Syntactic Translation Patterns from a Parallel Treebank

Mihaela Colhon
University of Craiova, Romania
Departament of Computer Science
A.I.Cuza street, no. 13, code 200585
mcolhon@inf.ucv.ro

ABSTRACT

The goal of the presented parallel phrase extraction algorithm is to provide rich and robust set of translation syntactic patterns. To make this approach feasible, we consider the phrase-to-phrase alignments of a bilingual treebank annotated with syntactic constituents. For the intended purpose, the extracted phrasal nodes are encoded by the syntactical information of their components, highlighting some special constructs such as the functional words.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*; I.2.6 [**Artificial Intelligence**]: Learning—*Parameter learning*

Keywords

Parallel syntactic patterns, phrase-based translation

1. INTRODUCTION

Parallel corpora can be used in order to generate extremely valuable linguistic knowledge such as they can support automatic identification of segments of texts that represent reciprocal translations [13]. Two segments of texts from a *bitext* (parallel corpora) which represent reciprocal translations make a *translation unit* [13]. The translation units that correspond to syntactic phrases can be used to generate other sentences in the target language of a Machine Translation system: instead of generating translation of individual words in the source language, generate translations of the phrases and assemble the final translation by a permutation of these [14].

Methods for Machine Translation (MT) have increasingly leveraged not only the formal machinery of syntax but also linguistic tree structures of either the source language, the target language or both. Phrase based statistical MT (PB-SMT) techniques for extracting phrases although not syntactically motivated, enjoy a very high coverage [1]. Basic PB-SMT systems work with phrase pairs that are consistent with the word alignment: the words of a phrase are contiguous strings consisting of words aligned to each other and not to words outside [8].

Machine translation based on syntactic trees has been extensively studied in the last years due to the general need of improving the performance of the state-of-the-art PB-SMT [2].

Alignment of the parse trees can offer structural alignment between two parallel sentences, more precisely, can help an experiment for testing the feasibility of the automatic cross-lingual transfer of syntactic constituents. Broadly speaking, a transfer component is a system of rules that relate words and structures in one language to words and structures of another language (the target language).

Traditionally, phrases are taken to be syntactic constituents of a sentence. Even if not all the words between two phrases are aligned, the phrases can still align very well [11]. By aligning the inner nodes of two parallel parse trees, the phrases represented by these nodes are put in correspondence as the subtrees of the syntactic analysis encode the structure of the represented syntactic phrases.

Such techniques have shown that starting with large syntactic phrase tables and preferring syntactic phrases when overlapping with non-syntactic ones allows the learning of ”translation knowledge”. They show improvements in decoding speeds and also improvement in translation quality that results from the precision of these syntax motivated phrases [1].

Most of the phrases identified in the parse trees are expected to be translated without interleaving with other phrases or words. In general, noun phrases tend to obey the above rule in a much greater degree. At the opposite corner, the verb phrases usually suffer modifications in structure during translation caused by the adjunct movement [4].

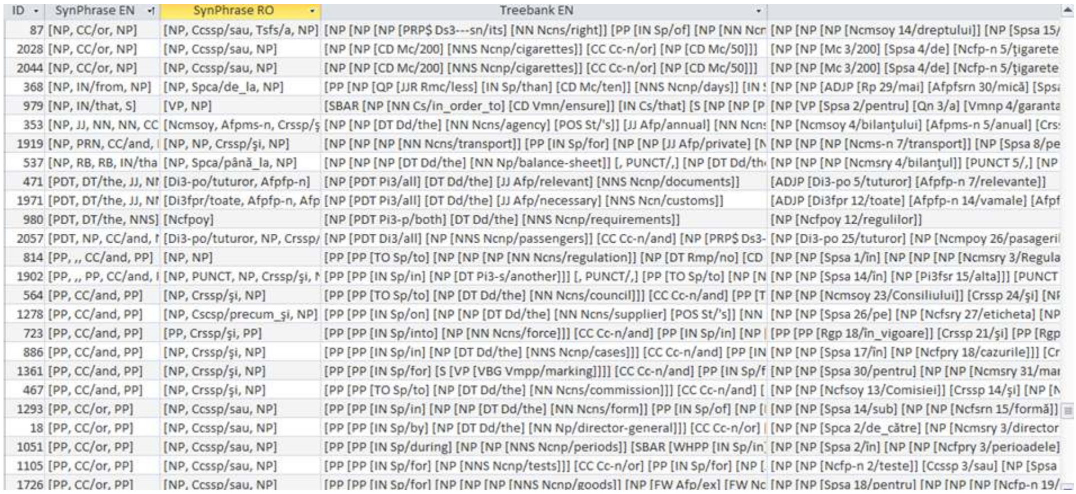
The goal of the presented algorithm for extracting parallel syntactic patterns from a bilingual treebank is to generate a set of good-quality translation patterns intended to be learned by a statistical Syntax-based Machine Translation.

The presented parallel syntactic sequences were extracted from a treebank with syntactic constituents, an English-Romanian Treebank [5]. The treebank was 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 based on the syntactic phrases of the English parallel texts automatically obtained by means of a syntactic parser, the Standford Parser [12]. The Romanian trees generation mechanism reuses and adjusts existing tools and algorithms for cross-lingual transfer of syntactic constituents and syntactic trees alignment.

The treebank was constructed upon 1420 sentences from an English-Romanian parallel corpus developed at A.I.Cuza

BCI'12, September 16–20, 2012, Novi Sad, Serbia.

Local Proceedings also appeared in ISBN 978-86-7031-200-5, Faculty of Sciences, University of Novi Sad.



ID	SynPhrase EN	SynPhrase RO	Treebank EN	
87	[NP, CC/or, NP]	[NP, Ccssp/sau, Tsfs/a, NP]	[NP [NP [NP [PRP$ Ds3---sn/its] [NN Ncns/right]] [PP [IN Sp/of] [NP [NN Ncr	[NP [NP [NP [Ncmsoy 14/dreptului]] [NP [Spsa 15/
2028	[NP, CC/or, NP]	[NP, Ccssp/sau, NP]	[NP [NP [CD Mc/200] [NNS Ncnp/cigarettes]] [CC Cc-n/or] [NP [CD Mc/50]]]	[NP [NP [Mc 3/200] [Spsa 4/de] [Ncfp-n 5/ţigarete
2044	[NP, CC/or, NP]	[NP, Ccssp/sau, NP]	[NP [NP [CD Mc/200] [NNS Ncnp/cigarettes]] [CC Cc-n/or] [NP [CD Mc/50]]]	[NP [NP [Mc 3/200] [Spsa 4/de] [Ncfp-n 5/ţigarete
368	[NP, IN/from, NP]	[NP, Spca/de_la, NP]	[PP [NP [QP [JJR Rmc/less] [IN Sp/than] [CD Mc/ten]] [NNS Ncnp/days]] [IN S	[NP [NP [ADJP [Rp 29/mai] [Afpfsrn 30/mică] [Spsa
979	[NP, IN/that, S]	[VP, NP]	[SBAR [NP [NN Cs/in_order_to] [CD Vmn/ensure]] [IN Cs/that] [S [NP [NP [P	[NP [VP [Spsa 2/pentru] [Qn 3/a] [Vmnp 4/garanta
353	[NP, JJ, NN, NN, CC	[Ncmsoy, Afpms-n, Crssp/ş	[NP [NP [DT Dd/the] [NN Ncns/agency] [POS St/'s]] [JJ Afp/annual] [NN Ncns	[NP [Ncmsoy 4/bilanţului] [Afpms-n 5/anual] [Crs
1919	[NP, PRN, CC/and,	[NP, NP, Crssp/şi, NP]	[NP [NP [NP [NN Ncns/transport]] [PP [IN Sp/for] [NP [NP [JJ Afp/private] [N	[NP [NP [NP [Ncms-n 7/transport]] [NP [Spsa 8/pe
537	[NP, RB, RB, IN/tha	[NP, Spca/până_la, NP]	[NP [NP [NP [DT Dd/the] [NN Np/balance-sheet]] [, PUNCT/,] [NP [DT Dd/th	[NP [NP [NP [Ncmsry 4/bilanţul]] [PUNCT 5/,] [NP
471	[PDT, DT/the, JJ, N	[Di3-po/tuturor, Afpfp-n]	[NP [PDT Pi3/all] [DT Dd/the] [JJ Afp/relevant] [NNS Ncnp/documents]]	[ADJP [Di3-po 5/tuturor] [Afpfp-n 7/relevante]]
1971	[PDT, DT/the, JJ, N	[Di3fpr/toate, Afpfp-n, Afp	[NP [PDT Pi3/all] [DT Dd/the] [JJ Afp/necessary] [NNS Ncn/customs]]	[ADJP [Di3fpr 12/toate] [Afpfp-n 14/vamale] [Afpf
980	[PDT, DT/the, NNS]	[Ncfpoy]	[NP [PDT Pi3-p/both] [DT Dd/the] [NNS Ncnp/requirements]]	[NP [Ncfpoy 12/regulilor]]
2057	[PDT, NP, CC/and,	[Di3-po/tuturor, NP, Crssp/	[NP [PDT Di3/all] [NP [NNS Ncnp/passengers]] [CC Cc-n/and] [NP [PRP$ Ds3-	[NP [Di3-po 25/tuturor] [NP [Ncmpoy 26/pasageril
814	[PP, ,, CC/and, PP]	[NP, NP]	[PP [PP [TO Sp/to] [NP [NP [NP [NN Ncns/regulation]] [NP [DT Rmp/no] [CD	[NP [NP [Spsa 1/în] [NP [NP [NP [Ncmsry 3/Regula
1902	[PP, ,, PP, CC/and,	[NP, PUNCT, NP, Crssp/şi,	[PP [PP [IN Sp/in] [NP [DT Pi3-s/another]]] [, PUNCT/,] [PP [TO Sp/to] [NP [N	[NP [NP [Spsa 14/în] [NP [Pi3fsr 15/alta]]] [PUNCT
564	[PP, CC/and, PP]	[NP, Crssp/şi, NP]	[PP [PP [TO Sp/to] [NP [DT Dd/the] [NN Ncns/council]]] [CC Cc-n/and] [PP [T	[NP [NP [Ncmsoy 23/Consiliului]] [Crssp 24/şi] [N
1278	[PP, CC/and, PP]	[NP, Cscsp/precum_şi, NP]	[PP [PP [IN Sp/on] [NP [NP [DT Dd/the] [NN Ncns/supplier] [POS St/'s]] [NN	[NP [NP [Spsa 26/pe] [NP [Ncfsry 27/eticheta] [NP
723	[PP, CC/and, PP]	[PP, Crssp/şi, PP]	[PP [PP [IN Sp/into] [NP [NN Ncns/force]]] [CC Cc-n/and] [PP [IN Sp/in] [NP	[PP [PP [Rgp 18/în_vigoare]] [Crssp 21/şi] [PP [Rgp
886	[PP, CC/and, PP]	[NP, Crssp/şi, NP]	[PP [PP [IN Sp/in] [NP [DT Dd/the] [NNS Ncnp/cases]]] [CC Cc-n/and] [PP [IN	[NP [NP [Spsa 17/în] [NP [Ncfpry 18/cazurile]]] [Cr
1361	[PP, CC/and, PP]	[NP, Crssp/şi, NP]	[PP [PP [IN Sp/for] [S [VP [VBG Vmpp/marking]]]] [CC Cc-n/and] [PP [IN Sp/f	[NP [NP [Spsa 30/pentru] [NP [NP [Ncmsry 31/ma
467	[PP, CC/and, PP]	[NP, Crssp/şi, NP]	[PP [PP [TO Sp/to] [NP [DT Dd/the] [NN Ncns/commission]]] [CC Cc-n/and] [	[NP [NP [Ncfsoy 13/Comisiei]] [Crssp 14/şi] [NP [N
1293	[PP, CC/or, PP]	[NP, Ccssp/sau, NP]	[PP [PP [IN Sp/in] [NP [NP [DT Dd/the] [NN Ncns/form]] [PP [IN Sp/of] [NP [	[NP [NP [Spsa 14/sub] [NP [NP [Ncfsrn 15/formă]]
18	[PP, CC/or, PP]	[NP, Ccssp/sau, NP]	[PP [PP [IN Sp/by] [NP [DT Dd/the] [NN Np/director-general]]] [CC Cc-n/or]	[NP [NP [Spca 2/de_către] [NP [Ncmsry 3/director
1051	[PP, CC/or, PP]	[NP, Ccssp/sau, NP]	[PP [PP [IN Sp/during] [NP [NP [NNS Ncnp/periods]] [SBAR [WHPP [IN Sp/in]	[NP [NP [Spsa 2/în] [NP [NP [Ncfpry 3/perioadele]
1105	[PP, CC/or, PP]	[NP, Ccssp/sau, NP]	[PP [PP [IN Sp/for] [NP [NNS Ncnp/tests]]] [CC Cc-n/or] [PP [IN Sp/for] [NP [	[NP [NP [Ncfp-n 2/teste]] [Ccssp 3/sau] [NP [Spsa
1726	[PP, CC/or, PP]	[NP, Ccssp/sau, NP]	[PP [PP [IN Sp/for] [NP [NP [NP [NNS Ncnp/goods]] [NP [FW Afp/ex] [FW Nc	[NP [NP [Spsa 18/pentru] [NP [NP [NP [Ncfp-n 19/

Figure 1: A screen shot of the English-Romanian Syntactic Patterns Dataset.

University of Iaşi by the Natural Language Processing Group of Faculty of Computer Science. The corpus is XML encoded obeying a simplified form of the XCES standard [10]. For the bilingual corpus construction, the English and Romanian parts of the Acquis-Communitaire[1] corpus 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 come from MULTEXT-East morphosyntactic specifications (the latest version of these specifications is given in [7]).

2. PARALLEL PATTERNS WITH SYNTACTIC CONSTITUENTS

Following the method of Galley described in [9], the phrase extraction process is supported by the parallel parse trees of the constructed English-Romanian treebank. For each alignment between inner nodes of the syntactic trees, the descendents of aligned nodes are examined. According to the purpose for which the syntactic sequences are extracted, in the list of descendents, some specific words or constructions of certain structure can be highlighted.

For the presented article, the syntactic sequences are intended to provide information about the manner in which the functional words can affect translation. For this reason the functional words are given in the complete word-form accompanied with complete information about their morphosyntactic properties.

In any syntactic structure we can identify two major categories of words: **content words** which describe objects, entities, properties, relationships or events and which are syntactically represented by *nouns*, *adjectives*, *verbs* and *adverbs* and **functional words** that help putting words together in a correct structural sentence form. Also, the functional words can tell how the other components of the sentence are related to each other. The functional words can be *determiners*, *quantifiers*, *prepositions* or *connectives.*

The span of a node n of a syntactic tree is taken to be the subset of nodes that are reachable from n [9]. In a bottom-up fashion, the algorithm for extracting parallel syntactic patterns "visits" each English syntactic tree and expands all its inner nodes that are aligned with at least one node from the Romanian parallel syntactic tree. The spans of the aligned English and Romanian phrasal nodes are taken to be the parallel syntactic patterns of our study and therefore are stored in a database (see Section 2.2). The method is quick and easy enough to be used on large-scale data sets.

Here are some examples of the importance of parallel syntactic patterns from automatically learned translation rules point of view:

- simple lexical patterns for translating special words such as, functional words can be treated as examples of patterns in which optional modifiers are inserted
- patterns in which we found "lexical holes" determined by existence of one-to-zero alignment mapping between the words/tokens of the parallel sequences. For example, English noun phrases that contain the word "of" as separator.
- analyzing large sets of the parallel patterns, we can identify the "part of speech afinities"; it is usually known that translated words tend to keep their part of speech but when this is not the case, the resulted part-of-speech for the translation is not random.

From the English-Romanian Parallel Treebank with syntactic Constituents, 2120 English-Romanian syntactic patterns with functional word were extracted. The representation in which the patterns are stored can provide good enough descriptions of the domain of locality for the functional words.

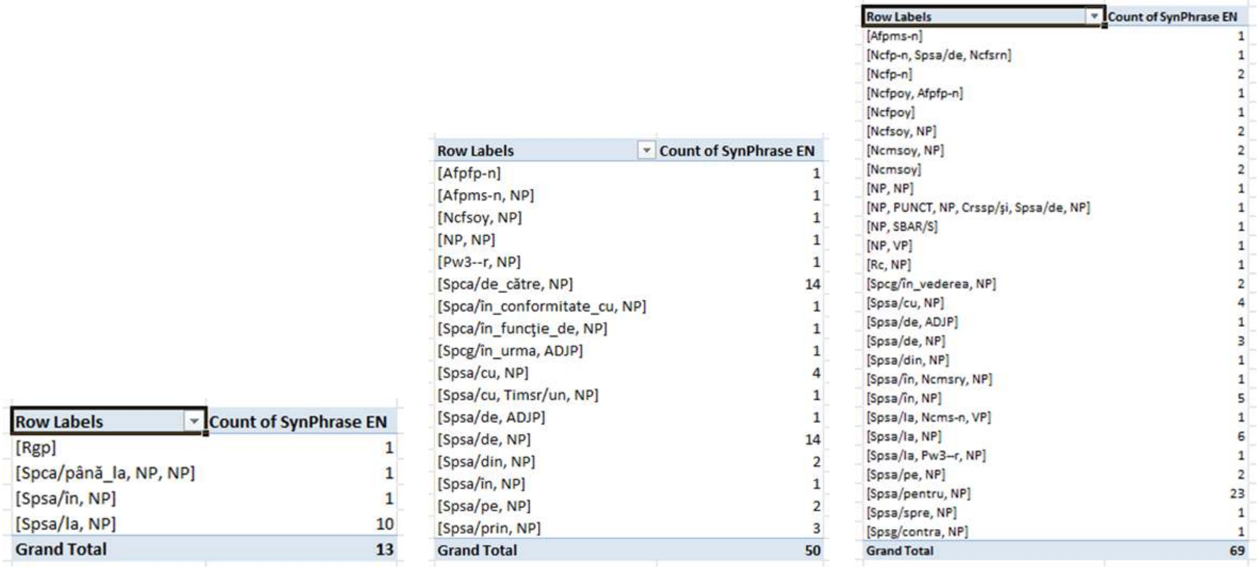
2.1 Representation Formalism

The representation for the English syntactic sequences with functional words is an ordered sequence of elements given in the following form:

$$[\ \{\ \text{Phrasal\_Tag}\ \}^*\ \text{Pos\_Tag/FW}\ \{\ \text{Phrasal\_Tag}\}^*\ ]$$

where by FW we note a functional word.

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



Row Labels	Count of SynPhrase EN
[Rgp]	1
[Spca/până_la, NP, NP]	1
[Spsa/în, NP]	1
[Spsa/la, NP]	10
Grand Total	13

(a) [IN/at, NP]

Row Labels	Count of SynPhrase EN
[Afpfp-n]	1
[Afpms-n, NP]	1
[Ncfsoy, NP]	1
[NP, NP]	1
[Pw3--r, NP]	1
[Spca/de_către, NP]	14
[Spca/în_conformitate_cu, NP]	1
[Spca/în_funcție_de, NP]	1
[Spcg/în_urma, ADJP]	1
[Spsa/cu, NP]	4
[Spsa/cu, Timsr/un, NP]	1
[Spsa/de, ADJP]	1
[Spsa/de, NP]	14
[Spsa/din, NP]	2
[Spsa/în, NP]	1
[Spsa/pe, NP]	2
[Spsa/prin, NP]	3
Grand Total	50

(b) [IN/by, NP]

Row Labels	Count of SynPhrase EN
[Afpms-n]	1
[Ncfp-n, Spsa/de, Ncfsrn]	1
[Ncfp-n]	2
[Ncfpoy, Afpfp-n]	1
[Ncfpoy]	1
[Ncfsoy, NP]	2
[Ncmsoy, NP]	2
[Ncmsoy]	2
[NP, NP]	1
[NP, PUNCT, NP, Crssp/şi, Spsa/de, NP]	1
[NP, SBAR/S]	1
[NP, VP]	1
[Rc, NP]	1
[Spcg/în_vederea, NP]	2
[Spsa/cu, NP]	4
[Spsa/de, ADJP]	1
[Spsa/de, NP]	3
[Spsa/din, NP]	1
[Spsa/în, Ncmsry, NP]	1
[Spsa/în, NP]	5
[Spsa/la, Ncms-n, VP]	1
[Spsa/la, NP]	6
[Spsa/la, Pw3--r, NP]	1
[Spsa/pe, NP]	2
[Spsa/pentru, NP]	23
[Spsa/spre, NP]	1
[Spsg/contra, NP]	1
Grand Total	69

(c) [IN/for, NP]

Figure 2: Statistics for English patterns consisting of a preposition and a noun phrase ([IN, NP]).

By parsing the English sentences with Stanford Parser, 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. In this annotation formalism, the functional words for the English texts can be considered as sentence tokens that in PENN POS tagset formalism have one of the following tags: **CC** (coordinating conjunction), **DT** (determiner) , **IN** (preposition/ subordinating conjunction), **MD** (modal), **PRP** (personal pronoun), **PP$** (possessive pronoun), **RP** (particle), **TO** (word *to*), **WDT** (*wh*-determiner), **WP** (*wh*-pronoun), **WP$** (possessive *wh*-pronoun), **WRB** (*wh*-adverb).

Here are some examples of English syntactic patterns:

- [**NP, PRP, CC/and, NP**] the syntactic phrase having this span is made of two noun phrases (NP) linked by a personal pronoun (PRP) and a functional word, the coordinating conjunction *and* (in this specific order). The two syntactic phrases NP are not expanded because each of them has its own alignment, and thus, their structure is given in other parallel syntactic sequence.
- [**RB, JJ, CC/and, JJ**] the syntactic phrase having this structure contains two adjectives (JJ) linked by a functional word (the conjunction *and*) and preceded by an adverb (RB).

Following the same representations, the corresponding Romanian syntactic sequences are encoded in a similar format.

The Romanian syntactic trees of the English-Romanian Treebank were automatically constructed by means of a bottom-up tree generation algorithm guided by the word alignments of the corpus ([5]). As a consequence, the annotations for the Romanian words preserve the MULTEXT-EAST words specifications of the corpus as these data include enough morphosyntactic details needed in any syntactic study, while for labeling the phrasal constituents, the PENN Treebank Phrasal tags are used.

As a direct consequence, the Romanian functional words are those tokens/words that in MULTEXT-EAST Tagset formalism have MSD tags with the following prefixes: **P**_ (pronoun) such as **Pd**_ (demonstrative pronoun), **Ps**_ (possessive pronoun), **Px**_ (reflexive pronoun), **D**_ (determiner), **T**_ (article), **S**_ (adposition), **C**_ (conjunction), **Q**_ (particle).

The representation for the Romanian syntactic sequences with functional words is an ordered sequence of elements given in the following form:

$$[\ \{\ \text{Phrasal\_Tag}\ \}^{*}\ \text{MSD Tag/FW}\ \{\ \text{Phrasal\_Tag}\}^{*}\ ]$$

where by FW we note a functional word and by MSD we note the morphosyntactic descriptions encoded in MULTEXT-East morphosyntactic specifications.

Here are some examples of Romanian syntactic patterns:

- [**Di3-po—e/altor, NP**] the syntactic phrase given by this sequence contains a determiner (a MSD tag starting with **D**) followed by a noun phrase
- [**VP, Crssp/şi, Tsfs/a, NP**] the syntactic phrase whose structure is encoded in this pattern is made of a verb phrase (VP) and a noun phrase (NP) liked by a conjunction (a **C**_ MSD tag) and an article (a **T**_ MSD tag).

2.2 Linguistic Resource with Syntactic Patterns

Each resulted parallel sequence is stored into a database record with four fields (see Figure 1). The $SynPhrase\ En$ field stores the span of an English syntactic phrase, while in the $SynPhrase\ RO$ field the span of the aligned Romanian syntactic phrase is given. The last two fields include the PENN syntactic subtrees rooted at the aligned syntactic phrases.

Indeed, $Treebank\ EN$ gives the bracket representation for the subtree rooted at the English phrase while $Treebank\ RO$ is the subtree corresponding to the Romanian phrase. Examples of some records of this linguistic resource are listed in Table 1.

Certain statistics about the translation of a particular English syntactic sequence into Romanian language can be easily obtained upon the constructed database table with the described information.

Table 1: Examples of English-Romanian Syntactic Patterns Together with Their Treebank Representations

Phrase En	*Phrase RO*	*Treebank EN*	*Treebank RO*
[IN/as, NP]	[Rw/cât, mai/Rp, ADJP]	[PP [IN Rsp/as] [NP [NP Afp/strict] [ADJP [RB Cs/as] [JJ Afp/possible]]]]	[PP [Rw 14/cât] [Rp 15/mai] [ADJP [Afpfp-n 16/stricte] [ADJP [Rgp 17/posibil]]]]
[IN/at, NP]	[Spsa/la, NP]	[PP [IN Sp/at] [NP [NP [DT Dd/the] [NN Ncns/end]] [PP [IN Sp/of] [NP [DT Dd/the] [JJ Afp/financial] [NN Ncns/year]]]]]	[NP [Spsa 1/la] [NP [NP [Ncfsry 2/încheierea]] [NP [Ncmsoy 3/exerciţiului] [ADJP [Afpms-n 4/financiar]]]]]

From the statistics illustrated in Figure 2, one can observe that the translation in Romanian for the English syntactic pattern [IN/at, NP] do not change the order between the noun phrase and the preceding preposition and replace the preposition "at" with the Romanian preposition "la". The preposition "by" from the English pattern [IN/by, NP] is equally translated with Romanian prepositions "de" and with the Romanian prepositional collocation "de_către", while the preposition "for" from the English sequence [IN/for, NP] is translated with the Romanian preposition "pentru".

3. CONCLUSIONS

Statistical Machine Translation systems that use syntactical information in the translation process must be trained with syntactic patterns that correspond to reciprocal translations in the languages of the MT system. Such training can help the translation not only with the structural differences between the translations but also with the re-ordering problems at the target sentence words [3].

Even if the lexical coverage of the used corpus, the Acquis Communitaire corpus, is not representative, a MT system can still benefit from the translations similar in structure and semantics that exist between the parallel sentences of the corpus.

Also the meanings of some special words, such as functional words, can be easily explored by analysing the changes during the translation suffered by syntactic patterns consisting of this kind of words. In the way it is constructed now, the resource focuses on the importance the functional words have in a translation process. But the syntactic patterns can be generated in order to highlight other constructions, for example, the polylexicals units of a natural language phrase.

As a future work, we intend to enlarge the size of the bilingual treebank in order to permit generation of a larger set of parallel syntactic patterns.

4. ACKNOWLEDGMENTS

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.

5. REFERENCES

[1] V. Ambati, A. Lavie, and J. Carbonell. Extraction of syntactic translation models from parallel data using syntax from source and target languages. In *MT Summit XII*, 2009.

[2] J. G. Araùjo and H. M. Caseli. Alignment of portuguese-english syntactic trees using part-of-speech filters. `http://www.cs.famaf.unc.edu.ar/~laura/nlpw/nlpw/papers/Araujo_Caseli.pdf`. Online; accessed 19-June-2012.

[3] A. Ceauşu. Rich morfo-syntactic description for factored machine translation with highly inflected languages as target. In *Workshop on Machine Translation and Morphologically-rich Languages, University of Haifa*, 2011.

[4] M. Colhon. A contrastive study of syntactic constituents in english and romanian texts. In *Proc. of the Workshop "Language Resources and Tools with Industrial Applications"*, pages 11–20, 2011.

[5] M. Colhon. Language engineering for syntactic knowledge transfer (submitted). *Computer Science and Information Systems Journal*, 2012.

[6] D. Cristea and C. Forăscu. Linguistic resources and technologies for romanian language. *Computer Science Journal of Moldova*, 14(1(40)), 2006.

[7] T. Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of LREC'10*. ELRA, 2010.

[8] M. K. G. Wenniger and K. Sima'an. A toolkit for visualizing the coherence of tree-based reordering with word-alignments. In *Proc. of the 5h MT-Marathon*, pages 97–104, 2010.

[9] M. Galley, M. Hopkins, K. Knight, and D. Marcu. What's in a translation rule? In *Proc. of HLT-NAACL 2004*, pages 273–280. ACL, Boston, USA, 2004.

[10] N. Ide, P. Bonhomme, and L. Romary. Xces: An xml-based encoding standard for linguistic corpora. In *Proc. of the 2nd LREC, Paris: ELRA*, 2000.

[11] R. Ion, R. Ceauşu, and D. Tufiş. Dependency-based phrase alignment. In *Proc. of the 5th LREC*, pages 1290–1293, 2006.

[12] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proc. of the 41st Annual Meeting of ACL*, pages 423–430, 2003.

[13] D. Tufiş and R. Ion. Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. In *Proc. of the 4th SPED*, 2007.

[14] K. Yamada and K. Knight. A syntax-based statistical translation model. In *Proc. of ACL*, pages 523–530, 2001.