UDK: 81’322.4

ACQUIRING SYNTACTIC TRANSLATION RULES FROM A PARALLEL TREEBANK

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

Abstract
The presented study addresses the issue of machine translation system development,as it makes use of a parallel syntactic phrase extraction algorithm in order to obtaina rich and robust set of syntactic translation patterns. To make this approach feasible, only phrase-to-phrase alignments from a bilingual treebank with syntactic constituents are considered.
The knowledge enclosed in this study is defined by means of syntactic motivated translation sequences, written in terms of the morpho-syntactic specifications developed during the MULTEXT-East Project (Erjavec, 2010).

Keywords
Parallel syntactic patterns, phrase-based translation.

1. Introduction

The basic approaches to Machine Translation (MT) are *direct translation*, *transfer-based translation* and *statistical translation*. In a direct MT, the translation is done word by word or phrase by phrase while the transfer-based approach does not pay attention to the lexical level, the translation being generated by means of a transfer mechanism, defined at a structural level, in order to transfer the syntactic representations of the source constructions to the corresponding structures in the target language. A statistical MT system consists of a model of the translation relation between two languages and the rules of translation are acquired automatically from the training resource (in this case, a bilingual corpus).

In Natural Language Processing (NLP), the corpus is used to provide an MT system with empirical and statistical data. 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 (Tufiş& Ion, 2007). Two segments of texts from a *bitext* (parallel text), which represent reciprocal translations, make a *translation unit*(Tufiş& Ion, 2007). The translation units that correspond to syntactic phrases can be used to generate other sentences in the target language of an MT system: instead of generating the translation of individual words in the source language, generate translations of phrases and assemble the final translation by a permutation of these (Yamada & Knight, 2001).

Parallel corpora can support any kind of MT system: in the direct approaches, corpora are used to extract information about lexical units (how a particular word is translated in a certain environment), in the transfer-based approaches, corpora are used to extract transfer rules and in the statistical approaches they are used to extract translation rules and to assign probabilities to possible translations.

A parallel treebank is a special type of parallel corpus in which the sentences are syntactically annotated. The syntax trees of a bitext from a parallel treebank are aligned ata sub-sentential level (word, phrase, or clause level) which is denoted in the literature as *phrase alignment*. Parallel treebanks can serve as a corpus for automatic derivation of transfer rules or extraction of bilingual dictionaries or, in general, for translation studies (Samuelsson& Volk. 1993).

Machine Translation methods 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 very high coverage (Ambati et al., 2009). Basic PB-SMT systems work with phrase pairs that are consistent with word alignment: the words of a phrase are contiguous strings, consisting of words aligned to each other and not to the words outside (Wenniger et al., 2010).

Machine translation based on syntactic trees has been extensively studied in recent years due to the general need of improving the performance of state-of-the-art PB-SMT (Araùjo&Caseli, 2010).

The node alignments of two parallel parse trees can offer structural alignment information about the represented pair of sentences, more precisely,they can help an experiment 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 in 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, phrases can still align very well. By aligning the inner nodes of two parallel parse trees, the phrases

represented by these nodes are put in correspondence and, as a direct result, the subtrees rooted in these nodes are also aligned.

The current practice in hierarchical phrase-based translation extracts regular phrases and hierarchical rules from word-aligned bitexts (Gispert et al., 2010). Even if alignment-based models are not suitable for direct use in translation, they can still provide a great deal of useful information, such as the statistics needed to build translation templates.

These techniques have shown that starting with large syntactic phrase tables and preferring syntactic phrases when overlapping with non-syntactic ones allow the learning of "translation knowledge". They show improvements in decoding speeds,as well as in translation quality that results from the precision of these syntax motivated phrases (Ambati et al., 2009).

Most of the phrases identified in the parse trees are expected to be translated without interleaving with other phrases/words. In general, noun phrases tend to obey the above rule to a much greater degree. Conversely, verb phrases usually suffer modifications in structure during translation, caused by adjunct movement (Colhon, 2011).

By employing node alignments between the tokens of parallel parse trees, hierarchical translation models for syntactic MT systems can be extracted. By gathering parallel syntactic patterns from a bilingual treebank, we can generate sets of good-quality translation patterns, intended to be learned by a statistical Syntax-Based Machine Translation. The main scope of the presented study is to create syntax motivated translation patterns.

1.1 English-Romanian Corpus. What's inside?

Machine Translation is a difficult and complex task. One way to improve the result of automatic translation is to limit the texts to be translated to a specific domain, containing similar content. This approach reduces the amount of ambiguity in the translation process, and thus the lexicon can be much smaller and more specific. MT systems using this approach are often corpus-based, which means they use translation data in the form of parallel corpora to construct their linguistic resources. Also, parallel corpora can be used to support contrastive inter-lingua studies.

In order to be used in an MT system for training and testing, the corpus has to be preprocessed. The preprocessing phase usually involves sentence splitting, tokenization, POS-tagging, lemmatization.

The POS tagset varies among different languages, but most tagset definitions can be translated from one language to another using simple translation tables. MULTEXT-East project (Erjavec et al., 2003) defined the *morphosyntactic descriptions* (MSD) as the same tagset for all the languages of the project.

The corpus that supports the presented study is called *JRC-Acquis*, the compiled part of parallel texts from the *Acquis Communautaire* legislative texts[1]. The *Acquis Communautaire*is the total body of European Union (EU) law applicable in the EU Member States. This collection changes continuously and currently comprises texts written between 1950s and 2008 in all languages of the EU Member States. The *Acquis Communautaire* is a collection of parallel texts in 22 official languages, including Romanian.

The *JRC-Acquis* corpus presents several advantages from the point of view of an MT usage:

- it is dedicated to a specific domain, the legislation of the European Union, but is not vocabulary restricted, as the included texts cover a variety of fields that fall under this legislation;

1 A significant part of the Acquis Communautaire texts has been compiled by the Language Technology Group of the European Commission into a parallel corpus, the JRC-Acquis corpus (Cristea& Forăscu, 2006).

- it is a consistent parallel corpus, at least in theory, because of the official nature of its documents which have to be in a very well organized structure (Iftene et al., 2010).

For the presented study, we have used an English and Romanian corpus that comes from English and Romanian parallel parts of the *JRC-Acquis* corpus. It consists of 1420 sentences and a total number of 12613 tokens in both languages.

1.2 Producing a Parallel Treebank

The presented study was implemented on an English-Romanian treebank (Colhon, 2012) constructed upon 1420 sentences from an English-Romanian corpus developed at the Alexandru Ioan Cuza University of Iaşi by the Natural Language Processing Group[2] of the Faculty of Computer Science. For this bilingual corpus construction, parallel English and Romanian parts of the *JRC-Acquis* corpus were used.

In order to make them useful, raw texts of the bilingual corpus were annotated at several levels encoded in XML, by adopting a simplified form of the XCES standard (Ide et al., 2000). Thus, the texts were segmented and lexical information added to them. As a result of that, the sentences have their boundaries marked and each token has its part of speech (POS), lemma, and morpho-syntactic information (gender, number, person, case, etc.) attached by running an automatic processing chain that includes: sentence segmentation, tokenisation, POS-tagging and lemmatisation (Simionescu, 2012).

The tagsets used to annotate the words in the bilingual corpus come from the MULTEXT-East morphosyntactic specifications. The latest version of these specifications is given in (Erjavec, 2010).

To build a dependency treebank, human annotators must decide for each word which is the one it depends on. Determining dependencies often involves a deep interpretation process and this is why for the same word sequence, sometimes, different dependency structures could be negotiated among annotators. In contrast, the decisions human annotators should take while building phrase-structure treebanks usually lead to much less ambiguity (Colhon&Cristea, 2012).

In order to construct a parallel treebank from the bilingual corpus, a broad-coverage statistical parser of the source language was needed and also word-alignments had to be provided. Because the alignment information is crucial in the treebank development process for the target language part of the corpus, we choose to specify it manually.

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 (Klein& Manning, 2003). The Romanian tree generation mechanism reuses and adjusts the existing tools and algorithms for cross-lingual transfer of syntactic constituents and alignment of syntactic trees. We have manually corrected the resulting treebank for completeness[3] (consistency[4] was ensured by the bilingual corpus construction).

The English-Romanian treebank used in this study is made of syntactic trees automatically obtained by running a well-known parser on the English part of the corpus and their structures projected on the corresponding Romanian texts. These projections are defined based on the word-alignments specified between the parallel sentences of the corpus.

Because of the implemented treebank gen-

2 Information about the NLP Group of the Alexandru Ioan Cuza University of Iaşi can be found at the web address: http://instrumente.infoiasi.ro/NLPTest/about.jsp

3 Completeness means that each token and each node is part of the syntactic tree (Samuelsson& Volk, 1993).

4 Consistency means that the same token sequence is annotated in the same way across the treebank(Samuelsson & Volk, 1993).

eration mechanism for Romanian texts, for each parallel pair of syntactic trees from the resulting English-Romanian treebank, the English syntactic constituents are implicitly aligned with theRomanian phrases (which are sequences of target words). From this hierarchical alignment information, transformational syntactic rules can be acquired, as it will be shown in the subsequent sections.

2. Parallel Patterns with syntactic constituents

Following the method of Galley described in (Galley et al., 2004), the phrase extraction process is supported by the parallel parse trees of the constructed English-Romanian treebank. For each alignment between the inner nodes of the syntactic trees, the descendants of the aligned nodes are examined. According to the purpose for which the syntactic sequences are extracted, in the list of descendants, some specific words or constructions of a certain structure can be highlighted.

For the presented article, the syntactic sequences are described in order to provide information about the manner in which functional words can affect translation. For this reason, functional words are given in the complete wordform, accompanied by complete information about their morpho-syntactic 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*.
- *functional words* that help putting words together in a correct structural sentence form. Also, functional words can tell how other components of the sentence are related to each other. 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* (Galley et al., 2004). 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 syntactic structure and the spans of the aligned English and Romanian phrasal nodes are taken to be the parallel 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 roles the parallel syntactic patterns have from the automatically learned translation rules point of view:

- simple lexical patterns for translating special words, such as *functional words*
- 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 made of two nouns linked by the functional word "of". In the corresponding Romanian translation the separator disappears.
- analyzing large sets of parallel patterns, we can identify "part of speech affinities"; it is usually known that translated words tend to keep their part of speech, but when this is not the case, the resulting part-of-speech ofthe translation is not random.

In this study we are interested in the variations or constraints determined by the functional words of source language constructions. We have generated a database consisting of several English and Romanian parallel sequences represented in different formats. Each representation format is intended to capture translation information concerning parallel sequences:

- the syntactic structure of the sequences,

- the word-alignments
- the MSD information of the words; for functional words we consider also the word form
-

2.1 Representation Formalism

One of the requirements for a generic rule induction framework from parallel data is to be able to incorporate any kind of syntactic information that may come from either side of the MT languages pair (Ambati et al., 2009). The proposed formalism works with constituentsyntax on both sides by paying attention to the importance that functional words have in the translation process.

The proposed extractor algorithm works in two phases in order to identify parallel sequences from the treebank. In the first phrase the nodes from the source tree that align with the nodes from the target tree are extracted together with their spans, defined in terms of morphosyntactic tags (MSD tags). The only lexical information that is taken into account addresses functional words. We have that the only word forms that appear in the proposed representation correspond to functional words. We call these parallel syntactic phrases *syntactic patterns*.

In the second phase, we extract the subtrees-from two parallel syntactic trees for which the root nodes are aligned and we write their structure following the bracketing scheme for syntactic trees used in the Penn Treebank project[5]. In what follows, this representation will be called *hierarchical patterns*.

From the English-Romanian treebank with syntactic constituents, 4389 English-Romanian parallel patterns (syntactic and hierarchical) were extracted. The representation in which the patterns are stored can provide good enough descriptions of the domain of locality for functional words, but is not restricted to that.

In any pattern-based approaches, it is important to choose properly the pattern description formalism and the pattern structures (Kurc et al., 2010). The knowledge enclosed in this study is represented by syntactic patterns defined in terms of morpho-syntactic specifications, developed during the MULTEXT-East Project (Erjavec, 2010) and in terms of POS and phrasal tag specifications as they were introduced in the Penn Treebank project[6].

2.1.1 English Syntactic Sequences

In the formalism we propose here, each pattern encodes the syntactic structures of English natural language constructions in one of the two representations: syntactic and hierarchical.Both representations pay special attention to the functional words that appear in the represented sequence, as we consider that this kind of words is very important from a MT system point of view.

Here isthe syntactic representation of English syntactic patterns:

[Phrasal_Tagtag(c_1) … tag(c_n)]

wherePhrasal_Tag is the label of the phrase in the Penn Treebank Formalism and tag(c_1) … tag(c_n) are the used notation for the direct constituents c_1 … $c_n$of the represented phrasewhere:

$$tag(c_i) = \begin{cases} Penn\ Phrasal\ Tag(c_i)\ \ c_i - a\ phrase\ constituent \\ Penn\ POS\ Tag(c_i) / c_i, c_i - a\ functional\ word \\ Penn\ POS\ Tag(c_i)\ \ c_i - a\ content\ word \end{cases}$$

Because the English part of the treebank was generated with Stanford Parser, PENN Treebank parse trees were generated. As a direct consequence, the English wordsof the English treebankare annotated with PENN POS tags, as this

5 The Penn Treebank is a structually annotated corpus that consists of a sequence of sentences (over 1 million) taken from the*Wall Street Journal*. Each sentence of the corpus has been annotated for part-of-speech labels and phrase-structures.

6 The web address of the project is http://www.cis.upenn.edu/~treebank/.

is the tagging standard used by Stanford Parser. In this annotation formalism, the functional words for the English texts can be considered to be sentence tokens that in the 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).

For this type of representation, we consider two versions: one that labels the phrasal nodes with their indexes and the constituents of phrasal nodes with the indexes of their Romanian alignments, and another that removes all this index informationby concentrating on the syntactic structuresof phrasal nodes.

The hierarchical pattern representation for the English sequences follows the bracket representation for syntactic trees with constituents (Taylor, 1996). Because the bracket representation for a syntactic constituent specifies the word forms of the phrase, as well such a formalism can provide lexical information for the represented phrase, along with the whole structure representation of the phrase.

In what follows we give two examples of representations English syntactic sequences.

Example 1. The construction "the preparation" taken from the corpus will be represented as follows:

- syntactic pattern:
  [NP-512[7] {3}:DT/the {3}:NN][8](with index information)
  [NP DT/the NN](without index information)
  This representation pays attention to the syntactic structure of an English phrase by highlighting the functional words that are identified not only by their syntactic tag, but also by their word form (see the construction DT/the).
  For the representation with index information, the numbers in braces that precede the syntactic tags, represent the alignments in the parallel Romanian phrase that correspond to this English phrase noted withNP-512. From the attached index information, one can observe that the whole construction encoded byDT/the NN is aligned with a single token in the Romanian phrase which has the index 3.
- hierarchical pattern:
  [NP-512 [[DT/the] [NN preparation]]]
  In this representation, the whole structure including the terminal nodes – the sequence words - corresponding to the syntactic phraseNP-512 is given. This phraseis made up of a determiner [DT/the] followed by thenounchunk [NN preparation].

Example 2. The construction "the proposal from the commission" is represented as follows:

- syntactic pattern:
  [NP-517 {506}:NP {507}:PP] (with index information)
  [NP NP PP](without index information)
  This representation encodesthe structure of an English noun phrase (noted with NP-517 in the formalism with index information) that consists of a noun phrase which is aligned with the constituent of the aligned Romanian phrase having the index 506 and a prepositional phrase aligned with the constituent having the index 507in the Romanian parallel phrase.
- hierarchical pattern:
  [NP-517
  [[NP-508 [DT/the] [NN proposal]] [PP-516 [IN/from] [NP-515 [DT/the]

7 For the inner nodes of a parse tree, we choose to node them by their phrasal labels together with their index values. In this manner, we can distinguish between two inner nodes of a parse tree that have the same phrasal label.

8 See Appendix A for a description of the Penn Treebank tags used in this paper.

[NN commission]]]]]

In this representation, the noun phrase NP-517 is given here with the whole subtreethat is governed by this phrasal node.

Following the same representations, the corresponding Romanian syntactic sequences are encoded in a similar format, with only one difference: for the Romanian words we take into account the MULTEXT-EAST morpho-syntactic annotations, as we consider these specifications to be appropriate for the inflectional richness of the Romanian language.

2.1.2 Romanian Syntactic Sequences

The Romanian syntactic trees of the treebank were automatically constructed by means of a bottom-up tree generation algorithm guided by the word-alignments of the corpus (Colhon, 2012). From the bilingual corpus, upon which the treebank was made, we preserve the annotations the MULTEXT-East word specifications of the corpus, as these data include enough morpho-syntactic details needed by any syntactic study, while for the labeling of phrasal constituents, the PENN Treebank Phrasal tags are used.

As a direct consequence, the Romanian functional words are those tokens/words that in MULTEXT-EastTagset 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_ (apposition), C_ (conjunction), Q_ (particle).

The Romanian syntactic patterns follow the same representation formalism used for the English syntactic patterns, with only a single difference, that the words are annotated with MSD tags instead of Penn POS tags. We have that a Romanian syntactic pattern is represented as a list:

[Phrasal_Tag tag(c1) … tag(cn)]

where Phrasal_Tag is the label of the phrase in the Penn Treebank Formalism and tag(c_1) … tag(c_n) arethe used notation for the direct constituentsc$_1$ … c$_n$ of the phrase where:

$$tag(c_i) = \begin{cases} Penn\,Phrasal\,Tag(c_i),\ c_i - a\ phrase\ constituent \\ MSD\,Tag(c_i)/c_i, c_i - a\ functional\ word \\ MSD\,Tag(c_i),\ c_i - a\ content\ word \end{cases}$$

The representation for the Romanian hierarchical sequences follows the bracket representation for the syntactic trees with constituents (Samuelsson &Volk. 1993).

In **Examples 3** and **4**,the Romanian phrases corresponding to the English phrases given in **Examples 1** and **2** are specified together with their representation patterns.

Example 3. The Romanian construction "pregătirea" that is aligned with the English sequence "the preparation"has the following representations:

- syntactic pattern:
  [NP-513 {3}:Ncfsry][9](with index information)
  [NP Ncfsry](without index information)
  This representation encodes anoun phrase that consists of a noun having the index 3 in the sentence withNcfsryas an MSD specification.
- hierarchical pattern:
  [NP-513 [Ncfsrypregătirea]]
  This representation gives a syntactic subtree corresponding to the node labeled byNP-513.

Example 4. The construction "propunerea Comisiei" is represented as follows:

- syntactic pattern:
  [NP-508 {506}:NP {507}:NP](with index information)
  [NP NP NP](without index information)
  This representation encodes the structure of the noun phrase labeled with NP-508 that consists of two noun phrases that have the

9 See Appendix A for a description of the MSDtags used in this paper.

indexes 506 and 507, respectively, in the Romanian parse tree.

- hierarchical pattern:
  [NP-508 [[NP-506 [Ncfsry propunerea]] [NP-507 [Ncfsoy Comisiei]]]]
  This representation encodes a syntactic subtree that is rooted at the node labeled with NP-508.

2.2 Linguistic Resource with Syntactic Patterns

The resulting English-Romanian parallel sequences taken from the bilingual Treebank are stored in a database (see **Figure 1**) that can be further interrogated in order to provide valuable information from the syntactic transfer point of view between English and Romanian phrases.

The records of the database correspond to English-Romanian pairs of aligned phrases taken from the treebank and encode their structural information in six records. The syntactical patterns are represented in four fields (English and Romanian, with and without index information) and two fields include the hierarchical patterns of English phrases and, Romanian phrasesrespectively. The examples of syntactic patterns from this linguistic database are given in **Table 1** and **Table 2**.

Figure 1. The database with parallel English-Romanian patterns

This kind of resource can be used in a rule and pattern-based approach for an English-Romanian Machine Translation. One major difficulty in the syntactic transformation task is ambiguity. Therefore a transformational model usually has to explore lexical information of the source and target language.

The parallel syntactic patterns can be used in order to train a rule-based machine translation while the hierarchical patterns can be used to train a statistical machine translation that can take into account,not only the hierarchical structure of each parallel English-Romanian sequence,but alsotheirspans, that is, the lexical information.

3. Linguistic analysis upon the bilingual patterns

The importance of the hierarchical patterns in any MT system depends on the information they can provide about phrase reordering and how particular words (e.g. functional words) of the source language affect phrase structures during translation.

Also, the developed English-Romanian database with syntactic patterns can be used in order to obtain various statistics, such as:

- *what are the possible Romanian patterns for a given English syntactic pattern?*(see **Figure 2** and **3**)
- *which English patterns have the maximum number of translation patterns in Romanian?* More precisely, the issue here is to determine the English structures that can pose big problems during translation in the target language
- *what English patterns correspond to unique Romanian patterns?* In other words, the problem can be stated as follows: what are the English patterns that have unique pattern translations in the Romanian language?

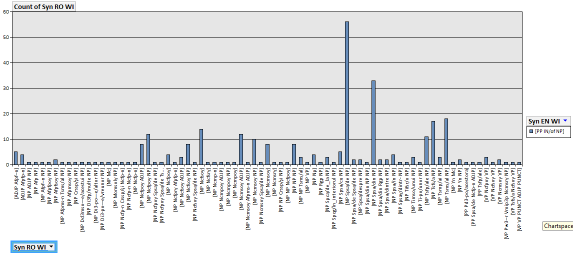

Figure 2. Romanian syntactic patterns for the [PP IN/of NP] English pattern

If a comparative study about the syntactic structure of source and target parallel sequences is needed, in order to develop a general transfer schema for an MT system, then the syntactic pattern representation without index information could be used. In **Table 1** are given some parallel English and Romanian syntactic patterns without index information.

From the data presented in **Table 1**, one can observe that, for example, the translation of the English adverbial pattern [ADVP IN/for] has a single pattern in which the preposition "for" is replaced by the Romanian preposition "pentru", while the pattern for the English prepositional phrase [PP IN/for NP] gathers more than 10 Romanian patterns from the treebank.

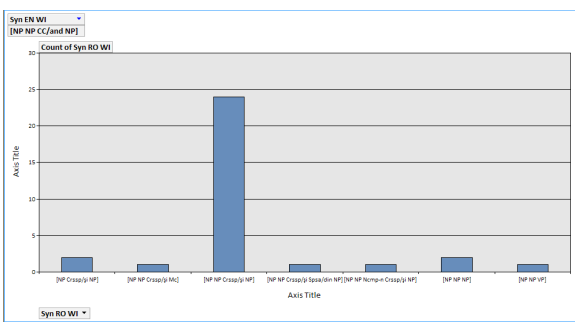

Figure 3. Romanian syntactic patterns for the [NP NP CC/and NP] English pattern

The syntactic patterns given in the form presented in **Table 1** are useful if general translation patterns are needed for developing a certain phase of the translation process. But, usually, more detailed information is needed, like token/word reordering or alignment.

Table 1. English-Romanian parallel syntactic patterns without index information

English patterns	Romanian patterns
[ADJP JJ CC/ and JJ]	[ADJP Afpms-n Crssp/şi Afpms-n]
[ADJP JJ PP]	[ADJP Afpfp-n NP]
[ADJP JJ]	[ADJP Afpfp-n], [ADJP Afpfsrn], [ADJP Pd3fsr/CEEA], [ADJP Rgp], [NP Ncmsry], [VP Vmis3s]
[ADVP IN/ below]	[ADVP Spsa/de RpRgp]
[ADVP IN/by NP]	[NP Spca/în_conformitate_cu NP]
[ADVP IN/ for]	[ADVP Spsa/pentru]
[CONJP CC/ but RB]	[CONJP Ccssp/ci Crssp/şi], [CONJP Ccssp/darSpsa/fără]
[NP ADJP NNS]	[NP ADJP Spsa/înNcms-n], [NP Ncfp-n ADJP], …
[NP CD NN]	[NP Mc Ncmp-n], [NP Mc Yn], [NP Mc], [NP Tifsr/o Ncfsrn]
[NP DT/a JJ NN]	[NP NcfsrnAfpfsrn], [NP Ncms-n Spsa/de Ncfsrn], [NP Tifso/ uneiAfpfsonNcfson], [NP Tifsr/o Di3fsr---e/altăNcfsrn], [NP Tifsr/o NcfsrnAfpfsrn], [NP Timso/ unuiNcms-n Afpms-n], [NP Timsr/un Di3ms/anumitNcms-n], [NP Timsr/ un Ncms-n Afpms-n], [NP Timsr/un Ncms-n]
[NP NN NN]	[NP Ncfp-n Afpfp-n], [NP Ncfp-n Spsa/de Ncfsrn], [NP Ncfp-n], [NP NcfsoyAfpfson], [NP NcfsrnSpsa/de Ncfsrn], [NP NcfsrnSpsa/înNcmsry], [NP Ncms-n Afpms-n], [NP Ncms-n], [NP NcmsryNcms-n]
[PP CC/either PP CC/or PP]	[NP Ccssp/fie NP Ccssp/fie NP], [NP Ccssp/fie NP NP]
[PP IN/as NP]	[PP Cscsp/precum_şi NP], [NP Rc NP], [NP Rgp NP], [NP Spsa/pentru NP], [NP Spsd/conform NP],
[PP IN/for NP]	[ADJP Afpms-n], [NP Ncfp-n], [PP Spcg/în_vederea NP], [PP Spsa/ cu NP], [PP Spsa/de NP], [PP Spsa/ din NP], [PP Spsa/în NP], [PP Spsa/ la NP], [PP Spsa/pe NP], [PP Spsa/ pentru NP], [PP Spsa/spre NP], [PP Spsg/contra NP], [PP Spsa/de ADJP]
[VP MD/shall VP]	[VP Px3--a--------w/se VP], [VP Vaip3p VP], [VP Vmip3p NP]

In the last scenario, the representations with index information for syntactic patterns are more suitable. By adding such information,one can exploit the linguistic relationships between words and phrases of the translator's source and target languages. The examples of parallel syntactic patterns with index information are given in **Table 2**.

Table 2. An example of parallel syntactic patterns with index information corresponding to the [NP DT/a JJ NN] English structure

English pattern	Romanian pattern
[NP-515 {5}:DT/a {6}:JJ {5}:NN]	[NP-508 {5}:Ncms-n {6}:Ncms-n]
[NP-524 {9}:DT/a {11}:JJ {10}:NN]	[NP-553 {9}:Timso/ unui {10}:Ncms-n {11}:Afpms-n]
[NP-522 {7}:DT/a {8}:JJ {9}:NN]	[NP-513 {7}:Tifso/ unei {8}:Afpfson {9}:Ncfson]
[NP-549 {12}:DT/a JJ {13}:NN]	[NP-517 {12}:Timsr/un {13}:Ncms-n]
[NP-586 {29}:DT/a {31}:JJ {30}:NN]	[NP-543 {29}:Timsr/ un {30}:Ncms-n {31}:Afpms-n]
[NP-530 {14}:DT/a {16}:JJ {15}:NN]	[NP-525 {14}:Tifsr/o {15}:Ncfsrn {16}:Afpfsrn]
[NP-555 {19}:DT/a {20}:JJ {21}:NN]	[NP-534 {19}:Timsr/ un {20}:Di3ms/anumit {21}:Ncms-n]
[NP-513 {3}:DT/a {4}:JJ {5}:NN]	[NP-552 {3}:Tifsr/o {4}:Mofsrln {5}:Ncfsrn]
[NP-591 DT/a {43}:JJ {41}:NN]	[NP-553 {41}:Ncms-n Spsa/de {43}:Ncfsrn]
[NP-603 DT/a {31}:JJ {30}:NN]	[NP-555 {30}:Ncms-n {31}:Afpms-n]

One way to filter alternative Romanian structures for a given English pattern that has several translation possibilities is to eliminate the structures that were identified in the corpus once and to take into account the structures that have more than one appearance in the corpus (see **Figure 4**). In this manner, we can eliminate "the noise" in the database that comes from incorrect word alignments.

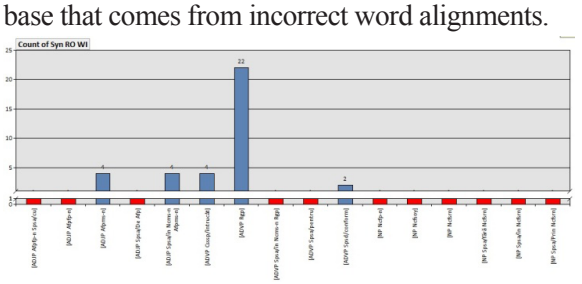

Figure 4. "Noise elimination" for the Romanian patterns corresponding to the [ADVP RB] English pattern

4. Conclusions

This paper describesthe work on the creation of a collection of richly annotated parallel syntactic sequences that are aligned at multiple levels. In the proposed formalism, the target language is an inflectional language and, as a direct consequence, special attention is paid to the morphological information that could be derived from the processed target construction of the used corpus. Such a collection is intended to be used in an MT system development with the scope of moving from words to phrases as the basic unit of translation.

In order to develop a Machine Translation system, parallel syntactic and lexical sequences in the source and target language have to be provided to the system in order to train it. The more numerous and accurate the sequences, the better the translation performance.

We intend to apply this resource to the development of a syntax-based MT system.The current results suggest that for this kind of data setup, high recall alignments are preferable to high precision alignments in producing better MT results. For this reason, the patterns generated by the presented study, more precisely the Romanian patterns, have to be generalized in order to capture more instances of variation of the morpho-syntactic features.

5. 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.

The author wants to thank the Natural Language Processing Group of the Faculty of Computer Science, Alexandru Ioan Cuza University of Iaşi, for providing the English-Romanian corpus upon which the presented study was made. Special thanks of gratitude go to Prof. Dr. Dan Cristea for his constant support and advice.

References

Ambati, Vamshi, Alon Lavie, and Jaime Carbonell. 2009. Extraction of Syntactic Translation Models from Parallel Data using Syntax from Source and Target Languages. In *MT Summit XII Proceedings of the twelfth Machine Translation*.

Araùjo, Josué G. and Helena M. Caseli. 2010. Alignment of Portuguese-English syntactic trees using part-of-speech filters. In *Workshop in Natural Language Processing and web-based technologies*, 31-40.

Colhon, Mihaela and Dan Cristea. 2012. Automatic extraction of syntactic patterns for dependency parsing in NP chunks. In *International Simpozion "The Syntax and Semantics of Specificity"*, Faculty of Foreign Languages and Literatures, University of Bucharest, Romania. (accepted).

Colhon, Mihaela. 2011. A Contrastive Study of Syntactic Constituents in English and Romanian Texts. In *Proceedings of the Workshop "Language Resources and Tools with Industrial Applications"*, eds. Iftene A., Trandabăţ D.-M., 11-20.

Colhon, Mihaela. 2012. Language Engineering for Syntactic Knowledge Transfer. *Computer Science and Information Systems Journal*, 9(3):1231-1247, ISSN 1820-0214

Cristea, Dan and Corina Forăscu. 2006. Linguistic Resources and Technologies for Romanian Language. *Computer Science Journal of Moldova*, 14(1).

Erjavec, Tomaž, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić and Duško Vitas.2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages, In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, 25-32, Budapest, Hungary.

Erjavec, Tomaž. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC'10*), European Language Resources Association (ELRA), ISBN 2-9517408-6-7.

Ide, Nancy, Patrice Bonhomme and Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association (ELRA), 2000.

Iftene, Adrian, Diana Trandabăţ, Alex-Mihai Moruz and Maria Husarciuc. 2010. Question Answering on Romanian, English and French Languages. In-*Notebook Paper for the CLEF 2010 LABs Workshop*, Padua, Italy,ISBN 978-88-904810-0-0, ISSN 2038-496322-23.

Galley, Michel, Mark Hopkins, Kevin Knight and Daniel Marcu. 2004. What's in a Translation Rule? In *Proceedings of HLT-NAACL 2004*, 273-280, Boston: Association for Computational Linguistics.

Gispert, Adrià de, Juan Pino and William Byrne. 2010. Hierarchical Phrase-based Translation Grammars Extracted from Alignment Posterior Probabilities. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 545-554.

Klein, Dan and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 423-430.

Kurc, Roman, Maciej Piasecki and Stan Szpakowicz. 2010. Corpus-based extraction of morpho-syntactic patterns for the automatic acquisition of hypernymy. *Intelligent Information Systems*, 77-90.

Samuelsson, Yvonne, and Martin Volk. 1993. Alignment Tools for Parallel Treebanks. *ACM Trans. Program. Lang. Syst*. 15(5):795-825.

DOI= http://doi.acm.org/10.1145/161468.16147.

Simionescu, Radu. 2012. Romanian deep noun phrase chunking using graphical grammar studio. In *Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language"*, eds. M. A. Moruz, D. Cristea, D. Tufiş, A. Iftene, H. N. Teodorescu, 135-143.

Taylor, Ann. 1996. Bracketing Switchboard: An Addendum to the Treebank II Guidelines, http://www.seas.upenn.edu/~jmott/prsguid2.pdf.

Tufiş, Dan and Radu Ion. 2007.Parallel Corpora,

Alignment Technologies and Further Prospects in Multilingual Resources and Technology Infrastructure. In*Proceedings of SPED 2007*.

Wenniger, Gideon Maillette de Buy, Maxim Khalilov and Khalil Sima'an. 2010. A Toolkit for Visualizing the Coherence of Tree-based Reordering with Word-Alignments. In *Proceedings of the Open Source Convention at the Fifth Machine Translation Marathon* (*MT-Marathon*), Le Mans (France), 97-104.

Yamada, Kenji and Kevin Knight. 2001. A Syntax-based Statistical Translation Model. In *Proceeding of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, 523-530.

Appendix A — glossary of notation

The following table gives the notation used in this paper:

Table 3. The meaning of MSD notations according to the MULTEXT-East lexical specifications

MSD tag	**The meaning of the notation**
Afpfp-n	Adjective qualifier positive feminine plural -definiteness
Afpfson	Adjective qualifier positive feminine singular oblique –definiteness
Afpfsrn	Adjective qualifier positive feminine singular direct –definiteness
Afpms-n	Adjective qualifier positive masculine singular –definiteness
Crssp	Conjunction *r* simple simple positive
Ccssp	Conjunction coordinating simple simple positive
Di3fsr	Determiner indefinite third feminine singular direct
Di3ms	Determiner indefinite third masculine singular
Mc	Numeral cardinal
Mofsrln	Numeral ordinal feminine singular direct letter -definiteness
Ncfsoy	Noun common feminine singular oblique +definiteness
Ncfson	Noun common feminine singular oblique -definiteness
Ncfsrn	Noun common feminine singular direct –definiteness
Ncfsry	Noun common feminine singular direct +definiteness
Ncfp-n	Noun common feminine plural -definiteness
Ncmp-n	Noun common masculine plural -definiteness
Ncmsoy	Noun common masculine singular oblique +definiteness
Ncmsry	Noun common masculine singular direct +definiteness
Ncms-n	Noun common masculine singular -definiteness
Pd3fsr	Pronoun demonstrative third feminine singular direct
Pp3fso-	Pronoun personal third feminine singular oblique
Px3--a----------w	Pronoun reflexive third accusative weak
Rc	Adverb *c*
Rgp	Adverb general positive
Rp	Adverb particle
Spca	Adposition preposition compound accusative
Spcg	Adposition preposition compound genitive
Spsa	Adposition preposition simple accusative
Spsd	Adposition preposition simple dative
Tifso	Article indefinitefemininesingular oblique
Tifsr	Article indefinite feminine singular direct
Timso	Article indefinite masculine singular oblique
Timsr	Article indefinite masculine singular direct
Tsms	Article possessive masculine singular
Vaip3p	Verb auxiliary indicative present third plural
Vmip3p	Verb main indicative present third plural
Vmis3s	Verb main indicative past third singular
Yn	Abbreviation nominal

Table 4. The meaning of the Penn Treebank notations

Penn Treebank tag	The meaning of the notation
ADJP	Adjectival Phrase
ADVP	Adverbial Phrase
CONJP	Conjunction Phrase
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
IN	Preposition or subordinating conjunction
JJ	Adjective
NN/NNS	Noun, singular or mass/ Noun, plural
MD	Modal
NP	Noun Phrase
PP	Prepositional Phrase
RB	Adverb
VP	Verb Phrase