# Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment

**Kiril Simov**
LMD, IICT-BAS
kivs@bultreebank.org

**Petya Osenova**
LMD, IICT-BAS
petya@bultreebank.org

**Laska Laskova**
LMD, IICT-BAS
laska@bultreebank.org

**Aleksandar Savkov**
LMD, IICT-BAS
savkov@bultreebank.org

**Stanislava Kancheva**
LMD, IICT-BAS
stanislava@bultreebank.org

## Abstract

The paper describes the basic strategies behind the word and semantic level alignment in the Bulgarian-English treebank. The word level alignment has taken into consideration the experience within other NLP groups in the context of the Bulgarian language specific features. The semantic level alignment builds on the word level alignment and is represented in the framework of the Minimal Recursion Semantics.

## 1 Introduction

Manually created aligned bi- or multilingual corpora have proven to be useful resources in variety of tasks, e.g. for the development of automatic alignment tools, but also for lexicon extraction, word sense disambiguation, machine translation, annotation transfer and others.

In this paper we describe the word level alignment of the Bulgarian-English Parallel HPSG Treebank (BulEngTreebank) and its connection to the semantic level alignment. The aim of constructing such a treebank is to use it as a source for learning of statistical transfer rules for Bulgarian-English machine translation along the lines of (Bond et al. 2011 to appear). The transfer rules in this framework are rewriting rules over MRS (Minimal Recursion Semantics) structures. The basic format of the transfer rules is:

$$[C:]\ I\ [!F] \rightarrow O$$

where $I$ is the *input* of the rule, $O$ is the *output*. $C$ determines the *context* and $F$ is the *filter* of the rule. $C$ selects positive context and $F$ selects negative context for the application of a rule. For more details on the transfer rules consult (Oepen 2008). This type of rules allows for the extremely flexible transfer of factual and linguistic knowledge between the source and the target languages. Thus the treebank has to contain parallel sentences, their syntactic and semantic analyses and correspondences on the level of MRS.

In the development of such a parallel treebank we rely on the Bulgarian HPSG resource grammar BURGER, and on a dependency parser (Malt Parser – Nivre et al. 2006), trained on the BulTreeBank data. Both parsers produce semantic representations in terms of MRS. The treebank is a parallel resource aligned first on a sentence level. Then the alignment is done on the level of MRS. This level of abstraction makes possible the usage of different tools for producing these alignments, since MRS is meant to be compatible with various syntactic frameworks. The chosen procedure is as follows: first, the Bulgarian sentences are parsed with BURGER. If it succeeds, then the produced MRSes are used for the alignment. In case BURGER fails, the sentences are parsed with Malt Parser, and then MRSes are constructed on the base of the dependency analysis. The latter MRSes are created via a set of transfer rules (see Simov and Osenova 2011). In both cases we keep the syntactic analyses for the parallel sentences.

With respect to the MRS alignments, a very pragmatic approach has been adopted – namely, the MRS alignments originated from the word level alignment. This approach is based on the following observations and requirements:



*Proceedings of Recent Advances in Natural Language Processing*, pages 29–38,
Hissar, Bulgaria, 12-14 September 2011.

- Both approaches for generation of MRS over the sentences are lexicalized;
- Non-experts in linguistics can do the alignments successfully on word level;
- Different rules for generation/testing are possible.

Both parsers (for Bulgarian and English), which we use for the creation of MRSes, are lexicalized in their nature. Thus, they first assign elementary predicates to the lexical elements in the sentences, and then, on the base of the syntactic analysis, these elementary predicates are composed into MRSes for the corresponding phrases, and finally of the whole sentence.

Our belief is that having alignments on word level, syntactic analyses and the rules for composition of MRS, we will be able to determine correspondences between bigger MRSes than only lexical level MRSes, using the ideas of (Tinsley et al, 2009). They first establish the mapping on word level (automatically), then for candidate phrases they calculate the rank of the correspondences on the base of the word level alignment. Thus, our idea is to score the correspondences between two MRSes on the base of involved elementary predicates as well as the syntactic structure of the parallel sentences.

As it was mentioned, the alignment on word level allows us to do more reliable alignments using annotators who are non-experts in linguistics. Currently, the inter-annotator agreement is 92 %. Also this kind of alignment does not require any initial knowledge of MRS from the annotators. Another advantage is that the result might be used for training tools for automatic word alignment, and thus automatic extension of the treebank can be performed. Additionally, the word level alignment might be done before the actual analysis of the sentences. This is especially useful in case of Bulgarian, where the BURGER grammar is underdeveloped in comparison with the English grammar.

The paper is structured as follows: the next section discusses the related works on word alignment strategies. Section 3 focuses on the basic principles behind the word alignment between Bulgarian and English. Section 4 describes the level of MRS alignments. Section 5 outlines the conclusions.

## 2 Previous Work on Word Level Alignment

The annotation guidelines for Bulgarian-English word alignment, presented here, gained from the tradition established by the guidelines used in similar projects, aiming at the creation of golden standards for different language pairs, such as the Blinker project for English-French alignment (Melamed 1998), the alignment task for the Prague Czech-English Dependency Treebank 1.0 (Kruijff-Korbayová et al. 2006), the Dutch parallel Corpus project (Macken 2010), among others.

As Lambert et al. (2006) point out, the alignment decisions presented in the guidelines reflect different tasks. There are projects such as ARCADE (Vèronis, 2000) and PLUG (Ahrenberg et al., 2000), which aim at building a reference corpora with word, not sentence pairs, and have a different annotation strategy in contrast to those that focus on sentence level. Different linguistic theoretical backgrounds appear to be another source of divergence that affects the rules of phrase alignments as well as the specific grammatical techniques. This holds especially in correspondences between synsemantic words (like prepositions, determiners, particles, auxiliary verbs) and synsemantic and/or autosemantic words (Macken 2010). In addition, some tools for manual word alignment, e.g. HandAlign[1], allow the user to link both phrases and their elements with different kind of links, which might be simulated in other tools, which are more restrictive. Finally, the use of the so called possible (also ambiguous, fuzzy or weak) links that signal correspondence between semantically and/or structurally nonequivalent words or phrases is also a matter of dispute. While some argue that alignment with possible links should be determined by unambiguous rules, formulated with consideration of inter-annotation agreement, others (Lambert et al. 2006) allow for different decisions to be kept, which is true to the role originally ascribed to this kind of links: "P (possible) alignment which is used for alignments which might or might not exist" (Och and Ney 2000).

## 3 Word Level Alignment

The word level alignment was performed by the **WordAligner**[2] – a web-based tool for word alignment, built on top of the word alignment interface developed by C. Callison-Burch. It allows the user to provide parallel input of non-aligned text through the interface or to upload file(s) with sentence level aligned texts. Editing and/or completion of alignments is also supported. Each pair

[1] Available at http://www.cs.utah.edu/~hal/HandAlign/

[2] http://www.bultreebank.bas.bg/aligner/index.php

of sentences is represented as a grid of squares (Fig. 1). For convenience English is considered to be the *source* and Bulgarian – the *target* language, but that has no implications for the translation direction. Correspondence between two tokens is marked by clicking on a square – once (black square) or twice (dark grey square). Originally, the two colours were introduced to allow the annotator to mark his/her degree of certainty about the alignment decision: *sure link* (S link, black) or *possible link* (P link, dark grey). It is worth noting that in an alignment there can be only one type of link between two tokens or, more precisely, there is no distinction between phrase and word levels.

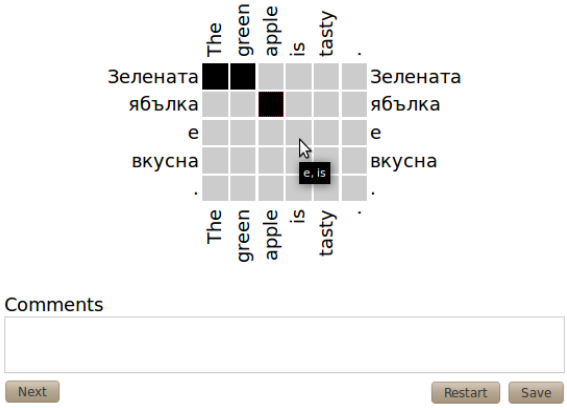



Fig 1. *Aligner interface. Mapping is done by clicking on the squares.*

Subsequently the colours were used to distinguish between *strong* and *weak* alignment (Kruijff-Korbayová et al. 2006), thus P link (dark grey) represents either *weak* alignment, or that the annotator is *uncertain* about the pairing, or both. S link (black) represents either *strong* alignment, or that the annotator is *certain* about the pairing, or both.

**General rules**

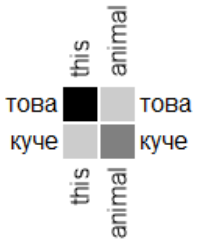
We adopt the general rules that have proven to be shared by the different annotation tasks and alignment strategies. The number of corresponding tokens to be aligned can be estimated by following these two rules (Veronis 1998, Merkel 1999, Macken 2010):

1. Mark as many tokens as necessary in the source and in the target sentence to ensure a two-way equivalence.
2. Mark as few tokens as possible in the source and in the target sentence, but preserve the two-way equivalence.

If a token or a phrase has no corresponding counterpart in the other language and bears no structural and/or semantic significance, it should be left unlinked (NULL link, square with no fill) (Melamed 1998).

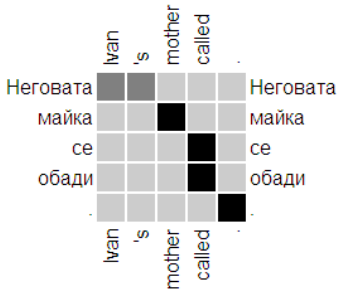
Idioms and free translations present a special case. If two autosemantic words or phrases refer to the same object, but do not share the same meaning, they are aligned with a P link, e.g.:

(1) *this animal*
*това куче* [‘this dog’]




The same rule holds when there is a synsemantic – autosemantic correspondence:

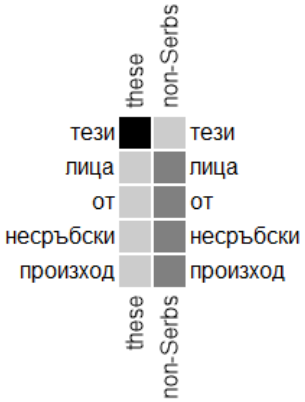
(2) ***Ivan 's*** *mother called.*
***Неговата*** *майка се обади.* [‘His mother called.’]




P link: *Ivan 's ~ Неговата*

P link is used when a lexical item is paraphrased in the other language:

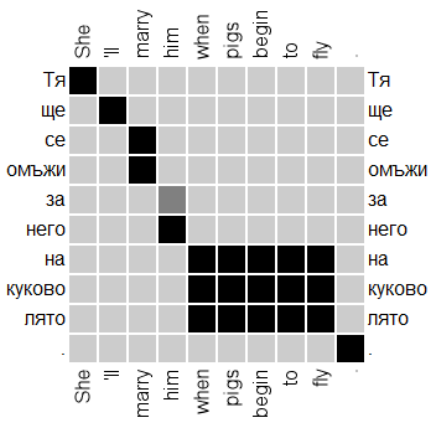
(3) *these non-Serbs*
*тези лица от несръбски произход* [‘persons from a non-Serbian origin’]




P link: *non-Serbs ~ лица от несръбски произход*

Idioms are linked with an S link; each token from the idiom in the source sentence is aligned with each token from the idiom in the target sentence.

(4) *She'll marry him* ***when pigs begin to fly****.*
*Тя ще се омъжи за него* ***на куково лято****.*




S link: *when pigs begin to fly ~ на куково лято*

**Specific rules**

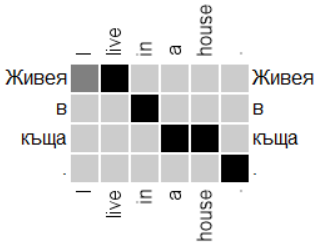
These rules are primarily language specific and their subjects are predominantly function words (prepositions, determiners, auxiliary verbs and the like). We give preference to the semantic equivalence where possible.

**Noun phrases**

**Determiners. Articles, demonstratives and possessive pronouns**

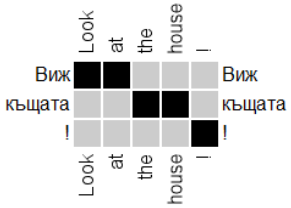
a) English determiners like *a(n)* or *the* correspond either to Bulgarian determiners *един* [one] (always in preposition, see example (7), or bare NP (5), or to the so called full/short definite article (6). In both languages they are attached to the first modifier of the NP, if there is one, regardless of its position[3].

(5) *I live in **a house**.*
*Живея в **къща**.*



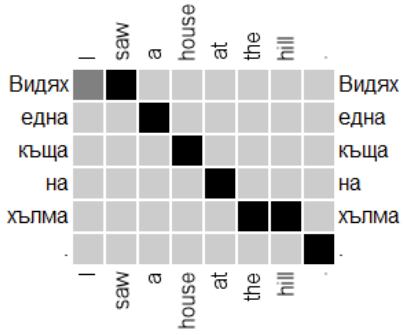

S link: *a house ~ къща*

(6) *Look at **the house**!*
*Виж **къщата**!*




S link: ***the** house ~ къща**та***

(7) *I saw **a house** at the hill.*
*Видях **една къща** на хълма.*




S link: *a ~ една*
S link: *house ~ къща*

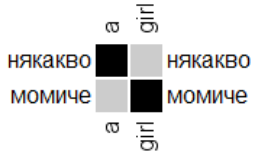
b) Usually if one of the two corresponding NPs has no modifier, the determiner and the head of the phrase are aligned together to the head of the other phrase (compare for example the rules presented in Kruijff-Korbayová 2006 or Macken 2010). Since in Bulgarian the article could be a morpheme attached to the first modifier (8), we decided to link both the article and the modifier from the English sentence to the corresponding Bulgarian modifier with an S link.

(8) ***the** lovely old house*
*хубава**та** стара къща*


S link: *the lovely ~ хубава**та***
S link: *house ~ къща*

c) We follow (Kruijff-Korbayová 2006) in linking determiners from different word classes, based on the similarity in their function. Thus the correspondence between indefinite articles and indefinite pronouns is marked with an S link (9).

(9) *a girl*
*някакво момиче*

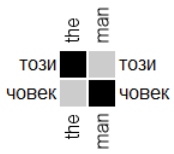



S link: *a ~ някакво*
S link: *girl ~ момиче*

d) English definite articles and Bulgarian demonstrative pronouns are also aligned with an S link (10).

(10) *the man*
*този човек*

[3] There are some exceptions in Bulgarian, e.g. *хубави **едни** дечица* ('pretty ones children' – some pretty children). In this case *едни* and *some* should be surely aligned.




S link: *the ~ този*
S link: *man ~ човек*

e) We use P link to align *the* with definite forms of full possessive pronouns (11) because the possesive.

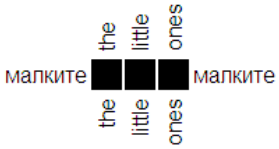
(11) *I heard **the** words,*
*Чух **техните** думи.*


P link: *the ~ техните*
S link: *words ~ думи*

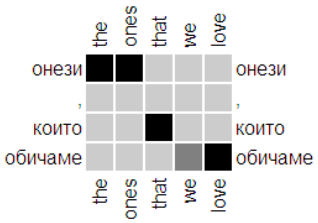
**Substitution with o*ne(s)***

Both lexical substitution and nominalization with the numeral *one(s)*, which are typical for English, have no structural and semantic analogy in Bulgarian. They should be aligned to the Bulgarian lexical unit that correspond to the premodifier of *one* (12), or, if there isn't any, to the coreferential Bulgarian pronoun (13).

(12) *the little ones*
*малките*




(13) ***the ones** that we love*
***онези**, които обичаме*




**Prepositional phrases**

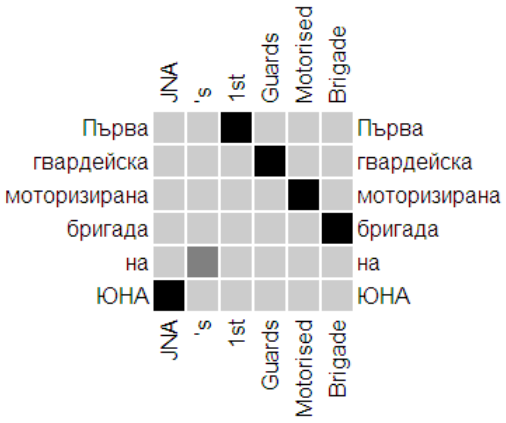
a) Very often English noun premodifiers are translated into prepositional phrases in Bulgarian (14). If that is the case, the preposition is aligned with a P link to the head noun, for example:

(14) ***Justice** Minister Cemil Cicek*
*Министърът **на** **правосъдието** Джемил Чичек*


S link: *Justice ~ правосъдието*
P link: *Justice ~ на*

b) English possessive noun forms are translated into Bulgarian either with *на* prepositional phrase (*John**'s** – **на** Иван*), or with an adjective that has possessive meaning (*John**'s** – Иванов*). In case of PP translation, the preposition itself is aligned to the possessive ***'s*** (for singular) or **'** (for plural) marker with an P link to reflect the fact that the two possessive markers are morpho-syntactically different (15).

(15) ***JNA's** 1st Guards Motorised Brigade*
*Първа гвардейска моторизирана бригада **на ЮНА***




S link: *JNA ~ ЮНА*
P link: *'s ~ на*

**Verb forms**

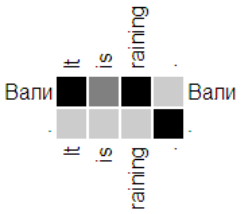
We follow the rules as they were first formulated in (Melamed 1998): link main verb to main verb and auxiliary verb(s) to auxiliary verb(s) if possible. Whenever the auxiliary form is not present or different in the source or target phrase, it should be aligned to the main verb (see for example (19), weakly or the two verb forms should be phrase aligned (21).

**Expletive subject and pro-drop**

a) Expletive subjects (*it, there*) usually have no correspondence in Bulgarian sentences, but they are obligatory for English. That is why we decided to link them with an S link to all Bul-

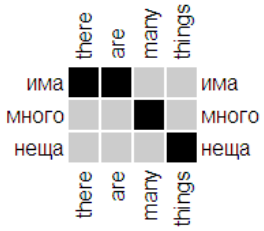
garian verb components, i.e. to the whole verb complex.

(16) *It is raining.*
*Вали.*




S link: *It ~ Вали*

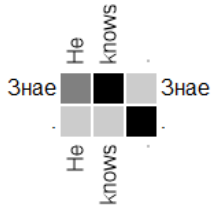
(17) ***there are*** *many things*
***има*** *много неща*




S link: *there are ~ има*

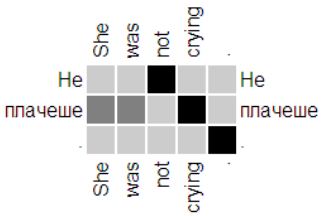
b) Bulgarian language is a pro-drop language. If the subject is unexpressed (18, 19, 20), then the English subject should be linked with a P link to all Bulgarian verb components that express one of the agreement categories: person, gender, number, and the main verb form itself. This decision is similar to the decision described in (Lambert et al. 2006) concerning the correspondences between English and Spanish verb phrases with omitted subjects.

(18) *He knows*
*Знае*




P link: *He ~ Знае*

(19) *She was not crying.*
*Не плачеше.*




P link: *She ~ плачеше*

(20) *They would not dare.*
*Не биха посмели.*


P link: *They ~ биха*
P link: *They ~ посмели*

**Reflexive pronouns in a verb complex**

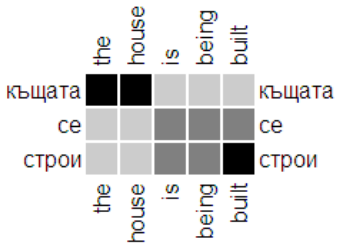
a) Reflexive Bulgarian *се* and *си* particles may be part of the verb lemma (21, 22). If that is the case, they should be aligned with an S link to the non-reflexive English verb form.

(21) *had* ***met*** *earlier*
*бяхме* ***се срещнали*** *по-рано*


S link: *met ~ се срещнали*

b) In contrast to the rules construed for Czech-English alignments (Kruijff-Korbayová 2006), if the reflexive particle is used to form a passive voice construction, it is aligned to the English verb phrase as a whole with a P link. The difference is due to the fact that although we also align the verb forms as phrases, we try to mark separately the correspondence between the main verbs.

(22) *the house* ***is being built***
*къщата* ***се строи***




S link: *is being ~ се*
S link: *built ~ строи*

***To*** **and** ***да*** **particles**

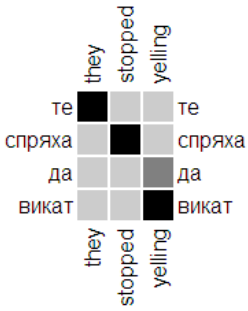
a) The correspondence between *to* and *да* is usually pretty straightforward.

(23) *the decision* ***to*** *stay*
*решението* ***да*** *остана*


S link: *to ~ да*

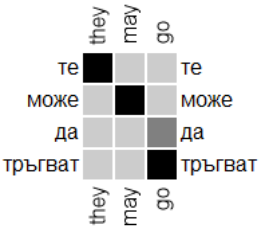
b) In the case when *to* is not present in the source sentence, *да* should be linked with a P link to the English verb that is aligned to the Bulgarian verb following the particle. Not surprisingly this rule resembles the rule for aligning Dutch (*om*)*...te* constructions (Macken 2010) with English full infinitive or *-ing* forms – as a infinitival particle Bulgarian *да* occupies similar syntactic positions and has similar functions.

(24) *they stopped **yelling***
*те спряха **да викат***



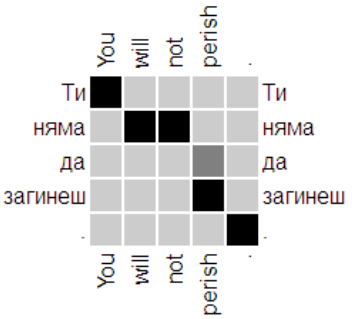

P link: *yelling ~ да*
S link: *yelling ~ викат*

(25) *they may **go***
*те може **да тръгват***




P link: *go ~ да*
S link: *go ~ тръгват*

(26) *You will not **perish**.*
*Ти няма **да загинеш**.*

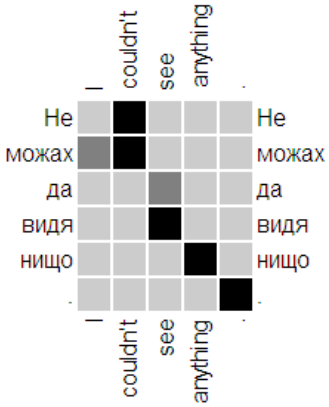



P link: *perish ~ да*
S link: *perish ~ загинеш*

**Double negation**

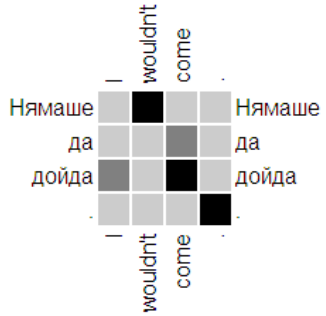
a) Double negation is typical for Slavic languages like Czech and Bulgarian, but not for English. In Czech the verb itself has a morphologically marked negative form that is weakly aligned with the positive form in English (Kruijff-Korbayová 2006). In Bulgarian the negative marker is not a morpheme, but a particle (*не,* 27) or an auxiliary verb with negative meaning (*няма, нямаше* 28). Often it is the case that one or more negative pronouns from the Bulgarian sentence correspond to indefinite English pronouns (27). They should be mapped with a P link.

(27) *I **couldn't** see **anything**.*
***Не можах** да видя **нищо**.*




S link: *could'nt ~ не можах*
S link: *anything ~ нищо*

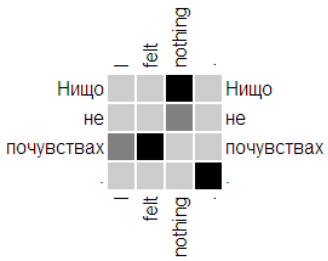
(28) *I **wouldn't** come.*
***Нямаше** да дойда.*




S link: *would n't ~ Нямаше*

If it is the English verb, that doesn't have negative form, then we use a P link to align the Bulgarian negative particle to the English word that bares negative meaning.

(29) *I felt nothing.*
*Нищо не почувствах.*

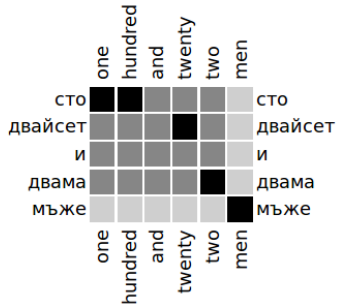



S link: *nothing ~ Нищо*
P link: *nothing ~ не*

**Numerals**

Cardinal and ordinal multiword numerals are treated as compound nouns and thus they are aligned as a block within which one-to-one correspondences are sure aligned (see for alternative decision Graça et al. 2008).

(30) ***one hundred and twenty two*** *men*
***сто двайсет и двама*** *мъже*




## 4 MRS Level Alignment

As it was mentioned above, we use the word level alignment in order to establish alignment on the level of MRS. For both languages the phrases are assigned an MRS structure which represents the semantic value of the phrase (in the case of dependency parse this MRS incorporates the semantic values of all dependent elements). The intuition behind our approach is that the lexical data of each structure in the syntactic analysis for a pair of sentences are aligned on word level. Then we assume that their MRS structures are equivalent modulo the meaning of the language specific elementary predicates. We exploit this intuition in constructing the semantic alignment in our treebank.

MRS is introduced as an underspecified semantic formalism (Copestake et al, 2005). It is used to support semantic analyses in HPSG English grammar – ERG (Copestake and Flickinger, 2000), but also in other grammar formalisms like LFG. The main idea is the formalism to rule out spurious analyses resulting from the representation of logical operators and the scope of quantifiers. Here we will present only basic definitions from (Copestake et al, 2005). For more details the cited publication should be consulted. An MRS structure is a tuple <*GT*, *R*, *C*>, where *GT* is the top handle, R is a bag of EPs (elementary predicates) and *C* is a bag of handle constraints, such that there is no handle *h* that outscopes *GT*. Each elementary predication contains exactly four components: (1) a handle which is the label of the EP; (2) a relation; (3) a list of zero or more ordinary variable arguments of the relation; and (4) a list of zero or more handles corresponding to scopal arguments of the relation (i.e., holes). Here is an example of an MRS structure for the sentence "*Every dog chases some white cat.*"

<h0, {h1: every(x,h2,h3), h2: dog(x), h4: chase(x, y), h5: some(y,h6,h7), h6: white(y), h6: cat(y)}, {}>

The top handle is *h0*. The two quantifiers are represented as relations *every(x, y, z)* and *some(x, y, z)* where *x* is the bound variable, *y* and *z* are handles determining the restriction and the body of the quantifier. The conjunction of two or more relations is represented by sharing the same handle (*h6* above). The outscope relation is defined as a transitive closure of the immediate outscope relation between two elementary predications – EP immediately outscopes EP' iff one of the scopal arguments of EP is the label of EP'. In this example the set of handle constraints is empty, which means that the representation is underspecified with respect to the scope of both quantifiers. Here we finish with the brief introduction of the MRS formalism.

First we establish correspondences on lexical level. Each two lexical items in the corresponding analyses are made equivalent on the basis of word alignment. Special attention is paid to the analytical verb forms and clitics. The next step is to traverse the trees in bottom-up manner. For each phrase or head for which the components are aligned, a correspondence on the MRS level is established. It should be explicitly noted that a correspondence on a sentence level is also established. Here we present an example:

Let us consider the following pair of sentences from the English Resource Grammar datasets:

| | | | |
|---|---|---|---|
| Kucheto | na | Braun | lae. |
| Dog-the(neut) | of | Browne | barks. |

*Browne's dog barks.*

The word level alignment is:

(***Kucheto*** = ***dog***)
(***na*** = ***'s***)
(***na Braun*** = ***Browne 's***)
(***lae*** = ***barks***)
(***Braun*** = ***Browne***)

Here are the MRS structures assigned to both sentences by ERG and BURGER. Some details are hidden for readability:

**ERG:**

```
<h1, {  h3: proper_q_rel(x3,h4,h6),
        h7: named_rel(x5,"Browne"),
        h8: def_explicit_q_rel(x10, h9, h11),
        h12: poss_rel(e13,x10,x5),
        h12: dog_n_1_rel(x10),
        h14: bark_v_1_rel(e2,x10)},
        { h4 qeq h7    h9 qeq h12 }>
```

**BURGER:**

```
<h1, {  h3:   kuche_n_1_rel(x4),
        h3:   na_p_1_rel(e5,x4,x6),
        h7:   named_rel(x6, "Braun"),
        h8:   exist_q_rel (x6, h9, h10),
        h11: exist_q_rel (x4, h12, h13),
        h1:   laya_v_rel (e2,x4)},
        { h12 qeq h3    h9 qeq h7 }>
```

The result of correspondences between MRS on the basis of word level establishes the following mappings of elementary predicates lists:

(**m1**)
(***Braun*** = ***Browne***)
{ h3: proper_q_rel(x5, h4, h6),
h7: named_rel(x5, "Browne") }
to
{ h7: named_rel(x6, "Braun"),
h8: exist_q_rel(x6, h9, h10) }

(**m2**)
(***na*** = ***'s***)
{ h12: poss_rel(e13, x10, x5) }
to
{ h3: na_p_1_rel(e5, x4, x6) }

(**m3**)
(***na Braun*** = ***Browne 's***)
{ h3: proper_q_rel(x5, h4, h6),
h7: named_rel(x5, "Browne"),
h8: def_explicit_q_rel(x10, h9, h11),
h12: poss_rel(e13, x10, x5) }
to
{ h3: na_p_1_rel(e5, x4, x6),
h7: named_rel(x6, "Braun"),
h8: exist_q_rel(x6, h9, h10) }

(**m4**)
(***Kucheto*** = ***dog***)
{ h12: dog_n_1_rel(x10) }
to
{ h3: kuche_n_1_rel(x4),
h11: exist_q_rel(x4, h12, h13) }

(**m5**)
(***lae*** = ***barks***)
{ h14: bark_v_1_rel(e2, x10) }
to
{ h1: laya_v_rel(e2, x4) }

As we mentioned above, our goal is to have MRS alignment not just on word level, but also on phrase level in the sentence. Thus, using the correspondences described in the previous section and the syntactic analyses of both sentences we can infer the following mapping:

(**m6**)
(***Kucheto na Braun*** = ***Browne 's dog***)
{ h3: proper_q_rel(x5, h4, h6),
h7: named_rel(x5, "Browne"),
h8: def_explicit_q_rel(x10, h9, h11),
h12: poss_rel(e13, x10, x5),
h12: dog_n_1_rel(x10) }
to
{ h3: na_p_1_rel(e5, x4, x6),
h7: named_rel(x6, "Braun"),
h8: exist_q_rel(x6, h9, h10),
h3: kuche_n_1_rel(x4),
h11: exist_q_rel(x4, h12, h13) }

Additionally, such correspondences might be equipped with similarity scores on the basis of word alignment types involved in the corresponding phrase, as well as the type of the phrase itself. For example, if the word alignment of two corresponding phrases involves only sure links, then the MRS alignment for these phrases also is assumed to be sure. Respectively, if on word level there are unsure links, then the MRS alignment could be assumed to be unsure. This idea could be developed further depending on the application. Also, in some cases the MRS level alignment could be assumed to be sure, although it includes some unsure links on word level. For example, in case of analytical verb forms many elements will be aligned only by possible links, but the whole forms are linked as a sure correspondence. We believe that such pairs of sentences with appropriate syntactic and semantic analyses and word alignment are a valuable source for construction of alignments on semantic level.

In our project, the mappings (explicit or inferred) are used for definition of a procedure for generating transfer rules as outlined in the introductory section.

## 5 Conclusion

In this paper we presented the alignment strategies behind the Bulgarian-English parallel treebank. The focus was on word and MRS level. On the base of each word alignment, an MRS alignment is produced together with the corresponding elementary predicates.

Although the current interannotator agreement on the word level is promising - 92 %, we will continue with the development of the guidelines in parallel to the alignment process.

The language specific features, which are likely to influence the transfer of information from Bulgarian to English, are as follows:

- Similarly to English and in contrast to other Slavic languages, Bulgarian is analytic language with a well-developed temporal system;
- Unlike English and similarly to other Slavic languages, Bulgarian has a relatively free word order and is a pro-drop language;
- Like other Slavic languages, Bulgarian verbs encode the aspect lexically;
- Being part of the Balkan Sprachbund, Bulgarian has clitics and clitic reduplication;
- Like other Slavic languages, Bulgarian has a double negation mechanism;
- Bulgarian polar questions are formed with a special question particle, which has also a foculizing role;
- Like other Slavic languages, the modification is mostly done by the adjectives (garden dog (EN) vs. gradinsko kuche (BG, 'garden-adjective dog')).

We hope that the MRS alignment in the treebank provides a good abstraction over the language specific features of Bulgarian as well as adequate equivalences to the English linguistic phenomena.

## Acknowledgments

This work has been supported by the European project EuroMatrixPlus (IST-231720).

## References

Ahrenberg L., Merkel M., Hein A.S., Tiedemann J. 2000. *Evaluation of Word Alignment Systems*. In: Proc. of the 2nd International Conference on Linguistic Resources and Evaluation (LREC). Athens, Greece, Vol. III: pp. 1255–1261.

Bond F., Oepen S., Nichols E., Flickinger D., Velldal E. and Haugereid P. 2011 (to appear). Deep open source machine translation. In Machine Translation Journal.

Copestake A., Flickinger D., Pollard C., and Sag I. 2005. *Minimal Recursion Semantics: An Introduction.* Research on Language and Computation, 3(4), pp. 281–332.

Copestake A. and Flickinger D. 2000. Open source grammar development environment and broad-coverage English grammar using HPSG. In Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 591–598.

Graça, J., Pardal J. P., Coheur L., Caserio D. 2008. *Multi-Language Word Alignments Annotation Guidelines (version 0.9)*. Spoken Language Systems Laboratory ($L^2F$). May 25, 2008. http://www.inesc-id.pt/pt/indicadores/Ficheiros/4734.pdf

Kruijff-Korbayová I., Chvátalová K., and Postolache O. 2006. *Annotation Guidelines for Czech-English Word Alignment*. In: Proceedings of the Fifth Language Resources and Evaluation Conference (LREC). http://www.coli.uni-saarland.de/~korbay/Publications/lrec06align.pdf

Lambert P., De Gispert A., Banchs R., and Mariño, J. B. 2006. *Guidelines for Word Alignment Evaluation and Manual Alignment*. In: Language Resources and Evaluation 39: 267–285. http://www.springerlink.com/content/dg2x327940442t12/

Macken, L. 2010. *Annotation Guidelines for DutchEnglish Word Alignment*. Version 1.0. TR, Language and Translation Technology Team, Faculty of Translation Studies, University College Ghent. http://webs.hogent.be/~lmac139/publicaties/SubsententialAnnotationGuidelines.pdf

Melamed, D. 1998. *Annotation Style Guide for the Blinker Project*. Version 1.0.4. Philadelphia. http://repository.upenn.edu/ircs_reports/53/

Merkel, M. 1999. *Annotation Style Guide for the PLUG Link Annotator*. http://www.ida.liu.se/~magme/publications/pluglinkannot.pdf

Nivre J., Hall J., Nilsson J. 2006. *MaltParser: A data-driven parser-generator for dependency parsing*. In Proc. of LREC-2006, pp. 2216-2219.

Och F.J., Ney H. 2000. *A Comparison of Alignment Models for Statistical Machine Translation*. In: Proc. of the 18th Int. Conf. on Computational Linguistics. Saarbrucken, Germany, pp. 1086–1090.

Oepen, S. 2008. *The Transfer Formalism. General Purpose MRS Rewriting*. Technical Report LOGON Project. University of Oslo.

Simov, K. and Osenova, P. 2011. Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing. In Proceedings of RANLP 2011.

Tinsley, J., Hearne, M. and Way, A. 2009. *Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation*.CICLing'09. 318-331.

Vèronis, J. 1998. *Arcade. Tagging guidelines for word alignment. Version 1.0*. http://aune.lpl.univ-aix.fr/projects/arcade/2nd/word/guide/index.html