of the matrix and the subordinate clause in the two languages – in (9a) *Croatia* is the object of the main clause, whereas its counterpart *Harvatska* is the subject of the subordinate clause in (9b).

Example 9

(a) *[N European Parliament urges* ***Croatia****] [S to fully cooperate with the Tribunal.]*

(b)	[N	*Evropeyskiyat*	*parlament*	*nastoyava*]	[S	***Harvatska***
		European-the;DEF	*Parliament*	*insist-PRS;SG*		*Croatia*

da	*satrudnitchi*	*napalno*	*na*	*tribunala.*
to	*cooperate-PRS;SG*	*fully*	*to*	*tribunal-the;DEF.*

Another frequent example is the different order of the clauses in a sentence. For instance, in Example (10), the English clauses N_S (10a) are in reverse order as compared with the Bulgarian translation – _SN (10b).

Example 10

(a) *[N She had to make a detour] [S* ***to*** *get to the stove.]*

(b)	[S	***Za da***	*stigne*	*do*	*pechkata,*]
		In order to	*get-PRS;SG*	*to*	*stove-the;DEF*

[N	*tya*	*tryabvashe*	*da*	*mine*	*pokray*	*tyah.*]
	she	*must-PST;SG*	*to*	*go-PRS;SG*	*past*	*they-ACC;PL.*

Translation asymmetries represent a systemic phenomenon and account for the inter-lingual variations in grammatical structure, lexicalisation patterns, etc. At the

same time, they often give rise to wrong alignments, mistranslations, and other errors. Therefore, the successful identification of such phenomena and their proper description and treatment is a prerequisite for improving the accuracy of alignment and translation models.

5 Conclusion and applications

The development of the Bulgarian-English Sentence- and Clause-Aligned Corpus is a considerable advance towards establishing a general framework for syntactic annotation and multilingual alignment, as well as for building significantly larger parallel annotated corpora. The manual annotation and/or validation has ensured the high quality of the corpus annotation and has made it applicable as a training resource for various NLP tasks. As the goal was to explore the influence of clause alignment, further levels of alignment were only partially attempted as a technique enhancing the alignment method.

The quality of the manual clause splitting, relation type annotation and alignment was guaranteed by inter-annotator agreement. Each annotator made at least two passes of each Bulgarian and English file, one performed after the final revision of the annotation conventions. Clause segmentation was additionally validated at the stage of clause alignment.

The NLP applications of the BulEnAC encompass at least three interrelated areas: (i) developing methods for automatic clause splitting and alignment; (ii) developing methods for clause reordering to improve the training data for SMT [6]; (iii) word and phrase alignment. These lines of research will facilitate the creation of large-scale syntactically and semantically annotated corpora. In the field of the humanities the corpus is a valuable resource for studies in lexical semantics, comparative syntax, translation studies, language learning, cross-linguistic studies.

The BulEnAC will be made accessible to the scholarly community through the unified multilingual search interface of the Bulgarian National Corpus[11].

6 Acknowledgements

The present paper was prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the *Human Resources Development* Operational Programme 2007-2013 co-financed by the European Social Fund of the European Union. The Institute for Bulgarian Language takes full responsibility for the content of the present paper and under no conditions can the conclusions made in it be considered an official position of the European Union or the Ministry of Education, Youth and Science of the Republic of Bulgaria.

[11] `http://search.dcl.bas.bg`

References

[1] B. Cowan, I. Kucerova, and M. Collins. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney*, pages 232–241, 2006.

[2] C.-L. Goh, T. Onishi, and E. Sumita. Rule-based reordering constraints for phrase-based SMT. In *Proceedings of the 15th International Conference of the European Association for MT, May 2011*, pages 113–120, 2011.

[3] J.-D. Kim, T. Ohta, and J. Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10), 2008.

[4] C. Kit, J.J. Webster, K. Kui Sin, Pan H., and H. Li. Clause alignment for bilingual Hong Kong legal texts: A lexical-based approach. *International Journal of Corpus Linguistics*, 9(1):29–51, 2004.

[5] S. Koeva and A. Genov. Bulgarian language processing chain. In *Proceedings of Integration of Multilingual Resources and Tools in Web Applications. Workshop in conjunction with GSCL 2011, University of Hamburg*, 2011.

[6] S. Koeva, B. Rizov, E. Tarpomanova, Ts. Dimitrova, R. Dekova, I. Stoyanova, S. Leseva, H. Kukova, and A. Genov. Application of clause alignment for statistical machine translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), Korea*, 2012.

[7] S. Piperidis, H. Papageorgiou, and S. Boutsis. From sentences to words and clauses. In J. Veronis, editor, *Parallel Text Processing, Alignment and Use of Translation Corpora*, pages 117–138. Kluwer Academic Publishers, 2000.

[8] A. Ramanathan, P. Bhattacharyya, K. Visweswariah, K. Ladha, and A. Gandhe. Clause-based reordering constraints to improve statistical machine translation. In *Proceedings of the 5th International Joint Conference on NLP, Thailand, November*, pages 1351–1355, 2011.

[9] K. Sudoh, K. Duh, H. Tsukada, T. Hirao, and M. Ngata. Divide and translate: improving long distance reordering in statistical machine translation. In *Proceedings of the Joint 5th Workshop on SMT and Metrics MATR*, pages 418–427, 2010.