DOI:10.2298/CSIS120130032C

Language Engineering for Syntactic Knowledge Transfer

Mihaela Colhon

Departament of Computer Science, A.I.Cuza street 13,
200585 Craiova, Romania1231
mcolhon@inf.ucv.ro

Abstract. In this paper we present a method for an English-Romanian treebank construction, together with the obtained evaluation results. The treebank is built upon a parallel English-Romanian corpus word-aligned and annotated at the morphological and syntactic level. The syntactic trees of the Romanian texts are generated by considering the syntactic phrases of the English parallel texts automatically resulted from syntactic parsing. The method reuses and adjusts existing tools and algorithms for cross-lingual transfer of syntactic constituents and syntactic trees alignment.

Keywords: parallel treebank, syntactic phrase alignment, bilingual corpus, word-alignments.

1. Introduction

Probably the most important trend in linguistics in the last decade is the massive use of large natural language corpora [7]. In any Natural Language Processing system (NLP system), corpora is often used to provide empirical and statistical data [8]. Typically, NLP applications that use corpora as basic linguistic resource are Word Sense Disambiguation (WSD) programs [19] and all types of parsers. Machine Translation (MT) represents the usage of computers as tools for translating texts from a source language to a target language [35]. The vast majority of current approaches to MT systems are also corpus-based. Among these, Phrase-Based Statistical MT (PBSMT) are by far the most dominant paradigm [24]. In this case, the linguistic resource is in the form of pairs of aligned parallel texts in the Source Language (SL) and Target Language (TL).

Current practice in phrase-based translation extracts regular phrases and translation rules from word-aligned parallel texts [13] as it is well-known that more and more researchers have devoted themselves to syntax-based MT systems [12], [18], [37]. Parallel treebanks are useful not only for syntax-based MT or example-based MT but also can be exploited in statistical approaches of translation. More precisely, by providing alignments between the syntactic tree of two corresponding sentences on a sub-sentential level

(word, phrase and/or clause level) automatic derivation of syntactic transfer rules, very important in any translation study, can be obtained.

The most common type of linguistic annotation is Part-Of-Speech (POS) tagging or, more accurately, morphosyntactic tagging, that is the procedure of assigning to each word token appearing in a text its morphosyntactic description [10]. Many studies consider that POS tags contain enough syntactical information to support word abstraction in any NLP system training. For example, the search space of a translation rules database can be greatly reduced by focusing only on POS tags instead of real words [35]. A treebank is a corpus that has been grammatically annotated in order to identify and label different syntactic components [15].

The treebank generation mechanism presented in this article automatically constructs a syntactic annotated parallel corpus from a bilingual word-aligned corpus with morphosyntactic annotations. The corpus was manually word aligned, tokenized, POS-tagged and lemmatized. The English texts were processed with one of the existing English syntactic parsers[1] while, for Romanian texts, a tree generation algorithm guided by the word-alignments of the corpus was implemented. As a consequence, the algorithm for Romanian syntactic tree generation depends greatly on the word-alignments of the bilingual corpus as will be shown in the following sections.

Parallel treebanks, like the treebank described in the present paper, are successfully used in various NLP applications but their main scope is to enhance syntax-based translation performance of the corpora language pairs. Also, based on the alignment mechanism encoded in the treebank annotations, rich and robust set of translation rules for the corpus languages can be identified.

As noted by competent linguists, Romanian language is morphologically rich and relatively flexible word order language [5]. The term Morphologically Rich Languages refers to languages in which substantial grammatical information, i.e., information concerning the arrangement of words into syntactic units or cues to syntactic relations, are expressed at word level. Because of its rich morphology, the morphological markers themselves could serve as strong cues for identifying the syntactic relations between the words in the sentence. But in languages with free or flexible word-order, such as Romanian, constituency-based representations are overly constrained, this fact causing word-order choice to influence the complexity of the syntactic analysis.

The method we present here is not restricted to the pair of languages chosen for the current implementation, which are English and Romanian languages. As it will be shown, the involved methodology for the treebank

[1] Some of the well known English syntactic parsers are: Stanford Parser (web page: http://nlp.stanford.edu/software/lex-parser.shtml), Link Parser (web page: http://www.link.cs.cmu.edu/link/) or Minipar (web page: http://webdocs.cs.ualberta.ca/~lindek/minipar.htm).

generation works on the abstract level of syntactic components and thus, all the particular information about the two languages lexicon is discarded. Also, the presented method does not make use of any particular information regarding the grammatical rules of any of the two involved languages.

The aim of the performed experiment is to test whether it is possible to reuse the syntactic constituents of one language texts in order to annotate their translations into another language. A study of this type was successfully performed for the same pair of languages but with the intention of testing the import of syntactic relations contracted by verbs [24].

2. Parallel Treebank from Bilingual Corpus

While monolingual treebanks are widely available thanks to large-scale annotation projects (NEGRA Treebank [28], Penn Treebank [30], Prague Dependency Treebank [31], Swedish Treebank [32]), bilingual parallel corpora with syntactic tree-based annotation on both sides, so-called parallel treebanks, are quite rare.

Despite of their enormous importance, the manually generation of such linguistic resources usually implies huge efforts. Manual construction is an expensive, time-consuming and error-prone process which requires linguistic expertise in both languages in question. For this reason, there has been a lot of research on automatic generation, basically using tree-to-string MT models, (e.g. [39]), while the development of tree-to-tree based MT models, despite their potential, has suffered.

The treebank generation algorithm presented in this paper is guided by the word alignments existing between the parallel sentences of a bilingual corpus. For this reason, the generation process is strongly dependent on the quality and quantity of the word-alignments, as accordingly to the Blinker annotation guidelines [23]: "*if a word is left unaligned on the source side of a sentence pair, this implies that the meaning it carries was not realized anywhere in the target side*". From the MT usage point of view, this implies that the meaning together with all morphosyntactic information of the source word to be lost. Therefore, the more accurate the word alignments are, the better the quality of the induced syntax tree for the target part of the resulted treebank will be.

2.1. Treebank Linguistic Resources

Treebanks, as large collections of syntactically parsed sentences, are considered valuable resources not only for computational tasks such as grammar induction and automatic parsing, but also for traditional linguistic and philological pursuits as well [17].

Syntactic annotation is the practice of adding syntactic information to a text by incorporating into it markers indicating syntactic dependencies relations. In order to obtain a parallel treebank from the bilingual corpus each sentence

has to be annotated with POS data. This kind of annotation is usually resulted with a POS Tagger tool. Another type of syntactic annotations consists of syntactic phrase labels for the both parts of a bilingual corpus which are aligned usually by following the word-alignments of the corpus.

English language is the best supported language, at this moment there are many large corpora, syntactic trees resources and testing language processing tools.

Although not to the extent of the languages with greater electronic visibility, efforts have been invested by researchers in different places (Romanian, Republic of Moldova, Unites States, United Kingdom, Germany, Italy, etc.) to develop Romanian linguistic resources such as corpora, dictionaries, wordnets and collections of linguistic data in both symbolic and statistical form [6].

From the available parallel corpora, the *Acquis Communautaire* linguistic resource represents the biggest parallel corpus existent at this moment, taking into account both its size and the number of covered languages [10]. The corpus includes the total body of European Union (EU) law applicable in the EU Member States. It is available in 22 official languages (including Romanian) of the European Union. A significant part of these parallel texts have been compiled by the Language Technology Group of the European Commission into an aligned parallel corpus, called JRC-Acquis Multilingual Parallel Corpus [18]. In most bilingual corpora derived from JRC-Acquis corpus, we find English paired with a European language.

In order to make a bilingual corpus with POS annotations an appropriate linguistic resource for the presented treebank generation method, word-alignments have to be provided (manually or with automatic tools such GIZA++ [27]).

A great progress has been done in the MT development from manually crafted linguistic models to empirically learned statistical models, from word-based models to phrase-based models and from string-based to tree-based models [21].

Two segments of texts from a bitext which represent reciprocal translations make a *translation unit* [39]. A translation unit may contain, in one or both the paired languages, one or more textual units (paragraph, sentence, phrase, word). Traditionally, phrases are taken to be syntactic components of a sentence. These units can be used to generate more complex constructions in that language and based on them a new phrase-based strategy was employed in MT: instead of generating translation of individual words from the source language, generate translations of the phrases and assemble the final translation by a permutation of these [39].

In literature, there are several translation theories formalized on parallel corpora with word-level alignments. In [11] is defined a generative process by means of which a symbol tree over a target language is derived from a string of source symbols. In order to distinguish between good and bad derivations, the notion of alignment is implemented. The triples

(*source_string*, *target_tree*, *word_alignment*)

are depicted in special structures named *alignments graphs* from which the set of derivation rules are inferred. Many translation models pay little attention to the context and to the syntactic structures of the translated phrases. The theory of alignments, spans and crossings is discussed in [11] where the phrasal coherence across two languages is studied. It is proved that incorporating syntactic information into translation models presents several advantages by the fact that syntactic phrases in one language tend to say together (i.e. cohere) during translation. Also, several studies have reported alignment or translation performance for syntactically augmented translation models.

The mechanism described in this article was designed in order to test the feasibility of the automatic cross-lingual transfer of syntactic phrases being built upon an English-Romanian parallel corpus developed at *Alexandru Ioan Cuza University of Iaşi* by the Natural Language Processing Group from Faculty of Computer Science. The corpus is XML encoded obeying a simplified form of the XCES standard [16]. For the bilingual corpus construction, the English and Romanian parts of the *Acquis-Communitaire* corpus[2] were used.

All the words of this English-Romanian corpus are annotated with lemmas, morphosyntactic information (gender, number, person and case) and Part of Speech markers. The tagsets used to annotate the words of the English-Romanian corpus comes from MULTEXT-East morphosyntactic specifications, version 3 (these specifications can be found at [25]). The latest version of these specifications, version 4 called “MondiLex”, is available at [26].

The MULTEXT-East project, developed for a large number of mainly Central and Eastern European languages (including Romanian) defines tagsets not only for Part of Speech data (POS data), but also includes the EAGLES-based morphosyntactic specifications, defining the features that describe word-level syntactic annotations [9].

The proposed algorithm works only on parallel sentences that are in 1:1 correspondence, meaning that every English sentence is translated into a single Romanian sentence. Best results are obtained for parallel sentences that are as closed as possible with respect to the syntactic realization of their content.

3. The Treebank Generation Algorithm

In this section we describe the Treebank Generation algorithm used to construct the parallel treebank with syntactic constituents from an English-Romanian corpus word-aligned and annotated at the morphological and

[2] Acquis Communitaire corpus contains about 12,000 Romanian documents and 6,256 parallel English-Romanian documents [6].

syntactic level. The resulted treebank is intended to set up a MT rule-based transfer system. Thus, instead of manually designing the rules, we could derive them from the generated treebank structures.
Because of the intended purpose, the algorithm works in the following scenario:

- one language of the bilingual corpus, the source language for the MT system must have a well-known syntactic parser by means of which the parse trees corresponding to this language texts could be obtained
- the part of the bilingual corpus corresponding to the target language of the MT system must have POS annotations or there must be available a POS tagger for the target language
- the bilingual corpus upon which the treebank is constructed must be word-aligned.

Following these requirements, the English sentences of the corpus were processed with Stanford Parser [20] in order to generate the English part of the treebank.

Stanford Parser is a natural language parser developed by Dan Klein and Christopher D. Manning from the Stanford Natural Language Processing Group. By parsing the English sentences with this tool, PENN Treebank parse trees were generated. As a direct consequence, the English texts are annotated with PENN Phrasal tags as this is the tagging standard used by Stanford Parser.

The parse trees labeled with PENN tagsets [30] consist of words in leaves, POS tags for the preterminal nodes and phrase tags for the next levels. The inner nodes denote grammatical constituents (for example NP for *noun phrase*, VP for *verb phrase*, etc.). Abstraction of words in syntactic trees represents almost no informational loss from syntactic point of view.

The implemented method can be summarized as follows: given a parse tree T_s for a source language sentence noted with *s*, and its target sentence, noted with *t* (that is, the translation of the source sentence in the target language) together with the word-level alignments, the parse tree of the sentence *t*, noted with T_t has to be constructed.

The algorithm generates the target tree T_t in a bottom-up fashion by mapping constituents of T_s onto contiguous substrings of *t* via lexical alignments.

For a lexical alignment, the most frequent alignment category is 1:1 such that one word in the source text is translated exactly to one word in the target text. However, there are other alignment categories, such as *omissions* (0:1 or 1:0), *expansions* (*n*:*m*, with $n<m$, $n,m\geq 1$), *contractions* (*n*:*m*, with $n>m$, $n,m\geq 1$) or unions (*n*:*n*, with $n>1$) [3].

A very popular way for visualization the parse tree of a source language sentence ending in leaf nodes – the sentence words connected by alignment links to the target sentence's words, is the *alignment graph*. An alignment link is a function $A(n) \rightarrow \{1,\ldots,m\}^*$ that maps a source leaf node *n* of the parse tree T_s to a set of zero or more of target leaf nodes.

Based on the lexical alignments and of the source tree structure decoded in the alignment graph, the corresponding target tree structure is generated. As it is described in Algorithm 1, the tree structure of T_t created from the source parse tree T_s following the word-alignments, noted with *WA*, by means of a bottom-up mechanism.

Indeed, the algorithm starts by constructing the set of the leaf nodes, *leaf*(T_t) together with the set corresponding to the first level of non-terminal nodes, *nonT*(T_t, 1), where the POS tags corresponding to the leaf node are included.

For the next levels, each non-terminal node $n_t \in nonT(T_t, level)$, $level \geq 2$, is considered to be the root of a subtree $tree \in T_t$ and labeled with the phrase tag of a non-terminal node n_s from T_s if the span of n_s is the frontier of the target subtree *tree*. The Treebank Generation algorithm is given in the next section.

3.1. The Algorithm

The alignment of syntactic trees is the process of finding the correspondences between internal and leaf nodes of two parsing trees representing parallel sentences in different languages. For example, Prime Factorization and Alignment (PFA) algorithm assigns prime numbers to terminal nodes and spreads them to the rest of the tree from the leaf nodes towards to the roots by assigning the product of child values to their fathers [1].

For two parallel syntactic trees: T_s corresponding to a source language sentence and T_t for the target language translation, a non-terminal node $n_s \in nonT(T_s, level_s)$ is aligned with a non-terminal node $n_t \in nonT(T_t, level_t)$, $level_s$, $level_t > 1$, if:

$$\text{span}(t_s) = \text{leaf}(tree)$$

where *tree* is a subtree of T_t, $tree \in T_t$ and $n_t = root(tree)$.

Because the target parse tree T_t is constructed taking into consideration the structure and the nodes of T_s the Treebank Generation algorithm also includes the alignments or correspondences between internal and leaf nodes of the two parallel trees, so it could be also considered as an alignment algorithm.

Algorithm 1. The Treebank Generation algorithm

Input:

a bilingual source language-target language corpus[3], the word alignments WA, a source language parser[4] and POS annotations for words in the target language

[3] An English-Romanian morphosyntactic annotated corpus was used.

[4] Stanford Natural Language Processing Group, Stanford Parser, http://nlp.stanford.edu:8080/parser/

Output: syntactic trees for target sentences

```
1. APPLY WA on the corpus C to obtain the word alignments
2. FOR each pair of parallel SL-TL sentences (s, t) of C
3.   FOR each word s_i of s =(s_1, ..., s_m)
4.     FOR each word t_j of t = (t_1, ..., t_n)
5.       IF s_i IS ALIGNED WITH t_j
6.         aligns(i,j) ← 1
7.       ENDIF
8.   APPLY STANFORD PARSER for sentence s
9.   LET T_s the parse tree of s
10.   LET leaf(T_s) the leaf nodes set of T_s
11.      leaf(T_s) ← {s_i| 1≤i≤m}
12.      LET nonT(T_s,lvl)←non-terminal nodes set in T_s,lvl≥2
13.         nonT(T_s, 1) ←{POS(s)| s ∈ leaf(T_s)}
14.   nonT(T_s, lvl) ←{parent(n)|n∈ nonT(T_s,lvl-1), lvl≥2}
15.   LET T_t the parse tree of t
16.   LET leaf(T_t) the leaf nodes set of T_t
17.      leaf(T_t) ← {t_j| 1≤j≤n}
18.   LET nonT(T_t, lvl) ← non-terminal nodes set in T_t,lvl≥2
19.   nonT(T_t, 1) ← {POS(n) | n ∈ leaf(T_t)}
20.   nonT(T_t, lvl) ← ∅, lvl≥2
21.   FOR each node s_i IN leaf(T_s)
22.      span(s_i) ← {t_j| aligns(i,j)=1}
23.   FOR each node N in nonT(T_s, lvl), lvl≥2
24.      span(N)←{span(s_i)|s_i∈leaf(T_s) ∧ s_i∈ descendant(N)}
25.   FOR each node t_j IN leaf(T_t)
26.      parent(t_j)←POS(t_j)
27. lvl ← 2
28. WHILE (lvl ≤ max_level(T_s))
29.   nonT(T_t,lvl)← ∅
30.   FOR each node t_j IN leaf(T_t)
31.     IF ∃ N∈ nonT(T_s, lvl): t_j∈ span(N)
32.        parent(last_parent(t_j))← N
33.        nonT(T_t,lvl)← nonT(T_t,lvl)∪{N}
34.     ENDIF
35.  lvl← lvl+1
36. ENDWHILE
```

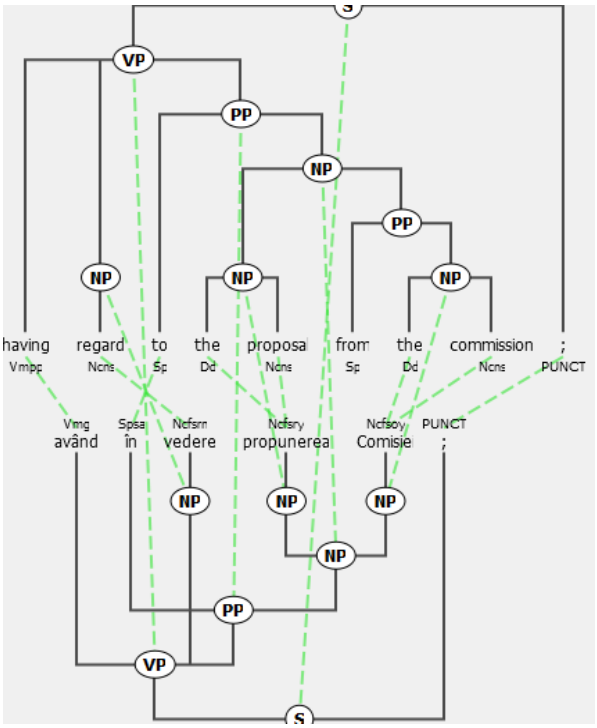

Figure 1. Cross word-alignments generate overlapping phrase-structures[5]

In any translation process, the one-to-zero lexical alignment is undesirable. One-to-zero means the lack of equivalent lexical translation in the target language, this phenomena being called “lexical hole”. The unaligned Romanian words can be resolved using specific grammatical information relative to the Romanian language but such a study does not make the subject of this article.

The time complexity of our algorithm relative to each pair of parallel Source Language-Target Language sentences (*s*,*t*) is $O(n^{max\_level(Ts)})$ because of the *while* loop that includes a linear browsing of the leaf nodes from *Tt*, where *n* is the length of the sentence in the Target Language (this means that sentence *t* has *n* words, property that is given by $t = (t_1, \ldots, t_n)$ in the algorithm notations).

Algorithm 1 assumes that all the spans of the non-terminal nodes of *Ts* are continuous lists of nodes and does not resolve the crossing alignments between each pair of English-Romanian parallel sentences. These issues will be discussed in the next section.

[5] The treebank alignments are loaded in Stockholm TreeAligner program [36].

3.2. The Alignments of the Treebank Parallel Components

The most important feature of the developed algorithm consists in finding the translation equivalence between two syntactic phrases of each bitext. Basically, translation equivalence rely only on the lexical tokens (words, phrases) paired by an alignment link. Even if not all the words between the two phrases are aligned, the phrases can still align very well.

The word alignments were drawn manually between the parallel sentences of the English-Romanian corpus. Although the syntactic structures in the two languages are not similar, some alignments can still be identified in order to support the syntactic equivalences. For all that, aligning two words with the same meaning but with different part of speech is not desirable from this kind of study point of view because, in this case, even if the alignment is semantically correct it can't help the phrases equivalence.

Crossing alignments

In any translation process, lexical mapping is inevitable. Crossing lexical alignments between a source sentence and a target sentence happen when the order between the source words and their translations is not preserved.

In Figure 1 it is shown one crossing among the word-alignments links, indicating one instance of reversing syntactic constituents during translation process. This particular crossing involves reversal of the prepositional word with the noun word. Depending on how often this reversal is encountered, in a translation process we could consider to invert all the TO constituents that appear before NP constituents.

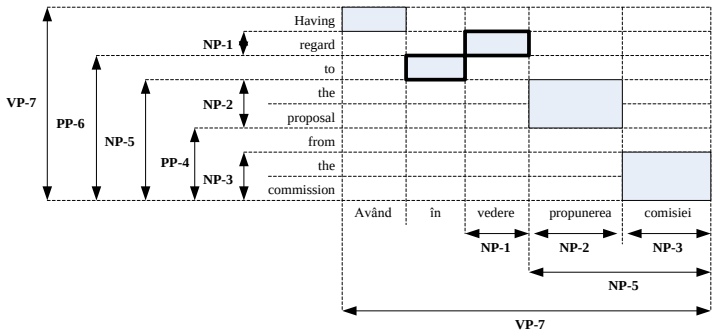

Figure 2. The alignment matrix[6]

[6] The crossing alignments are marked with a thick border.

The crossing alignments can be easily identified using the `aligns` matrix constructed in Algorithm 1.

The alignment matrix that corresponds to a pair of English-Romanian parallel sentences (*s*, *t*) from the *JRC-Acquis* corpus, for:

s = (*Having*, *regard*, *to*, *the*, *proposal*, *from*, *the*, *commission*)

t = (*Având*, *în*, *vedere*, *propunerea*, *comisiei*)

is illustrated in Figure 2.

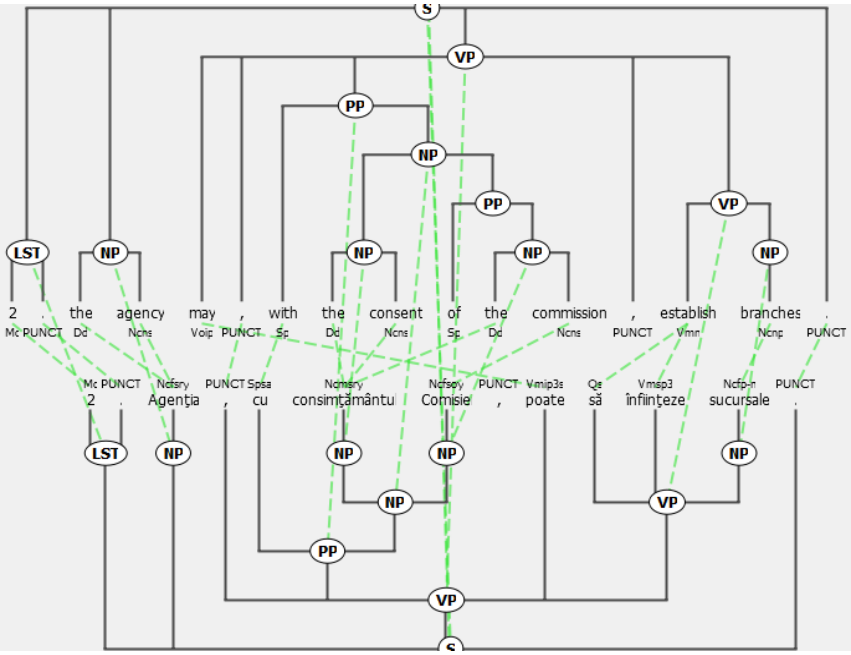

Figure 3. Cross word-alignments do not necessarily generate overlapping phrase-structures[7]

Usually, the crossing alignment problem implies reordering of the source tree such that the lexical order of the leaf nodes matches the order of the target sentence. But resolving this issue implies particular studies that address the particularities of the target language with respect to the particularities of the source language. Such studies do not make the subject of the present article, being left for a future work.

Still, not all crossing word-alignments determine overlapping between the target tree syntactic components as it is exemplified in Figure 3.

[7] The treebank alignments are loaded in Stockholm TreeAligner program [27].

3.3. Parallel Treebank Annotations

After the parse trees were generated for both parts of the corpus, the hierarchical language representations for the parse trees have to be flattened into linear string representations, which can be easily input to many feature-like probabilistic models. Thus, during model-training, these string representations together with the alignment information can generate statistics needed to build translation grammars. Our goal is to extract rich and robust set of English-Romanian translation rules.

In line with the PENN parse tree format used by the Stanford Parser we propose a format in which the aligned phrase tags for the inner nodes of the trees are indexed by the same number. The common notation for the phrase nodes accompanied by the lexical alignments for the leaf nodes make easier to find the alignments between the parallel parse trees.

The annotations of the treebank preserve, from the used English-Romanian corpus, the MULTEXT-EAST words specifications as these data include all the morpho-syntactic details needed for any syntactic study, while for the phrasal constituents the PENN Treebank Phrasal Tags are used.

In order to evaluate the Phrasal tags for Romanian sentences resulted from the Treebank Generation algorithm, the corpus annotations with syntactic chunks for the Romanian words are compared with the Phrasal tags sequences "inherited" from the parallel English sentences parse trees.

The chunks annotations of the corpus were generated by means of a simple regular expression chunker in order to mark the syntactic constituents that form a given sentence. More precisely, two separate English and Romanian grammars were implemented for generating PERL regular expressions over sequences of POS tags for English and Romanian types of phrases founded in the corpus sentences [24]. Using the languages regular expressions defined over the tagsets, the chunker accurately recognizes the (non-recursive) syntactic phrases both for Romanian and English.

The chunk parser detects the chunks of a text, like *noun phrases* (*NPs*), *prepositional phrases* (*PPs*) or *verb phrases* (*VPs*). Chunks are non-overlapping spans of text, usually consisting of a head word (such as noun) and the adjacent modifiers and function words [6]. The chunk annotations are the references in the evaluation process based on which the performance of the Treebank Generation algorithm is measured.

4. Experimental Results and Evaluation

The Treebank Generation algorithm for the Romanian sentences was tested by taking into account the chunker annotations for the Romanian part of the previously mentioned bilingual corpus. More precisely, for every word of a Romanian sentence, each syntactic phrase determined by the Treebank Generation mechanism that correctly matches within the syntactic chunks annotations of that word adds to the mechanism precision.

Resuming, we have that the PENN phrase tags identified for the Romanian words by the Treebank Generation algorithm are compared with the sequences of syntactic chunks specified for the Romanian words of the English-Romanian corpus.

Table 1. The corpus annotations and the corresponding PENN Phrasal Tags

Corpus annotations	Ap adjective phrase/ adverb phrase	Np noun phrase	Pp prepositional phrase	Vp verb phrase
Corresponding PENN Phrasal Tags	ADJP ADVP PP WHADVP	NP WHNP	PP WHPP	VP

As it is given in Table 1, the *NP* and *WHNP* PENN Phrasal tags are considered the equivalent PENN notations for the *Np* chunk annotations, a *VP* tag matches only with a *Vp* chunk while the *PP* and *WHPP* tags match only with *Pp* chunks. In the case of the *Ap* chunk a discrimination algorithm had to be implemented in order to correctly evaluate this notation according with its corresponding meaning.

Table 2. Example of parallel sequences of treebank tags and chunker annotations together with their matching degrees

Token(word)	Treebank tags/ chunker annotations	Number of matches
vot	Ncms-n VP VP NP VP VP S Np Pp	no match[8]
de_asemenea	Rgp **ADVP** VP S ROOT **Ap**	one match
economic	Afpms-n **ADJP NP** NP VP … **Ap Np** Pp	two matches
dividende	Ncfp-n **NP PP** VP S ROOT **Np Pp**	two matches
în	Spsa **PP VP PP** S **Ap Vp Pp**	three matches

Indeed, because a single *Ap* notation is used by the chunker for both the *adjectival phrase* and the *adverbial phrase*, the evaluation mechanism has to

[8] The Romanian noun "*vot"* inherits different PENN tags from the alignment mechanism because it was aligned with a word with different Part of Speech, more precisely it was aligned with a verb.

discriminate among the cases when the *Ap* means *ADJP*, that is adjectival phrasal tag or *ADVP*, the adverbial phrasal tag or *WHADVP*, wh-adevarb phrase or even *PP* prepositional phrase tag. This discrimination is done upon the part of speech of the annotated token/word.

More precisely, if a Romanian word, annotated with the *Ap* chunk, is:

- an adjective, then *Ap* is considered the correspondent of the *ADJP* tag in the PENN Phrasal format
- an adverb then its *Ap* annotation will match only with *ADVP* or *WHADVP* PENN tags
- a preposition then the *Ap* annotation is considered equivalent with the *PP* PENN tag.

The number of matches between the tags of a PENN phrasal sequence of a Romanian word and the chunks of the corpus annotations for that word is counted for the transfer precision. In Table 2 we exemplify the manner in which the transfer precision is determined. Because we evaluate the knowledge transfer degree, it is obviously that only the sequences of PENN phrasal tags that correspond to Romanian words with non-null alignments in the English parallel part of the corpus will be considered.

The performance of the Treebank Generation algorithm is measured in terms of *Precision* and *Recall*, such that:

- *Precision* is the fraction of correctly identified Phrasal tags with respect to the total number of generated Phrasal tags
- *Recall* is the fraction of correctly identified Phrasal tags with respect to the total number of Phrasal tags specified in the chunker annotation sequences for the words of the corpus

The resulted scores of the evaluation process are given in Table 3. Analyzing the numbers of Table 3, one can observe that the scores for Precision and Recall do not critically depend on the size of the data sets (200 sentences of the first data set vs. 1420 sentences for the second data set).

Table 3. Data sets and the resulted precision and recall numbers

Corpus size	Number of tokens (words)	Precision	Recall
200 sentences	3433	0.8691	0.8411
1420 sentences	22345	0.8542	0.8225

5. Conclusions

The proposed mechanism provides a way to generate syntactic representations for a language without many parsing tools (like Romanian) by reusing tools of an intense studied language (English) to which word-alignments could be provided. It is well-known that the lexical alignments influence greatly the alignment of internal nodes in two parallel syntactic trees.

From our description it can be easily deduced that wrong or incomplete alignments can affect greatly the knowledge transfer. Also, the quality of the source representations has a great impact on the target induced representations.

Nevertheless, the method has to be improved in order to deal with specific constructions for the target language, which do not have any correspondence in the source language. As the authors say in [24], in order to achieve better results, the target language specific syntactic structures require a pre- and post- processing of the data.

It is important to say that although the treebank generation mechanism presented in this paper was carried out on a specific language pair, that is English and Romanian, it is so far language independent.

Acknowledgements. The author M. Colhon has been funded for this research by the strategic grant POSDRU/89/1.5/S/61968, Project ID 61986 (2009), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007-2013.

Also, the author would like to thank the Natural Language Group of Faculty of Computer Science, *Al. I. Cuza University of Iaşi*, Romania, for providing the English-Romanian corpus upon which the presented treebank generation mechanism was developed and also evaluated.

References

1. J.G. Araújo and H.M. Caseli, “Alignment of Portuguese-English syntactic trees using part-of-speech filters”. In Workshop on Natural Language Processing and web-based technologies (IBERAMIA-2010), Bahía Blanca,1-10. (2010)
2. D. Bamman, M. Passarotti, G. Crane, and S. Raynaud, “Guidelines for the Syntactic Annotation of Latin Treebanks” (v. 1.3), Technical report, Tufts Digital Library, Medford, 2007, http://nlp.perseus.tufts.edu/syntax/ treebank/1.3/ docs/guidelines.pdf
3. H. de Medeiros Caseli, A. M. de Paz Silva and M. das Graças Volpe Nunes, “Evaluation of Methods for Sentence and Lexical Alignment of Brazilian, Portuguese and English Parallel Texts”. In Brazilian Symposium on Artificial Intelligence - SBIA, 184-193. (2004)
4. A. Ceauşu, “Rich morpho-syntactic description for factored machine translation with highly inflected languages as target”. In Workshop on Machine Translation and Morphologically-rich langauges, University of Haifa. (2011)
5. A. Ceauşu, D. Tufiş, “Addressing SMT Data Sparseness when Translating into Morphologically-Rich Languages”, NLPSC 2011, Special Issue Human-Machine Interaction in Translation, August 2011, Copenhagen, Denmark. (2011)
6. D. Cristea, C. Forăscu, “Linguistic Resources and Technologies for Romanian Language”, Computer Science Journal of Moldova, vol. 14, no. 1(40). (2006)
7. J. Cuřín, M. Čmejrek, J. Havelka, V. Kuboň, “Building a Parallel Bilingual Syntactically Annotated Corpus”, in K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, 168–176. (2005)
8. R. Edqvist, “Developing a Core Lexicon for a Corpus-based Machine Translation System”, Master’s thesis in Computational Linguistics, Uppsala University, Department of Linguistics and Philology. (2005)

9. T. Erjavec, “MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora”, Lexicons and Corpora. (2010)
10. T. Erjavec, B. Sárossy, “Morphosyntactic Tagging of Slovene Legal Language”, Informatica(30), 483-488. (2006)
11. H. J. Fox. “Phrasal cohesion and statistical machine translation”. In Proceedings of. EMNLP-02, 304–311. (2002)
12. M. Galley, M. Hopkins, K. Knight and D. Marcu, “What’s in a translation rule?,” In Proceedings of HLT-NAACL 2004, Publisher: Association for Computational Linguistics, Boston, USA, 273-280. (2004)
13. M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, I. Thayer, “Scalable Inference and Training of Context-Rich Syntactic Translation Models”. In: ACL, 961–968. (2006)
14. A. de Gispert, J. Pino, W. Byrne, “Hierarchical Phrase-based Translation Grammars Extracted from Alignment Posterior Probabilities”. In Proceedings of EMNLP'2010, 545-554. (2010)
15. A. Göhring, “Spanish Expansion of a Parallel Treebank”, Ph.D Thesis, University of Zürich, Switzerland. (2009)
16. N. Ide, P. Bonhomme and L. Romary, “XCES: An xml-based encoding standard for linguistic corpora”. In Proceeding of the Second International Language Resources and Evaluation Conference, Paris: European Language Resources Association. (2000)
17. E. Irimia, “EBMT Experiments for the English-Romanian Language Pair”. In Recent Advances in Intelligent Information Systems, ISBN 978-83-60434-59-8, 91-102
18. JRC-Acquis, Available: http://langtech.jrc.it/JRC-Acquis.html
19. E. F. Kelly, Philip J. Stone, “Computer Recognition of English Word Senses”, North-Holland, Amsterdam. (1975)
20. D. Klein, C. D. Manning, “Accurate Unlexicalized Parsing”, In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430. (2003)
21. Y. Liu, Y. Huang, Q. Liu, S. Lin, “Forest-to-string statistical translation rules”, in: ACL, 704–711. (2007)
22. M. P. Marcus, B. Santorini and M. A. Marcinkiewicz, “Building a Large Annotated Corpus of English: The Penn Treebank”. In COMPUTATIONAL LINGUISTICS, vol. 19(2), 313-330. (1993)
23. Melamed, I.D.: Manual Annotation of Translational Equivalence: The Blinker Project. In IRCS Technical Reports Series, University of Pennsylvania. (1998)
24. V.B. Mititelu and R. Ion, “Automatic Import of Verbal Syntactic Relations Using Parallel Corpora”. In Proceedings of Recent Advances in Natural Language Processing, Borovets, Bulgaria. (2005)
25. MULTEXT-East version 3 specifications, http://nl.ijs.si/ME/V3/msd
26. MULTEXT-East version 4 specifications, http://nl.ijs.si/ME/V4/
27. S. De Neefe and K. Knight, “Synchronous Tree Adjoining Machine Translation”. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing EMNLP 2009 vol. 2. (2009)
28. NEGRA Treebank, http://www.coli.uni-sb.de/sfb378/negra-corpus/
29. F.J. Och, and H. Ney, A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29, 19-51. (2003)
30. Penn Treebank, http://www.cis.upenn.edu/~treebank/
31. Prague Dependency Treebank, http://ufal.mff.cuni.cz/pdt2.0/
32. Swedish Treebank: http://stp.ling.uu.se/~nivre/swedish_treebank
33. J. Tinsley, M. Hearne and A. Way, “Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation”. In K. De Smedt, J. Hajič and S.

Kübler (Eds.), Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. (2007)
34. D. Tufiş, R. Ion, "Parallel Corpora, Alignment Technologies and Further Prospects in Multilingual Resources and Technology Infrastructure", in C. Burileanu and H-N Teodorescu (Eds.), Proceedings of the 4th Conference on Speech Technology and Human-Computer Dialogue, SpeD 2007, Iaşi, Romania. (2007)
35. D. Tufiş, S. Koeva, T. Erjavec, M. Gavrilidou, and C. Krstev, "Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages", in M. Tadić, M. Dimitrova-Vulchanova and S. Koeva (eds.) Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008), 145-152, Dubrovnik, Croatia, ISBN 978-953-55375-0-2. (2008)
36. M. Volk, J. Lundborg, and M. Mettler, "Alignment tools for parallel treebanks", in Proc. of The Linguistic Annotation Workshop at the Association for Computational Linguistics (LAW-ACL). (2007)
37. J. Vičič, and A. Broddnik, "Parse Tree Based Machine Translation for Less-used Languages", Available: http://mrvar.fdv.uni-lj.si/pub/mz/m25.1/abst/vicic.htm
38. G. Maillette de Buy Wenniger, M. Khalilov and K. Sima'an, "A Toolkit for Visualizing the Coherence of Tree-based Reordering with Word-Alignments", in The Prague Bulletin of Mathematical Linguistics no. 94, 97–106. doi: 10.2478/v10108-010-0024-4. (2010)
39. K. Yamada, K. Knight, "A syntax-based statistical translation model". In Proceedings of the 39th Meeting of the Association for Computational Linguistics ACL 2001, 523-530, Toulouse, France. (2001)
40. D. Zhang, M. Li, C.-h. Li, M. Zhou, "Phrase reordering model integrating syntactic knowledge for SMT", in: EMNLP/CoNLL, 533–540. (2007)

Mihaela Colhon (born Mihaela Ghindeanu): Since 2005: Assistant professor at Departament of Computer Science, University of Craiova, Romania; Since 2009: PhD in Computer Science, Department of Computer Science, Faculty of Mathematics and Computer Science, University of Pitesti, Romania; Competence domains: Logic programming (Prolog and Lisp programming languages), Knowledge Representation, Knowledge Bases, Expert Systems, Natural Language Processing (syntactic analysis); Teaches: Algorithms and Data structures at Departament of Computer Science, University of Craiova and Artificial Intelligence at Departament of Computer Science, University of Craiova, Romania and Departament of Computer Science, University of Bucharest, Romania; Representative articles: 1) Florentina Hristea, Mihaela Colhon, Feeding Syntactic Versus Semantic Knowledge to a Knowledge-lean Unsupervised Word Sense Disambiguation Algorithm with an Underlying Naive Bayes Model, FUND INFORM (2012), 2) Ion Iancu, Nicolae Constantinescu, Mihaela Colhon: Fingerprints Identification using a Fuzzy Logic System, INT J COMPUT COMMUN (2010), 3) Nicolae Tandareanu, Mihaela Ghindeanu, Sergiu Nicolescu: Hierarchical Distributed Reasoning System for Geometric Image Generation , INT J COMPUT COMMUN (2009).

Received: January 30, 2012; Accepted: May 28, 2012.