Exploiting Aligned Parallel Corpora in Multilingual Studies and Applications

Dan Tufiş

Research Institute for Artificial Intelligence, Romanian Academy, 13, "13 Septembrie", 050711, Bucharest, Romania
tufis@racai.ro

Abstract. Parallel corpora encode extremely valuable linguistic knowledge, the revealing of which is facilitated by the recent advances in multilingual corpus linguistics. The linguistic decisions made by the human translators in order to faithfully convey the meaning of the source text can be traced and used as evidence on linguistic facts which, in a monolingual context, might be unavailable to (or overlooked by) a computer program. Multilingual technologies, which to a large extent are language independent, provide a powerful support for systematic and consistent cross-lingual studies and allow for easier building of annotated linguistic resources for languages where such resources are scarce or missing. In this paper we will briefly present some underlying multilingual technologies and methodologies we developed for exploiting parallel corpora and we will discuss their relevance for cross-linguistic studies and applications.

Keywords: alignment, annotations, collocations, cross-language studies, disambiguation (POS, WSD), encoding, parallel corpora, multilingual technologies, tagging, wordnets.

1 Introduction

The "world of knowledge", as the virtual space of the internet has rightfully been called, is the conceptual framework where the notion of "digital-divide" has been coined. This phenomenon results from the unequal application of, and access to, information and communication technologies. Narrowing the knowledge gaps between different communities of the world has been, and continues to be a top priority not only for local authorities but for major international organizations as well. For instance, in its 32nd Session (30 September-17 October 2003) the UNESCO General Conference adopted a highly relevant document "RECOMMENDATION ON THE PROMOTION AND USE OF MULTILINGUALISM AND UNIVERSAL ACCESS TO CYBERSPACE" where it is shown that "*multilingualism in cyberspace is of vital and strategic importance to ensure the right to information and cultural diversity"* and that *"everyone and every nation must have an equal opportunity to benefit from cultural*

diversity and scientific progress, which must remain, more than ever, a basic human right in the emerging information society"[1].

Yet, mere access to the internet does not open the gates to the "world of knowledge". The e-content is expressed in many languages and this is natural to be so. "Language constitutes the foundation of communication between people and is also part of their cultural heritage. For many people, language carries far-reaching emotive and cultural associations and values embedded in vast literary, historical, philosophical and educational heritage. For this reason the users' language should not constitute an obstacle to accessing the multicultural human heritage available in cyberspace" (ibid.).

Among others, the multilingual language technologies are expected to be most instrumental in lowering as much as possible the language and cultural barriers for harmonious and collaborative development of the information society.

The web language services [3], [13], linguistic grids [9], [14], multilingual collaborative and distributed platforms [12], [27], automatic translation, are seen as key technologies, the most promising in fostering the cross-cultural cooperation. Identifying opinions and emotions expressed in texts, one of the hottest current research area, revealed cross-cultural similarities but also disparities which should be very carefully considered for a smooth intercultural communication [18], [25].

In this paper we will describe another extremely useful technology for multilingual processing, namely the word alignment. Word alignment is not a goal in itself, but an enabling technology which serves all the higher level multilingual applications. After describing the necessary text pre-processing and the alignment procedure, we will exemplify a few cases of text alignment exploitation: lexical semantics knowledge acquisition and validation, cross-lingual studies, and transfer of linguistic (syntactic and semantic) annotations in a multi-cultural cooperation program.

2 Parallel Corpora and Textual Alignment

A bitext is a pair of texts in two languages, so that the texts can be considered reciprocal translations. They are called translation equivalents. By extension, a multi-text is a set of multiple language texts, so that each pair of texts represents a bitext. A large collection of bitexts or multi-texts is called a parallel corpus. Knowing that two or more texts are reciprocal translations is useful, but much more useful is detecting the translation equivalence at finer grained levels.

The automatic identification in a parallel corpus of the segments of texts that represent reciprocal translations is a prerequisite for taking advantage of the implicit linguistic and cultural knowledge embedded into the translations. This problem, known as parallel corpus alignment, can be defined at various levels of text segmentation granularity (paragraph, sentence, phrase, word) with different degrees of difficulty. Two segments of texts from a bitext which represent reciprocal translations make a translation unit. A translation unit may contain, in one or both paired languages, one or more textual units (paragraph, sentence, phrase, word) and one distinguishes between the 1:1 and non-1:1 alignment translation units. While at

[1]http://portal.unesco.org/ci/en/ev.php-URL_ID=4969&URL_DO=DO_TOPIC&URL_SECTION=201.html

the paragraph granularity level the non-1:1 alignments are exceptional (most aligners assume the number of paragraphs to be the same in the two sides of a bitext, and thus only 1:1 alignments are considered), at the sentence or phrase level they are quite rare (usually no more than 5-10% of the total number of translation units). At the word level, the non-1:1 alignments are more frequent and their number strongly depends on the language pair and on the type of translation (literal versus free translation). Another source of increased difficulties for fine-grained alignments is that while at the paragraph and (to a large extent) sentence level the ordering of the textual units is preserved in both sides of a bitext (discourse coherence requirement), at the finer grained level this is not true in general (the word or phase ordering being ruled in each language by its syntax).

Depending on the alignment granularity, required accuracy, and the purpose of the alignment, the input textual data might need pre-processing steps in all languages of the parallel corpus (e.g. sentence splitting, tokenization, POS-tagging and lemmatization) or at least in one of the languages of the corpus (e.g. chunking, dependency linking/parsing, and word sense disambiguation).

3 Preprocessing Steps

Text segmentation. The first pre-processing step in most NLP systems deals with text segmentation. In our processing chain this step is achieved by a modified version (much faster) of the multilingual segmenter developed within the MULTEXT project which has tokenization resources for many western European languages, further enhanced in the follow up MULTEXT-EAST project with corresponding resources for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. Our segmenter is able to recognize paragraphs, sentence and clause boundaries, dates, numbers and various fix phrases, and to split clitics or contractions (where the case). We significantly updated the tokenization resources for Romanian and English (the languages we have been most interested in lately).

Sentence alignment. We developed a sentence aligner [6] inspired by Moore's program [21] which removes its 1:1 alignment restriction, the assumption on the monotonic ordering of the sentences in the two languages, as well as the upper limit on the number of sentence-pairs that can be aligned. It has a comparable precision but a better recall than Moore's aligner.

The sentence aligner consists of a hypothesis generator which creates a list of plausible sentence alignments from the parallel corpus and a filter which removes the improbable alignments. The hypothesis generator uses a character based preliminary sentence aligner, similar to Church and Gale's CharAlign [11], which creates a list of possible alignments, represented as pairs of sentence identifiers <N, M>[2]. Each pair of this list is supplemented by pairs obtained by local variations of the sentence indexes

[2] M is the index of the m^{th} sentence in one part of the bitext which is presumably aligned to the the n^{th} sentence (of index N) in the other part of the bitext.

<M±k, N±k> with k ranging from 1 to a user-defined upper limit (the default value is 1[3]).

The filter is an SVM binary classifier [8] initially trained on a Gold Standard. The features of the initial SVM model are: the word sentence length, the number of non-word tokens, and the rank correlation for the first 25% of the most frequent words in the two parts of the training bitext. This model is used to preliminary filter alignment hypotheses generated from the parallel corpus. The set of the pairs of sentences that remained after this filtering is used as the input for an expectation maximisation algorithm which builds a word translation equivalence table by a similar approach to the IBM model-1 procedure [4]. The SVM model is rebuilt (again from the Gold Standard) this time including, as an additional feature, the number of word translation equivalents existing in the sentences of a candidate alignment pair. This new model is used by the SVM classifier for the final sentence alignment of the parallel corpus.

POS-tagging. It is generally known that the accuracy of POS-tagging depends on the quality of the language model underlying the morpho-lexical processing, which, on its turn, is highly dependent on the quality and quantity of the training data and on the tagset of the language model. For languages with a productive inflectional morphology the morpho-lexical feature-value combinations may be very numerous, leading to very large tagsets with unavoidable training data sparseness threat. The lack of sufficient training data affects the robustness of the language models which, consequently, will generate an increased number of tagging errors at the run time. To cope with the tagset cardinality problem we developed the tiered-tagging methodology [34] and implemented it using the TnT trigram HMM tagger [2]. The methodology involves the use of a reduced hidden corpus tagset, automatically constructed from the large targeted lexical tagset, and a procedure to map back the reduced tagset into the large one, used in the final annotated text. The two tagsets (the lexical and corpus tagsets) are related by a subsumption relation. When the reduction of the cardinality of the large tagset is information lossless (that is just redundancy elimination) the mapping from the reduced tagset to the large one is deterministic and it is simply ensured by looking up a wordform dictionary. For tagset reduction with information loss, which ensures a much significant reduction of the lexical tagsets, the recovering of the left out morpho-lexical information, although to a large extent deterministic, requires an additional preprocessing to solve some non-deterministic cases. In the previous version of the tiered tagging approach we used several hand-crafted rules (regular expressions defined over the reduced tagset, with a span of ±4 tags around the ambiguously mapped tags).

Recently, we have re-implemented the tiered tagging methodology, by relying on a combination between an HMM tagger, called TTL [15], which produces also the lemmatization, and a maximum-entropy tagger [5].The HMM tagger works with the reduced tagset while the ME-tagger ensures the mapping of the first tagset onto the large one (the lexical tagset) dispensing on the hand-written mapping rules.

Lemmatization is in our case a straightforward process, since the monolingual lexicons, developed according to MULTEXT-EAST morpho-lexical specifications [7], contain for each word, its lemma and the morpho-lexical tag. Knowing the word-

[3] For this case, besides the pair <N, M> the alignment candidates list will also contain the pairs <N-1,M>, <N+1,M>, <N,M-1>, <N,M+1>, <N-1,M+1>, <N+1,M-1>,.

form and its associated tag, the lemma extraction is simply a matter of lexicon lookup for those words that are in the lexicon. For the unknown words, which are not tagged as proper names, a set of lemma candidates is generated by a set of suffix-stripping rules induced from the word-form lexicon. A four-gram letter Markov model (trained on lemmas in the word-form dictionary) is used to choose the most likely lemma.

Chunking. By means of a set of language dependent regular expressions defined over the tagsets, our chunker accurately recognizes the (non-recursive) noun phrases, adjectival/adverbial phrases, prepositional phrases and verb complexes (analytical realization of tense, aspect mood and diathesis and phrasal verbs) both for Romanian and English.

Word Alignment. The word alignment [23], [24] of a bitext is an explicit representation of the pairs of words $<w^i_{L1}\, w^j_{L2}>$ occurring in the same translation units that represent mutual translations (called translation equivalence pairs). Either of w^i_{L1} or w^j_{L2} may be NULL (this is the case of *null alignments* where one word in one part of the bitext is not translated in the other part). When w^i_L, w^j_{L2} or both appear in several pairs of the same translation unit they correspond to *multi-word expression alignments*.

The input raw texts, pre-processed as described in the previous section, are fed into the word alignment engine, called COWAL [31], [32] which is a wrapper of two stand-alone aligners (YAWA and MEBA). COWAL merges the alignments produced by each stand-alone aligner and then uses a trained SVM classifier to prune the unlikely alignment links. The classifier is based on the LIBSVM kit [8] used with the default parameters (C-SVC classification and radial basis kernel function). The classifier was trained with positive and negative hand-validated examples of word alignment links.

The usefulness of the aligner combination has been convincingly demonstrated on the occasion of the Shared Task on Word Alignment organized by the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond" [20]. We participated (on the Romanian-English track) with the two standalone aligners and the combined one [32]. Out of the 37 competing systems, MEBA was rated the 20th and TREQ-AL, (the former version of YAWA), was rated the 21st, but COWAL, their combination, was the winning system.

Table 1. Combined alignment.

Aligner	Precision	Recall	F-measure
YAWA	88.80%	74.83%	81.22%
MEBA	92.15%	73.40%	81.71%
COWAL	87.26%	80.94%	83.98%

Meanwhile, both stand-alone aligners have been improved (see Table 1) in various ways and trained on more data, but the combined aligner still performs better than both of them.

COWAL is now embedded into a larger platform, called MTkit that incorporates the tools for bitexts pre-processing, a graphical interface that allows for comparing and editing different alignments, as well as a word sense disambiguation module (described in the next section).

4 Exploiting the Alignments

In the following, there will be shown a few examples of how we used the word alignments. The most obvious application of the word alignment is building translation lexicons [30]. The aligned corpora we worked with were the Ro-En sub-corpus of the „1984" multilingual [7] corpus (about 110,000 tokens per language), a partial translation in Romanian of SemCor2.0 (about 177,000 tokens per language), the journalistic parallel corpus Ro-En) used in the word-alignment competition at ACL2005 [20] (about 1,000,000 words per language) and the four-language (English-Romanian-French-German) sub-corpus of the 21-language parallel corpus Acquis Communautaire[4] (about 8 million tokens per language). In our experiments we were interested only in open-class words, the alignment of which is significantly more accurate than the „all words" alignment. This is to say that most of the alignment errors in the evaluation shown in Table 1 were related to functional words and punctuation. From the parallel En-Ro corpora we extracted a large translation lexicon (about 500,000 entries) used in various applications, some of which will be mentioned in the next sections.

4.1 Aligned Wordnets Validation

Once the translation equivalents identified, it is reasonable to expect that the words of a translation pair $<w^i_{L1}, w^j_{L2}>$ share at least one conceptual meaning stored in an interlingual sense inventory. In the BalkaNet project [35] we used the Princeton WordNet (PWN) [10] as an interlingual index [37]. Based on the interlingually aligned wordnets, obtaining the sense labels for the words in a translation pair is straightforward [16]:

a) one has to identify for w^i_{L1} the synset S^i_{L1} and for w^j_{L2} the synset S^j_{L2} so that S^i_{L1} and S^j_{L2} are projected over the same concept. The index of this common interlingual concept (ILI) is the sense label of the two words w^i_{L1} and w^j_{L2}.
b) if no common interlingual projection will be found for the synsets to which w^i_{L1} and w^j_{L2} belong, the senses of the two words will be given by the indexes of the most similar interlingual concepts corresponding to the synsets of the two words. The semantic-similarity score is computed as SYM (ILI_1, ILI_2) = 1/1+k where k is the number of PWN links from ILI_1 to ILI_2 or from both ILI_1 and ILI_2 to the nearest common ancestor.

In case none of the two cases above holds, then it is very likely that there are some problems which can be categorized as follows:

i) the translation pair is wrong (either because of human translator or because of the word aligner), so it is natural not to find any ILI matching for the two words of the pair;
ii) one or both words do not have implemented the relevant senses;
iii) one of both words are missing from the relevant existing synsets;
iv) one or both synsets to which the words of the current translation pair belong are not correctly linked to the relevant ILI;

[4] http://langtech.jrc.it/JRC-Acquis.html

v) the two words in the current translation pair have different POS. Since all the BalkaNet wordnets were aligned to PWN version 2.0 preserving the POS of the synsets, all the cross-pos translations will fit this case.

For the semantic validation of the wordnets created during the BalkaNet project the cases i) and v) were not relevant. The alignment pairs of the ii), iii) and iv) types extracted from the various bilingual sub-corpora of the "1984" corpus were validated by native speakers, with very good command of English. As a result, many synsets were extended with missing literals, the missing synsets were added, and wrong interlingual projections were corrected. The final report of the BalkaNet[5] project gives a detailed quantitative and qualitative account of the errors and incompletenesses that were detected by this procedure (and corrected by each partner).

Since the BalkaNet project finished, we have consistently extended the Romanian wordnet [36] (currently it contains more than 39,000 synsets, and this number is steadily growing). We repeated the semantic validation procedure several times until we haven't noticed any problematic case in the Ro-En sub-corpus of "1984".

Recently, we have applied a slightly modified variant of this procedure, this time using a partial translation in Romanian of the SemCor2.0. The two sets of documents were sentence aligned (8276 aligned sentence pairs), word aligned and the alignment correctness for the closed-class categories was estimated (based on a random set of links) at 98.5%. With this accuracy, the i) problem was not expected to have a relevant contribution to the number of problems that could (and did) show up.

With the word senses available in the English part, the validation procedure of the Romanian wordnet proceeded the following way:

For each pair of aligned words we checked whether the ILI number of the sense marked on the English word corresponded to the one of any synsets containing the Romanian translation equivalent or to any hypo/hypernyms of it.

Out of the 178499 tokens in the English part[6], only 79595 tokens were sense marked-up (the rest were functional words or punctuation) but 3535 were not translated into Romanian and 3694 translation pairs had different part of speech. For the remaining 72366 POS-preserving translated English content words, the sense correspondence was found for 48392 Romanian tokens (66.87%). We thoroughly analyzed the 23934 cases of the unmatched senses and found two main situations:

a) Romanian translation equivalent was not in the Romanian wordnet: 11930 cases
b) The Romanian synset with the ILI of the English word exists, but does not include the Romanian word (incomplete synset) : 12044 cases

While the a) situation was somehow expected (given that the Romanian wordnet contains much less synsets and literals than PWN, leaving room for long-time development efforts for our team), the second situation was worrying and we concentrated our evaluation on that case. As expected, a large number (more than 4000) of the translation equivalence pairs coming under this rubric were alignment errors but still, numerous incomplete synsets (more than 7000) were detected. The extension of all these synsets with the missing literals is one of our current and future activities.

[5] http://www.ceid.upatras.gr/Balkanet/resources.htm
[6] The Romanian part contains 175603 tokens.

4.2 WordNet-Based Sense Disambiguation.

The WSD task can be stated as being able to associate to an word (w), ambiguous in a text or discourse, the sense (s_k) which is distinguishable from other senses ($s_1 \ldots s_{k-1}$ $s_{k+1} \ldots s_n$) and is prescribed for that word by a reference semantic lexicon. The task of word sense disambiguation (WSD) requires one reference sense inventory, in terms of which the senses of the target words will be labeled. A meaningful discussion of the performances of a WSD system cannot dispense of clearly specifying the sense inventory it uses, and the comparison between two WSD systems that use different sense inventories is frequently more confusing than illuminating. Essentially, this is because the differences in the semantic distinctions (sense granularities), as used by different semantic dictionaries (sense repositories), make the difficulty of the WSD task range over a large spectrum. For instance, the discrimination of homographs (more often than not having different parts of speech), is much simpler than the metonymic distinctions.

It is straightforward to turn the previous validation procedure into a WSD engine and we claim [33] that state-of-the-art (and even better) performances can be achieved in sense disambiguation of the words in a parallel text, provided the respective bitext is word aligned and aligned wordnets exists for the considered languages.

As mentioned, we used the PWN version 2.0 (PWN2.0) synset identifiers as the reference sense inventory. Through the 1-1 mapping, existing between the synsets in the Romanian wordnet and PWN2.0, the SUMO/MILO [22] and DOMAINS [19] labels became available in our wordnet (as in all the other wordnets which are aligned to PWN2.0). Since both SUMO/MILO and DOMAINS synset labeling is POS insensitive, their use is extremely helpful in assigning senses to the words in a cross-pos translation equivalence pair. For instance, if we consider the En-Ro translation equivalence pairs <*fire*-verb *concediat*[7]-adj> the POS preserving ILI-based alignment between PWN2.0 and RoWN is not really helpful. However, if one uses instead of synset identifiers the SUMO labels, a match would be found since the senses *fire*:4 for the verb and *concediat:1* for the adjective belong to synsets labeled by the same SUMO concept *TerminatingEmployment.* An extra-bonus is that one can infer that the first sense of the adjective *fired* (the translation of *concediat*) is derived from the fourth sense (and no other one) of the verb *fire*. This is a useful type of information which is not yet encoded in any version of PWN.

4.3 Annotation Transfer as a Cross-lingual Collaboration Task

Having a parallel corpus aligned at the word and phrase level may be the starting point on significant geographical and cross-lingual distribution of the tasks aimed at creating a multiple layer annotated corpus. One can imagine a cross-cultural initiative, which agreed on some parallel corpora (e.g. AcquisCom) containing the languages of interest, and where each partner is willing to annotate his/her language part of the multilingual corpora with the information for which the appropriate tools exists (POS, multi-word expressions, sense labels, parse tree annotations,

[7] concediat (Ro) = fired (En)

argument/frame structures, etc). Based on the assumption that correcting annotations is easier and cheaper than creating them from scratch, word alignment technology could be used to transfer information from the tokens in one language to their translation equivalents in the other language. The transfer could be controlled by language specific rules, the writing of which is certainly less demanding than the direct annotation.

An example at hand is transferring the word senses. Hand word sense disambiguation of a large text is an extremely labor intensive work, prone to human errors and extremely expensive.

SemCor2.0 is an English corpus, with (most of) the content words being sense disambiguated and carefully validated. It is not surprising that several research teams (see for instance http://multisemcor.itc.it/index.php) decided to translate as much SemCor documents as possible and then to transfer the senses into the translations. As we have shown in section 4.1, we partially translated the SemCor2.0 and used word-alignments to check-out the completeness of our wordnet. We showed that more than 12,000 Romanian words translating the English sense annotated words were absent from our wordnet. By sense transfer, we can extend the target wordnet with these 12,000 synsets. It is true that these synsets are partial (they are mono-literals) and would certainly require extensions with other synonyms but much effort is already saved. Similar considerations apply for the automatic extension of the 7000 incomplete synsets discovered by the experiment we described in section 4.1.

Another type of annotation transfer, more difficult and less reliable, but extremely useful, refers to the syntactic/semantic relations annotated in the source language. When the same type of syntactic/semantic annotation exists in both languages, the annotation transfer allows for annotation validation in one or both languages of the bitext, or provides evidence for corpus-supported comparative/contrastive studies. In [1] it is described an experiment aimed at assessing the possibility of statistically inducing a dependency grammar for Romanian by semi-automatic transfer of the dependency relations from a parsed English text. The major assumption to be evaluated was the so-called Direct Correspondence Assumption (DCA): checking whether a dependency relation that holds between two English words remains valid between the Romanian translation equivalents in the aligned bitext. In what follows we present the main lines of the strategy employed in this experiment and some preliminary results.

The aligned English-Romanian bitext used for this study was extracted from the *1984* parallel corpus; the English part of the bitext was parsed with the FDG parser [28] by a partner in Wolverhampton and validated by another partner in Iaşi.

From the entire bitext only 1537 sentence pairs (about 25%) were retained for the proper experiment. We discarded very long sentences, the non-1:1 translation units, and those translation units with fewer aligned words than an empirical threshold or containing slang and non-grammatical language (*proles* language).

The relations between English words with a NULL translation equivalent were not taken into account for the evaluation of the transfer accuracy (in Table 2 below, these are counted in the Lost column), while the rest of relations were mechanically transferred into the Romanian part of the bitext.

Table 2. Percentage of correctly transferred relations.

No.	Rel.	RO	Lost	EN	Acc. (%)	No.	Rel.	RO	Lost	EN	Acc (%)
1	qn	10	0	12	83.33	11	cc	94	2	155	61.44
2	neg	10	0	13	76.92	12	pm	44	1	75	59.46
3	oc	3	0	4	75.00	13	obj	79	2	137	58.52
4	dat	3	0	4	75.00	14	mod	114	1	201	57.00
5	cnt	8	0	11	72.73	15	ha	41	0	74	55.41
6	ad	25	0	35	71.43	16	cla	8	0	15	53.33
7	pcomp	218	9	316	71.01	17	tmp	23	0	46	50.00
8	det	126	173	355	69.23	18	man	16	0	32	50.00
9	comp	70	1	112	63.06	19	subj	121	72	319	48.99
10	attr	151	4	245	62.66	20	v-ch	35	48	143	36.84

Two experts independently evaluated the validity of the transferred relations with disagreements negotiated and, agreed one way or another.

There were identified three types of relation transfer: for the first type, the transfer is possible and correct without amendments; the second type refers to correct link transfer but incorrect labeling of the links; it needs mapping rules for switching the names of the correct link dependencies (e.g. the rule responsible for an active voice construction in English, translated by a passive voice construction in Romanian switches the obj and subj labels in the target language sentence); the third category of transfers refers to the "lexicalized" dependencies (relations whose governor (rarely the dependent) is instantiated by a specific word) where the transfer is always wrong (both the dependency link and its name), due to the different behavior of corresponding predicates in the considered languages (e.g. like/plăcea).

Table 2 gives information on the correctness of the unconditional transfer for the relations[8] from the source part of the bitext. Due to the enclitic definite articulation in Romanian, half of the English determiners (the occurrences of the) are not explicitly translated and consequently, half of the det relations are lost. The large number of the subj relations that are lost is due to the pro-drop nature of Romanian. One also may notice that this relation has a low correct transfer figure (48.99%) which is correlated with the low correct transfer figure of the relation obj. The simple mapping rule, mentioned before, for dealing with passive/active voice alternation in the aligned sentences would improve with almost 50% the success rate.

We consider these results extremely encouraging, and one of our future research topics will be the design of a set of transfer rules for correcting the role assignment for the dependency links which were correctly transferred (second type, see above). The "lexicalized" dependencies will be collected as they would be detected and stored (with the correct transfer information) as exceptions from the general transfer procedure.

A similar transfer experiment, but this time involving the valency frames existing in the Czech wordnet of the Balkanet project was carried on with Romanian wordnets as target. We used 601 valency frames, kindly offered by the Czech partner in Balkanet and the Czech-Romanian aligned sub-corpus of the *"1984"* parallel corpus. The manual validation of the automatic transfer of the Czech valency frames from the

[8] The phr relation is not included being specific for English phrasal verbs.

Czech verbs to their Romanian translation equivalents revealed a surprisingly high matching (80%), given the differences between Slavic and Romance languages.

5 Collocations analysis in a parallel corpus

Once a parallel corpus has been word-aligned a very interesting cross-lingual study can be achieved in the area of multilingual terminology, multiword expressions and collocational patterns. Within the context of a trilateral project (University Marc Bloch from Strasbourg, IMS Stuttgart University and RACAI) we experimented on a large four-language parallel corpus (En-Ro-Fr-Ge) extracted from the Acquis Communautaire (AcqCom) multilingual corpus. For the word-alignment we used the English text as a hub language and after generating the En-Ro, En-Fr and En-Ge alignments, by transitivity we computed the alignments Ro-Fr, Ro-Ge and Fr-Ge. These last three alignments were combined, as discussed in Section 3, with the corresponding Ro-Fr, Ro-Ge and Fr-Ge alignments directly generated from the parallel corpus, thus obtaining more accurate alignments.

For the texts in each language in this parallel corpus the collocations were independently extracted (RACAI did it for Romanian and English, IMS for German and University Marc Bloch of Strasbourg for French). Our collocation extraction algorithm is similar to Smadja and McKeown's approach [26]. Based on the word alignment of the different bitexts, one could extract the translations in one language of the collocations detected in the other languages. At the time of the writing of this paper, we performed the partial analysis of the collocations in Romanian and English with respect to their translations in English and Romanian respectively. We selected the best scored 20.000 independently extracted collocations in English ($COLLOC_{EN}$) and Romanian ($COLLOC_{RO}$). Then, by translation equivalence relations, found by the word aligner, we identified the translations into Romanian of the English collocations (TR_{RO}-$COLLOC_{EN}$) and the translations into English of the Romanian collocations (TR_{EN}-$COLLOC_{RO}$).

Given that the corpus contains specific uses and specialized language most of the collocations represent specific multi-word terms and most of them have word by word translation (Member State = Stat Membru, administrative transparency = transparenţă administrativă, act of accession = act de aderare, enter into force = intra în vigoare, etc.). The vast majority of these collocations were found in the intersection of the sets:

$$\text{SURE-COLLOC}_X = \text{COLLOC}_X \cap \text{TR}_X\text{-COLLOC}_Y \quad (1)$$

with X=English &Y=Romanian or X=Romanian &Y English respectively.

However, the most interesting collocations were those not found in the previous intersection sets:

$$\text{INTERESTING-COLLOC}_Z = \text{COLLOC}_Z \setminus \text{SURE-COLLOC}_Z \quad (2)$$

with Z=English or Romanian

The multiword expressions in the lists INTERESTING-$COLLOC_Z$ were hand validated for termhood, cleaned-up and classified into three major cases:

a) aligner failure to detect the equivalence, due to preprocessing error and its imperfect RECALL (ex: "in vitro", "in vivo"; in Romanian these words were both wrongly tagged and lemmatized)

b) aligner failure to detect the equivalence due to a free human translation of the original text
c) aligner failure to detect the equivalence due to a non-word-by-word translation of the terms (especially those containing light-verbs).

An example of the case b) is given by the following original English text:
„Whereas under Article 6 of the abovementioned Regulation the time when a transaction is carried out is considered as being the date on which occurs event, as defined by Community rules or, in the absence of and pending adoption of such rules, by the rules of the Member State concerned, in which the amount involved in the transaction becomes due and payable." , which was translated as:
„Întrucât, conform art. 6 din regulamentul menționat anterior, se consideră ca moment al realizării operației data la care intervine faptul generator de creanță legată de valoarea aferentă operației respective, așa cum este el definit de reglementarea comunitară sau, dacă aceasta nu există şi urmează a fi adoptată, de reglementarea statului membru interesat."
In this example the scattered English text: *„(the) event.. in which the amount.. becomes due and payable"* coresponds to the Romanian term *„situație generatoare de creanțe*" (a literal translation would be „(a) situation generating dues")

The last category is the most interesting as it outlines the multiword expressions which, due to their structural differences, are the hardest to translate by a simple-minded word-by-word approach. They range from legalese jargon (e.g. adversely affect <-> *a aduce atingere*[9]; legal remedy <-> *cale de atac*[10]; to make good the damage <-> *a compensa daunele*[11] etc.) to constructions which are language and culture specific: to shake hands <-> *a da mâna*[12]*;* piece of cake <-> *floare la ureche*[13]; Failing to use the exact wording of such a multiword expression, usually, is the major error source for language comprehension/production by language learners, as well as for other human beings in need to communicate but constrained to use a foreign language.

An interesting preliminary contrastive report on the light-verbs based collocations, with a case study of the verb *a face* (to do/make) for the French-Romanian AcqCom data is presented in [29].

We plan to develop a multilingual collocation dictionary, placing the major emphasis on the "hard" collocations (those existing in at least one inventory INTERESTING-COLLOC$_Z$ with Z one of the project languages) providing structural descriptions, translations in all considered languages, morpholexical restrictions on constituents (such as obligatory definiteness/indefiniteness, singular/plural, obligatory case, etc). We aim at a unified description of the collocational patterns in the four languages (with a perspective to extend our work to all the languages represented in the Acquis Communautaire corpus) and the development of a comprehensive multilingual dictionary, essential for dealing with the hard topic of collocation translation.

---

[9] A mot-a-mot translation would be *to bring a touch*
[10] A mot-a-mot translation would be *way to attack*
[11] A mot-a-mot translation would be *to compensate the damages*
[12] A mot-a-mot translation would be *to give the hand*
[13] A mot-a-mot translation would be *a flower at the ear*

6 Web Services

We implemented a NLP web-services platform which currently ensures the basic preprocessing steps (tokenization-including multiword expression recognition, tiered tagging, lemmatization, language identification, sentence alignment, wordnet browsing) for English and Romanian corpora, as well as a search engine for the English-Romanian parallel corpora.

The services are implemented using standard technology (SOAP/WSDL/UDDI) on a dedicated bi-processor server with a reasonable high-speed internet connection (100Mb/s). The NLP web-services will be continuously extended with new services (word alignment, collocation extraction, translation, QA in open domains, summarization, etc). Although most of the present (and near-future) services are available only for Romanian and English, we plan to add as many new languages as possible. The CLARIN initiative (http://www.clarin.eu/), recently included into the European Roadmap for Research Infrastructures[14], has been adhered by more than 60 institutions from 30 European countries and it is supposed to be the major collaborative work environment that will create, adapt and maintain the language resources and tools we could add to the NLP web services platform.

The current web-services were already used for mass data processing by various remote project partners in the ROTEL project (http://rotel.racai.ro), LT4L project (http://www.lt4el.eu/) and the cross-lingual (Romanian-English) CLEF 2006 QA task (http://www.clef-campaign.org/2006/working_notes/CLEF2006WN-Contents.html).

7 Conclusions

The recent advances in NLP technology, demonstrate the tremendous benefits of collecting and adequately encoding large parallel corpora and multilingual semantic lexicons and ontologies. Building, in a concerted way, this kind of resources, for as many languages and as large as possible, should be a constant objective for an internationally established research infrastructure. Initiatives like Global WordNet (http://www.globalwordnet.org/), furthered by Wordnet Grid proposal [9], Language Grid (http://langrid.nict.go.jp/), the previously mentioned CLARIN initiative and a few others are pioneering this ever increasing need.

The multilingual tools we discussed in this paper have been tested on several languages and showed that when large and good quality language resources are available, rewarding results can be obtained with limited cross-linguistic expertise. Although our experiments considered only Indo-European languages, we are confident that even with more distant pairs of languages, provided the adequate resources are available, the alignment system and the linguistic knowledge induction applications should work reasonably well.

[14] http://cordis.europa.eu/esfri/roadmap.htm

References

1. Barbu-Mititelu, V., Ion, R.: Cross-language Transfer of Syntactic Relations Using Parallel Corpora. Proceedings of the Workshop on Cross-Language Knowledge Induction, EUROLAN'2005, Cluj-Napoca,Romania, (2005) 46-51.
2. Brants, T.: TnT a statistical part-of-speech tagger. Proceedings of the 6th ANLP Conference, Seattle, WA., (2000) 224-231
3. Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C-R., Hsieh, S-K., Marchetti, A., Tesconi, M.: Fostering Intercultural Collaboration: a Web Service Architecture for Cross-Fertilization of Distributed Wordnets. In [17] (2007)
4. Brown, P. F., Della Pietra, S.A., Della Pietra, V. J., Mercer, R. L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2) (1993) 263–311
5. Ceauşu, Al.: Maximum Entropy Tiered Tagging. Proceedings of the Eleventh ESSLLI Student Session, Malaga, Spain (2006) 173-179
6. Ceauşu, Al., Ştefănescu, D., Tufiş, D.: Acquis Communautaire sentence alignment using Support Vector Machines. Proceedings of the 5th LREC Conference, Genoa, Italy (2006) 2134-2137
7. Erjavec T.: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the 4th LREC Conference, Lisbon, Portugal (2004) 1535 - 1538
8. Fan, R.-E., Chen, P.-H. and Lin, C.-J.:. Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University, www.csie.ntu.edu.tw/ ~cjlin/papers/quadworkset.pdf (2005)
9. Fellbaum, C., Vossen, P.: Connecting the Universal to the Specic: Towards the Global Grid. In [17] (2007)
10. Fellbaum, Ch. (ed.): WordNet: An Electronic Lexical Database, MIT Press (1998)
11. Gale, W.A., Church, K.W.: A Program for Aligning Sentences in Bilingual Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, USA (1991) 177–184
12. Hashimoto, C., Bond, F., Flickinger, D.: The Lextype DB: A Web-based Framework for Supporting Collaborative Multilingual Grammar and Treebank Development. In [17] (2007)
13. Hayashi, Y.: Conceptual Framework of an Upper Ontology for Describing Linguistic Services. In [17] (2007)
14. Inaba, R., Murakami, Y., Nadamoto, A., Ishida, T.: Multilingual Communication Support Using the Language Grid. In [17] (2007)
15. Ion, R.: Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română, PhD Thesis, Romanian Academy, Bucharest , Romania, (2006) 145 pp.
16. Ion, R., Tufiş, D.: Multilingual Word Sense Disambiguation Using Aligned Wordnets. Romanian Journal on Information Science and Technology, Tufiş D. (ed.) Special Issue on BalkaNet, Romanian Academy, vol7, no. 2-3 (2004) 198-214
17. Ishida, T., Fussell, S.R., Vossen, P.T.J.M. (eds): Intercultural Collaboration I. Lecture Notes in Computer Science, Springer-Verlag (2007)
18. Koda, T.: Cross-cultural Study of Avatars' Facial Expressions and Design Considerations within Asian Countries. In [17] (2007)
19. Magnini B. Cavaglià G.: Integrating Subject Field Codes into WordNet. In Proceedings of LREC2000, Athens, Greece (2000) 1413-1418.
20. Martin, J., Mihalcea, R., Pedersen, T.: Word Alignment for Languages with Scarce Resources. In Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"., Ann Arbor, Michigan, June, Association for Computational Linguistics (2005).65–74

21. Moore, R. C.: Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users. Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California, Springer-Verlag, Heidelberg, Germany (2002) 135-244
22. Niles, I., and Pease, A.: Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine (2001) 17-19.
23. Och, F., J., Ney, H.: Improved Statistical Alignment Models, Proceedings of ACL2000, Hong Kong, China (2000) 440-447
24. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, 29(1) (2003) 19-51
25. Schneider, P., Mattenklott A: Emotion Eliciting Events in the Workplace: An Intercultural Comparison. In [17] (2007)
26. Smadja, F. A., McKeown, K. R.: Automatically extracting and representing collocations for language generation. In Proceedings of the 28th annual meeting on Association for Computational Linguistics, Pittsburgh, Pennsylvania (1990) 252-259
27. Sornlertlamvanich, V., Charoenporn, T., Robkop, K., Isahara, H.: Collaborative Platform for Multilingual Resource Development and Intercultural Communication. In [17] (2007)
28. Tapanainen, P., Järvinen, T.: A dependency parser for English. Technical Report no. TR-1, Department of General Linguistics, University of Helsinki, Finland (1997)
29. Todiraşcu, A.: Towards an automatic extraction of collocations; the case of the verb MAKE/DO (în Romanian). In Proceedings of the National Workshop on Romanian Language Processing, Iaşi, 3-4 November, (2006) 95-101
30. Tufiş, D., Barbu, A., Ion, R.: Extracting Multilingual Lexicons from Parallel Corpora, Computers and the Humanities, Volume 38, Issue 2 (2004) 163 – 189
31. Tufiş, D., Ion, R., Ceauşu, Al., Ştefănescu, D.: Improved Lexical Alignment by Combining Multiple Reified Alignments. In Proceedings of the 11th Conference of the European Association for Computational Linguistics (EACL), Trento, Italy (2006) 153-160
32. Tufiş, D., Ion, R. Ceauşu, Al., Stefănescu, D.: Combined Aligners. In Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan, June, Association for Computational Linguistics (2005) 107-110.
33. Tufiş, D., Ion, R., Ide, N.: Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In Proceedings of the 20th COLING Conference, Geneva (2004) 1312-1318
34. Tufiş, D.: Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence, Vol. 1692. Springer-Verlag, Berlin Heidelberg New-York (1999) 28-33.
35. Tufiş, D. Cristea, D., Stamou S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal on Information Science and Technology, Tufiş D.(ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2-3 (2004) 9-34
36. Tufiş, D., Barbu-Mititelu, V., Bozianu, L., Mihăilă, C.: Romanian WordNet: New Developments and Applications. Proceedings of the 3rd Conference of the Global WordNet Association, Jeju, Republic of Korea (2006) 337-344
37. Vossen P. (ed.): A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht (1998)