Programme

14:30- Opening Remarks

14:35 – 15:00 Overwiew of Multilingual ressources in Central and (south)-Easter Europe

15:00 – 15: 30 *South-East European Times: A parallel corpus of Balkan languages*, Francis Tyers and Murat Serdar Alperen

15:30- 16:00 *Verbal Valency in Multilingual Lexica,* Oleg Kapanadze

16:00 – 16:30 Coffee Break

16:30 – 17:00 Augmenting a statistical machine translation baseline system with syntactically motivated translation examples, Elena Irimia and Alexandru Ceausu

17:00 – 17:30 Linguistic Resources for Factored Phrase-based Statistical Machine Translation Systems, Mirabela Navlea and Amalia Todirascu

17:30 – 18:00 Sense Disambiguation-"Ambiguous Sensation"? Evaluating Sense Inventories for verbal WSD in Hungarian , Judit Kuti, Enika Héja and Bálint Sass

18:00 -18:30 Building and exploiting Romanian corpora for the study of Differential Object Marking , Anca Dinu and Alina Tigau

18:30 -19:00 *Towards the Construction of Language Resources for Greek Multiword Expressions: Extraction and Evaluation*, Evita Linardaki, Carlos Ramisch, Aline Villavicencio and Aggeliki Fotopoulou

Organiser(s)

•

- Stelios Piperidis, ILSP, Natural Language and Knowledge Extraction Department, Athens Greece
- Milena Slavcheva, Bulgarian Academy of Science, Institute for Parallel Processing, Linguistic Modeling Department, Sofia, Bulgaria
- Cristina Vertan, University of Hamburg, Research Group„Computerphilologie“, University of Hamburg, Germany

Programme Committee

- Galja Angelova(University of Sofia, Bulgaria)
- Bogdan Babych (Centre for Translation Studies,Leeds)
- Damir Cavar (University of Zadar, Croatia)
- Tomaz Erjavec (Jozef Stefan Institute, Slovenia)
- Maria Gavrilidou (ILSP Athens, Greece)
- Walther von Hahn (University of Hamburg)
- Cvetana Krstev (University of Belgrad)
- Vladislav Kubon (Charles University Prague)
- Petya Osenova (University of Sofia, Bulgaria)
- Stelios Piperidis (ILSP Athens, Greece)
- Gabor Proszeky (Morphologic)
- Adam Przepiórkowski (IPAN, Polish Academy of Sciences)
- Milena Slavcheva (Bulgarian Academy of Sciences)
- Kiril Simov (Bulgarian Academy of Sciences)
- Dan Tufis (Romanian Academy of Sciences)
- Cristina Vertan (University of Hamburg)
- Dusko Vitas (University of Belgrade, Serbia)

Table of Contents

Author Index

Foreword

The reconciliation of differences in the availability of language resources and tools for more intensively and less intensively spoken languages has been the main concern of several European initiatives. Central and (South-) Eastern European languages can be the subject of a case study in that respect: integration of diverse languages into a broad language community.

The main result of these initiatives was the increased production of language resources and especially language technology tools for Central, Eastern and Southern European languages in recent years. While monolingual systems have achieved performances comparable to those for intensively studied languages, still a lot of work has to be invested in multilingual tools for applications such as machine translation or cross-lingual information retrieval. At least three major issues have critical influence on the performance of such systems:

- the availability of the appropriate quantities of data for training and evaluation
- the analysis of structural linguistic differences among languages so as to be able to improve statistical methods with targeted linguistic knowledge;
- the availability of knowledge bases for incorporation into language processing systems.

The identification of key aspects of linguistic modelling and resource supply for multilingual technologies involving Central, Eastern and Southern European languages can have impact not only on the local improvement of such systems but also on the overall development of multilingual technologies. The same holds for well established or emerging linguistic knowledge representation frameworks, which can only benefit from embedding components for Central, Eastern and Southern European languages.

The current workshop, organised in conjunction with LREC, that is, the most important conference addressing the Language Resources, aims at capturing the attention of the Language Technology community by presenting specific features of multilingual resources for Central and (South-)Eastern European Languages, as well as language idiosyncrasies, which raise burning questions in general.

The accepted papers reflect exactly the issues in question: four papers deal with multilingual resources and machine translation, the rest three papers focus on particular language problems like verb sense disambiguation or multiword expression extraction. Twelve Central and (South-)Eastern European languages are considered. The fact that besides English, French and German are also involved in the multiligual experiments, makes a good impression on the reader.

We hope that the included papers as well as the emerging discussions during the workshop will give a new impulse to the research and production of language ressouruces for Central and (south)-Eastern European languages.

Stelios Piperidis,
Milena Slavcheva
Cristina Vertan

Augmenting a statistical machine translation baseline system with syntactically motivated translation examples

Elena Irimia, Alexandru Ceauşu
Research Institute for Artificial Intelligence, Romanian Academy
Calea 13 Septembrie no 13, Bucharest, Romania
{elena,aceausu}@racai.ro

Abstract

This paper describes a series of machine translation experiments with the English-Romanian language pair. The experiments were intended to test and prove the hypothesis that syntactically motivated long translation examples added to a base-line 3gram statistically extracted phrase table improves the translation performance in terms of the score BLEU. Extensive tests with a couple of different scenarios were performed: 1) simply concatenating the "extra" translations example to the baseline phrase-table; 2) computing and taking into account perplexities for the POS-string associated to the translation examples; 3) taking into account the number of words in each member of a translation example; 4) filtering the "extra" translation examples by taking into account a score that appreciates the correctness of their lexical alignment. Different combinations of the four scenarios were also tested. Also, the paper presents a method for extracting syntactically motivated translation examples using the dependency linkage of both the source and target sentence. To decompose the source/target sentence into fragments, we identified two types of dependency link-structures - super-links and chains - and used these structures to set the translation example borders.

1. Introduction

Corpus-based paradigm in machine translation has seen various approaches for the task of constructing reliable translation models,

- starting from the naïve "word-to-word" correspondences solution which was studied in the early works (Gale and Church, 1991; Melamed, 1995).
- continuing with the chunk-bounded n-grams (Kupiec, 1993; Kumano and Hirakawa, 1994; Smadja at al., 1996) which were supposed to account for compounding nouns, collocations or idiomatic expressions,
- passing through the early approach of the bounded-length n-grams IBM statistical translation models and the following phrase-based statistical translation models (Och et al, 1999; Marcu and Wong, 2002; etc.),
- exploring the dependency-linked n-grams solutions which can offer the possibility of extracting long and sometimes non-successive examples and are able to catch the structural dependencies in a sentence (e.g., the accord between a verb and a noun phrase in the subject position), see (Yamamoto and Matsumoto, 2003),
- and ending with the double-sided option for the sentence granularity level, which can be appealing since the sentence boundaries are easy to identify but brings the additional problem of fuzzy matching and complicated mechanisms of recombination.

Several studies were dedicated to the impact of using syntactical information in the phrase extraction process over the translation accuracy. Analyzing by comparison the constituency-based model and the dependency based model, Hearne et al. (2008) concluded that "using dependency annotation yields greater translation quality than constituency annotation for PB-SMT". But, as previous works (Groves and Way, 2005; Tinsey et al., 2007) have noted, the new phrase models, created by incorporating linguistic knowledge, do not necessarily improve the translation accuracy by themselves, but in combination with the "old–fashioned" bounded-length phrase models.

The process of extracting syntactically motivated translation examples varies according to the different resources and tools available for specific research groups and specific language pairs. In a detailed report over the syntactically-motivated approaches in SMT, focused on the methods that use the dependency formalism, Ambati (2008) distinguishes the situations when dependency parsers are used for both source and target languages from those in which only a parser for the source side is available. In the latter case, a direct projection technique is usually used to do an annotation transfer from the source to the target translation unit. This approach is motivated by the direct correspondence assumption (DCA, Hwa et al., 2005), that states that dependency relations are preserved through direct projection. The projection is based on correspondences between the words in the parallel sentences, obtained through the lexical alignment (also called word alignment) process. Obviously, the quality of the projection is dependant of the lexical alignment quality. Furthermore, Hwa (2005) notes that the target syntax structure obtained through direct projection is isomorphic to the source syntax structure, thus producing isomorphic translation models. This phenomenon is rarely corresponding to a real isomorphism between the two languages involved.

In the experiments we describe in this paper, we had the advantage of a probabilistic non-supervised dependency analyzer which depends on the text's language only through a small set of rules designed to filter the previously identified links. As both source and target dependency linking analysis is available, there is no need of direct projection in the translation examples extraction and the problem of the "compulsory isomorphism" is avoided.

2. Research background

In previous experiments with an example-based approach on machine translation for the English-Romanian

language pair, we developed a strategy for extracting translation examples using the information provided by a dependency-linker described in (Ion, 2007). We then justified our opting for the dependency-linked n-grams approach based on the assumption in (Cranias et al., 1994) that the EBMT potential should rely on exploiting text fragments shorter than the sentence and also on the intuition that a decomposition of the source sentence in "coherent segments", with complete syntactical structure, would be "the best covering" of that sentence.

The dependency-linker used is based on Yuret's Lexical Attraction Model (LAM, Yuret, 1998), in whose vision the lexical attraction is a probabilistic measure of the combining affinity between two words in the same sentence. Applied to machine translation, the lexical attraction concept can serve as a mean of guaranteeing the translation examples usefulness. If two words are "lexically attracted" to one another in a sentence, the probability for them to combine in future sentences is significant. Therefore, two or more words from the source sentence that manifest lexical attraction together with their translations in the target language represent a better translation example than a bounded length n-gram.

The choice for the Yuret's LAM as the base for the dependency analyzer application was motivated by the lack of a dependency grammar for Romanian. The alternative was to perform syntactical analysis based on automatically inducted grammatical models. A basic request for the construction of this type of models is the existence of syntactically annotated corpora from which machine learning techniques could extract statistical information about the ways in which syntactical elements combine. As no syntactically annotated corpus for Romanian was available, the fact that Yuret's method could use LAM for finding dependency links in a not-annotated corpus made this algorithm a practical choice.

LexPar (Ion, 2007), the dependency links analyzer we used for the experiments described in this paper, is extending Yuret's algorithm by a set of syntactical rules specific to the processed languages (Romanian and English) that constraints the link formation. It also contains a simple generalization mechanism for the link properties, which eliminates the initial algorithm inadaptability to unknown words. However, the LexPar algorithm does not guarantee a complete analysis, because the syntactic filter can contain rules that forbid the linking of two words in a case in which this link should be allowed. The rules were designed by the algorithm's author based on his observations of the increased ability of a certain rule to reject wrong links, with the risk of rejecting good links in few cases.

In our research group, significant efforts were invested in experimenting with statistical machine translation methodologies, focused on building accurate language resources (the larger the better) and on fine-tuning the statistical parameters. The aim was to demonstrate that, in this way, acceptable MT prototypes can be quickly developed and the claim was supported by the encouraging Bleu scores we obtained for the Romanian<->English translation system. The translation experiments employed the MOSES toolkit, an open source platform for development of statistical machine translation systems (see next section). The major rationale for selecting this environment was its novel decoding component that facilitates the usage of multiple (factored) translation models.

One of the goals of this paper is to report our findings on the impact of incorporating syntactic information in the translation model by means of a probabilistic dependency link analyzer. Although the non-supervised nature of the analyzer is affecting its recall, using this tool brings the advantage of having syntactic information available for translation without the need of syntactically annotated corpora. We feed the Moses decoder with the new translation model and we compare the translation results with the results of the baseline system.

3. A baseline Romanian-English Machine Translation System

The corpus. The Acquis Communautaire is the total body of European Union (EU) law applicable in the EU Member States. This collection of legislative text changes continuously and currently comprises texts written between the 1950s and 2008 in all the languages of EU Member States. A significant part of these parallel texts have been compiled by the Language Technology group of the European Commission's Joint Research Centre at Ispra into an aligned parallel corpus, called JRC-Acquis (Steinberger et al., 2006), publicly released in May 2006. Recently, the Romanian side of the JRC-Acquis corpus was extended up to a size comparable with the dimensions of other language-parts (19,211 documents)).

For the experiments described in this paper, we retained only 1-1 alignment pairs and restricted the selected pairs so that none of the sentences contained more than 80 words and that the length ratio between sentence-lengths in an aligned pair was less than 7. Finally, the Romanian-English parallel corpus we used contained about 600,000 translation units.

Romanian and English texts were processed based on the RACAI tools (Tufiş et al, 2008) integrated into the linguistic web-service platform available at http://nlp.racai.ro/webservices. After tokenization, tagging and lemmatization, this new information was added to the XML encoding of the parallel corpora. Figure 1 shows the representation of the Romanian segment encoding for the translation unit displayed in Figure 1. The tagsets used were compliant with the MULTEXT-East specifications Version3 (Erjavec, 2004) (for the details of the morpho-syntactic annotation, see http://nl.ijs.si/ME/V3/msd/).

```
<tu id="3936">
    ...
        <seg lang="ro">
            <s id="31985L0337.n.83.1">
                <w lemma="informație"
 ana="Ncfpry">Informațiile</w>
                <w lemma="culege"
 ana="Vmp--pf">culese</w>
                <w lemma="conform"
 ana="Spsd">conform</w>
                <w lemma="art." ana="Yn">art.</w>
                <w lemma="5" ana="Mc">5</w>
                <c>,</c>
                <w lemma="6" ana="Mc">6</w>
```

```
                <w lemma="şi" ana="Crssp">şi</w>
                <w lemma="7" ana="Mc">7</w>
                <w lemma="trebui"
        ana="Vmip3s">trebuie</w>
                <w lemma="să" ana="Qs">să</w>
                <w lemma="fi" ana="Vasp3">fie</w>
                <w lemma="lua"
        ana="Vmp--pf">luate</w>
                <w lemma="în" ana="Spsa">în</w>
                <w lemma="considerare"
        ana="Ncfsrn">considerare</w>
                <w lemma="în cadrul" ana="Spcg">în
        cadrul</w>
                <w lemma="procedură"
        ana="Ncfsoy">procedurii</w>
                <w lemma="de" ana="Spsa">de</w>
                <w lemma="autorizare"
        ana="Ncfsrn">autorizare</w>
                <c>.</c>
            </s>
        </seg>
    ...
</tu>
```

Figure 1: Linguistically analysed sentence (Romanian) of a translation unit of the JRC-Acquis parallel corpus

Based on the monolingual data from the JRC-Acquis corpus we built language models for each language. For Romanian we used the TTL (Ion, 2007) and METT (Ceauşu, 2006) tagging modelers. Both systems are able to perform tiered tagging (Tufiş, 1999), a morpho-syntactic disambiguation method that was specially designed to work with large (lexical) tagsets.

In order to build the translation models from the linguistically analyzed parallel corpora we used GIZA++ (Och and Ney, 2000) and constructed unidirectional translation models (EN-RO, RO-EN) which were subsequently combined. After that step, the final translation tables were computed. The processing unit considered in each language was not the word form but the string formed by its lemma and the first two characters of the associated morpho-syntactic tag (e.g. for the wordform "informațiile" we took the item "informație/Nc"). We used for each language 20 iterations (5 for Model 1, 5 for HMM, 1 for THTo3, 4 for Model3, 1 for T2To4 and 4 for Model4). We included neither Model 5 nor Model 6, as we noticed a degradation of the perplexities of the alignment models on the evaluation data.

The MOSES toolkit (Koehn et al., 2007) is a public domain environment, which was developed in the ongoing European project EUROMATRIX, and allows for rapid prototyping of Statistical Machine Translation systems. It assists the developer in constructing the language and translation models for the languages he/she is concerned with and by its advanced factored decoder and control system ensures the solving of the fundamental equation of the Statistical Machine Translation in a noisy-channel model:

$$Target^* = \arg\max_{Target} P(Source|Target)*P(Target) \quad (1)$$

The P(Target) is the statistical representation of the (target) language model. In our implementation, a language model is a collection of prior and conditional probabilities for unigrams, bigrams and trigrams seen in the training corpus. The conditional probabilities relate lemmas and morpho-syntactic descriptors (MSD), word-forms and lemmas, sequences of two or three MSDs. The P(Source|Target) is the statistical representation of the translation model and it consists of conditional probabilities for various attributes characterizing equivalences for the considered source and target languages (lemmas, MSDs, word forms, phrases, dependencies, etc). The functional argmax is called a decoder and it is a procedure able to find, in the huge search space P(Source|Target)*P(Target) corresponding to possible translations of a given Source text, the Target text that represent the optimal translation, i.e. the one which maximizes the compromise between the faithfulness of translation (P(Source|Target)) and the fluency/grammaticality of the translation (P(Target)). The standard implementation of a decoder is essentially an A* search algorithm.

The current state-of-the-art decoder is the factored decoder implemented in the MOSES toolkit. As the name suggests, this decoder is capable of considering multiple information sources (called factors) in implementing the argmax search. What is extremely useful is that the MOSES environment allows a developer to provide the MOSES decoder with language and translation models externally developed, offering means to ensure the conversion of the necessary data structures into the expected format and further improve them. Once the statistical models are in the prescribed format, the MT system developer may define his/her own factoring strategy. If the information is provided, the MOSES decoder can use various factors (attributes) of each of the lexical items (words or phrases): occurrence form, lemmatized form, associated part-of-speech or morpho-syntactic tag. Moreover, the system allows for integration of higher order information (shallow or even deep parsing information) in order to improve the output lexical items reordering. For further details on the MOSES Toolkit for Statistical Machine Translation and its tuning, the reader is directed to the EUROMATRIX project web-page http://www.euromatrix.net/ and to the download web-page http://www.statmt.org/moses/.

4. Extracting translation examples from corpora (ExTrAct)

In our approach, based on the availability of a dependency-linker for both the source and the target language, the task of extracting translation examples from a corpus contains two sub-problems: dividing the source and target sentences into fragments and setting correspondences between the fragments in the source sentence and their translations in the target sentence. The last problem is basically fragment alignment and we solved it through a heuristic based on lexical alignments produced by GIZA++. The remaining problem was addressed using the information provided by LexPar, the dependency linker mentioned above. With a recall of 60,70% for English, LexPar was considered an appropriate starting point for the experiments (extending or correcting the set of rules incorporated as a filter in LexPar can improve its recall).

Using MtKit, a tool specially designed for the visualization and correction of lexical alignments adapted

to allow the graphical representation of the dependency links, we could study the dependency structures created by the identified links inside a sentence and we were able to observe some patterns in the links' behavior: they tend to group by nesting and to decompose the sentence by chaining. Of course, these patterns are direct consequences of the syntactical structures and rules involved in the studied languages, but the visual representation offered by MtKit simplified the task of formalization and heuristic modeling (see Fig. 2).

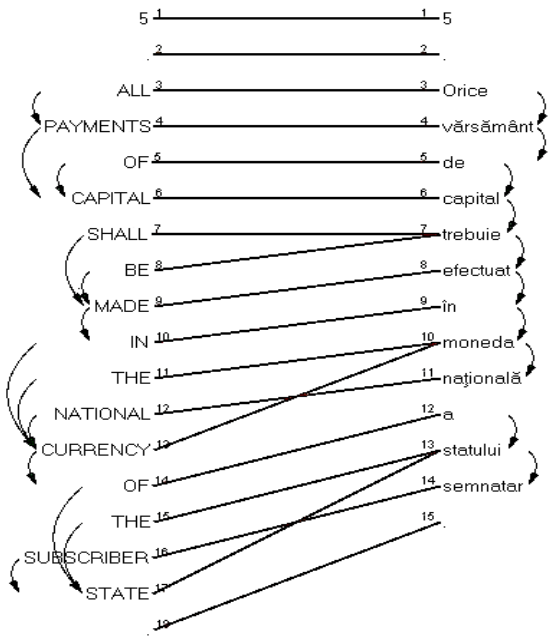

Figure 2. MtKit visualisation of the alignments and links for an English-Romanian translation unit. An arrow marks the existence of a dependency link between the two words it unites. The arrow direction is not relevant for the dependency link orientation.

These properties suggest more possible decompositions for the same sentence, and implicitly the extraction of substrings of different length that satisfy the condition of lexical attraction between the component words.

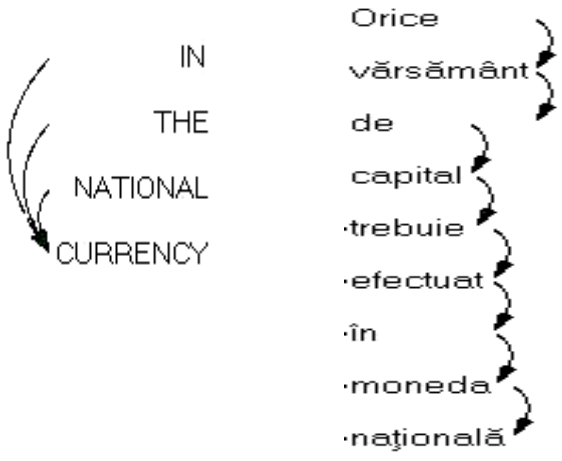

Fig. 3. Superlink structure

Fig 4. Chain structure.

Example 1: in Figure 1, from the word sequence "made in the national currency" the flowing subsequences can be extracted: "national currency", "the national currency", „in the national currency", „made in the national currency". The syntactically incomplete sequences and those susceptible of generating errors (like "the national", "in the", "made in the national") are ignored.

The patterns observed above were formalized as superlinks (link structures composed of at least two simple links which nest, see Figure 3) and as chains (link structures composed of at least two simple links or superlinks which form a chain, see Figure 4).

As input data, ExTract (the application that extracts translation examples from corpora) receives the processed corpus and a file containing the lexical alignments produced by GIZA++ (Och and Ney, 2000). We will describe the extracting procedure for a single translation unit U in the corpus, containing Ss (a source sentence) and its translation Ts (a target sentence). In each member (source or target) of the translation unit we identify and extract every possible chaining of links and superlinks, with the condition that the number of chain loops is limited to 3.The limitation was introduced to avoid overloading the database. Subsequent experiments showed that increasing the limitation to 4 or 5 chains did not significantly improve the BLEU score of the translation system. Two list of candidate sentence fragments, from Ss and Ts, are extracted.

Every fragment in both sentences is projected through lexical alignment in a word string (note that this is not the direct syntactical structure projection discussed above, but a surface string projection) in a fragment of the correspondent sentence. *Example: In Figure2, the projection of the target structure "in the national currency" to the source sentence does not involve the dependency link structure's transfer to the source fragment: "în moneda naţională". The projection means only a translation correspondence between the source/target word sequences, identified by means of the lexical alignment.*

A projected string of a candidate fragment in Ss is not necessarily part of the list of candidate sentence fragments Ts, and vice versa (sometimes, LexPar is not able to identify all the dependency links in a sentence and the lexical alignments are also subject to errors). But if a fragment candidate from Ss projects to a fragment candidate from Ts (as is the case in our example: "in the national currency" is a superlink candidate from Ss, while "*în moneda naţională*" is a chain candidate form Ts), the pair has a better probability of representing a correct translation example. In this stage, the application extracts all the possible translation examples (<source fragment candidate, projected word string>, <projected word string, target fragment candidate>) but distinguish between them, associating a "trust" flag f="2" to the translation examples of the form <source fragment candidate, target fragment candidate>, and a flag f="1" to all the other. Thereby, it is possible to experiment with translation tables of different sizes and different quality levels.

5. Experiments and results

Taking into account results from previous works (Ambati, 2008; Hwa et al., 2005) that proved that dependency-based translation models give improved performance in combination with a phrase-based translation model, we decided to conduct our experiments in a mixed frame: we extracted from the dependency-based translation model only the translation examples longer than (3 source words <-> 3 target words), creating a reduced dependency-based translation model and we combined it with the phrase-based translation model generated with the Moses toolkit.

Starting from the reduced D-based translation model, we can develop two different translation tables, based on the "trust" flags we introduced before:

- a trustful D-based translation table (if we keep only the examples with the flag f="2")
- a relaxed D-based translation table (if we accept all the examples, irrespective of the flags).

For the filtering of the D-based translation model we also implemented a heuristic to evaluate the lexical alignment correctness of each translation example. This brought an increase of around 1% (from 52% to 53% for English-Romanian) in the BLEU score.

In an effort to assure the correctness of the examples used by the Moses decoder from the D-based translation model, we introduced a perplexity score which evaluates the MSD-sequence associated to a string against a MSD-language model. Perplexities were computed for both the English and Romanian side of the translation example database. Nor introducing the perplexity scores as translation factors in the decoder, neither filtering the examples in the D-based translation model produced significant difference in the translation performance.

We also wanted to test if we can increase the performance by introducing a score that favors the longer translation examples in the sentence decomposition. Unexpectedly, the results were not improved: the score BLEU was a little bit lower (e.g. a decrease of around 0.3% for English-Romanian, with no statistic relevance). We think this can be explained by the idea that the longer word sequences in the translation are breaking the integrity of the surrounding sequences: the entire sentence translation performance is remaining similar, since the improvements brought by the longer sequences are balanced by the translation errors coming for the shorted sequences. We also assume this effect is noticeable only for the systems in which the base-line translation model already produces good or very-good translations (in our case, a BLEU score of 0.53 for the Moses table is a very good performance).

As we previously mentioned, the initial working corpus contained around 600,000 translation units. From this number, 600 were extracted for tuning and testing. The tuning of the factored translation decoder (the weights on the various factors) was based on the 200 development sentence pairs and it was done using MERT (Och, 2003) method. The testing set contains 400 translation units.

The evaluation tool was the last version of the NIST official mteval script[1] which produces BLEU and NIST scores. For the evaluation, we lowered the case in both reference and automatic translations. The results are synthesized in the following table, where you can notice that our assumption that the trustful table would produce better results than the relaxed one was contradicted by evidence. We thus learned that a wider range of multi-word examples is preferable to a restricted one, even if their correctness was not guaranteed by the syntactical analysis.

		English to Romanian	**Romanian to English**
Moses phrase table	Nist	8.6671	10.7655
	Bleu	*0.5300*	*0.6102*
Dependency trustful table	Nist	8.4998	10.3122
	Bleu	*0.5006*	*0.5812*
Dependency relaxed table	Nist	8.5978	10.3080
	Bleu	*0.5208*	*0.5921*
D-filtered alignment table	Nist	8.6900	10.3235
	Bleu	***0.5334***	***0.6191***
D-filtered align + ppl table	Nist	8.6827	10.1432
	Bleu	*0.5312*	*0.6050*
D-fitered align+ ppl+length table	Nist	8.5000	10.2910
	Bleu	*0.5306*	0.6083

Table 1: Evaluation of the dependency translation table compared with the translation table generated with Moses (on unseen data).

As the scores in the previous table differ only in a superficial manner, we wanted to look closer at the translation results and study how the augmenting of the translation table with new, longer examples, actually affected the translation quality. In a set of 400 sentences, only 93 (~25%) were translated using 1 or more sequences longer than 3 words (i.e. sequences form D-based translation model). When we examined these sentences, we found out that:

- in 15% of them, using the D-based sequences had a negative impact on the BLEU score (but not necessary on the quality of the translation as assessed by a human evaluator);
- in 50% of the cases, the final form of the translation didn't change (no effect on the performance) ;
- in 35% of the cases, the quality of the translation improved in terms of both the BLEU score and the human evaluator opinion.

Example 2: In the following example the reader can notice a case in which the n-gram matching (and consequently the BLEU score) between the translation and the Romanian reference are improved in the D-based model BEST TRANSLATION (see the bolded words). The example contains also a case in which the performance is not affected by the use of a longer translation example (see the second italic text fragment in the TRANSLATION HYPOTHESIS DETAILS[2]).

English reference (source):
member states shall adopt the measures necessary to comply with this directive within six months of its notification and shall forthwith inform the commission thereof .

[1] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v12.pl

[2] Translation hypothesis details as outputted by the Moses decoder.

Romanian reference (target):
*statele membre **iau măsurile necesare** pentru a se conforma prezentei directive în termen de şase luni de_la data notificării acesteia şi informează imediat comisia cu_privire_la aceasta .*

Moses BEST TRANSLATION: *statele membre adoptă măsurile necesare pentru a se conforma prezentei directive în termen de şase luni de_la notificarea acesteia şi informează de_îndată comisia cu_privire_la aceasta.*
TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..2] member states shall
TRANSLATED AS: statele membre
SOURCE: [3..3] adopt
*TRANSLATED AS: **adoptă***
SOURCE: [4..5] the measures
TRANSLATED AS: măsurile
SOURCE: [6..7] necessary to
TRANSLATED AS: necesare pentru a
SOURCE: [8..10] comply with this
TRANSLATED AS: se conforma prezentei
SOURCE: [11..12] directive within
TRANSLATED AS: directive în termen
SOURCE: [13..14] six months
TRANSLATED AS: de şase luni
SOURCE: [15..17] of its notification
TRANSLATED AS: de_la notificarea acesteia
SOURCE: [18..18] and
TRANSLATED AS: şi
SOURCE: [19..21] shall forthwith inform
TRANSLATED AS: informează de_îndată
SOURCE: [22..24] the commission thereof
TRANSLATED AS: comisia cu_privire_la aceasta
SOURCE: [25..25] .
TRANSLATED AS: .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2 3 4-5 6-7 8-9-10 11-12 13-14 15-16-17 18 19-20-21 22-23-24 25
TARGET: 0-1 2 3 4-5-6 7-8-9 10-11-12 13-14-15 16-17-18 19 20-21 22-23-24 25

D-based model BEST TRANSLATION: *statele membre iau măsurile necesare pentru a se conforma prezentei directive în termen de şase luni de_la notificarea acesteia şi informează de_îndată comisia cu_privire_la aceasta.*
TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..5] member states shall adopt the measures
*TRANSLATED AS: statele member **iau** măsurile*
SOURCE: [6..7] necessary to
TRANSLATED AS: necesare pentru a
SOURCE: [8..10] comply with this
TRANSLATED AS: se conforma prezentei
SOURCE: [11..12] directive within
TRANSLATED AS: directive în termen
SOURCE: [13..14] six months
TRANSLATED AS: de şase luni
SOURCE: [15..17] of its notification
TRANSLATED AS: de_la notificarea acesteia
SOURCE: [18..18] and
TRANSLATED AS: şi
SOURCE: [19..23] shall forthwith inform the commission
TRANSLATED AS: informează de_îndată comisia
SOURCE: [24..25] thereof .
TRANSLATED AS: cu_privire_la aceasta .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2-3-4-5 6-7 8-9-10 11-12 13-14 15-16-17 18 19-20-21-22-23 24-25
TARGET: 0-1 2-3 4-5-6 7-8-9 10-11-12 13-14-15 16-17-18 19 20-21-22 23-24-25

Example 3: This example presents a case in which the score Bleu of the D-based translation is decreased (a case in the 15% of negative impact mentioned before), but the translation remains very good for the human evaluator's perspective (see the bolded fragments).

English reference (source):
for the purpose of determining entitlement to benefits in kind pursuant to article 22 (1) (a) and article 31 of the regulation , " member of the family " means any person regarded as a member of the family under the law on the public health service .
Romanian reference (target):
*în scopul determinării dreptului la prestații în natură în aplicarea art. 22 alin. (1) lit. (a) și a art. 31 din regulament , " membru de familie " reprezintă orice persoană considerată membru de familie **în_conformitate_cu legea privind serviciul public de sănătate** .*

*Moses BEST TRANSLATION: în scopul determinării dreptului la prestații în natură în temeiul articolului 22 alineatul (1) litera (a) și articolul 31 din regulament , " membru de familie " reprezintă orice persoană considerată membru de familie **în_conformitate_cu legea privind sănătatea_publică** .*

TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..2] for the purpose
TRANSLATED AS: în scopul
SOURCE: [3..5] of determining entitlement
TRANSLATED AS: determinării dreptului
SOURCE: [6..7] to benefits
TRANSLATED AS: la prestații
SOURCE: [8..9] in kind
TRANSLATED AS: în natură
SOURCE: [10..12] pursuant to article
TRANSLATED AS: în temeiul articolului
SOURCE: [13..13] 22
TRANSLATED AS: 22
SOURCE: [14..15] (1
TRANSLATED AS: alineatul (1
SOURCE: [16..17]) (
TRANSLATED AS:) litera (
SOURCE: [18..19] a)
TRANSLATED AS: a)
SOURCE: [20..22] and article 31
TRANSLATED AS: și articolul 31
SOURCE: [23..25] of the regulation
TRANSLATED AS: din regulament
SOURCE: [26..28] , " member
TRANSLATED AS: , " membru
SOURCE: [29..31] of the family

TRANSLATED AS: de familie
SOURCE: [32..32] "
TRANSLATED AS: "
SOURCE: [33..35] means any person
TRANSLATED AS: reprezintă orice persoană
SOURCE: [36..38] regarded as a
TRANSLATED AS: considerată
SOURCE: [39..39] member
TRANSLATED AS: membru
SOURCE: [40..42] of the family
TRANSLATED AS: de familie
SOURCE: [43..43] under
TRANSLATED AS: în_conformitate_cu
SOURCE: [44..45] the law
TRANSLATED AS: legea
SOURCE: [46..47] on the
TRANSLATED AS: privind
SOURCE: [48..49] public health
TRANSLATED AS: sănătatea_publică
SOURCE: [50..51] service .
TRANSLATED AS: .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2 3-4-5 6-7 8-9 10-11-12 13 14-15 16-17 18-19 20-21-22 23-24-25 26-27-28 29-30-31 32 33-34-35 36-37-38 39 40-41-42 43 44-45 46-47 48-49 50-51
TARGET: 0-1 2-3 4-5 6-7 8-9-10 11 12-13-14 15-16-17 18-19 20-21-22 23-24 25-26-27 28-29 30 31-32-33 34 35 36-37 38 39 40 41 42

D-based model BEST TRANSLATION: în scopul determinării dreptului la prestații în natură în temeiul articolului 22 alineatul (1) litera (a) şi articolul 31 din regulament , " membru de familie " reprezintă orice persoană considerată membru de familie , ***conform legislației în domeniul sănătății_publice*** *.*

TRANSLATION HYPOTHESIS DETAILS:
SOURCE: [0..5] for the purpose of determining entitlement
TRANSLATED AS: în scopul determinării dreptului
SOURCE: [6..9] to benefits in kind
TRANSLATED AS: la prestații în natură
SOURCE: [10..12] pursuant to article
TRANSLATED AS: în temeiul articolului
SOURCE: [13..13] 22
TRANSLATED AS: 22
SOURCE: [14..15] (1
TRANSLATED AS: alineatul (1
SOURCE: [16..17]) (
TRANSLATED AS:) litera (
SOURCE: [18..20] a) and
TRANSLATED AS: a) și
SOURCE: [21..21] article
TRANSLATED AS: articolul
SOURCE: [22..25] 31 of the regulation
TRANSLATED AS: 31 din regulament
SOURCE: [26..28] , " member
TRANSLATED AS: , " membru
SOURCE: [29..31] of the family
TRANSLATED AS: de familie
SOURCE: [32..32] "
TRANSLATED AS: "
SOURCE: [33..35] means any person
TRANSLATED AS: reprezintă orice persoană
SOURCE: [36..38] regarded as a
TRANSLATED AS: considerată
SOURCE: [39..39] member
TRANSLATED AS: membru
SOURCE: [40..42] of the family
TRANSLATED AS: de familie
SOURCE: [43..45] under the law
TRANSLATED AS: , conform legislației
SOURCE: [46..47] on the
TRANSLATED AS: în
SOURCE: [48..49] public health
TRANSLATED AS: domeniul sănătății_publice
SOURCE: [50..51] service .
TRANSLATED AS: .

SOURCE/TARGET SPANS:
SOURCE: 0-1-2-3-4-5 6-7-8-9 10-11-12 13 14-15 16-17 18-19-20 21 22-23-24-25 26-27-28 29-30-31 32 33-34-35 36-37-38 39 40-41-42 43-44-45 46-47 48-49 50-51
TARGET: 0-1-2-3 4-5-6-7 8-9-10 11 12-13-14 15-16-17 18-19-20 21 22-23-24 25-26-27 28-29 30 31-32-33 34 35 36-37 38-39-40 41 42-43 44

6. Conclusion

We briefly presented only a small part of the various machine translation experiments done in the last year in our research group (including both statistical and dependency-based translation models, the language pair English-Romanian and other languages like Greek and Slovene). We tried to look for solutions to improve the already very good performance of the baseline system on the Romanian-English pair, but in terms of the automatic evaluation method we used (the BLEU/NIST score), the results were not convincing. We analyzed and discovered that the performance increasing impact of adding longer dependency-motivated translation examples can be observed in 5% percent of the translated sentences. We assume that the expected important increasing in the system's performance was not to be seen because the translation quality offered by the baseline MOSES configuration was already very good. Future experiments should address other domains and literary registries, with lesser baseline performances, to check our assumption.

7. Acknowledgements

The work reported here is funded by the STAR project, financed by the Ministry of Education, Research and Innovation under the grant no 742 and by the ACCURAT project funded from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

8. References

Ambati, V. 2008. Dependency Structure Trees in Syntax Based Machine Translation, 11-734 Spring 2008, Survey Report, http://www.cs.cmu.edu/~vamshi/publications/DependencyMT_report.pdf

Ceausu Al. 2006. Maximum Entropy Tiered Tagging. In Janneke Huitink & Sophia Katrenko (editors), Proceedings of the Eleventh ESSLLI Student Session, pp. 173-179

Cranias, L., H. Papageorgiou and S. Piperidis. 1994. A Matching Technique in Example-Based Machine Translation. In Proceedings of the 15th conference on Computational linguistics - Volume 1, Kyoto, Japan 100–104.

Erjavec, T. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris, pp. 1535 – 1538

Gale, W. and K. Church, 1991. Identifying Word Correspondences in Parallel Texts. In Proceedings of the 4th DARPA Speech and Natural Language Workshop, Pacific Grove, CA., pp. 152-157.

Groves D. & Way A. 2005. Hybrid Example-Based SMT: the Best of Both Worlds? In Proceedings of ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, p. 183–190, Ann Arbor, MI.

Hearne, M., S. Ozdowska, and J. Tinsley. 2008. Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. In Proceedings of TALN '08, Avignon, France.

Hwa, R., Ph. Resnik, A. Weinberg, C. Cabezas and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. Nat. Lang. Eng., 11(3):311–325, September.

Ion, R. 2007. Word Sense Disambiguation Methods Applied to English and Romanian, PhD thesis (in Romanian), Romanian Academy, Bucharest, 138 p.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Wade, S., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. MOSES: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Kumano, A. and H. Hirakawa. 1994. Building an MT dictionary from parallel texts based on linguistic and statistical information. In COLING-94: Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan, pages 76-81.

Kupiec, J. 1993. An Algorithm for Finding Noun Phrases Correspondences in Bilingual Corpora. In 31st Annual Meeting of the Association for Computational Linguistics, Columbus, OH., pages 23-30.

Marcu D. and W. Wong. 2002. A Phrased-Based, Joint Probability Model for Statistical Machine Translation. In Proceedings Of the Conference on Empirical Methods in Natural Language Processing (EMNLP 02); pages 133-139, Philadelphia, PA, July.

Melamed, I.D. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best translation lexicons. In proceedings of the Third Annual Workshop on Very Large Corpora, Cambridge, England, pp. 184-198.

Och, F. J. 2003. Minimal Error Rate Training in Statistical Machine Translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167.

Och, F.J., Ney H. 2000. Improved Statistical Alignment Models. In Proceedings of the 38th Conference of ACL, Hong Kong, pp. 440-447

Och, F.-J., Ch. Tillmann and H. Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 99), pages 20–28, College Park, MD, June.

Smadja, F., K.R. McKeown and V. Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics 22(1):1-38.

Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, pp. 2142-2147

Tinsley J., Hearne M. & Way A. 2007. Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In Proceedings of The Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07), Bergen, Norway.

Tufiş, D. 1999. Tiered Tagging and Combined Language Models Classifiers. In Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, Text, Speech and Dialogue (TSD 1999), Lecture Notes in Artificial Intelligence 1692, Springer Berlin / Heidelberg,. ISBN 978-3-540-66494-9, pp. 28-33.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008). RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.

Yamamoto, K. and Y. Matsumoto. 2003. Extracting translation knowledge from parallel corpora. In: Michael Carl & Andy Way (eds.) Recent advances in example-based machine translation (Dordrecht: Kluwer Academic Publishers, 2003); pages 365-395.

Yuret, D. 1998. Discovery of linguistic relations using lexical atrraction. PhD thesis, Department of Computer Science and Electrical Engineering, MIT.

Building and exploiting Romanian corpora for the study of Differential Object Marking

Anca Dinu, Alina Tigau
University of Bucharest, Faculty of Foreign Languages and Literatures
Edgar Quinet 5-7 Bucuresti
Email:{anca_d_dinu, alina_mihaela_tigau}@yahoo.com

Abstract

The motivation for this work is that in Romanian the uses of the accusative marker "pe" with the direct object in combination or not with clitics involve mechanisms which are not fully understood and seeming messy for the non-native speaker: sometimes the accusative marker is obligatory, sometimes it is optional and even forbidden at times. The Differential Object Marking parameter draws a line between languages such as Spanish, Romanian, Turkish, or Russian which show a propensity for overtly marking those objects which are considered to be 'prominent', i.e. high in animacy, definiteness or specificity and other languages, such as German, Duch and English, where such a distinction between types of direct objects is not at stake (they rely mostly on word order to mark the direct object).
In order to find empirical evidences for the way DOM with accusative marker "pe" is interpreted in Romanian, we semi-automatically constructed a corpus of Romanian phrases. We manually annotated the direct objects from the corpus with semantically interpretable features we suspected, based on previous studies, are relevant for DOM, such as [±animate], [±definite],[±human]. We present a systematic account for these linguistic phenomena based on empirical evidence from the corpus.

1. Introduction

The motivation for this work is that in Romanian the uses of the accusative marker "pe" with the direct object in combination or not with clitics involve mechanisms which are not fully understood and seeming messy for the non-native speaker: sometimes the accusative marker is obligatory, sometimes it is optional and even forbidden at times. The Differential Object Marking parameter draws a line between languages such as Spanish, Romanian, Turkish, or Russian which show a propensity for overtly marking those objects which are considered to be 'prominent', i.e. high in animacy, definiteness or specificity and other languages, such as German, Duch and English, where such a distinction between types of direct objects is not at stake (they rely mostly on word order to mark the direct object). Thus, this research tackles a specific linguistic difference among those languages. It presents a systematic account for these linguistic phenomena based on empirical evidence present in corpora. Such an account may be used in subsequent studies to improve statistical methods with targeted linguistic knowledge.

2. Building the corpus

In order to find empirical evidences for the way DOM with accusative marker "pe" is interpreted in Romanian, we semi-automatically constructed a corpus of Romanian phrases. The construction of the corpus was straightforward: we only included the phrases containing the word "pe" from a given set. The only problem was to manually detect and delete from the corpus the occurrences of "pe" which lexicalized the homonym preposition meaning on. By doing so, we obtained 960 relevant examples from present day Romanian: 560 of these were automatically extracted from publically available news paper on the internet; the other 400 examples (both positive and negative) were synthetically created, the majority of which are made up due to the fact that we needed to test the behavior of the direct object within various structures and under various conditions, which made such sequences rare in the literature.
We manually annotated the direct objects from the corpus with semantically interpretable features we suspected, based on previous studies, are relevant for DOM, such as [±animate], [±definite],[±human].
We also assembled a corpus containing 779 examples from XVI-th and the XVII-th century texts (approx. 1000 pages of old texts were perused), in order to study the temporal evolution of DOM in Romanian. In what the XVIth century is concerned, we used Catehismul lui Coresi (1559) (Coresi's Cathehism), Pravila lui Coresi (1570) (Coresi's Code of Laws) as well as various prefaces and epilogues to texts dating from the XVI-th century: Coresi: Tetraevanghel (1561) (The Four gospels), Coresi: Tîlcul evangheliilor (1564) (Explainig the Gospels), Coresi: Molitvenic(1564) (The Prayer Book), Coresi: Psăltire Romînească (1570) (The Romanian Psalm Book), Coresi: Psăltire Slavo-Romînă (1570) (The Slavic-Romanian Psalm Book), Coresi: Evanghelie cu învăţătură (Gospel with Advice), Palia de la Orăştie (1582) (The Old Testament from Orăştie). To these texts we have added a number of documents, testaments, official and private letters. The texts dating from the XVII century were basically chronicles – we had a wider choice of texts as we moved along the centuries. We have studied the following works: Istoria Ţării Româneşti de la octombrie 1688 până la martie 1718 (The History of Ţara Românească from October 1688 until March 1718),

Istoriile domnilor Ţării Rumâneşti. Domnia lui Costandin – vodă Brâncoveanu (Radu Popescu) (The Lives of the Rulers of Ţara Românească. The reign of Costandin Brâncoveanu (Radu Popescu)), Istoria ţării rumâneşti de când au descălecat pravoslavnicii creştîni(Letopiseţul Cantacuzîno)(The Hystory of Ţara Românească since the Advent of the Christian Orthodox Believers)(The Cantacuzino Chronicle), Letopiseţul Ţării Moldovei (Ion Neculce) (The Chronicle of Moldavia by Ion Neculce).
From this old Romanian corpus we noticed that prepositional PE came to be more extensively employed in the XVII-th century texts and by the XVIII-th century it had already become the syntactic norm. It seems that the Accusative was systematically associated with P(R)E irrespective of the morphological and semantic class the direct object belonged to. This is in line with the results arrived at by Heusinger & Onea (2008) who observe that the XIX-th century was the epitome in what the employment of DOM is concerned. This evolution was then reversed around the XIX-th –XX-th centuries so that the use of PE today is more restrained than it was two centuries ago, but more relaxed if we were to compare it to the XVI-th century.

3. Previous accounts of DOM in Romanian

We started our analysis of DOM in Romanian, considering a range of former accounts of the prepositional PE such as the studies of Aissen (2003), Cornilescu (2000), Farkas & Heusinger (2003) in an attempt to isolate the exact contribution of the marker PE on various types of direct objects.
Apparently, DOM in Romanian is affected both by the scale of animacy and by the scale of definiteness (Aissen 2003), as it is largely restricted to animate–referring and specific objects i.e. it is obligatory for pronouns and proper names but optional for definites and specific indefinites. In order to solve this puzzle, Aissen crosses the two scales and comes up with a partial ranking.

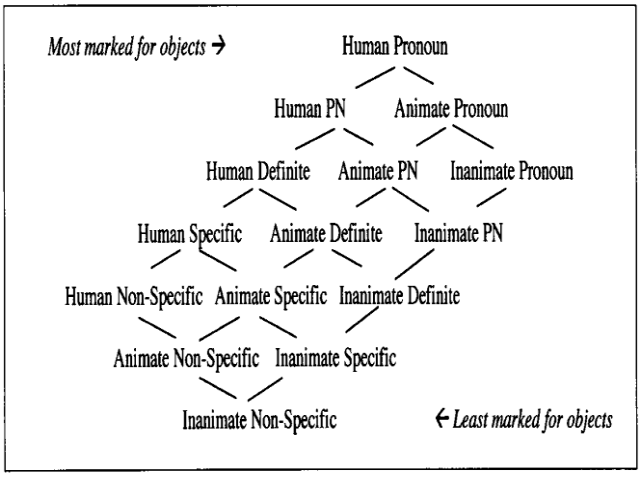

Figure 1: Partial ranking on animacy and definiteness

Thus, as one can see above, pronouns referring to humans outrank (universally) all other types of expressions due to the following reasons: pronouns are the highest on the definiteness scale, outranking all other types of expressions just like the feature [+ human] which outranks everything on the animacy scale. However, there seems to be a problem when it comes to comparing animate pronouns and human determiner phrases (DPs) as the former outranks the latter in terms of the definiteness scale whereas the latter outranks the former with respect to the animacy one. Aissen holds that in this case it is up to the grammar of a particular language to set the ranking.

In Romanian the definiteness scale seems to override the animacy one in that pronouns will always be overtly case marked as opposed to definite DPs whose case marking is optional.
Although Aissen's analysis seems to account for several important general facts about Romanian e.g. why are personal pronouns overtly case-marked as opposed to non-specific indefinites, it does not account for the optionality of overtly case-marking definite DPs and specific indefinites, nor does it explain how the choice when it comes to elements ranked on the same level with the complex scale is made (e.g. human indefinites (optionally case-marked) as opposed to inanimate-referring proper names which are not overtly case-marked).
Cornilescu's (Cornilescu 2000) proposal is that PE is a means of expressing semantic gender - which distinguishes between non-neuter gender (personal gender) and neuter gender (non-personal gender). One advantage of such an account is that it would explain the optionality of PE in certain cases. Thus, while grammatical gender is necessarily marked on the noun's morphology i.e. it is an obligatory feature, semantic gender on the other hand is only sometimes marked formally by PE, 'when it is particularly significant, because the intended referent is prominent. Thus, PE is optional even for nouns denoting person. Furthermore, Cornilescu points to the fact that semantic gender is related to individualization because individualized referents are granted "person" status. Thirdly, it appears that the presence of PE places constraints on the denotations of the overtly case-marked DPs. Thus, the DPs which get overtly case-marked always have an object-level reading and as for their specific denotations, these DPs always select only argumental denotations i.e. <e> (i.e. object) or <<e,t>t> (i.e. generalized quantifier). On the other hand, these DPs never have a property reading i.e. <e,t>, nor do they ever get a kind interpretation which is related to the property reading.
Our analysis is developed within the Discourse Representational Theory (DRT) as it is put forth by Kamp & Reyle (1993) and developed by Farkas & de Swart (2001) and Farkas (2002). DRT is a theoretical semantic-pragmatic framework with the aim of bridging sentence-level semantics and dynamic, discourse level aspects of semantic interpretation. Within this framework, the interpretation process involves updating the semantic

representation with material that affects the truth conditional and the dynamic aspects of discourse. Thus, each new sentence is interpreted with respect to the contribution it makes to an already existing piece of (already) interpreted discourse. The interpretation conditions for sentences act as instructions for the updating the representation of the discourse. The most important tenets of this approach that we employed and along which all distinctions between DPs with respect to DOM were provided, were that each argumental DP contributes a discourse referent (or a value) and a condition on it.

The idea underlying our analysis and which we adopted from Farkas (2002) is that DPs differ one with respect to another on account of the value conditions they contribute. Also on account of the value conditions these DPs introduce, we developed the analysis of DOM in Romanian sentences. The core notion we employed in this respect was that of 'determined reference' which seems to be the underlying parameter organizing DPs along the definiteness scale provided by Aissen (2003). DPs with determined reference are obligatorily marked by PE. (The few exceptions will be accounted for). The animacy scale of Aissen (2003) remains an important factor when it comes to differentially marking the object DP and can sometimes override the parameter of determined reference.

4. Empirically grounded accounts of DOM in Romanian

We give here our findings based on corpus analysis, for the three classes of DPs: proper names and definite pronouns, definite descriptions, indefinite descriptions.

Proper names and definite pronouns differ from definite descriptions in that only the former but not the latter are obligatorily marked by means of PE. This difference was captured in terms of the conditions on how variables introduced by DPs are assigned values. Thus, proper names and definite pronouns contribute equative conditions on the variable they introduce – in virtue of the equative value conditions these DPs contribute, the variables they introduce meet the determined reference requirement. Hence these DPs are obligatorily marked by PE. The only exception in this case is that [- animate] proper names are not marked by means of PE, nor is the relative pronoun ce 'what'. Consider:

1.a.Deseori(o)văd*(pe)Ioana stand la fereastră. [+human]
Often (her.cl.) see PE Ioana sitting at widow.
'I often see Ioana sitting by the window.'

b. Îl chem *(pe) Lăbuş dar s-a ascuns şi aşteaptă să-l găsesc. [-human, +animate]
Him.cl. call.I PE Lăbuş but refl. has hidden and wait SĂ him.cl. find.I.
'I call Lăbuş but he is hiding somewhere waiting for me to find him.'

c. Am traversat (*pe) Parisul pe timp de noapte uitându-ne temători împrejur la tot pasul.
Have we crossed (*PE) Paris during night looking-refl. fearful around every step.
'We crossed Paris during the night, fearfully peering around all the time.'

Thus, proper names acquire PE as a consequence of the interaction between two parameters: determined reference and the animacy scale. The former parameter requires the obligatory use of PE, hence all proper names should be marked in this respect. However, the latter parameter overrides the parameter of determined reference when it comes to [- animate] proper names because these DPs may not receive DOM.

2. a. Îi aşteptam *(pe) ei cu sufletul la gură, dar nu eram prea încântat că vor veni şi ele.
The.cl. waited PE them (masculine) with soul at mouth but not were.we too thrilled that will come and they. (feminine)
'I could hardly wait for the boys' coming but I was not too thrilled that the girls were coming too.' (personal pronoun).

b. Vă strigă pe dumneavoastră, domnule Dinică.
You call PE you Mr. Dinică.
'It is you that they call, Mr. Dinică.' (pronoun of politeness)

c. Babele stăteau toate roată pe lângă poartă doar-doar s-a prinde vreo veste.
Old ladies sat all around near the gate so as to catch any news.
Altele, mai curajoase, stăteau la pândă pe după casă. *(Pe) acestea din urmă le- am speriat de moarte.
Others, more courageous sat in waiting behind the house. PE these latter them.cl. have.I frightened to death. (demonstrative pronoun)

Unlike definite pronouns and proper names, **definite descriptions** contribute a predicative condition on the variables they introduce. This condition does not fix the reference of the variable in question in the way equative conditions do therefore this difference with respect to the nature of the value conditions could be taken to account for the optionality of DOM with definite descriptions. Nevertheless, as pointed out by Farkas (2002), there are some cases of special definite descriptions which may acquire determined reference i.e. if the NP denotes a singleton set relative to the model or a contextually restricted set of entities According to Farkas (2002), this can be achieved in several ways: if the NP is a superlative (e.g. 'the first man on the moon'), if it points to unique referents in relation to the model relative to which the discourse is interpreted (e.g. 'the moon').

Now, if these special types of definite DPs may acquire determined reference, our expectation with respect to their marking by means of PE was for DOM to be obligatory with such DPs. Our corpus analysis proved, however, that this is only partially true as only [+human, + determined reference] definite descriptions were obligatorily marked by means of PE. We needed therefore to weaken our initial hypothesis so as to correspond to the facts present in corpus.

3. a. L-am văzut *(pe) ultimul supravieţuitor de pe

Him.cl. have.I seen PE last survivor on
Titanic şi m-au impresionat foarte tare amintirile lui.
Titanic and me.cl.have impressed very much memories his.
'I have seen the last survivor from the Titanic and I was very impressed with his memories.'

b. Nu am văzut-o (pe) prima căţea care a ajuns pe
Not have.I seen it.cl. PE first dog which reached the
lună, dar ştiu că o chema Laica.
moon, but know.I that it.cl. called Laica.
'I haven's seen the first dog which reached the moon but I know her name was Laica.'

c?Nu-l stiu pe primul obiect gasit in piramida lui Keops
Not him.cl know.I PE first object found in pyramid of Keops
dar trebuie sa fi fost foarte pretios.
but must have been very precious.
'I don't know which was the first object they found in Keops's pyramid but it must have been very precious.'

Thus, the parameter of determined reference still imposes obligatoriness of DOM on those DPs that have determined reference. Nevertheless, in the case of definite descriptions, this parameter is overridden by the animacy scale of Aissen (2003). This accounts for both the obligatory nature of DOM with [+human, + determined reference] definite descriptions (normally DOM is optional with [+ human, - def] definite descriptions) and for the behavior of [- human, +/- animate, + determined reference] definite DPs. The results concerning the interaction between the two parameters are summarized below:

4. a. [+ determined reference] – obligatory DOM
[+ human] – the highest on the animacy scale – preference for DOM
Result: obligatory DOM

b. [+ determined reference] – obligatory DOM
[- human, + animate] – lower on the animacy scale, optional DOM
Result: optional DOM

c. [+ determined reference] – obligatory DOM
[- human, -animate] – lowest on the animacy scale, no DOM
Result: no DOM

As for the definite descriptions having indetermined reference, it proved from the corpus data that, in all these cases where definite DPs had a kind-generic interpretation, (hence they could not acquire determined reference), the use of DOM was prohibited. As it seems, the fact that these DPs could not acquire determined reference was reason enough to disallow the employment of DOM. Consider the example below containing kind denoting definite descriptions (‚fel' kind, or ‚tip' type) – these DPs may not acquire determined reference therefore we expect DOM to be at best optional (if not impossible). In fact, it proves impossible.

5. a. Mihai nu agreează tipul ăsta de fete.
Mihai not like type.the this of girls.
'Mihai does not like this type of girls.'

Furthermore, verbs like 'a iubi' (to love), 'a urî' (to hate), 'a respecta' (to respect), 'a admira' (to admire) range among those verbs which allow a 'kind' reading for the DP occupying their object position. As the examples below point out, PE-DPs (in the plural) are not allowed with these verbs. On the other hand, definite DPs in the plural that are not accompanied by PE can occur in the object position of these verbs and can receive a 'kind' reading as well.

6. a. Ion iubeste femeile.(generic)
Ion loves women.the.

b. ?Ion le iubeste pe femei.(generic).
Ion them.loves PE women.
'Ion loves women'.

Finally, we reverted our attention to **indefinite DPs** and to their behavior with respect to DOM. Since these DPs contribute a discourse referent and a predicative condition on this value, we would not expect them to acquire determined reference, hence the lack of obligatoriness with DOM. However, following the lines of Farkas (2002) we linked the issue of variation in value assignments with indefinites comes with specificity. As it seems, when indefinites are specific (scopally specific or epistemically specific) they may be marked by means of PE as also pointed out by Carmen Dobrovie-Sorin (1994).

7. a. Fiecare parlamentar asculta un cetatean.
Every member of parliament listened a citizen.
'Every member of parliament listened to a citizen.

b. Fiecare parlamentar îl asculta pe (anumit) un cetatean.
Every member of parliament him.cl listened PE (certain) a citizen.
'Every member of parliament listened to a citizen.'

Thus, sentence 7.a. above is ambiguous. It may have a quantificational reading i.e. when the variable introduced by the indefinite is within the scope of the universal quantifier (dependent indefinite i.e. the variable introduced by the indefinite is dependent on the variable introduced by the quantifier). On the other hand, the indefinite may also be outside the scope of the quantifier and point to a certain citizen. If one applies the preposition PE to the indefinite in this case, the interpretation is no longer ambiguous and the balance will be tilted in favor of a referential reading.

Nevertheless, the facts should not be taken at face value: all the examples we provided, the indefinite object was marked by PE but was also resumed by clitic pronoun in the same time. Therefore the specific reading the indefinite DP acquires in these examples may also be due to the presence of the clitic. Another problem which remains unsolved at this point is that concerning the optionality of DOM with these DPs. Thus, indefinite DPs may acquire a specific reading in the absence of DOM (the presence thereof however tilts the balance towards a clearcut specific interpretation). This optionality may reside with the speaker who might play a bigger role in DOM assignment than forseen so far. As it seems, further research is necessary in this respect.

Lastly, we extracted from the corpus some cases where the DOM was impossible: PE can never occur with mass

nouns, bare plurals and incorporated DPs. This was a confirmation for the theoretical premises we had assumed: all these DPs fail to contribute a discourse referent let alone a condition on it.

We have requested the help of 42 native speakers of Romanian who kindly accepted to pass judgments and evaluate our synthetically created corpus of positive and negative 400 examples. Their judgments massively support our empirically grounded findings. The inter-subject agreement was high, i.e. r = 0.89.

5. Conclusions

In order to find empirical evidences for the way DOM with accusative marker "pe" is interpreted in Romanian, we semi-automatically constructed a corpus of Romanian phrases. We manually annotated the direct objects from the corpus with semantically interpretable features we suspected, based on previous studies, that are relevant for DOM, such as [±animate], [±definite],[±human]. Although rather small, these annotations could make such a corpus attractive to be subsequently used to study other linguistic phenomena at semantic and pragmatic level.

Proper names and pronouns (Personal pronouns, pronouns of politeness, reflexive pronouns, possessive pronouns and demonstrative pronouns) are obligatorily marked by means of PE irrespective of the status of the referent on the animacy scale. For proper names the use of PE is conditioned by the animacy scale which overrides the parameter of determined reference: it is obligatory with proper names pointing to [+ human] Determiner Phrases and optional with [+ animate] DPs, and ungrammatical with [-animate] proper names.

Definite descriptions are optionally marked by means of PE; the parameter of determined reference still imposes obligatoriness of DOM on those DPs that have determined reference. Nevertheless, in the case of definite descriptions, this parameter is overridden by the animacy scale. This accounts for both the obligatory nature of DOM with [+human, + determined reference] definite descriptions (normally DOM is optional with [+ human, - def] definite descriptions) and for the behaviour of [- human, +/- animate, + determined reference] definite DPs.

Indefinite Description: Only specific Indefinite Descriptions are optionally marked by means of PE. The others cannot be marked.

Based on the empirical evidence present in the corpus, we proposed a systematic account for DOM with accusative marker "pe" in terms of determined reference (cf. Farkas (2002)) and the animacy scale (Aissen (2003)). Thus, we argue that DOM is triggered both by the semantic nature of the object DP (in terms of how referentially stable it is) and by parameters such as 'animacy'.

6. Acknowledgements

This work was supported by CNCSIS –UEFISCSU, project number PNII – IDEI 228/2007.

References

Aissen, Judith .2003. Differential Object Marking: Iconicity vs. Economy. Natural Language and Linguistic Theory 21: 435-483.

Cornilescu. 2002. At the Romanian Left Periphery. Bucharest Working Papers in Linguistics 4: 88-106.

Enç, Mürvet .1991. The Semantics of Specificity. Linguistic Inquiry 22: 1-25.

Dobrovie-Sorin, Carmen.1994. The Syntax of Romanian. Berlin: Mouton de Gruyter.

Farkas, D & Heusinger, von Fintel. 2003. Stability of reference and object marking in Romanian. Ms. Workshop on Direct reference and Specificity ESSLLI Vienna

Farkas, Donka. 2002a. Specificity Distinction. Journal of Semantics 19: 213-243.

Farkas, D. and H. de Swart. 2001. The Semantics of Incorporation. Stanford, California: CSLI

Franco, J. 2000. Agreement as a Continuum. In Beukema Frits, den Dikken Marcel (eds), Clitic Phenomena in European Languages, 146-189. Amsterdam: Benjamins.

von Heusinger, Klaus & Onea, Edgar 2008. Triggering and blocking effects in the diachronic development of DOM in Romanian, Probus 20:71-116.

Kamp, H. and U. Reyle. 1993. From Discourse to Logic. Dordrecht: Kluwer.

Stan, C. Complementul direct. unpublished ms.

Tigau, Alina.2010. The Syntax of the direct object in Romanian. Unpublished PhD thesis.

Verbal Valency in Multilingual Lexica

Oleg Kapanadze

OK'OMPLEX – Innovative Information and Language Technologies

V. Beridze, 1

0118 Tbilisi, Georgia

E-mail: ok@caucasus.net

Abstract

As it is well known, detailed valency patterns are highly language-specific. This suggests that an optimal multilingual representation will not necessarily be optimal with respect to the individual languages involved - especially when these languages belong to different language families. In this paper we discuss the results of our research on the generalization of syntactic and semantic valency patterns for a mid-size multilingual Georgian, Russian, English, German verbal valency NLP lexicon. From the developed different approaches to verb valency classification we suggest semantic field-oriented multiple hierarchies. In this approach, the verbal material is first grouped with respect to a number of semantic fields. Then, a syntactico-semantic classification is carried out for each of the fields. These hierarchies capture fine-grained differences between subcategorization frames and allow for a direct relation between case frames and corresponding subcategorization frames. For the implementation, we used the DATR-formalism.

1. Introduction

The subcategorization (subcat) information and the thematic role frame information related to a lexical unit (LU) are commonly referred to as ***valency patterns*** (or simply ***valency***) of this LU. Although certain (first of all predicative) nouns and adjectives may also possess a valency, the notion of valency is most naturally related to verbs. Therefore, most of the research on the representation of valency (including ours) and most of the valency lexica are concerned with verbal valency.

The majority of the traditional valency dictionaries and of NLP valency lexica is monolingual. Both, dictionaries and lexica tend to capture only subcategorization information, providing an exhaustive enumeration of subcategorization ***frames*** for each lemma in the alphabetically ordered list. This technique is suitable in the case of intra- and interlinguistic idiosyncratic subcategorization information. Thus, German **GRATULIEREN** "congratulate" subcategorizes for $\mathbf{NP_{nom}}$ $\mathbf{NP_{dat}}$ ***zu*** $\mathbf{NP_{dat}}$ (i.e., for an NP in the *nominative*, an NP in the *dative* and a PP with the preposition *zu* "to", which requires a *dative*) while its near synonym **BEGLÜCKWÜNSCHEN** subcategorizes for $\mathbf{NP_{nom}}$ $\mathbf{NP_{acc}}$ ***zu*** $\mathbf{NP_{dat}}$. Obviously, such an idiosyncrasy must be represented. Also, Russian

доставлять
[dostavljat'] ("deliver")

subcategorizes, among others, for an indirect complement ***na*** "on" $\mathbf{NP_{gen}}$, its German equivalent **LIEFERN** does not; rather, it subcategorizes for ***auf*** "on" $\mathbf{NP_{acc}}$.

However, this technique is certainly suboptimal for any user - be it man or machine - in the case of "regular valency", where a correlation between the semantics of a lemma and its valency exists, and, subsequently, lemmas with similar semantics share at least some of their valencies. In this case, as far as the human user is concerned, the repetition of valency information is either (i) superfluous (as in the case of **DRINK** and **EAT**) or (ii) obscure in that it hides valuable (especially to language learners) hints where generalization and thus learning of valency by analogy is possible. As far as the machine is concerned, regular valency repetition violates the principle of efficient information representation. That is, a more appropriate representation of valency information would make the idiosyncratic valencies explicit, meaningfully generalize valencies where possible, but also provide the means to factor out all valencies of a LU if required. It would also include semantic valency, i.e., case or thematic role patterns as, e.g., in the FrameNet Project (Baker et al. 1998). Each case frame can be associated a set of subcategorization frames by which it can be realized. Semantic valency specified along with subcat information provides thus an explicit link between semantic and syntactic levels of

linguistic representation. It is thus essential, e.g., for NLP-applications that map semantic structures onto surface or operate on semantic structures which are then mapped onto surface structures, but also for meaningful generalization.

In this paper, we discuss an approach to the representation of multilingual valency information that subscribes to the following principles:

- it attempts to capture both the syntactic and the semantic valency;

- it makes explicit the correlation between the semantics of a lemma and its valency;

- it builds up a hierarchical structure in which valencies shared by several LUs are extracted and inherited then to each of these LUs (and has thus an efficient representation not only intralinguistically, but also interlinguistically);

- it is formalized to serve as resource for different NLP applications, but is also thought to provide a mechanism that maps the formal representation onto the SGML format and can be translated into a conventional dictionary.

This approach has been developed in the framework of the GREG Project (a project funded by the EU in the INTAS program). In order to demonstrate the practicality of our approach, we compiled a midsize multilingual valency lexicon (with about 1200 citation forms for each language). The languages involved are English, German, Russian and Georgian (i.e., two Germanic, one Slavic and one South Caucasian/Kartvelian language). The internal formalism used for representation is DATR (Evans & Gazdar 1996).

2. Approaches to the Representation of Valency Information

Nonredundant representation of lexical items means, as a rule, construction of an inheritance hierarchy in which information common to several items is extracted and placed higher in the hierarchy so as to be inherited to all items that posses this information. If an item inherits information that is not compatible with its patterns, this information is overridden; if an item possesses local information, this information is added to the information inherited. This suggests that the problem of the representation of valency information must be considered from two angels: (i) the way LU's that share some or all valency patterns can be grouped together and (ii) the way valency information can be inherited.

The vast majority of valency lexica is monolingual. Since we are concerned with multilingual valency, we first address the question how the representation of multilingual valency differs from the representation of monolingual valency. Then, we review the possible representations of monolingual and multilingual information.

2.1. Monolingual vs. Multilingual Valency Representation

Syntactic valency tends to be dominant in monolingual valency lexica. This might suggest a syntactic classification in which all LUs possess:

- Surface-oriented sharing of valency info
- Semantically-oriented sharing of valency info

However, for multilingual representation it should be clarified to what extent do the translation equivalents possess the same or equivalent valency info?

Several cases are to be discussed:

- adding details to the inherited patterns (e.g., where in English we have **[NP NP]**, in Georgian, German and Russian, we would have something like $\mathbf{[NP_{case\_1}\ NP_{case\_2}]}$.

- sharing of valency across languages (thus, in German and Russian and sometimes also Georgian cases can be observed, where the valency patterns are identical (incl. grammatical cases in the subcat frames),

- standard correspondence (e.g., The Georgian *narrative* - a certain sort of the commonly termed *ergative* - is by default realized in German and Russian by a *nominative*; the Russian *instrumental* is realized in German by default by a *dative*, in certain cases that can be traced back to different case frames, by an *accusative*).

2.2. Exhaustive Listing of Valency Information

As mentioned above, traditionally, in valency dictionaries and NLP valency lexica exhaustive lists of subcat frames are provided for each lemma.

The amount and variation of valency information differs largely from lexeme to lexeme. Thus, the valency of the

English verb **EAT1** and its French equivalent **MANGER1** consists of a single thematic role frame and a single subcat frame (the number attached to a citation form denotes here and henceforth a specific reading, or sense, of this form):

AGENT PATIENT/OBJECT
NP NP

and so does their German equivalent:

AGENT PATIENT/OBJECT
NP_{nom} NP_{acc}

In contrast, the German verb **LIEFERN** "deliver" possesses, among others, the following valency patterns:

NP_{nom} NP_{acc} PP [*aus* NP_{dat}]
NP_{nom} NP_{acc} PP [*von* NP_{dat}]
NP_{nom} NP_{acc} PP [*zu* NP_{dat}]
NP_{nom} NP_{acc} PP [*an* NP_{acc}]

As pointed out above, an exhaustive listing of valency information is not appropriate in the case of regular valency - especially in cases when a number of lemmas shares a considerable number of thematic role frames and subcategorization frames. In the remainder of this section, we analyse the strategies that allow for a more appropriate representation - strategies that imply inheritance. Such an analysis can be carried out either from the "technical" angle or from the linguistic angle. The following subsection focuses on the former, and the next one on the latter.

2.3. Subbpatern "Concatenation" Inheritance

A quick glance at an exhaustive list of valency for a (even small) number of lemmas reveals that lemmas much more often share valency with respect to selected individual actants rather than with respect to all actants. For instance, nearly all lemmas in a Russian lexicon would subcategorize for **NP_{nom}** with respect to the first actant, less would subcategorize for **NP_{acc}** with respect to the second actant, and still less would subcategorize for **za $NP_{instr(umental)}$** with respect to the third actant. This suggests to establish a hierarchical structure in which parts of valency patterns rather than whole patterns are inherited actant-wise. The inherited parts are then concatenated to a complete pattern. Thus, **NP_{nom}** for the first actant is inherited by all classes of of verbs. The class of transitive verbs adds then the specification of the subcategorization information of the second actant, i.e., **NP_{acc}**. This approach is pursued e.g., by (Kilgarriff 93). However, while seemingly attractive because it reduces the redundancy of information, it turns out to be problematic in the case of a relatively large number of detailed valency patterns.

2.4. Complete Pattern Inheritance

In the second approach, individual valency patterns are inherited as a whole. For instance, in German, the subcategorization frame **NP_{nom} NP_{acc}** is inherited by **FEIERN** '[to] celebrate', **KALKULIEREN** '[to] calculate', **VERWIRKLICHEN** '[to] realize', etc., which belong to the same class in the semantic (mental) field and to all other verbs in the same field whose members possess this pattern. In this approach which has been adopted for the German part of GREG, the verbal material was first grouped with respect to a number of semantic fields. Then, a syntactico-semantic classification was carried out for each of the fields.

3. Terminology and Resources Used in GREG

Let us, before we dwell into the details on the framework adapted in GREG, introduce the conventions for the notation of valency information, the repertoire of thematic roles and subcategorization patterns, and the resources we drew upon in the GREG project.

A considerable amount of work has already been done on the compilation of valency resources. Especially the automatic acquisition of subcategorization patterns has been popular during the last decade (Brent 1993, Manning1993, Wauschkuhn 1999, Korhonen at al. 2006, Korhonen at al. 2009). Therefore, we did not start from the scratch. For English and German already subcategorization lexica were available for use in GREG. These were the IMSLex (Lezius et al. 2000) and LTAG (Joshi et al. 1975).

- The source material in GREG has been compiled starting from the list of the most common 1000 verbs in Georgian.
- The English, German and Russian parts of lexicon have been obtained by translating the Georgian originals and

adding the most common 300 verbs of the respective languages.

- The major part of the German subcategorization frames stem from the lexicon provided by the IMS, University of Stuttgart (Lezius et al. 2000).
- The English language data have been extracted from LTAG (Joshi et al. 1975) and a methodological approach partly from PolyLex, a trilingual lexicon developed for English, German and Dutch (Cahill & Gazdar 1999) .
- For Georgian and Russian, subcat lexica have been compiled.
- No thematic role lexica were available for any of the languages involved.

The set of thematic roles used in GREG to describe semantic valency have been compiled from the lists worked out in Systemic Linguistics (Halliday 1985), Frame Semantics (Fillmore 1982), and Cognitive Linguistics (Chafe 1970, Anderson 1971).

A: **Agent** (or Actor in Systemic Linguistics)
AD: Addressee
AT: Attribute
A/P: Agent/Patient
BL: Human body location
BN: Beneficiary
C: Cause
CA: Carrier
E: Existent
G: Goal
I: Instrument (something /someone by means of which/whom the process is carried out)
L: Location (including Source and Destination)
O: Object
P: Patient
PH: Phenomenon
SN: Senser
SG: Saying
SR: Sayer
T: Token
TH: Theme (something or someone involved in the process, but not directly affected)
V: Value

In some cases, more detailed names (such as “Source” or “Destination” instead of the more general “Location”) are used.

4. The GREG Framework

The following four features are most characteristic of the GREG framework:

1) lemma per lexeme,
2) semantically determined argument structure (and thus the number and kind of thematic roles and subcategorization realizations) of a lexeme,
3) exhaustive coverage of all possible valency patterns of a lexeme,
4) distinction between three levels of representation,
5) hierarchical structure that is based on inheritance.

Unlike many traditional valency dictionaries and the majority of valency lexica for NLP, we advocate a “lemma per lexeme” structure. In other words, for each sense of a word (i.e. lexeme), a separate lemma is introduced. This is in line with our assumption that

(i) the GREG resources should be application neutral and thus as discriminatory as possible (for instance, for text generation a discrimination of senses and the separate representation of the realization information with respect to each sense is essential).

(ii) different senses possess different thematic role patterns and there is a correlation between semantic and syntactic valency in the sense that thematic role patterns choose for corresponding subcategorization.

Thus, we identify for the Georgian verb,

შედგომა
[Šedgoma]

at least 5 different senses, which are represented by the following subcategorization frames:

1. NP_{nom} NP_{dat}
A O
იგი საქმეს შეუდგა
[igi sak’mes Šeudga]
(lit. “He get started doing the job”)

2. NP_{nom} NP_{dat} PP [$NP_{dat\text{-}post}$]
A P BL

იგი	მას	მხარში	შეუდგა
[igi	mas	mxarši	šeudga]

(lit. “He backed his shoulder”)

3. NP_{nom}

E

კრება შედგა

[ķreba šedga]

(lit. “A meeting took place”).

4. NP_{nom} NP_{abl}

A/P G

იგი ბერად შედგა

[igi berad šedga]

(lit. “He got as a monk [to cloister]”)

5. NP_{nom} NP_{ins} NP_{dat}

A P I

ფარაონი დამარცხებულ მტერს ფეხით შედგა

[p’araoni damarcxebul mṭers p’exit’ šedga]

(lit. “The Pharaoh stepped by his feet a defeated enemy”)

5. Levels of Representation

There are three levels for the GREG lexicon representation:

1. “ground” lexical entries
2. lexical descriptions
3. delivery representation

The ‘ground lexical entries’ are what the GREG lexicon describes. Given that GREG aims to provide a lexicon that is useful for language engineering, the form of the lexical entries are one which ‘customers’ for the GREG lexicon are likely to find useful. LE users want to use the GREG lexicon for analysis and generation, so the lexical entries should be of a form that is suited to the dominant paradigms for parsing and/or generation. The GREG lexicon specifies these ground lexical entries by providing lexical description of them, or more precisely a single description of the entire lexicon.

The formalism for describing the GREG lexicon is DATR (Evans et al. 2003). DATR is a knowledge description language explicitly designed for the concise and elegant description of lexicons. (Evans & Gazdar 1996) present the language and demonstrate its suitability for the task, also comparing it to the available alternatives. By using DATR, GREG has the resources of a powerful description language and is able to benefit from a substantial existing body of work on multilingual lexical architectures (Cahill & Gazdar 1999) and valency lexicons (Evans et al. 1995), (Carroll et al. 1998), (Smets & Evans 1999). It provides a framework for describing the GREG lexicon as an inheritance network with overridable default, and supports lexical rules and other indirectly specified lexical relationships.

In GREG (and indeed in general) we consider it methodologically important to make a sharp distinction between descriptions and objects described (that is level 2 and 1 formalism, respectively). Whereas the choice of formalism for (1) is dictated by considerations external to GREG, the choice of (2) is ideally transparent to the users of the GREG lexicon, and is dictated by the scientific goals of the project: the formalism selected is best suited to stating generalisations over Multilingual Valency Lexicon entries.

The distinction between (1) and (2) can be thought of as follows: (2) provides a concise and maintainable way of describing a lexicon expressed in (1); in principle (2) provides sufficient information to fully expand the GREG lexicon to the set of ground lexical entries, at which point knowledge of (2) would not be required to use them.

Level (3) is the formalism for web delivery of lexical entries. Lexicons are large resources, and may be held and maintained at different locations to those where entries are used. The web currently uses HTML as its lingua franca, but in the near future, web browsers will be equipped to operate with XML, that is more powerful than HTML, but more constrained than the somewhat unwieldy full SGML language. XML has the advantage of being a practical and supported data interchange language for any LE application and being directly deliverable over the GREG lexicon.

There is also another advantage of using XML in GREG. Independently of this project, UoB have developed a version of DATR with XML syntax, and are currently developing an XML-based DATR server. This means that the entire GREG lexicon, both level 2 description and level 1 ground lexical entries, can be delivered in the same underlying representation, and made directly and dynamically available over the web. In addition, this architecture allows the final formatting of the output lexical entries to also be controlled via DATR embedded in XML, making it easy to support different formatting conventions.

Consider, for illustration, "a printable" sample of GREG lexicon entry for a lemma ***address*** in figures 1, 2 and 3.

citation: **address**
key: 2
gloss: The employees addressed their complaints to the court
sign
syntax
category: **verb**
subcat
first
syntax
category: **np**
case: **nom**
semantics = <sayer>
rest
first
syntax
category: **np**
case: **acc**
semantics = <saying>
rest
first
syntax
category: **pp**
form: **locative**
semantics = <*addressee*>
semantics
predicate: **address**
thematic
addressee
sayer
saying

Figure 1.

citation: **გაგზავნა**
key: **2**
gloss: **მან წერილი მის მისამართზე გაგზავნა**
sign
syntax
category: **verb**
subcat
first
syntax
category: **np**
case: **nom/erg/dat**
semantics = <*agent*>
rest
first
syntax
category: **np**
case: **dat/nom**
semantics = <*patient*>
rest
first
syntax
category: **pp**
post_p: - **ზე**
semantics = <*location*>
semantics
predicate: **address**
thematic
agent
location
patient

Figure 2.

citation: **адресовать**
key: 2
gloss: Папа Римский адресовал всем верующим мира свое послание
sign
syntax
category: **verb**
subcat
first
syntax
category: **np**
case: **nom**
semantics = <sayer>
rest
first
syntax
category: **np**
case: **acc**
semantics = <addressee>
rest
first
syntax
category: **pp**
form: **locative**
semantics = <*saying*>
semantics
predicate: **address**
thematic
addressee
sayer
saying

Figure 3.

6. Conclusions

The work in the GREG Project demonstrated that detailed valency patterns are highly language specific. This suggests that an optimal multilingual representation will not necessarily be optimal with respect to the individual languages involved – especially when these languages belong to different language families as Georgian, Russian and English/German. As the practice showed, a considerable redundancy is still encountered and a certain percentage of collected subcategirization frames need different classifications. Duo to the mentioned observation, future work will be dedicated to the investigation to what extent monolingual valency information can be generalized, while retaining the ability to represent true lexical idiosyncrasies between the languages involved in the project. Finally, the GREG lexicon with 1258 lemmata is still relatively small and for broad coverage NLP, a considerably larger lexicon is needed. Therefore, another

important part of future work will consist in enlarging the lexicon.

Acknowledgements

I would like to thank the EC for supporting the work described in this paper which has been funded under the contract number INTAS Georgia '97, 1921. The partners involved in the GREG project, besides the State University of Tbilisi of Georgia, were: University of Stuttgart / Institute of Informatics (Leo Wanner, Stefan Klatt), University of Brighton/Information Technology Research Institute (Adam Kilgarriff, Roger Evans, and Lynne Cahill). I would like to thank the partners from University of Stuttgart and University of Brighton who have developed the GREG lexicon formal framework and contributed to the lexicon compilation. My special thanks are due to Dr. Leo Wanner for his invaluable efforts that made possible the appearance of this contribution.

References

Anderson, John M. (1971). The Grammar of Case: Towards Localistic Theory. Cambridge: Cambridge University Press.

Baker, C.F., C.J. Fillmore & J.B. Lowe. (1998). The Berkeley FrameNet Project. In: *Proceedings of COLING/ACL 1998*, Montreal.

Brent, M. (1993). From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. In: *Computational Linguistics*. 19(3) pp. 243:262.

Cahill, L.J. & G. Gazdar. (1999). The polylex architecture: multilingual lexicons for related languages. In: *Traitement automatique des langues*, 38 (1).

Carroll, J., Nicolov N., Shaumyan O., Smets M., and D. Weir. (1998). The lexsys project. In *Proceedings of the Fourth International Workshop on Tree Adjoining Grammar and Related Frameworks*, Philadelphia.

Chafe, W.L. (1970). Meaning and the structure of language. University of Chicago Press, Chicago.

Evans, R., Gazdar, G., and D. Weir. (1995). Encoding lexicalized tree adjoining grammars with a nonmonotonic inheritance hierarchy. In *ACL95*, Morgan Kaufmann, San Francisco, pp. 77-84.

Evans, R. & G. Gazdar. (1996). DATR: A Language for Lexical Knowledge Representation. In *Computational Linguistics*, 22(2), pp. 167-216.

Evans, R. et al. (2003). *The GREG Framework for Multilingual Valency Lexicons*, Technical Report, ITRI, University of Brighton, Brighton, U.K.

Fillmore, C.J. (1982). Frame Semantics. In *Proceedings of the Conference Linguistics in the Morning Calm.* Hanshin Publishing Co., Seoul, pp. 111-137.

Joshi, A.K., et al. (1998). Tree Adjunct Grammars. In *Journal of Computer and System Sciences* 10.

Halliday, M.A.K. (1985). Introduction to Functional Grammar. Edward Arnold, London.

Kilgarriff, A. (1993). Inheriting verb alternations. In *Proceedings of the 6th European ACL Meeting*, pp. 213-221.

Korhonen, A., Krymolowski, Y., Briscoe, E. J. (2006). A Large Subcategorization Lexicon for Natural Language Processing Applications. In: *Proceedings of the 5th LREC*, Genova, Italy.

Korhonen, A., Preiss, J., Briscoe, T. (2009). A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In: *Proceedings of the 45th Annual Meeting of the ACL*, pp. 912–919.

Lezius, W., S. Dipper & A. Fitschen, (2000). IMSLex – Representing morphological and syntactical information in a relational database. In: *Proceedings of the 9th EURALEX International Congress*, Stuttgart, Germany, pp. 133-139.

Manning, Ch., D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In: *Proceedings of the 31st ACL*, pp. 235-242.

Smets M. & R. Evans. (1998). A compact encoding of a dtg grammar, in: *Proceeding of the Fourth International Workshop on Tree Adjoining Grammar and Related Framework*, Philadelphia, pp. 164-167.

Wauschkuhn O. (1999). Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. Shaker Verlag, Aachen, Germany.

Sense Disambiguation—"Ambiguous Sensation"?
—Evaluating Sense Inventories for verbal WSD in Hungarian—

Judit Kuti, Enikő Héja, Bálint Sass
Research Group for Language Technology, Research Institute for Linguistics
of the Hungarian Academy of Sciences
33. Benczúr str., 1068 Budapest
E-mail: kutij@nytud.hu, eheja@nytud.hu, sass.balint@nytud.hu

Abstract

The present case study is a first attempt to evaluate the applicability of three resources used as sense inventories in machine-performed WSD in Hungarian. For this purpose we conducted an experiment focusing on inter-annotator agreement (ITA) among human annotators relying on these databases when determining verb senses in context. The chosen resources, one of which is language-independent in its construction method, represent three points in the spectrum ranging from introspection-based to distribution-based databases. Our goal was on the one hand to see whether a reliable ceiling for machine-performed WSD (in terms of an acceptably high ITA-value) can be obtained using any of the databases, on the other hand to test Véronis' claim that the distribution-based construction of sense inventories proves to be more consistent and thus, more reliable than the introspection-based one. Our results show that none of the available databases for Hungarian can in its present stage form the basis of an ITA-value that could serve as ceiling for machine-performed WSD. Our results do not confirm Véronis' claim (but do not refute it, either), which might be due to the restricted capability of our distribution-based database to handle meanings, since it is not specifically designed for WSD-purposes, and needs further targeted improvement.

1. The challenge

One of the central tasks of language technology is to provide usable systems for word sense disambiguation (in the following WSD). WSD is an inevitable task in several HLT applications, such as machine translation, information extraction and information retrieval. We can divide the task of WSD into two basic steps:[1] (1) choosing a sense inventory, (2) assigning the meanings listed in the sense inventory to the word form in question, according to some algorithm (Ide & Véronis, 1998). WSD-related research usually focuses on the latter one: it looks at what algorithms may be used for obtaining the best result using an existing sense inventory (Latent Semantic Analysis (LSA), Hyperspace Analogue to Language (HAL)).

As opposed to this, choosing the right sense inventory and checking its quality receives minimal attention (all-words WSD tasks typically concentrate on finding the best algorithm to an existing database – see e.g. the SemEval contests 2007 or 2010).

The most various subtasks involving WSD (target word disambiguation, automatic keyword extraction, labelling thematic roles etc.) rely to a large percentage on some version of the WordNet, while the use of other sense inventories (FrameNet, VerbNet) is largely neglected. More than half of the WSD test corpora used in the Senseval competitions were annotated with some version of WordNet. Although it is well-known that these databases were not specifically created for the purpose of WSD, the organising principles of databases necessary for WSD are far less discussed in the related literature than the plethora of potential algorithms. (It is probably not by chance that the subtitle of Agirre & Edmonds' book *Word Sense Disambiguation* is "Algorithms and Applications".)

Neither disambiguating word senses as a complex task nor determining (verb) senses as such in an exact manner can be regarded as a resolved problem:[2] "Wordsense tagging is one of the hardest annotation tasks."[3] The sense distinctions in dictionaries often differ as to where they draw boundaries of senses, and are often too fine-grained for annotators to relate to (sometimes even for fellow lexicographers). This problem has been highlighted by the SENSEVAL initiatives, in particular with respect to WordNet. It was in this framework that Véronis (2003) tried to show the problematic nature of the intuition-based meaning-definition and the inapplicability of enumerative lexicons in WSD through two experiments. In the first experiment he showed that the inter-annotator agreement was relatively low even regarding the question whether a certain word form was monosemous or polysemous (in the case of verbs the ITA-value was 0.37 – using Cohen's κ as coefficient. In the second experiment 3724 occurrences of 60 words had to be assigned a sense from the *Petit Larousse* explanatory dictionary, which fitted the given context. The inter-annotator agreement was relatively low, in this case, too (0.41 in the case of verbs) – which underpins the difficulty of the task.

Two main types of solution have been proposed to this problem: one of these (Kilgariff, 1999) proposes sticking to professional lexicographers and arbitration, while the other the clustering together of dictionary senses to

[1] Naturally, the above two steps do not apply to unsupervised WSD methods.

[2] See: "... explicit WSD has not yet been convincingly demonstrated to have a significant positive effect on any application." Agirre & Edmonds (2007), Introduction.

[3] The direct quote as well as the subsequent train of thoughts is taken from Artstein & Poesio, 2008.

coarser-grained units (see Buitelaar, 1998; Bruce & Wiebe, 1998; Palmer, Dang and Fellbaum, 2007) that naïve annotators can relate to more easily, as well. However, none of the two methods avoid relying on intuition – even if lexicographers' intuition – at some point or another. Véronis (2003) claims that it is the inconsistencies of the editing process of such lexicons due to the heavy reliance on intuition that is responsible for this relative unresolved nature of the WSD challenge. He claims that the only way to avoid inconsistencies is to rely primarily on observable distributional phenomena when distinguishing between senses. Accordingly, the applicability of enumerative lexicons for purposes of WSD is questionable, as well.

In our case study presented in the following we carried out a similar experiment to Véronis' above mentioned second experiment, for Hungarian verbs. The motivation for the work was on the one hand to perform a first evaluation of three available sense inventories for Hungarian in WSD. We looked at whether a reliable ceiling for machine-performed WSD can be obtained using any of the sense inventories, in terms of an acceptably high inter-annotator agreement. On the other hand, through this, we wanted to test Véronis' claim regarding the primacy of the distribution-based construction method of sense inventories over the introspection-based one.

2. The resources

The resources were meant to represent three points in the spectrum ranging from introspection-based to distribution-based databases.

The human intuition-based end of the scale is represented by the Hungarian Explanatory Dictionary (henceforward HED), which is the official reference work as a Hungarian monolingual dictionary, with approx. 75000 entries, including collocations and idioms. The HED organizes its entries into main senses and their sub-senses. The main senses may be regarded as sense clusters that were decided on introspectively by the lexicographers compiling the dictionary. When carrying out our experiment we decided to take advantage of this construction and take both the main sense distinctions and the sub-senses into account (see Section 4.).

The other two resources used in the case study do not only exist for Hungarian, but are either readily available in most European languages (wordnets) or are language-independent as to their construction method.

The Frequency Dictionary of Verb Phrase Constructions (see Sass & Pajzs, 2009) represents the distribution-based end of the imaginary scale between introspection-based and distribution-based databases, and is the product of an ongoing research on automatic collection of verb + noun phrase constructions of different specificity – ranging from verb subcategorization frames to complex verbs and idioms (verb + case suffix / postposition + most frequent lemmas). The algorithm generating the database selects constructions on a strictly distributional basis and it also determines whether an argument is free or bound by its distribution. The database does not contain definitions (and can be regarded as a 'meaningless dictionary' in the sense of Janssen (2008)), but for every construction it gives corpus-derived examples, which do serve to indicate meaning.

It is important to note, however, that the primary purpose of the algorithm extracting the verb phrase constructions from the corpus was not to distinguish between verb senses, but to arrive at typically occurring verbal constructions. Nevertheless, a partial aim of our case study was to see to what extent these syntactically delineateable verbal constructions fall together with distinguishable verbal meanings.

Furthermore, it is important that since the algorithm is language-independent, and has been demonstrated to work for different European languages (see e.g. Sass, 2009 for Danish), among others, Serbian[4], it could provide an important alternative to man-made sense inventories not only in Hungarian, but other languages.

An entry in FDVC consists of the following information: the verb in question, its frequency in the Hungarian National Corpus (see Váradi, 2002) (in square brackets in the Figure 1., below), and its automatically collected phrasal construction together with their frequencies. The lexically bound elements are indicated with boldface, the case suffixes in square brackets. Since the following example serves only illustration, it is given in English translation (the approximate English translation of idioms is given in parenthesis):

EMEL [18635]
emel{ACC} [1745]
emel **ár**{ACC} [164]
emel {SUB} {ACC} [521]
emel **vád**{ACC} {ellen} [359]

Figure 1: Illustration of an entry in the FDVC

LIFT
lift{ACC}
lift **price**{ACC} (raise the price of sg)
lift{SUB}{ACC} (lift sg onto sg)
lift **accusation** [against] (charge sy with sg / accuse sy)

Figure 2: Translation of the entry in Figure 1.

The Hungarian WordNet (henceforward HuWN) (see Miháltz et al., 2008) is a lexical database developed between 2005 and 2007, modeled partly upon the Princeton WordNet 2.0 (PWN) for English, but also on other wordnets developed for European languages (the fact that extensive verb frame information is given in the verbal synsets links the HuWN to its Czech counterpart). The basic unit of HuWN, as of all wordnets, is a concept (called synset) and not that of traditional dictionaries, i.e.

[4] Demonstrated on 26.02.2010 at a meeting between the Institute for Slavic Studies of the Eötvös Loránd University and the Research Group for Linguistics of the Hungarian Academy of Sciences.

a word / lexeme. Apart from encoding the semantic relationships of PWN for about 3000 verbal concepts, the Hungarian verbal WordNet helps to indicate the aspectual characteristics of verbs, too, by introducing some new relations. It is also important to note that when deciding on what verb senses should be incorporated into the Hungarian verbal WordNet, automatically extracted information about argument constructions was taken into account, as well. Therefore, the sense distinctions in HuWN are neither based on introspection alone nor on the sense distinctions in PWN.[5]. Accordingly, from a methodological point of view the verbal HuWN can be placed between the entirely introspection-based HED and the automatically obtained FDVC.

So far none of the above mentioned resources has been evaluated as for its applicability in word sense annotation. Nonetheless, the HED was used by Vincze, Szarvas and their colleagues (see Vincze et al., 2008; Szarvas et al., 2007) to construct a WSD test corpus for Hungarian. The choice of the explanatory dictionary might seem surprising, since WordNet "is becoming a de facto standard for sense annotation" (Véronis, 2003). At the time of their work, however, the HuWN was still in its construction phase, which led to the choice of the HED as a sense inventory for this work.

3. The case study

3.1. The annotation task

In our case study, based on Véronis' second experiment we focused solely on verbal meaning. We wanted to look into how strong an agreement may be achieved among human annotators, and into whether the type of sense inventory used makes a difference to the degree of inter-annotator agreement. We also hoped to find out more about what a sense inventory developed specifically for sense disambiguation should be like.

We had 30 occurrences of 15 verbs in the context of one sentence annotated by 5-5 annotators for senses from the three respective databases. Apart from choosing the appropriate category labels defined by the given database, the annotators could choose the category 'no matching sense' and 'I don't know.' We explicitly asked the annotators to choose one single category label, which best fitted the verb in the given context, if possible. The distinction between the 'I don't know' and 'no matching sense' labels proved to be relevant, since many annotators who answered 'I don't know' remarked that they had several possible senses in mind but were unable to chose, probably because of overlaps between the senses. This was not the case with the 'no matching sense' labels, which indicated a clear lack of match.

The annotators were either first-year students of Applied Linguistics of Eötvös Loránd University (ELTE) in Budapest, or volunteers, interested in the task. Every annotator was between 20 and 30 years of age. The annotators did not receive any financial reward for their contribution. However, the annotation work was accepted as the completion of 30% of the class assignment in a computational linguistics class at ELTE. There was no time limit set for the annotation task, the average time needed to accomplish the task was 4 hours. The annotators received a written guideline as to how they should proceed in the task, and containing information about the database they were going to use. Every annotator had to work on his / her own during the task, but when it was necessary, we did, of course, answer their questions. It was made explicit that there was no 'absolutely good' answer, only their intuition counted. Nevertheless, the annotators had the possibility of making a remark to each sentence they were annotating.

3.2. The data

Since our goal was to focus on sense tagging highly polysemous words, we chose verbs to be the subject of our investigation.[6] It is widely acknowledged that verbs are in general more polysemous than nouns and adjectives (e.g. Palmer, Dang and Fellbaum, 2007). The average number of senses a verb has in the Hungarian Explanatory Dictionary is 1,87, this number for nouns is 1,36 and 1,43 for adjectives. Within verbs we decided to treat a relatively small amount of words, since a larger scale experiment would have exceeded the practical limits of our current initiative.

We considered three main factors when choosing the verbs for our case study: (1) Frequency of the selected verbs: The selected verbs had to be among the 500 most frequent verbs in the Hungarian National Corpus, in order to have ample different contexts for our test sentences. (2) Collocability of the selected verbs: we selected verbs that had at least eight constructions listed in the FDVC with a frequency of at least 100 in the HNC. (3) In order for the results to be comparable it was important that all the selected verbs be included in all three databases. Since the Hungarian verbal WordNet is the smallest one among the three databases (approx. 3000 verbal concepts), this imposed a practical constraint upon our choice.

Among the 15 verbs chosen there are 2 that have verbal prefixes ('fel | tesz' (put (sg. on sg.)) and 'meg | old' (untie)), and 13 that do not, i.e., they are verb stems. Verbs with prefixes are in general underrepresented within the 500 most frequent verbs. The reason for this is probably the following: verbal prefixes in Hungarian are often form-identical to adverbials of place and can separate from the verb stem under certain syntactic conditions. In such cases the morphological analyser identifies only the verb stem in a sentence, which results in the apparent dominance of "prefixless" verb stems among the most frequent verbs automatically collected.

[5] For the methodological considerations behind building the database see Kuti et al., 2007.

[6] Verbs have long been, partly for this reason, in the focus of interest of the authors, anyway.

3.3. The ITA-measure used

We determined inter-annotator agreement (ITA) using Fleiss's multi π. In the following we present a detailed description of the chosen coefficient based on Artstein and Poesio (2008).

One simple approach to determine ITA is the percentage agreement or observed agreement (A_o) which is the number of items on which the coders agree divided by the total number of items.

$$\text{Percentage agreement: } A_o = \frac{1}{i}\sum_{i \in I} agr_i$$

However, this measure of ITA is not corrected for chance agreement, that is, it does not account for cases where agreement is due to chance. One of the two factors that influences chance agreement is the number of categories used in the annotation task: the fewer categories are used to classify a certain phenomenon the higher agreement by chance might be expected. Since various sense-inventories contain diverse divisions of senses, our measure has to handle such cases to be able to compare the usefulness of sense-inventories above chance.

The calculation of the chance-corrected inter-annotator agreement coefficients start by giving some estimations to the chance agreement (A_e). The coefficient is calculated on the basis of this value and on the basis of the observed agreement (same as percentage agreement defined above) as follows:

$$\frac{A_o - A_e}{1 - A_e}$$

$1 - A_e$ measures how much agreement over chance agreement is attainable, whereas $A_o - A_e$ tells us how much agreement over chance agreement was actually found.

It therefore shows where the observed agreement value is to be found along the continuum specified by the expected agreement (0) and unanimous agreement (1). In cases where agreement is lower than expected, this measurement unit can take a negative value. The closest the obtained value is to 1, the higher the possibility that the agreement between the annotators is not by chance.

One chance-corrected coefficient of ITA is Scott's π (1955), which measures the agreement between two annotators. Scott's π (1955) assumes that if coders were operating by chance alone, their assignment would be guided by the distribution of items among categories in the actual world, thus yielding the same distribution for each coder. Thus, as opposed to Cohen's κ which presumes separate distributions for each of the coders, A_e does not reflect individual annotator bias. The prior distribution is estimated on the basis of the observed assignments, i.e. the total number of assignments to the categories by both coders divided by the overall number of assignments where n_k stands for the total number of assignments to category k and i for the number of items to be assigned.

$$\text{Estimation of the prior distribution: } \frac{n_k}{2i}$$

Then, given the assumption that coders act independently, expected agreement is determined as follows, where K designates the set of categories:

$$A_e^{\pi} = \frac{1}{(2i)^2}\sum_{k \in K} n_k^2$$

However, being invented for two annotators, Scott's π is not apt to measure agreement among multiple coders. Therefore, we relied on Fleiss's multi-π (1971) throughout our analysis, which is a generalization of Scott's π for multiple coders. The basic idea behind this coefficient is that A_o cannot be thought of as the percentage agreement defined above. This is due to the fact that in the case of multiple annotators neccessarily there will be items on which some coders agree and others disagree. The proposed solution is to compute *pairwise agreement* as the proportion of agreeing judgment pairs and the total number of judgement pairs for that item. The overall A_o will be the mean of the pairwise agreement for all items. Here i stands for the number of items (30 in our case), c for the number of coders (5), and n_{ik} for the number of times an item is classified in category k. I denotes the set of items while K denotes the set of categories (the senses in the sense inventory).

$$A_o = \frac{1}{ic(c-1)}\sum_{i \in I}\sum_{k \in K} n_{ik}(n_{ik} - 1)$$

In the case of multiple coders A_e i.e. the agreement by chance might be concieved of as the probability that two arbitrary coders would make the same judgement for a particular item by chance. Holding the same presuppositions about the distribution of the judgements as Scott, A_e is calculated in the same way as in the two coder case, except for the fact that instead of 2 coders c coders make the assignments, that is c assignments need to be considered, when calculating the mean.

$$A_e^{\pi} = \frac{1}{(ci)^2}\sum_{k \in K} n_k^2$$

An additional advantage of Fleiss's multi π is that it is insensitive to categories that were never selected by any of the annotators, therefore the results do not reflect how many categories the annotators could originally choose from.

4. Results

Table 1. below shows how many senses the respective verbs had in the three dictionaries – i.e. how many categories the annotators could chose from. Although this value is not reflected in Fleiss's multi π, it might be interesting to draw some tentative conclusions on the basis of these data later on. The additional column, "HED-clusters", refers to the reading of the HED which

only takes into consideration the basic sense-clusters of the given verbs.

The selected verbs	HED	HED-clusters	HuWN	FDVC
emel / lift	13	5	10	16
feltesz / put (sg. on sg.)	14	7	7	8
fizet / pay	12	5	1	23
használ / use	8	4	2	22
köt / bind	29	12	21	19
lép / step	12	7	11	31
megold / untie	6	4	2	12
mutat / show	13	5	4	27
okoz / cause	2	2	3	26
rendelkezik / order	6	4	3	15
segít / help	7	5	4	19
szolgál / serve	15	7	8	16
tárgyal / negotiate	3	3	2	16
választ / choose	6	4	2	24
vállal / undertake	6	3	3	26

Table 1: Number of senses for the selected verbs in the respective sense inventories

The selected verbs	HED	HuWN	FDVC
emel / lift	0.450	0.753	0.170
feltesz / put (sg. on sg.)	0.493	0.693	0.265
fizet / pay	0.157	0.61	0.259
használ / use	0.210	0.954	0.336
köt / bind	0.449	0.637	0.237
lép / step	0.346	0.595	0.443
megold / untie	0.137	0.197	0.255
mutat / show	0.187	0.153	0.284
okoz / cause	0	0.59	0.286
rendelkezik / order	0.195	0.469	0.471
segít / help	0.112	0.371	0.434
szolgál / serve	0.279	0.516	0.548
tárgyal / negotiate	0.840	0.543	0.407
választ / choose	0.452	0.935	0.444
vállal / undertake	0.207	0.311	0.275
average Fleiss' multi π	**0.300**	**0.483**	**0.340**

Table 2: Average value of Fleiss' multi π for the three databases[7]

Table 2. shows the obtained ITA-values specified according to the selected 15 verbs, and according to the databases used as sense inventory.

Table 3. focuses on the different ITA-values obtained when considering all sense distinctions in the HED (the default choise) as opposed to only the main sense-clusters. The results shown in all tables were obtained by regarding the "no matching sense" answers (which can be regarded as equally relevant and valid as specified matches) as independent values. The "I don't know" answers (the ratio of these was between 2-6%) were dealt with similarly.

The selected verbs	HED	HED-clus.
emel / lift	0.450	0.848
feltesz / put (sg. on sg.)	0.493	0.745
fizet / pay	0.157	0.278
használ / use	0.210	0.611
köt / bind	0.449	0.535
lép / step	0.346	0.601
megold / untie	0.137	0.449
mutat / show	0.187	0.365
okoz / cause	0	0
rendelkezik / order	0.195	0.474
segít / help	0.112	0.173
szolgál / serve	0.279	0.509
tárgyal / negotiate	0.840	0.840
választ / choose	0.452	0.713
vállal / undertake	0.207	0.623
average Fleiss' multi π	**0.300**	**0.517**

Table 3. Comparing the two results from the two readings of HED

5. Interpretation of the results

On the basis of the above data the following conclusions can be drawn: the order of the inter-annotator agreement value is comparable to Véronis' results, in the case of all databases. When taking the usually accepted threshold value of 0.7-0.8 as a basis of comparison (see Arnstein & Poesio, 2008) we can see that none of the ITA-values obtained is high enough for any of the databases to form the basis of annotating a reliable WSD test corpus.

The type of the sense inventory used largely influences the ITA-value. In its present form it was the Hungarian verbal WordNet on the basis of which the best ITA-value was obtained, with the exception of the sense-cluster oriented evaluation of the Hungarian Explanatory Dictionary.

A comparison of Table 1. and Table 3. indicates that whereas the fine-grainedness of the sense distinctions highly influences the ITA-value (compare the two values obtained by two evaluation perspectives on the HED results), the polisemy of the verb alone does not. For example the verb *köt* (bind) is one of the most polysemous verbs among the selected ones, has one of the highest ITA-values on the basis of the HuWN, and the highest ITA-values on the basis of FDVC were obtained for verbs with above 15 constructions as categories.

Véronis' claim that the distribution-based construction method of sense inventories produces more consistent and thus, (probably both for humans and machines) more reliable databases than the introspection-based one, could, accordingly, so far not be confirmed on the basis of the current versions of the available Hungarian databases, but has not been refuted, either: the purely distribution-based FDVC can, in the present stage of development not replace the existing (at least partially intuition-based) sense inventories. However, by looking at the results presented in Table 2, it is clear, that even in its present

[7] The English translations of the Hungarian verbs are taken from Magay & Országh, 2005. In all cases we took the first translation provided (multiword units excluded).

form (i.e. without any targeted improvement) FDVC can provide an alternative to man-made sense inventories for verbs, as mentioned, not only in Hungarian, but other languages: comparing the ITA-value obtained for the three databases we can see that FDVC ranked second, with 0.340, as opposed to 0.300 for HED.

After the first qualitative analyses we assume that the weaknesses of FDVC in the given experiment can be ascribed to the fact that the FDVC has not been designed to distinguish between word senses, and for the purpose of the present study we have not applied any modifications to the database, either. The open interpretation possibilities of the verbal argument frames listed in the FDVC leave a choice to the annotator: some of them assigned the available constructions to the occurrences on a purely syntactic basis, while others interpreted single lemmas, postpositions or even case suffixes presupposing they represented semantic content. The annotations of the following two test sentences exemplify this:

(1) Ezek az eredmények pedig az érdekképviseletek presztízsét *emelik*.
[These results on the other hand *enhance* the prestige of the representative bodies.]

(2) A kipattanó labdát Makaay négy méterről a teljesen üres kapu fölé *emelte*.
[Makaay *lifted* the bouncing ball above the completely empty gate from four meters' distance.]

In the case of (1) all five annotators assigned a different verb phrase construction to the occurrence of *emel*:

emel **magas**{SUB}	*lift* **high**{SUB}
emel **magas**{SUB}	*lift* **high**{ILL}
emel **ár**{ACC}	*lift* **price**{ACC}
emel {ACC}	*lift*{ACC}
"no matching sense"	"no matching sense"

In the case of (2) three annotators chose the verb phrase construction '*lift* **above** [ACC]', one chose '*lift* [ILL]' and one chose the "no matching sense" option.

6. Conclusion and further work

In our case study presented we carried out a sense tagging experiment similar to the one described in Véronis (2003), for Hungarian verbs. On the one hand we wanted to perform a first evaluation of three available sense inventories for Hungarian in WSD. We looked at whether a reliable ceiling for machine-performed WSD can be obtained using any of the sense inventories, in terms of an acceptably high inter-annotator agreement. On the other hand, through this, we wanted to test Véronis' claim regarding the primacy of the distribution-based construction method of sense inventories over the introspection-based one.

Summarising our findings we can conclude that the order of magnitude of our results is the same as Véronis' results, in the case of all the databases used. Accordingly, the use of the currently available sense inventories for Hungarian need further WSD oriented work if they should be able to provide the basis for a reliable test corpus annotation. Since Véronis' results for adjectives and nouns are much in the range of the same order of magnitude as the results for verbs, we cannot assume to have significant differences in the results obtained for different part-of-speech categories in Hungarian, either. However, further – qualitative and quantitative – research should and could illuminate what criteria a database specifically designed for WSD-purposes should fulfill. Therefore extending the present experiment towards nouns and adjectives could be in order for purposes of collecting more data, in order to be able to carry out such further work.

Another line of further research should focus on corpus-driven rather than corpus-based approaches while constructing sense inventories. However, as the results obtained through the use of the FDVC show, corpus-driven techniques require a solution to the question of how to delineate the units obtained through the observation of distributional phenomena in a way that they reflect meanings.

References

Artstein, R., Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics,* 34(4), pp. 555—596.

Agirre, E., Edmonds, Ph. (2007). *Word sense disambiguation. Algorithms and Applications*. (Text, Speech and Language Technology), Springer-Verlag New York, Inc., Secaucus, NJ.

Bruce, R., Wiebe, J. (1998). Word-sense distinguishability and inter-coder agreement. In : *Proceedings of EMNLP,* pp. 53-60, Granada.

Buitelaar, Paul. 1998. CoreLex : Systematic Polysemy and Underspecification. Ph.D. thesis, Brandeis University, Waltham, MA. 2004.

Fellbaum, Ch. (1998). *WordNet. An Electronic Lexical Database.* Cambridge (MA): MIT Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. In: *Psychological Bulletin*, 76(5), pp. 378–382.

Ide, N., Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. In : *Computational Linguistics,* 24(1), pp. 2-40.

Janssen, M. (2008). Meaningless dictionaries. In: *Proceedings of the XIII EURALEX International Congress*, Barcelona : Institut Universitari de Linguistica Aplicada, Universitet Pompeu Fabra, pp. 409-420.

Kilgarriff, A. (1999). 95% replicability for manual word sense tagging. In : *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics,* Bergen, Norway. pp. 277-278.

Kipper, K. (2005). *A broad-coverage, comprehensive verb lexicon*. Doctoral dissertation at University of Pennsylvania.

Kuti, J., Varasdi, K., Gyarmati, Á., Vajda, P. (2007). Hungarian WordNet and representation of verbal event structure. In: *Acta Cybernetica,* 18(2), pp. 315-328.

Magay, T., Országh L. (2005). *Magyar-Angol Kéziszótár* [A Concise Hungarian-English Dictionary] Budapest: Akadémiai Kiadó.

Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T. (2008). Methods and Results of the Hungarian WordNet Project. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, Ch., Vossen, P. (eds.): *Proceedings of the IVth Global WordNet Conference*, pp. 311-321.

Palmer, M., Dang, H. T., Fellbaum, Ch. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. In: *Natural Language Engineering*, 13(2), pp. 137-163.

Sass, B. (2009). Verb Argument Browser for Danish. In: Jokinen, K., Bick, E. (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics*, NoDaLiDa, Odense, Denmark, pp. 263-266.

Sass, B., Pajzs, J. (2009). FDVC — Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions for Hungarian. In *Abstracts of the eLexicography in the 21st century Conference*, Louvain-la-Neuve, Belgium, pp. 183--186.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. In: *Public Opinion Quarterly*, 19(3), pp. 321–325.

Szarvas Gy., Hatvani, Cs., Szauter, D., Almás, A., Vincze, V., Csirik J. (2007). Magyar jelentés-egyértelműsített korpusz. [Hungarian Sense-disambiguated Corpus.] In: Tanács, A. & Csendes D. (eds.): V. Magyar Számítógépes Nyelvészeti Konferencia (MSzNy 2007) [5th Hungarian Computational Linguistics Conference]. Szeged : Szegedi Tudományegyetem.

Váradi, T. (2002). The Hungarian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC-2002) Las Palmas de Gran Canaria: European Language Resources Association, pp. 385-389.

Véronis, J. (2003). Sense tagging: does it make sense? In Wilson, A., Rayson, P. and McEnery, T. (Ed*.) Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Frankfurt: Peter Lang.

Vincze, V., Szarvas, Gy., Almási, A., Szauter D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J. (2008). Hungarian Word-sense Disambiguated Corpus. In: *Proceedings of 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco.

Towards the Construction of Language Resources for Greek Multiword Expressions: Extraction and Evaluation

Evita Linardaki[◇], **Carlos Ramisch**[♠♡], **Aline Villavicencio**[♡], **Aggeliki Fotopoulou**[◇]

[◇]Institute for Language and Speech Processing, Athens, Greece
[♠]GETALP, Laboratory of Informatics of Grenoble, University of Grenoble, France
[♡]Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
elinardaki@gmail.com, {ceramisch,avillavicencio}@inf.ufrgs.br, afotop@ilsp.gr

Abstract

Multiword Expressions have been posing problems for Natural Language Processing systems for many years. Their automatic identification has, as a result, been in the focus of NLP research for almost two decades now. The advances in the most widely spoken languages like English, French, German, etc. are remarkable and have been extensively documented. This paper presents our work towards the creation of a dictionary of Multiword Expressions for Greek using automatic extraction and human validation. We investigate the use of a knowledge-poor statistical approach based on four association measures. The results obtained by these measures on the Greek Europarl corpus are compared and contrasted with those obtained by the same measures using the web as a corpus. The manual evaluation of the results by Greek native speakers shows that the automatic approach performs well enough to help in the construction of a lexical resource, despite of the difficulty of the task.

1 Introduction

A Multiword Expression (MWE) can be defined as a word combination that presents some syntactic, semantic, pragmatic or statistical idiosyncrasies, i.e. an expression whose interpretation crosses the boundaries between words (Sag et al., 2002). This rough definition covers a very wide range of linguistic phenomena such as idioms (*sweep something under the rug*), compounds (*school bus*), phrasal verbs (*take off*), terminology (*central processing unit*), etc. MWEs are very frequent in languages other than English as well. Some examples indicating the wide range of linguistic structures classified as MWEs in Greek, for example, are: *κάλιο αργά παρά ποτέ* (*better late than ever* — idiom), *πλυντήριο πιάτων* (*washing machine* — compound), *οπτική ίνα* (*optical fiber* — terminology).

Even though native speakers rarely realize how frequently they employ MWEs, they are challenging for foreign language learners, since they are not only arbitrary to some extent, but also numerous and very important to confer naturalness and fluency to the discourse. Indeed, Jackendoff (1997) estimates that MWEs correspond roughly to half of the entries in the lexicon of a native speaker, but the proportion of MWEs in natural language may be even larger in specialized domains or sub-languages (Sag et al., 2002).

The main challenge for Natural Language Processing (NLP) systems is that many MWEs have to be treated as single units, since breaking them up into their parts could cause, for example, unrecoverable overgeneration on a syntactic level or loss of information on a semantic level, both in analysis and generation tasks. MWEs, however, are not easy to identify, mainly because of their flexible and heterogeneous nature. In some cases their internal morphosyntactic structure is fixed (*ad hoc*), while in others it is more flexible (*call_/s/ed somebody up* and *call_/s/ed up somebody*). On a semantic level MWEs range from fairly compositional (*traffic light*) to non-compositional (*kick the bucket*). As a consequence of this variability, MWEs are often at the root of errors like incorrect sentence parses and awkward literal translations.

Table 1 illustrates the importance of MWE treatment in the context of multilingual applications such as cross-lingual information retrieval and Machine Translation (MT). It shows a set of sentence fragments taken from the Greek portion of the Europarl corpus along with an English translation generated by a commercial MT system[1]. The corresponding reference translations from the English portion of the Europarl corpus show that the expected translations of the highlighted MWEs in the source text are clearly not equivalent to the actual output of the system. This example helps us to understand that automatic MWE acquisition without the use of extended linguistic information is no doubt a challenging task for the creation and adaptation of NLP systems and resources to new languages and language pairs.

If we ignore MWE entries, the manual construction of large-coverage dictionaries from scratch, independently of the language, is a very costly task that demands specialized knowledge and, often, a huge amount of time. Therefore, when taking MWEs into account in the context of multilingual lexica and MT systems, the unfeasibility of brute force approaches becomes more and more clear. If building a hand-crafted list containing a large number of MWEs for a given language would be very difficult and expensive, then building *bilingual* or *multilingual* lexica by hand for several pairs of languages becomes inconceivable. Among the underlying problems, one would need to chose an appropriate representation that allows to translate from (potentially flexible) multiple lexical units to (a) single words, (b) mul-

[1]The result of MT was obtained through Systran's online translation service, available at http://www.systranet.com/. The goal of this table is to show the importance of MWEs in multilingual applications. We do not intend to compare Systran with other MT systems or to evaluate its quality. This means that other MT systems could translate these examples correctly, as well as Systran could correctly translate other MWEs in different contexts.

Greek source text	Result of MT	English reference text	Frequency
..., όπως αυτό ορίζεται από την ανθρώπινη **οπτική γωνία**	*..., as this is fixed by the human* ***optical corner***	*..., as seen from the human* ***point of view***	131
Το **ξέπλυμα βρώμικου χρήματος** *αντιπροσωπεύει το 2 έως 5% ...*	*The* ***rinsing of dirty money*** *represents the 2 until 5% ...*	***Money laundering*** *represents between 2 and 5% ...*	21
Για τα **εργοστάσια ατομικής ενέργειας** *η Ευρωπαϊκή Ένωση έχει αναλάβει δράσεις για την υψηλότερη ασφάλεια,...*	*For the* ***factories of individual energy*** *the European Union has undertaken action for the higher safety,...*	***Nuclear power stations*** *in the European Union have the highest safety standards...*	8

Table 1: Example text fragments where MWEs can be at the root of translation problems. The source and reference texts were taken from the Europarl corpus. The last column shows the number of occurrences of the highlighted Greek MWE in the corpus.

tiple lexical units or (c) valid paraphrases[2]. Then, a very large amount of bilingual expert work would be required in order to list a considerable number of MWE pairs. Moreover, we argue that such methodology is not in line with the current trend in the MT field, where empirical or statistical approaches are rapidly taking over standard techniques (Koehn, 2009). Specially in the case of phrase-based statistical MT, linguistic information about MWE units could potentially help to improve the overall quality of translations (Bai et al., 2009; Stymne, 2009).

Thus, NLP researchers have been proposing techniques and tools that aid in the creation and exploitation of monolingual and multilingual resources (Preiss et al., 2007; Messiant et al., 2008) and that help linguists and domain experts to speed up lexicographic work. Nonetheless, when it comes to MWEs, the availability of such tools is still quite limited both in terms of effectiveness and of available languages/language pairs, contrasting with the wide and pervasive nature of MWEs.

Recently, however, intensive research efforts on the automatic identification of MWE have brought considerable advances in techniques, and their performance has been tested on the extraction of MWEs on languages like English and German with good results (Evert and Krenn, 2005; Villavicencio et al., 2007; Ramisch et al., 2008). As a consequence, the construction of wide-coverage MWE resources for languages like English, French, Spanish or German is picking up pace, whereas for languages like Greek, which is the focus of this work, computational approaches for the automatic or semi-automatic construction of such resources are still underexploited. Although some of the approaches employed are language dependent, or use tools or resources that are not widely available, some others are language independent and/or based on shallow processing of large quantities of text.

The main goal of this work is, therefore, to evaluate the effectiveness of a language-independent MWE identification approach for the automatic construction of MWE resources for Greek. We look at the effectiveness of some statistical measures of association and of morphosyntactic patterns for extracting MWE candidates, focusing on nominal cases such as κράτος μέλος (member state). For evaluation we use data from two corpora: the Greek portion of the *Europarl corpus*, henceforth EP, and the *World Wide Web* as a corpus, henceforth WWW. The EP corpus data is used for generating an initial set of candidates, and the WWW for providing a larger corpus for validating them. In order to do that, the corpus was preprocessed, with lemmatization and part-of-speech tagging, and all n-grams found in it following a set of predefined MWE morphosyntactic patterns formed an initial set of MWE candidates. For filtering these candidates we applied a set of statistical association measures using counts collected both from the corpus and from the Web. A subset of the final set of candidates was manually evaluated by Greek native speakers. Based on these judgements, we analyse the precise contribution of the different filters and information sources in terms of the number of correct MWEs retrieved. The results indicate that such methods can indeed be used for extending NLP resources with MWE information, and improving the quality of NLP systems that support Greek.

The remainder of this paper is organized as follows: in section 2 we discuss some related work on the construction of language resources for the Greek language, as well as some language-independent methods for the automatic acquisition of MWEs. In section 3 we analyse the types of morphosyntactic patterns we are interested in and show some examples. In section 4 we describe our experiments using `mwetoolkit` to extract and filter a list of MWE candidates from the Greek portion of the Europarl corpus. Finally, we summarize the main contributions of our work to the creation of language resources for Greek in section 5.

2 Related Work

Automatic MWE processing for English has been an active research area in the last decade. Some of the approaches proposed concentrate on specific types of MWE. Verb-particle constructions, for example, have been extensively analysed and methods were proposed for their identification (Baldwin, 2005) and semantic classification (Ramisch et al., 2008). Analogously, Nakov and Hearst

[2]When the target language does not allow the corresponding construction, for instance, when translating English verb-particle constructions such as *give [something] up* to Greek.

(2005) presented a method to interpret the internal boundaries of (long) compound nouns based on the association strength of its components. Although of widespread use in English, we underline, however, that both compound nouns and verb-particle constructions are not general MWEs that occur in all natural languages, but they are rather specific to the Germanic language family.

Even if work on MWEs in several languages has been reported — e.g. Dias (2003) for Portuguese and Evert and Krenn (2005) for German — work on English still seems to predominate, e.g. Nakov and Hearst (2005), Baldwin (2005) and Ramisch et al. (2008). As not all languages are as resource rich as English, there has been some effort in developing methods for language- and type-independent MWE identification. Evert and Krenn (2005) evaluate a set of statistical Association Measures (AMs) to filter MWE candidates and the results suggest that they are language-independent. Villavicencio et al. (2007) apply some of these measures to a list of type-independent MWE candidates and finds that their performance seems to be better for some types of MWE than others. However, some works also found that the same measure can behave differently for the same kind of construction extracted from two different languages. Seretan (2008) provides a detailed discussion on language- and type-independent AMs and notes that in spite of all the work, a single "best" measure for a given language or type has not yet been determined.

In this context, the Multiword Expression Toolkit, or `mwetoolkit` (Ramisch, 2009), implements language-independent MWE extraction techniques following a standard methodology, which consists of candidate extraction followed by filtering and validation. Originally conceived to extract multiword terminology from specialized corpora, the toolkit can also perform automatic identification of other types of MWEs. It extracts candidates based either on flat n-grams or specific patterns (of surface forms, lemmas, POS tags). Once the list of candidates is output, it is possible to filter them and/or calculate a set of features that range from simple ones, as the number of words, to sophisticated ones like AMs. Since the latter are based on corpus word and n-gram counts, the toolkit provides an indexation tool and integration with Web search engines. Additionally, it is possible to evaluate the resulting MWEs as well as feed them into a machine learning tool that allows the creation of supervised MWE extraction models if annotated data is available.

In the context of MT applications, Stymne (2009) proposes an approach to identify compounds using a dictionary. Then, the MWEs are split/joint when translating from/to languages like Swedish, Danish and German. The results show that the overall quality of the translations generated by a statistical MT system is improved when this extra preprocessing and postprocessing steps are applied. Analogously, Bai et al. (2009) show that the quality of the Cinese–English translations output by a statistical MT system improves when MWE translations are automatically generated using evidence from parallel corpora.

For Greek, in particular, considerable work has been done to study the linguistic properties of MWEs (Fotopoulou, 1993; Moustaki, 1995; Fotopoulou, 1997). However, published results about a purely computational treatment for Greek MWEs are still very limited. One of the few works concerning the extraction of MWEs for Greek is the one of Fotopoulou et al. (2008). Their approach combines grammar rules and statistical measures in an attempt to extract from a 142,000,000-word collection of Greek texts as many nominal MWEs as possible while at the same time assuring consistency of results. The said collection of texts is a combination of the Hellenic National Corpus and the Greek corpus maintained by the Université de Louvain. Once the corpus is tagged and lemmatized, the initial list of candidates is extracted based on a set of predefined part-of-speech patterns. This is then filtered using a set of more specific rules and word lists that identify possible, less likely and impossible MWE combinations. Depending on the type of list a given word belongs to, the candidate can either be rejected or marked to be assigned extra weight in the statistical analysis stage. During this final step the remaining candidates are hierarchically organized based on their log-likelihood scores.

Another approach is that of Michou and Seretan (2009). They describe a Greek version of the *FipsCo* system that is able to extract collocations from corpora. Their method uses a hand-crafted generative parser for Greek built upon the *Fips* framework to analyse the sentences of the Europarl corpus and then extract MWE candidates based on syntactic patterns. The candidates are further filtered according to their association strength through the log-likelihood measure. Their system also allows the potential extraction of bilingual Greek–French multiword expressions when parallel corpora is available.

Despite the methodology similarities, our approach differs from these works not only in the techniques used in each extraction step, but also in its goal: instead of building a hand-crafted specialized deep analysis tool aimed at the identification of Greek MWEs, we use the language-independent `mwetoolkit` to extract shallow MWE candidates and then we evaluate the effectiveness of several filtering measures implemented by the toolkit using both textual corpora and the World Wide Web as a corpus. We believe that a systematic evaluation of this technique is crucial for determining whether it can be used to help creating both mono- and multilingual language resources for Greek, that can be subsequently employed in an NLP application, such as automatic translation.

3 Types of MWEs

Calzolari et al. (2002) define MWEs as *a sequence of words that acts as a single unit at some level of linguistic analysis*, with some of the following characteristics:

1. have reduced syntactic and/or semantic transparency;
2. have reduced compositionality;
3. are more frozen;
4. violate general syntactic rules;
5. have a high degree of lexicalization;
6. have a high degree of conventionality.

The general characteristics of Greek MWEs fall into these categories. They also vary to a great extent in terms of the fixedness of their morphosyntactic structure and of their semantic interpretation, that can be more or less transparent depending on the type of MWE (idioms tend to be less transparent than specialized terms, for example). The decision to investigate nominal MWEs (as opposed to verbal) was largely based on the fact that they are less heterogeneous in nature and can, therefore, be more easily encoded (Mini and Fotopoulou, 2009).

The most common types of Greek nominal MWEs identified in the literature are[3] (Anastasiadi-Symeonidi, 1986; Fotopoulou et al., 2008):

- AJ-N: In this case we have an adjective followed by a noun which constitutes the head of the phrase, e.g. *φορητός υπολογιστής* (*laptop*), *ομφάλιος λώρος* (*umbilical cord*).
- N-N: MWEs of this type consist of two nouns that either:
  - carry the same weight and have the same case, e.g. *κράτος μέλος* (*member state*), *παιδί θάυμα* (*child prodigy*); or
  - the second is in genitive and modifies the first, e.g. *σύνοδος κορυφής* (*summit*), *Υπουργείο Εξωτερικών* (*ministry of foreign affairs*).
- N-DT-N: These MWEs have a noun phrase modifying a preceding noun, e.g. *κοινωνία της πληροφορίας* (*information society*), *μήλο της Έριδος* (*apple of discord*).
- N-P-N: In this case we have a prepositional phrase modifying a preceding noun, e.g. *σκλήρυνση κατά πλάκας* (*multiple sclerosis*), *φόνος εκ προμελέτης* (*premeditated murder*).
- P-N-N: MWEs in this category are very similar to those in the previous one in terms of their grammatical composition, the only difference being that the modifier precedes the noun it modifies, e.g. *δια βίου μάθηση* (*lifelong learning*), *κατά κεφαλήν εισόδημα* (*per capita income*).

In addition to these, we are going to examine two more categories:

- N-AJ-N: MWEs in this category consist of an adjectival phrase in the genitive case modifying a preceding noun, e.g. *ξέπλυμα βρώμικου χρήματος* (*money laundering*), *εμπόριο λευκής σαρκός* (*white slavery*).
- N-CJ-N: In this last category we come across phrases that consist of two conjoined nouns, e.g. *σάρκα και οστά* (*[take] shape*), *τελεία και παύλα* (*full stop*).

4 MWE Extraction

The candidate extraction process was carried out on the Europarl (EP) parallel corpus v3 (Koehn, 2005), which consists of extracts from the proceedings of the European Parliament during the period Apr/1996 – Oct/2006 in 11 languages[4]. The Greek portion of the corpus consists of 962,820 sentences and 26,306,875 words making it one of the largest Greek corpora widely available. Even though EP does not contain a great variation of text types, it can be assumed to constitute a relatively representative sample of general-purpose Greek language, mainly due to its size.

```
<CHAPTER ID=1>
Έγκριση των συνοπτικών πρακτικών της
προηγούμενης συνεδρίασης
<SPEAKER ID=1 NAME="Πρόεδρος" >
Τα συνοπτικά πρακτικά της χθεσινής συνεδρίασης
έχουν διανεμηθεί.
<P>
Υπάρχουν παρατηρήσεις
<P>
<SPEAKER ID=2 LANGUAGE"IT" NAME="Speroni">
Κύριε Πρόεδρε, χτες, στο τέλος της ψηφοφορίας
σχετικά
```

Figure 1: Extract from Greek EP from 17/12/1999.

4.1 Preprocessing the Corpus

All data is stored in one text file per day and each file contains document (`<CHAPTER id>`), speaker (`<SPEAKER id name language>`), and paragraph (`<P>`) mark-up on a separate line, as the example in figure 1.

In order to tag and lemmatize the corpus, we first had to remove the XML tags and split the text so that each file contained one sentence per line. Then, we used the Greek part-of-speech (POS) tagger developed at ILSP by Papageorgiou et al. (2000). Since Greek is a morphologically rich language, the tagset used for the description of the various morphosyntactic phenomena is very large compared to tagsets used by annotation schemata in other languages (584 vs. 36 tags in the Penn Treebank). These labels were reduced to simplified POS tags, as those in the example of figure 2. The word lemmata were determined using the ILSP morphological dictionary which contains around 80,000 lemmata corresponding to approximately 2,500,000 fully inflected entries.

The tagged corpus contains a relatively small number of errors like *πρακτικών* which has been misclassified as an adjective (*ο πρακτικός — practical*) rather than a noun (*τα πρακτικά — proceedings*). These errors may affect the extraction process since the patters for MWE candidate extraction are defined in terms of POS tags. In this context tagging errors imply in some candidates being incorrectly kept (false positives) while others are incorrectly removed (false negatives). This, however, cannot be avoided where large quantities of automatically POS tagged data are employed and their manual checking is not feasible, as is the case here.

[3] Where AJ stands for adjective, N for noun, DT for determiner, P for preposition and CJ for conjunction

[4] Europarl is publicly available at `http://www.statmt.org/europarl/`

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE patterns SYSTEM "mwttoolkit-patterns.dtd">
<patterns>
<pattern><w pos="AJ"/><w pos="N"/></pattern><!--φορητός υπολογιστής -->
<pattern><w pos="N"/><w pos="N"/></pattern><!--κράτος μέλος, Υπουργείο Εσωτερικών-->
<pattern><w pos="N"/><w pos="DT"/><w pos="N"/></pattern><!--φαινόμενο του θερμοκηπίου-->
<pattern><w pos="N"/><w pos="AJ"/><w pos="N"/></pattern><!--εμπόριο λευκής σαρκός-->
<pattern><w pos="N"/><w pos="P"/><w pos="N"/></pattern><!--σκλήρυνση κατά πλάκας-->
<pattern><w pos="P"/><w pos="N"/><w pos="N"/></pattern><!--κατά κεφαλήν εισόδημα-->
<pattern><w pos="N"/><w pos="CJ"/><w pos="N"/></pattern><!--τελεία και παύλα-->
</patterns>
```

Figure 3: XML file containing the description of the POS patterns we are interested in extracting.

```
      (SENT     <S>
1\1   TOK       Έγκριση     έγκριση      N
1\9   TOK       των         ο            DT
1\13  TOK       συνοπτικών  συνοπτικός   AJ
1\24  TOK       πρακτικών   πρακτικός    AJ
1\34  TOK       της         ο            DT
1\38  TOK       προηγούμενης προηγούμενος AJ
1\51  TOK       συνεδρίασης συνεδρίαση   N
1\62  PTERM_P   .           .      PTERM_P
1\63  CHUNK     _
      )SENT     </S>
```

Figure 2: Tagger output containing surface form, lemma and simplified POS tag.

4.2 Extraction Process

Once the corpus was cleaned, tagged and lemmatized, it was fed as input to the Multiword Expression Toolkit, which was used to extract and hierarchically organize the MWE candidates. The main advantage of this tool is that it is knowledge poor. This means that it does not require the creation of hand-crafted rules, dictionaries or specialized resources, but it is straightforward to apply on every language for which some shallow tools (lemmatizer and/or POS tagger) are available. `mwetoolkit` is a collection of Python scripts that take as input an XML tagged corpus and output either a list of n-grams or a list of word sequences of predetermined patterns. The seven POS patterns in figure 3 produced 526,012 word sequences and are defined on the basis of the types discussed in section 3. In order to reduce the effects of data sparseness and avoid computational overhead, we disregarded n-grams that occurred less than 10 times in EP. As a result, the size of the list of candidates reduced to 25,257 word sequences, which constitute our list of MWE candidates.

For each candidate entry, `mwetoolkit` gets the individual word counts both in EP and in WWW. These, combined with the n-gram joint count, are used to calculate four statistical Association Measures (AMs) for each MWE candidate: pointwise mutual information (`pmi`), maximum likelihood estimator (`mle`), Student's t score (`t-score`) and Dice's coefficient (`dice`). These measures are based on the observed cooccurence counts of an n-gram, i.e. of a sequence of n contiguous words (w_1 through w_n) denoted as $c(w_1 \dots w_n)$ in a corpus (in our case, EP or WWW) that contains N word tokens[5]. If words coocurred by chance, i.e. if we suppose that word occurrences are independent events, a given n-gram would be expected to have a count of E times, where $E(w_1 \dots w_n)$ is the expected joint count of the words:

$$E(w_1 \dots w_n) = \frac{c(w_1) \dots c(w_n)}{N^{n-1}}$$

The association measures are defined as:

- `pmi`: (pointwise) mutual information is a measure borrowed from information theory and is often used as a significance function for the identification of MWEs. The intuition behind it is to measure the amount of information shared by the occurrence of each word w_i in the n-gram. This method compares the observed number of cooccurence of the words with their expected count of cooccurence in the case they were independent events. It is defined as:

$$\texttt{pmi} = log_2 \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)}$$

- `mle`: maximum likelihood estimation is based on the assumption that the best estimate for the parameters examined is the one that maximizes the probability of the observed sample. No smoothing or discounting technique was applied, i.e. our measure relies on the closed world assumption:

$$\texttt{mle} = \frac{c(w_1 \dots w_n)}{N}$$

  Even though `mle` does not take into account the frequency of the individual words, it could be interesting to investigate how far simple joint frequencies would take us. Having said that, `mle` is not expected to be a strong measure of association by itself.

- `t-score`: Student's t-score is generally considered as a more reliable measure than `pmi` in low frequency

[5]For the WWW corpus, the approximate count of a given n-gram is estimated through the number of page hits in Yahoo!'s index. The total number of pages indexed by Yahoo!, estimated to 55 billion according to `http://www.worldwidewebsize.com/`, is used as the value of N.

```
<cand candid="13421">
  <ngram>
    <w lemma="αχίλλειος" pos="AJ" >
      <freq name="EP" value="14" /><freq name="WWW" value="16700" /></w>
    <w lemma="πτέρνα" pos="N" >
      <freq name="EP" value="14" /><freq name="WWW" value="49900" /></w>
    <freq name="EP" value="14" /><freq name="WWW" value="15400" />
  </ngram>
  <occurs>
    <ngram><w surface="αχίλλειος" lemma="αχίλλειος" pos="AJ" />
           <w surface="πτέρνα" lemma="πτέρνα" pos="N" />
           <freq name="EP" value="8" /></ngram>
    <ngram><w surface="Αχίλλειος" lemma="aq'illeios" pos="AJ" />
           <w surface="πτέρνα" lemma="pt'erna" pos="N" />
           <freq name="EP" value="1" /></ngram>
    <ngram><w surface="αχίλλειο" lemma="aq'illeios" pos="AJ" />
           <w surface="πτέρνα" lemma="pt'erna" pos="N" />
           <freq name="EP" value="5" /></ngram>
  </occurs>
  <features>
    <feat name="pos-pattern" value="AJ#S#N#S" /><feat name="n" value="2" />
    <feat name="mle-EP" value="7.4773e-07" /><feat name="pmi-EP" value="44.5092" />
    <feat name="t-EP" value="3.7416" /><feat name="dice-EP" value="1.0" />
    <feat name="mle-WWW" value="3.08e-07" /><feat name="pmi-WWW" value="55.3587" />
    <feat name="t-WWW" value="124.0966" /><feat name="dice-WWW" value="0.4624" />
  </features>
</cand>
```

Figure 4: Extract of the XML output file with MWE candidates and their AM scores.

ranges because it also takes into account the words' distribution variability:

$$\texttt{t-score} = \frac{c(w_1 \dots w_n) - E(w_1 \dots w_n)}{\sqrt{c(w_1 \dots w_n)}}$$

- `dice`: Dice's coefficient is a simple and popular similarity measure from information retrieval:

$$\texttt{dice} = \frac{n \times c(w_1 \dots w_n)}{\sum_{i=1}^{n} c(w_i)}$$

  It is considered by some as the most reliable AM, especially in the case of statistical machine translation of MWEs (Bai et al., 2009). However, we will consider all AMs in our evaluation since the effectiveness of each AM seems to depend on many different factors and there has not been an agreement on which one should be used in each case (Seretan, 2008).

The `mwetoolkit` outputs a file containing the following information on each MWE candidate: the lemma forms and POS tags of its individual words, the frequencies of these words as well as of the entire n-gram sequence both in EP and in the WWW, all the surface forms of each candidate together with their frequencies in the original corpus (EP) and a set of features that correspond to the candidate's score for each of the previously mentioned AMs. An example of an extracted candidate is showed in figure 4. Finally, the candidates are sorted into eight lists, according to each AM based on the EP and on the WWW counts. In the next section, the name of the corpus will be denoted as a subscript of the AM whenever it needs to be specified.

4.3 Evaluating the Results

Since there is, to our knowledge, no gold standard containing a considerable number of MWE entries in Greek, there is no way of automatically evaluating which are the real MWEs on the list of candidates. Consequently, evaluation was performed manually by three native speakers. Due to the size of the candidate list (25,257 candidates), it was not possible to perform exhaustive manual judgement of all the candidates. Instead, the human judges annotated a sample containing the first 150 candidates proposed by each measure (four measures from two corpora). From these, we manually removed the most striking cases of noise (introduced by the tagger) such as single words or candidates that appeared more than once based on a different grammatical classification. In short, each annotator classified around 1,200 entries in one of the following categories:

1. *mwe*: the candidate is a MWE, i.e. a true positive;
2. *maybe*: the candidate is ambiguous, but it may be considered as a MWE under certain assumptions;
3. *part*: the candidate includes a or is part of a MWE or;
4. *not*: the candidate is not a MWE, but a regular sequence of words with no particularity.

In the following evaluation steps, we considered MWEs to be those that were classified as such (*mwe*) by at least two out of three of our judges. This is a conservative evaluation scheme that does not take into account other categories such as *maybe* and *part*. Therefore, we also propose a scoring scheme that will be described later in this section to be used in the creation of the final dictionary of Greek MWEs.

The evaluation of NLP applications is usually based on precision and recall. Precision is defined as the percentage of candidates that were classified as MWEs, while recall is defined as the number of MWEs identified over the total number of MWEs. In order to calculate recall, however, we would need to know how many MWEs exist in EP, in the WWW, or more generally in the Greek language. Given that it is impossible to know and very difficult to estimate these values, our evaluation procedure will be based on precision only.

Our initial anticipation was that the `dice` coefficient or the `pmi` would be the first in rank, followed by `t-score` and then `mle`. As figure 5 shows this was not exactly the case. Considering only EP counts, the $\texttt{dice}_{\text{EP}}$ coefficient did indeed have the highest score, 81.08%. This level of precision surpassed all our expectations since it is one of the highest reported in the Greek literature. The second highest precision (58.21%) is achieved by the $\texttt{t-score}_{\text{EP}}$, followed by the $\texttt{mle}_{\text{EP}}$ at approximately the same levels (57.43%), leaving $\texttt{pmi}_{\text{EP}}$ behind with a precision of 52.66%. Most of these numbers agree with other studies in the literature. It is, of course, widely accepted that there is no single AM that always serves best. Some, however, are considered better than others in identifying idiosyncrasies in word relationships. Almost two decades ago, `pmi` was considered the most effective measure in MWE extraction (Smadja, 1993). More recent studies, however, have shown the `dice` coefficient and/or the `t-score` to be more effective (Evert and Krenn, 2005; Ramisch, 2009). The measure that best captures the relationship among the MWE components seems to depend on a number of factors like the size of the corpus, the type of texts it consists of, the language in question and others (Evert and Krenn, 2005; Villavicencio et al., 2007).

The most surprising result obtained, however, was the level of precision achieved by `mle`. As previously mentioned, this measure does not take into account individual word frequencies which led us to believe that it would be a very poor judge of MWEness. Surprisingly enough though, this did not turn out to be the case. Based on these findings, we intend to investigate the precision of this measure on other corpora as well, for consistency purposes.

The WWW-based precision for each AM other than the $\texttt{pmi}_{\text{WWW}}$ reached the same levels as the EP-based one. More precisely, the $\texttt{dice}_{\text{WWW}}$ coefficient yielded a precision of 79.43%, corresponding to a marginal decrease of approximately 2%. $\texttt{mle}_{\text{WWW}}$ and $\texttt{t-score}_{\text{WWW}}$, on the other hand, show an increase of $2.6\% - 2.7\%$, with their exact precision values being 58.99% and 59.71% respectively. These values seem to confirm our earlier assumption that EP, despite its lack of textual genre variation, can reasonably be assumed to contain a representative sample of the Greek language, mainly due to its size. The most striking result about the WWW-based results, however, is the dramatic decrease (almost 60%) in the precision the $\texttt{pmi}_{\text{WWW}}$ measure achieves (a highly unimpressive 21.62%).

These values seem to both verify and contradict some of the arguments presented in the literature about the use of the web as a corpus. The slight increase in the precision achieved by the `t-score` and the `mle` measures seems to indicate that the web's size makes it an invaluable tool for the MWE extraction process and possibly for other NLP applications as well. The magnitude of the precision decrease of `pmi`, however, seems to indicate that the threshold of 10 n-gram occurrences, which was more than satisfactory in the case of EP, turned out to be a serious underestimate in the case of the WWW, where almost all of the proposed candidates were wrong.

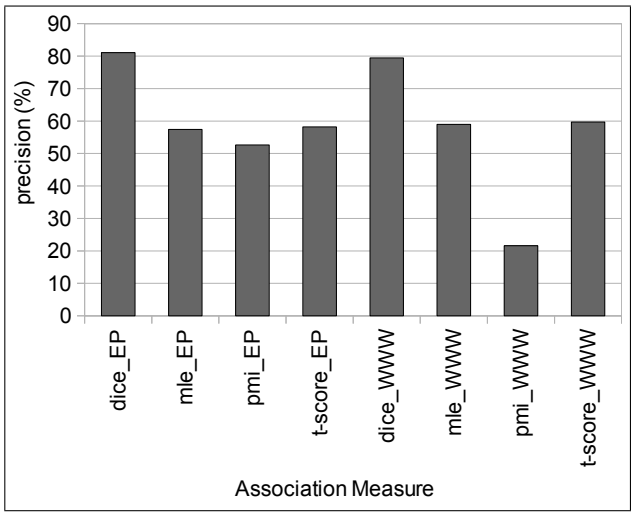

Figure 5: Precision based on the EP counts.

At the same time, `pmi` seems to overestimate the importance of the size of the word sequence since the candidate lists consisted entirely of three-word candidates both in the case of EP and WWW as opposed to, say, the `dice` coefficient whose candidate lists consisted of entirely of bigrams (something that can be attributed to their higher frequency of occurrence in general language use).

A large number of the candidates proposed by `pmi` included partial MWEs, which were not proposed as a unit by themselves, but in combination with some other word. To be more precise, out of the 148 candidates evaluated, 32 were classified as MWEs while another 50 included some MWE, which in the majority of cases was *εν λόγω* (*in question*). Indeed, some of the candidates classified as *part* or *maybe* should be manually analysed for deciding whether to include them as entries in a dictionary, as they could constitute interesting MWEs.

Therefore, to evaluate the MWE list, we propose a scoring scheme where each candidate is assigned a value s that depends on the number of judges $\#(C)$ that classified the candidate as an instance of a category C (*mwe*, *maybe* or *part*). The precise formulation of the scheme to be adopted depends on which criteria one wants to emphasize: precision or coverage. To emphasize precision, one could consider as genuine MWEs only those candidates classified as *mwe* by most judges. On the other hand, to emphasize coverage, one can also consider those candidates classified as *maybe* and *part*. In addition, a preference on the categories can also be taken into account in the scoring scheme, where each category could be assigned a specific weight depending on how much influence it has. For instance, for unambiguous MWEs to be given more weight than ambiguous or partial cases, *mwe*, *maybe* and *part* can be given decreasing

	mwe	maybe	part	not	κ	$s \geq 4$
$\texttt{dice}_{\text{EP}}$	78%	10%	3%	9%	40%	82%
$\texttt{mle}_{\text{EP}}$	55%	9%	3%	33%	65%	60%
$\texttt{pmi}_{\text{EP}}$	50%	9%	16%	26%	52%	57%
$\texttt{t-score}_{\text{EP}}$	56%	9%	3%	32%	61%	61%
$\texttt{dice}_{\text{WWW}}$	78%	8%	2%	12%	56%	84%
$\texttt{mle}_{\text{WWW}}$	58%	6%	1%	36%	74%	60%
$\texttt{pmi}_{\text{WWW}}$	21%	7%	36%	36%	63%	24%
$\texttt{t-score}_{\text{WWW}}$	58%	6%	1%	35%	70%	61%

Table 2: Inter-annotator agreement for each of the four categories and each evaluated AM in both corpora, as well as Fleiss' kappa coefficient (k) and proportion of true positives according to score $s \geq 4$.

weights.
The scoring scheme adopted in this work is:

$$s = 2 \times \#(\textit{mwe}) + \#(\textit{maybe}) + \#(\textit{part})$$

This scheme considers candidates which were classified by the judges as belonging to one of these 3 categories. For this evaluation, we consider as interesting MWE candidates those that have a score greater than or equal to 4, including cases which were classified as *mwe* by at least one judge and as ambiguous/partial by the others. We did not chose among the evaluated AMs, but combined the four EP-based lists into a single one since the candidate lists retrieved by each measure are very heterogeneous. The WWW-based results were, for the moment, disregarded, since they did not bring large performance improvements over EP-results (this does not mean that they could not be useful in the case of smaller corpora, for example).
As an additional evaluation, we quantified the difficulty of the classification task for the human judges. Therefore, we calculated the inter-judge agreement rate using Fleiss's kappa coefficient, which is best suited in the case of multiple annotators. The results for each analysed AM are summarized in table 2: the first four columns correspond to the individual agreement proportions for each of the categories while the last two columns of the table contain respectively the kappa value and the proportion of instances that were considered as true MWEs according to the scoring scheme proposed above. The values in the last column are slightly greater than the performance values showed in figure 5 mainly because the scoring scheme is less conservative than the majority vote used to perform the preliminary evaluation of each AM independently.
The agreement coefficients are very heterogeneous, ranging from $\kappa = 40\%$ to $\kappa = 74\%$. A coefficient of 40%, for example, means that there is a probability of 40% that this agreement was not obtained by chance. This explains such low κ values despite the high agreement for category *mwe*, which is also the most frequent in this data set. The coefficient is, therefore, unable to assign more importance to a given category. Moreover, there is no general agreement on how to interpret these results, but it is believed that kappa values should be above 70%. Our results show, however, that there is no high agreement among annotators according to this criterion. If we look in detail at the proportion of agreement for each category, we can see that is is quite easy for the annotator to identify true MWEs, whereas, for the other classes, the agreement is much lower (e.g. annotators cannot truly distinguish categories *maybe* from *part*). While, on one hand, this might be caused by ambiguous annotating guidelines, on the other hand, it is also an indicator of how difficult it is for a human annotator to identify and classify MWEs.
We also found out that there is high correlation ($r \approx .99$) between the agreement on category *mwe* and the precision of the method, i.e. it is easy to identify true MWEs in a high-quality list, whereas it is much more difficult to select useful MWEs when the list contains a lot of noise. While this might seem obvious, it corroborates the hypothesis that precise automatic methods can considerably speed up lexicographic work in the process of language resources creation. Additionally, the agreement is always higher when web-based AMs are analysed, and this is not in direct correlation to the AM's performance. At first glance, we can suppose that it is easier to interpret the results coming from a web-based method than the results from EP, even if the former does not necessarily improve precision. This issue, however, needs further investigation, since it is not clear to date what benefits one could take from the WWW combining with or replacing well-formed corpora like EP.

5 Conclusions

The ubiquity of multiword expressions in language makes them a key point for many NLP tasks and applications, such as machine translation. However, due to the limited number of MWE lexical resources for many languages, it is crucial to develop methods for semi-automatically extracting them from corpora. In this paper, we described an investigation of the performance of language-independent automatic MWE identification methods applied to Greek data.
We used the `mwetoolkit` to extract an initial list of MWE candidates from the Greek EP corpus. Then, we applied a set of morphosyntactic filters to remove noisy cases and ranked the remaining candidates according to some statistical measures of association. These were calculated based on counts retrieved from 2 corpora, EP and WWW, to verify the robustness of the results and minimize potential problems caused by data sparseness. The ranked lists produced by four AMs applied on two different corpora were

manually evaluated by native speakers. The κ inter-judge agreement scores give a good indication of the difficulty of the task.

From the AMs used, the one that produced better results for the Greek EP corpus was the `dice` coefficient, which significantly outperformed the other measures, followed by the `t-score`. The performance of the latter, however, is surprisingly similar to the performance of the `mle` measure, suggesting that sophisticated measures are not needed when enough data is available. We plan to investigate the possibility of bringing further improvement to the results by using AMs based on contingency tables too. Of course, these positive results must be considered in context, as due to the size of the candidate list, the evaluation focused on the first 150 candidates in each list. Even so, these results of the manual evaluation confirm that it is possible to obtain satisfactory results using simple and language-independent tools to filter and rank the candidate lists. This is an important outcome for a language like Greek, for which MWE resources are limited.

In relation to the use of the web as a corpus, it has a number of advantages over standard corpora, the most salient being its availability and accessibility, especially in the case of specialized domains, where large corpora are rarely available. However, in this paper, the results obtained with WWW counts did not bring considerable improvements over the EP corpus in general in terms of performance. On the other hand, WWW-based results seem to be easier to interpret given that they achieve higher inter-annotator agreement, independently of the particular AM. In short, we believe that further investigation is required in order to explain the variations in the WWW-based results among the various measures used, specially in the case of `pmi`. In addition, we plan to investigate the use of WWW-based counts in combination with corpus-based counts for the candidates.

The next steps of our research are towards the integration of the generated MWE dictionary into a machine translation system in order to investigate the benefits brought by the use of such information in the translation pipeline. Therefore, we will need to investigate effective ways of aligning the extracted expressions with corresponding counterparts in the target language. Another possibility would be to use the extracted MWEs in a language model or to rerank translation candidates when translating from another language to Greek. Finally, we believe that the main contribution of this paper is to show that simple MWE extraction tools can speed up the creation of language resources for the Greek language, that may be employed to improve the results of monolingual and multilingual NLP systems supporting this language.

Acknowledgements

This research was supported by the Greek State Scholarship Foundation (IKY). It has also been partly funded by FINEP/SEBRAE (COMUNICA project 1194/07).

6 References

Anna Anastasiadi-Symeonidi. 1986. *Neology in Modern Greek (in Greek)*. Ph.D. thesis, Aristotle University of Thessaloniki.

Ming-Hong Bai, Jia-Ming You, Keh-Jiann Chen, and Jason S. Chang. 2009. Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 478–486, Suntec, Singapore, August. Association for Computational Linguistics.

Timothy Baldwin. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4):398–414.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain, May. European Language Resources Association.

Gael Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 41–48, Sapporo, Japan, July. Association for Computational Linguistics.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4):450–466.

Aggeliki Fotopoulou, Giorgos Giannopoulos, Maria Zourari, and Marianna Mini. 2008. Automatic recognition and extraction of multiword nominal expressions from corpora (in greek). In *Proceedings of the 29th Annual Meeting, Department of Linguistics*, Aristotle University of Thessaloniki, Greece.

Agelikki Fotopoulou. 1993. *Une classification des phrases à compléments figés en grec moderne : étude morphosyntaxique des phrases figées*. Ph.D. thesis, Université Paris VIII.

Aggeliki Fotopoulou, 1997. *L' ordre des mots dans les phrases figées à un complément libre en grec moderne*. Saint-Cloud. INALF.

Ray Jackendoff. 1997. Twistin' the night away. *Language*, 73:534–559.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit(MT Summit 2005)*, pages 79–86, Phuket, Thailand, September. Asian-Pacific Association for Machine Translation.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge, UK.

Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. Lexschem : A large subcategorization lexicon for french verbs. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May. European Language Resources Association.

Athina Michou and Violeta Seretan. 2009. A tool for multi-word expression extraction in modern Greek using syntactic parsing. In *Proceedings of the Demonstrations*

Session at EACL 2009, pages 45–48, Athens, Greece. Association for Computational Linguistics.

Marianna Mini and A. Fotopoulou. 2009. Typology of multiword verbal expressions in modern greek dictionaries: limits and differences (in greek). Proceedings of the 18th International Symposium of Theoretical & Applied Linguistics, School of English, pages 491–503, Aristotle University of Thessaloniki, Greece.

Argyro Moustaki. 1995. *Les expressions figées* είμαι*/être Prép C W en grec moderne*. Ph.D. thesis, Université Paris VIII.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In Ido Dagan and Dan Gildea, editors, *Proceedings of the Ninth Conference on Natural Language Learning (CoNLL-2005)*, University of Michigan, MI, USA, June. Association for Computational Linguistics.

Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A unified pos tagging architecture and its application to greek. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1455–1462, Athens, Greece, May. European Language Resources Association.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, July. Association for Computational Linguistics.

Carlos Ramisch, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In Alex Clark and Kristina Toutanova, editors, *Proceedings of the Twelfth Conference on Natural Language Learning (CoNLL 2008)*, pages 49–56, Manchester, UK, August. Association for Computational Linguistics.

Carlos Ramisch. 2009. Multi-word terminology extraction for domain-specific documents. Master's thesis, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, France, June.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, volume 2276/2010 of *Lecture Notes in Computer Science*, pages 1–15, Mexico City, Mexico, February. Springer.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, University of Geneva.

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Sara Stymne. 2009. A comparison of merging strategies for translation of german compounds. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 61–69, April.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic, June. Association for Computational Linguistics.

Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems

Mirabela Navlea, Amalia Todiraşcu
LiLPa, Université de Strasbourg
22, rue René Descartes, BP 80010, 67084 Strasbourg cedex, France
mirabela_abe@yahoo.fr, todiras@unistra.fr

Abstract

We present an on-going project aiming at the development of linguistic resources for a French - Romanian factored phrase-based statistical machine translation system. Factored statistical machine translation systems apply several categories of linguistic properties and use sentence and word aligned corpora. We present the parallel corpus and the alignment method. Then, after aligning the corpus, we identify several alignment errors, due to language variance. We define some heuristic rules to improve the lexical alignment.

1. Introduction

The goal of our project is to build linguistic resources for a factored phrase-based statistical machine translation system for Romanian and French. As well, we study the influence of several categories of linguistic informations to the quality of translation provided by the system.

Our project adopts a methodology proposed by the SEE-ERA.net (Tufiş et al., 2008). Their goal was to develop factored statistical machine translation systems for Slavic and Balkan languages (Bulgarian, Greek, Romanian, Serbian, Slovenian) from and to English. For MT, this project developed and used sentence and word aligned parallel corpora. The corpora were tagged, lemmatized, and annotated with partial syntactic constituents.

The increasing number of documents, available in several languages, requires new methods to manage large multilingual document databases, to create multilingual web sites and applications, to improve cross-lingual information retrieval methods. Some of these methods use MT techniques that should be adapted to handle new languages. Indeed, most of the available linguistic resources for machine translation were developed considering English as source or target language.

The quantity of web pages available in other languages than Western European languages increases constantly and new resources should be developed for these new languages as South, Central and Eastern European languages. Monolingual resources as lexicons, annotated corpora or NLP tools are available for these languages, but bilingual resources were developed considering English as a target or source language. English morphology is quite simple, while other South and Eastern European languages are characterized by a rich morphology. The situation of Romanian resources is representative. If lexicons, taggers and annotated corpora are available for Romanian, most of the MT systems are available for English and Romanian (Marcu & Munteanu (2005); Irimia (2008); Ceauşu (2009); *Google Translate* system[1]). At the moment, only *Google Translate* supports Romanian - French translation. Even if MT systems improved their performances last years, erroneous output is still very important, due to the lack of resources as lexicons or grammars. Grass (2009) identifies thirteen frequent error categories provided by MT systems, as polysemy, homonymy, ambiguity (syntactic or referential), neologism identification. Specific problems occur for two rich morphology languages as Romanian and French. Indeed, the high

[1] http://translate.google.com/

number of inflected word forms increases the translation hypothesis. To avoid these errors, MT tools should use complex linguistic resources: synonymy dictionaries, collocation dictionaries, terminological databases or knowledge bases. To avoid the use of these expensive resources, other approaches focus on statistical techniques using aligned corpora.

If linguistically-based methods (*Systran*[2]) use expensive resources to obtain good quality translation, factored SMT systems (*EuroMatrix*, 2009) provide comparable results with linguistically-based systems. The resources used by these systems are annotated and aligned parallel corpora.

Factored SMT systems extend phrase-based statistical methods (Koehn, Och & Marcu, 2003) and use linguistic properties as lemma, tags and syntax. The systems are modular: several categories of linguistic properties might be used. Koehn and Hoang (2007) used morpho-syntactic properties, Avramidis and Koehn (2008) exploited syntactic information to improve translation quality. Our work aims to adapt a factored statistical machine translation system for two Romance languages, Romanian and French. We use several large monolingual tagged corpora to build language models for French and Romanian. As well, we build sentence and word aligned corpora to build translation models.

In the next section, we present the architecture of a factored SMT system and the linguistic informations used to build language and translation models. The corpus used is presented in section 3. The differences between Romanian and French are explained in section 4. The alignment process and the errors found in the word aligned corpus are described in section 5.

2. The Project

Our project aims at developing a factored statistical machine translation tool for Romanian and French. This system was initially implemented for English and Romanian[3] (Ceauşu, 2009). This system uses a sentence and word aligned parallel corpus. In addition, the system applies several categories of linguistic factors:

- word forms ;
- lemmas ;
- morpho-syntactic descriptors (the MSD tagset from Multext project (Ide & Véronis, 1994)) ;
- chunks (simple, non-recursive groups) ;
- collocations.

The system proves its efficiency (Ceauşu, 2009), mainly for law texts. The system uses *MOSES* decoder (Koehn et al., 2007), with an optimal configuration of linguistic parameters as MSD, lemmas, chunks.

In order to adapt *MOSES* to a new language pair, it is necessary to build a language model from monolingual corpus in target language and a factored translation model from annotated and aligned corpus. Then, the *MOSES* decoder is used to find the most probable translation from the language and translation models. Factored SMT systems use several kinds of linguistic informations: word forms, lemmas, POS tags etc. These informations might be combined in order to obtain optimal translation results.

As the system is not yet adapted for French and Romanian, our work focuses on building the resources necessary to obtain language and translation models. As well, we study the influence of language specific linguistic factors to the translation results.

To build the required resources, we follow several steps:

1) creation of parallel corpora;
2) preprocessing (tokenization, lemmatization, tagging, chunking);
3) sentence and word alignment ;
4) alignment errors correction, after a detailed analysis of these errors. These data are used to retrain the alignment module ;
5) building language models from monolingual corpora and translation models from parallel aligned corpora for French - Romanian;
6) system configuration with most relevant linguistic factors;
7) the linguistic analysis of translation errors, system reconfiguration and restarting at 4, in order to optimize the system.

In the next sections, we present the monolingual and parallel corpora and the preprocessing steps, as well as the alignment process and a complete set of alignment

[2] http://www.systransoft.com/
[3] http://www.racai.ro/webservices

errors identified from the French - Romanian word aligned corpus.

3. The corpus

We use a French - Romanian parallel corpus extracted from the *JRC-Acquis* (Steinberger et al., 2006). *JRC-Acquis* is based on the *Acquis Communautaire* multilingual corpus, aligned at paragraph level and available for all the 231 pairs of languages, obtained from the 22 official languages of EU. This corpus is in XML format and is freely available. *Acquis Communautaire* is composed of laws adopted by EU states member and candidates since 1950. For our project, we use a set of 228 174 pairs of 1-1 aligned sentences from the *JRC-Acquis* (5 828 169 tokens for French, 5 357 017 tokens for Romanian), selected from the common documents available in French and in Romanian.

We use also a French - Romanian parallel corpus extracted from the *DGT Translation Memory (DGT-TM)*. This is a freely available resource based also on the *Acquis Communautaire*, but most sentence alignment was corrected manually. The French - Romanian *DGT-TM* contains 490 962 aligned sentences pairs (9 953 360 tokens for French and 9 142 291 tokens for Romanian). This corpus is in TMX format.

Due to the fact that the *JRC-Acquis* and the *DGT-TM* are law corpora, we built other multilingual corpora. In order to test the system for other domains (politics, aviation), we selected several bilingual texts available in French and Romanian from several Web sites:

- European Parliament documents (263 788 tokens);
- Romanian airplane companies sites (63 353 tokens).

In order to resolve the missing diacritics problem of the most of Romanian texts collected from the Web, we used *Diac+*, a diacritics recovering system (Tufiş & Ceauşu, 2008).

For evaluation purposes, we use a small French - Romanian corpus of 1000 aligned sentences, developed for "Collocations in context" project (Todiraşcu et al., 2008). This corpus was derived from two corpus (English - Romanian and English - French), from JRC-Acquis, obtained by automatic word alignment. Then, through a derivation process (Tufiş & Koeva, 2007), we obtained a French - Romanian word aligned corpus, which was corrected manually.

To build language models, we use monolingual corpora. For Romanian we use several corpora [4] :

- Newspapers corpus (NAACL corpus) (800 000 tokens);
- L4TeL corpus (600 000 tokens);
- newspapers corpus (RoCo corpus) (7,5 millions of tokens).

For French, we use a corpus composed of:

- a law corpus, selected from *JRC-Acquis* (498 788 tokens);
- a newspaper corpus (*Le Monde*, 1980-1988) (488 543 tokens).

In order to exploit linguistic informations to improve word aligner's results and to build language models and factored translation models, we preprocess the corpora: we apply a tagger, TTL[5] (Ion, 2007) available for Romanian and French, as a Web service. This tagger tokenizes, lemmatizes and annotates the text with chunk information. The chunks are simple noun phrases, simple prepositional phrases or verbal phrases. The output of this tagger is in XCES format (Figure 1) and it uses the MSD from the MULTEXT tagset (Ide & Véronis 1994).

The quality of the POS tagging is crucial to obtain useful language models. If TTL has already language models for Romanian, we developed the linguistic resources for French tagging. The monolingual French corpus, manually corrected, was firstly used to train TTL. Then, we evaluated the output of the tagger on a small part of the corpus (100 000 tokens), due to the large volume of data. We identified several systematic tagging and chunking errors for French :

- confusion of plural, indefinite determiners (*des*) and of aggregates (*des = de+les*);
- impossibility to decide if the definite article *les* or *l'* 'the' are masculine or feminine;
- pronouns erroneoulsy tagged as definite articles, so the pronoun is not annotated as

[4] available on demand from the authors

[5] TTL = Tokenizing, Tagging and Lemmatizing free running texts

part of the verbal chunk;
- main verbs tagged as auxiliary verbs;
- wrong lemmas.

```
<seg lang="fr"><s id="ttlfr.3">
<w lemma="voir" ana="Vmps-s">vu</w>
<w lemma="le" ana="Da-fs" chunk="Np#1">la</w>
<w lemma="proposition" ana="Ncfs" chunk="Np#1">proposition</w>
<w lemma="de" ana="Spd" chunk="Pp#1">de</w>
<w lemma="le" ana="Da-fs" chunk="Pp#1,Np#2">la</w>
<w lemma="commission"ana="Ncfs"chunk="Pp#1,Np#2">Commission
</w>
<c>;</c>
</s></seg>
```

Figure 1: TTL output for French.

For systematic errors, we propose a base of correction rules. These correction rules complete the annotations (chunk-based level). The corrected output should be used to improve the existing language models.

4. Building aligned corpora

As we mentioned in the previous section, we use sentence and word aligned corpora to build translation models. The quality of the aligned corpus is essential to obtain good translation results. While we do not have resources for French - Romanian word alignment, we use the sentence aligned corpus (*Ceauşu et al, 2006*), which is also lemmatized, tagged and annotated at chunk level.

Then, we prepare the corpus in the input format required by *GiZA++* (Och & Ney, 2003), but also providing lemma and lexical category information. We apply the word alignment bidirectionnally and then we obtain the intersection of word alignments.

From this sentence aligned corpus, *GIZA++* builds a list of translation equivalents. Then, we build a list of cognates for Romanian and French, in order to filter the translation equivalents. To identify cognates, we apply an algorithm computing the longest common character sequence for two given words. If the length of the longest common sequence is at least 70% of the length of the shortest word, then the word pair is selected as cognate.

We use the filtered list of translation equivalents as a starting point to build word alignments. As proposed by (Tufiş et al., 2005), we apply a set of heuristics to align words:

- we define some POS equivalence classes (a noun could be translated to a verb or to an adjective);
- we align content words;
- we align chunks containing translation equivalents;
- we align the elements contained into a chunk, applying some heuristic rules.

To improve word aligner's results, we studied the systematic alignment errors. After error analysis, we propose several heuristic rules to repair these errors.

5. Alignment errors

We aligned a French - Romanian 1000 sentences corpus from *JRC-Acquis*, with *GIZA++*, as described in the previous section. Then, we analyzed the most frequent and the most systematic errors. Most of these errors are due to the fact that Romanian and French have specific properties, others are specific of the domain: collocations, domain terms.

5.1. French and Romanian

Even these languages are Romance languages and the grammar is similar to Latin grammar, various elements are quite different. French and Romanian are Romance languages characterized by a rich morphology. Syntactic structures are quite similar, but morphological properties are different. So, Romanian nouns and nominals are characterized by a case (nominative, accusative, genitive, dative, vocative) marked by an additional morpheme or a suffix. As well, Romanian nouns have a third gender (neuter). The defined determiner is a suffix in Romanian, while in French is a single, separate word. In Romanian, clitic pronouns might be used simultaneously with the direct or the indirect object. In French, the use of the clitic pronoun excludes surface realization of the direct or of the indirect object as a noun phrase. Other differences concern specific syntactic structures for each language (relative clauses), verb specific

constructions (additional morphemes for some modes or times in Romanian, specific auxiliary for some move verbs in French), and the supplementary morphemes from one language to another (possession relation).

5.2. Error analysis

In this section, we present several systematic errors identified in the word alignment output. These errors are explained by the morphosyntactic differences between the two languages.

To correct these errors, we defined heuristic rules to apply to *GIZA++* output in order to repair some problems.

We analyzed alignment errors of our corpus and we defined 27 morphosyntactic correction heuristic rules for the word aligner. These resources are used to proceed to the word alignment inside chunks.

We present some of the most frequent errors found in the *JRC-Acquis* corpus.

Possession relations. One of the frequent errors concerns the expression of the possession relation. Due to the differences of surface realization of this relation in the two languages, the markers of this relation are not aligned. In French, we express the possession by *de* "s' preposition, while in Romanian one of the possibilities to express this relation is to use a supplementary morpheme for genitive case *al, a, ai, ale* "s' followed by a genitive noun:

Example:

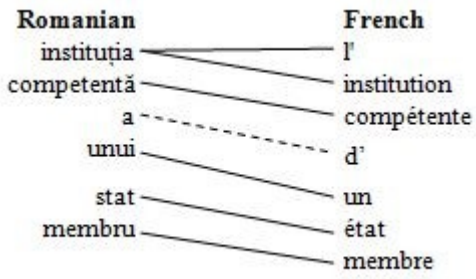

Figure 2: Possessive particles

We propose the following contextual heuristic rules, exploiting POS tagging, to avoid this error category in the Table 1:

Romanian	N (definite determiner) + ADJ + *al* | *a* | *ai* | *ale* + indefinite determiner (genitive form) + N genitive
French	definite determiner + N + ADJ + *de* + indéfinite determiner + N

Table 1: Possessive repairing rules

Relative clauses. Another category of frequent errors concerns the relative clauses. In Romanian relative clauses, the direct object is expressed twice by the relative pronoun *care* 'which' (accusative), preceded by the *pe* 'on' preposition, and *îl, -l 'him', o, îi, 'her' -i, le 'them'* personal pronouns. In French relative clauses, the direct object is expressed by the relative pronoun *que 'which'*. Due to this difference, the *que* French relative pronoun is not aligned with *pe* Romanian preposition and with *le* 'them' personal pronoun.

Example:

Figure 3: Relative pronouns alignment errors

Table 2 contains the contextual heuristic rule proposed to solve this type of error:

Romanian	N + *pe* + *care* (accusative) + *îl* | *-l* | *o* | *îi* | *-i* | *le* + V
French	N + *que* + V

Table 2: Contextual alignment rule for relative pronouns

Passive constructions. French uses frequently the passive constructions while Romanian shows a preference for the constructions with reflexive verbs in the studied corpus. Thus, the *être* 'to be' French

auxiliary verb is not aligned with the *se* 'himself' Romanian reflexive pronoun.

Example:

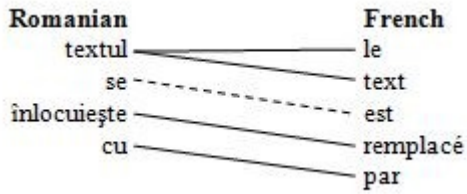

Figure 4: Reflexive verb vs. passive form

We propose the contextual heuristic rules to repair this type of alignment errors in the Table 3:

Romanian	N + *se* + V + *cu* (preposition)
French	N + *être* + V (past participle) + *par* (preposition)

Table 3: Repairing passive – reflexive forms

Infinitive. An infinitive French verb is often translated as a subjunctive form of the Romanian verb. In this case, the *să* Romanian particle of subjunctive is not aligned with the French verb in the infinitive.

Example:

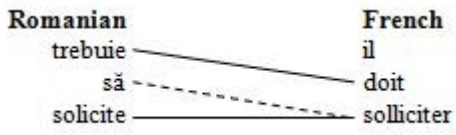

Figure 5: Subjunctive – infinitive alignement

The contextual heuristic rules for resolving this alignment error are in the Table 4:

Romanian	V + *să* + V
French	V + V (infinitive)

Table 4: Repairing subjunctive – infinitive alignment

Collocations. We repair collocations alignment problems. Collocations are polylexical expressions, composed of several words related by lexico-syntactic relations (Todiraşcu et al., 2008). These expressions might be translated as collocations, or as a single lexical unit. For example, one verbo-nominal collocation as *avoir le droit* (lit.' to have the right') has a Romanian equivalent the collocation *a avea dreptul*, but *procéder à l'examen* (lit. 'to proceed the exam') is translated by the verb *a examina* ('to examine') Moreover, one collocation as *mettre en application* (lit. 'to put into application', to apply) might be translated by its nominalisation *punerea în aplicare* (= *la mise en application*). We use a multilingual dictionary of collocations, containing their contextual morphosyntactic properties (Todiraşcu et al., 2008) to align these expressions. Collocations are used as main indices to build word alignment.

Collocations are not aligned together. Therefore, lexical units belonging to the collocations remain unaligned. In the example below, the *de* indefinite determiner of French collocation *ne pas prendre de mesures (*negative form of *take measures)*, in the negative form, is unaligned.

Example :

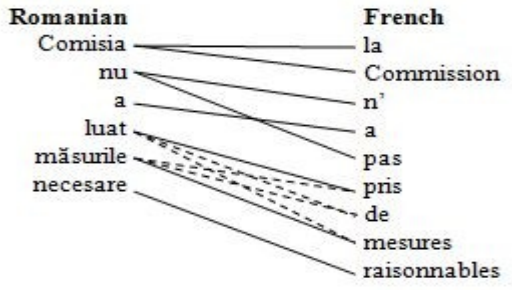

Figure 6: Repairing collocations alignment

We solve this type of alignment error by the use of a multilingual dictionary of collocations, containing their contextual morphosyntactic properties (Todiraşcu et al,. 2008). The dictionary contains 250 trilingual entries (Romanian, French, German). The entries are VN collocations. Each entry contains informations about the morphosyntactic properties of the verb, of the noun, but also the properties of the collocations (subcategorization properties, prepositions preferences). These informations are used to complete the alignment.

6. Conclusion

We present here an on-going project aiming at the development of linguistic resources for factored SMT systems for two languages with rich morphology: French and Romanian. We focus here on the presentation of resources used to build word aligned corpora. We analyzed the output of alignment module and we proposed heuristic rules to improve alignment results. Further work includes the evaluation of several combinations of linguistic informations (POS, MSD, lemma, chunks) in order to find the best parameters for Romanian and French. Furthermore, we will compare these results with a pure SMT system as proposed in (Gavrilă, 2009) for German – Romanian.

7. Acknowledgements

We thank Rada Mihalcea (University of Texas) for the NAACL corpus and Dan Cristea (University of Iaşi, Romania) for the L4TeL corpus. We thank Dan Tufiş (Romanian Academy, Bucharest) for the RoCo corpus.

8. References

Avramidis, E., Koehn, P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation, In *Proceedings of ACL-08: HLT,* Columbus, June 2008, pp. 763-770.

Ceauşu, A., Ştefănescu, D., Tufiş, D. (2006). Acquis Communautaire Sentence Alignment using Support Vector Machines, In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006, pp. 2134-2137.

Ceauşu, A. (2009). Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă, Ph.D. Thesis, Romanian Academy, Bucharest, april 2009, 123 p.

EuroMatrix, 2009 : The EuroMatrix Project: Statistical an Hybrid Machine Translation Between All European Languages, www.euromatrix.net/.

GIZA++, 2003 : Training of statistical translation models, www.fjoch.com/GIZA++.html.

Grass, T. (2009). A quoi sert encore la traduction automatique? *Les Cahiers du GEPE, Outils de traduction - outils du traducteur?*, n° 3, Strasbourg, 14 p..

Gavrilă, M. (2009) SMT experiments for Romanian and German using JRC-Acquis. In *Proceedings of RANLP-associated workshop: Multilingual resources, technologies and evaluation for central and Eastern European languages*, 17 September 2009, Borovets, Bulgaria.

Ide, N., Véronis, J. (1994). Multext (multilingual tools and corpora). In *Proceedings of the 15th CoLing*, Kyoto, August 5-9, pp. 90-96.

Ion, R. (2007). Metode de dezambiguizare semantică automată. Aplicatii pentru limbile engleză si română, Ph.D. Thesis, Romanian Academy, Bucharest, mai 2007, 148 p.

Irimia, E. (2008). Experimente de Traducere Automată Bazată pe Exemple, *Atelierul de Lucru Resurse Lingvistice Românesti si Instrumente pentru Prelucrarea Limbii Române*, Iasi, 19-21 novembre 2008, pp. 131-140.

Koehn, Ph., Och, F. J., Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003 Main Papers*, Edmonton, May-June 2003, pp. 48-54.

Koehn, Ph., Hoang H., (2007). Factored translation models, In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, June 2007, pp. 868-876.

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, Ch., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, Ch., Zens, R., Dyer, Ch., Bojar, O., Constantin, A. Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, June 2007, pp. 177-180.

Marcu, D., Munteanu, D. S. (2005). Statistical Machine Translation: An English - Romanian Experiment, *EUROLAN 2005*.

Och, F. J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, vol. 29, n° 1, March 2003, pp. 19-51.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages, In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006, pp. 2142-2147.

Todiraşcu, A., Heid U., Ştefănescu, D., Tufiş D., Gledhill C., Weller M., Rousselot F. (2008). Vers un dictionnaire de collocations multilingue. *Cahier de Linguistique*, vol. 33, n° 1, Louvain, août 2008, p. 161-186.

Tufiş, D., Ion, R., Ceauşu A., Ştefănescu, D. (2005) Combined Aligners. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 107-110, Ann Arbor, USA, Association for Computational Linguistics. ISBN 978-973-703-208-9

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In Ishida T., Fussell, S. R., Vossen P.T.J.M., eds., *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, Trento, Association for Computational Linguistics, pp. 153-160.

Tufiş, D., Koeva, S. (2007). Ontology-supported Text Classification based on Cross-lingual Word Sense Disambiguation. In Masulli, F., Mitra, S., Pasi, G.., eds., *Applications of Fuzzy Sets Theory. 7th International Workshop on Fuzzy Logic and Applications (WILF 2007)*, volume 4578 of Lecture Notes in Artificial Intelligence, September 2007, Springer - Verlag, pp. 447-455.

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., Krstev, C. (2008). Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Tadić(M.), Dimitrova-Vulchanova (M.) et Koeva (S.), eds, In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, Dubrovnik, September 25-28, pp. 145-152.

Tufiş, D., Ceauşu, A. (2008). DIAC+: A Professional Diacritics Recovering System. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May 2008.

South-East European Times: A parallel corpus of Balkan languages

Francis M. Tyers, Murat Serdar Alperen

Dept. Lleng. i Sist. Informàtics, Dept. Economics
Universitat d'Alacant, Ege University,
E-03071 Alacant (Spain), İzmir (Turkey)
ftyers@dlsi.ua.es, msalperen@gmail.com

Abstract

This paper describes the creation of a parallel corpus from a multilingual news website translated into eight languages of the Balkans (Albanian, Bulgarian, Croatian, Greek, Macedonian, Romanian, Serbian, and Turkish) and English. The corpus is then applied to the task of machine translation, creating 72 machine translation systems. The performance of these systems is then evaluated and thought is given to where future work might be focussed.

1. Introduction

The article has a twofold aim, the first is to describe the creation and status of a *free*[1] parallel corpus for the Balkan languages. The second is to describe the use of this corpus to create 72 machine translation (MT) systems between the Balkan languages and evaluate the differing challenges facing MT between these languages. It also presents the first published results for systems between, for example Macedonian and Albanian and gives some thought on where further research might be aimed.

This is a parallel corpus in the vein of EuroParl (Koehn, 2005) or JRC-Acquis (Steinberger et al., 2006), that is designed to be useful for machine translation and other multilingual natural language processing research, not necessarily useful for corpus-linguistic research due to uncertainty of which is the source and which is the target language of the translated sentences.

Aside from the ubiquitous English, the languages contained in this corpus fall into several linguistic groups, Turkic (in the case of Turkish), Slavic (Bulgarian, Croatian, Macedonian and Serbian), Hellenic (Greek), Romance (Romanian) and Albanic (Albanian).

A number of these languages also form what is known as the *Balkan Sprachbund*, or Balkan linguistic area. This is a group of languages which have similar lexical and grammatical features, but as a result of geographical proximity rather than genetic relationship (Lindstedt, 2000).

It has been shown that translating between genetically and typologically related languages is easier than languages with less relation. Homola and Kuboň (2004) for example discuss the relative ease of translating between genetically related Slavic languages, typologically related Baltic languages and English. Part of the purpose of this paper is to see if this holds for the languages of the Balkan Sprachbund.

Although this corpus is described as a corpus of the Balkan languages, it is worth noting that it is not comprehensive. It does not for example include the smaller regional and minority languages of the Balkans, such as Aromanian and Romani, nor does it include Slovenian.

This corpus has also been aligned before (Paskaleva, 2007). However, it was only aligned to English, not between Balkan languages. This paper is motivated by the fact that the corpus in Paskaleva (2007) was not made public, and by producing more aligned text between the Balkan languages, not just with respect to English.

2. Data preparation

The South-East European Times (http://www.setimes.com) website is a news site which covers current events in the Balkans in the languages of the Balkans and English. The text content of the site is released as *public domain*, meaning it can be used, modified and redistributed for any purpose without permission. Content has been published starting 2002 and is ongoing.

The website has four main sections: Features, which are mainly longer articles, News Briefs, which are usually shorter articles summarising the news, Articles, which contains news articles somewhat shorter than Features, and Round-up, which is usually a page of extracts of longer articles.

To download all the files we have derived a list of English files using the XENU web-spider.[2]

The unique URL structure of the website made it easy to derive the correspondent language, i.e. en_GB/.../2009/08/07/feature-02 has the corresponding Turkish version at tr/.../2009/08/07/feature-02. Not only could we easily locate the translation , we were also able to apply batch alignments without difficulty.

[1] Here we follow the use of free as defined by the Free Software Foundation; http://www.gnu.org/philosophy/free-sw.html, meaning free to *use*, *modify* and *redistribute* for any purpose – including commercial.

[2] http://home.snafu.de/tilman/xenulink.html

2.1. Collection

After collating the links, pages were downloaded with `httrack`[3] and stripped of HTML with `funduc`[4]. The encoding of the files (variously in ISO-8859-1, ISO-8859-9 and UTF-8) was normalised to UTF-8.

2.2. Sentence splitting

Sentences were split using the `SentParBreaker` splitter.[5] This splitter unfortunately only accepted input in ISO-8859-1 encoding, so a transliteration scheme was devised for languages which were not written in a script which was representable in this encoding (Bulgarian, Greek and Macedonian). Languages which used a different single-byte encoding (Serbian, Croatian, etc.) were transliterated to ISO-8859-1.

2.3. Sentence alignment

Once split, sentences were aligned pairwise between the languages using *hunalign* (Varga et al., 2005).
A preliminary eye-ball evaluation showed the sentence alignment accuracy to be less than perfect. We performed a more complete evaluation of the whole corpus by selecting from each of the alignments one hundred sentences semi-randomly.[6] These alignments were then checked manually and an accuracy figure calculated for each pair.
The results of this evaluation are presented in table 1. For comparison we applied the above method to the EuroParl corpus (Koehn, 2005) alignments for English to Spanish which received a comparable 93% accuracy. These results are not, however, entirely comparable as the SETimes corpus has a smaller number of sentences and the sentences tend to be of a shorter length.

2.4. Common test set

Unlike other parallel corpora covering many languages, the SETimes corpus does not contain all of the text in all of the languages, some translations on the site were missing in some of the languages. In creating the corpus we attempted to maximise the number of aligned sentences in all language pairs, so sentence pairs were included even if they were not translated into all of the languages.
However, in order to effectively evaluate the machine translation systems produced it was desirable to have a subset of sentences which translated into all of the languages as to make the results comparable.
An example sentence from this test set is given with all translations in figure 1.
For the common test set and training set, 1,000 sentences were extracted and the alignments were manually validated. These 1,000 sentences were split into 400 held out and 600 for testing.

[3] `http://www.httrack.com/`

[4] `http://www.funduc.com/search_replace.htm`

[5] `http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector`

[6] The standard program `unsort` was used for this purpose.

Language	Tokens	Types	Ratio
Turkish (`tr`)	34,246,226	139,412	0.40
Croatian (`hr`)	34,968,453	127,756	0.36
Serbian (`sr`)	37,989,711	133,073	0.35
Macedonian (`mk`)	37,623,521	113,393	0.30
Bulgarian (`bg`)	38,419,402	104,669	0.27
Greek (`el`)	41,599,313	105,221	0.25
Albanian (`sq`)	41,741,782	104,322	0.24
Romanian (`ro`)	41,501,934	94,268	0.22
English (`en`)	38,463,808	68,005	0.17

Table 3: Type-token ratio for each of the languages calculated from the raw corpora.

2.5. Corpus statistics

Some simple statistics were calculated over corpus as a whole, and over each of the pairwise alignments. Table 2 gives the number of words in the target language per pair, which is between four million and five million words, excepting pairs with Bulgarian. It is suspected that there was an error in the Bulgarian data that caused this anomaly, possibly due to a sentence in the middle of the data which had no alignment, causing the final sentence extraction script to stop processing.
Table 3 gives the type-token ratio for each of the languages, this presents some kind of measure of their morphological richness, with morphologically rich languages having a larger number of types per token.

3. Machine translation

For each of the language pairs we trained a phrase-based[7] statistical machine translation system using the Moses toolkit (Koehn et al., 2007). The training process followed the instructions for the baseline system in WMT09, the shared task in the ACL 2009 workshop on statistical machine translation (Callison-Burch et al., 2009) with the following changes: The IRSTLM (Marcello et al., 2008) toolkit was used for the target language model, MERT training was skipped due to time constraints, and text was not recased. The language model trained was a five-gram language model was trained on the target language side of the bilingual aligned text. The total training time for the 72 systems, including time to build the language models and binarise the phrase and reordering tables was around ten days.

4. Evaluation

We evaluated the system with BLEU (Papineni et al., 2002), an automatic metric which attempts to measure translation quality by comparing the source text with one or more pre-translated reference texts. While this has been shown to be

[7] The *phrases* in phrase-based statistical machine translation are not syntactic constituents and might be better termed segments or chunks, but here we follow the normal SMT nomenclature.

	bg	el	en	hr	mk	ro	sq	sr	tr
bg	-	97.26	87.14	93.50	98.76	93.15	91.89	89.74	91.54
el	98.63	-	90.27	90.78	97.56	93.50	98.70	92	97.10
en	93.42	92.75	-	100	92.85	98.63	98.59	98.66	98.63
hr	98.55	91.35	95.83	-	92	98.68	96	98.79	100
mk	95.94	97.18	93.84	94.28	-	84.375	87.67	84	97.22
ro	98.48	98.59	100	100	94.20	-	100	98.68	100
sq	92.40	91.54	98.66	97.61	90.41	100	-	91.30	100
sr	89.04	84.81	97.26	100	95.65	96.92	95.94	-	97.84
tr	90.66	90.90	97.46	100	91.54	100	100	98.48	-

Table 1: Percentage of correct alignments out of a semi-randomly selected one hundred for each of the language pairs, with lines containing only formatting excluded.

English:	Ivaylo Markov, 42, was shot dead in his underground parking garage.
Bulgarian:	42-годишният Ивайло Марков бе застрелян в подземния си гараж.
Macedonian:	42-годишниот Ивајло Марков беше убиен во неговата подземна паркинг гаража.
Croatian:	Ivaylo Markov, 42, ubijen je iz vatrenog oružja u svojoj podzemnoj garaži.
Serbian:	Ivajlo Markov, 42, ubijen je iz vatrenog oružja u svojoj podzemnoj garaži.
Greek:	Ο Ιβαίλο Μάρκοφ, 42 ετών, πυροβολήθηκε σε υπόγειο χώρο στάθμευσης.
Romanian:	Ivailo Markov, în vârstă de 42 de ani, a fost ucis prin împușcare în garajul său subteran.
Albanian:	Ivajlo Markov, 42 vjeç u qëllua për vdekje në garazhin e tij nëntokësor.
Turkish:	42 yaşındaki İvaylo Markov yeraltı otoparkında vuruldu.

Figure 1: An example sentence from the manually-validated aligned test set in all nine languages

problematic when comparing different systems (Callison-Burch et al., 2006; Labaka et al., 2007), we consider it to provide a reasonable measure for comparing the quality of translations output by models trained using the same system for different languages on comparable data.

Table 4 shows the results for all of the systems trained. The scores were calculated using the NIST `mteval-v13a` script,[8] and are presented *as output*. Tests for statistical significance have not been made.

It is interesting to note that the scores for Bulgarian are much worse than could be expected, this is probably due to the much lower number of aligned sentences. The lower number of sentences for Bulgarian is probably due to an error in the alignment process, although the alignment validation gave similar alignment quality. This is a subject for further investigation. Other scores are comparable with similar systems.

For an idea of how morphological richness affected translation quality, we calculated the type-token ratio[9] and plotted this against the average BLEU score for translating into the language (see Figure 4.). We consider the type-token ratio a measure of morphological richness, the higher the ratio, the richer the language. For translating into genetically unrelated languages, there is a good correlation between type-token ratio and BLEU score, there is also a good correlation when only considering translation between genetically related languages. For translating from a language, the picture is more clear, and genetic relatedness does not play so much a part, except for the case of Serbian and Croatian where the mean is skewed by the exceptionally high results between these two languages. In fact, considering that for some time, and still to a certain extent these two languages are considered as two written standards of the same language, we calculated the scores for using the Serbian source text as a translation of Croatian and vice-versa. Considering Croatian source text as a Serbian test text gives a score of 0.4114, while the other way around gives a score of 0.4112. This is comparable with the scores of the MT system for *translating* between other languages.

5. Discussion

We have presented, to our knowledge the first pan-Balkan parallel corpus. The corpus is available publically,[10] so that other researchers can reproduce and expand on our results. As the SETimes website continues to publish daily, we intend to continue adding text to the corpus. Targets for the next release will be fixing the Bulgarian data, and rerunning with a Unicode-aware sentence tokeniser.

As can be seen from the results, both translating to and from English gives the best scores for unrelated languages, this could be a result of a number of factors, one is that as English is a very weakly inflected language, the amount of distinct word forms will be lower making translation easier.

[8] Available for download from `http://www.itl.nist.gov/iad/mig/tools/`.

[9] The type-token ratio is the ratio between the number of unique tokens in the corpus and the number of tokens in the corpus.

[10] The full corpus, including the translation models trained can be downloaded from `http://elx.dlsi.ua.es/~fran/SETIMES/` and is mirrored at `http://www.statmt.org/setimes/`.

	bg	el	en	hr	mk	ro	sq	sr	tr
bg	-	3,907,720	4,179,847	3,774,060	3,407,459	4,521,592	4,549,607	4,092,154	4,244,756
el	2,962,995	-	4,920,462	4,454,243	4,822,344	5,360,872	5,385,591	4,857,574	4,463,723
en	2,810,557	5,072,596	-	4,376,267	4,531,170	5,232,684	5,259,457	4,618,166	4,229,978
hr	2,886,604	5,256,393	4,927,971	-	4,712,449	5,302,851	5,336,016	4,775,975	4,322,212
mk	2,193,720	5,005,785	4,498,495	4,085,597	-	4,892,898	4,899,071	4,643,575	4,091,469
ro	2,494,235	5,378,388	4,927,260	4,455,043	4,725,079	-	5,347,961	4,706,146	4,292,371
sq	2,505,685	5,376,278	4,922,668	4,447,139	4,709,619	5,322,065	-	4,715,860	4,296,759
sr	2,738,029	5,029,567	4,666,445	4,335,380	4,793,010	5,062,830	5,098,649	-	4,203,352
tr	3,466,788	5,212,746	4,780,958	4,366,332	4,604,138	5,133,404	5,164,828	4,685,690	-

Table 2: Total number of aligned words per language pair. Number of words words in target language and calculated with the standard `wc` command.

	Target language									
	bg	el	en	hr	mk	ro	sq	sr	tr	Mean from
bg	-	0.1508	**0.2898**	0.1537	0.2501	0.2284	0.2135	0.1560	0.1694	0.2015
el	0.1539	-	**0.4269**	0.2871	0.3979	0.3731	0.3617	0.3051	0.1757	0.3102
en	0.2151	*0.4055*	-	0.3162	***0.4506***	*0.4299*	*0.4464*	0.3477	*0.2090*	0.3526
hr	0.1297	0.3251	0.3952	-	0.4041	0.3478	0.3271	***0.6556***	0.1847	0.3462
mk	*0.2361*	0.3514	**0.4316**	0.3090	-	0.3654	0.3442	0.3376	0.1702	0.3182
ro	0.1735	0.3490	**0.4364**	0.3018	0.3970	-	0.3754	0.3263	0.1894	0.3186
sq	0.1722	0.3760	**0.4908**	0.3011	0.4147	0.4064	-	0.3269	0.1851	0.3341
sr	0.1382	0.3349	0.4016	***0.6524***	0.4118	0.3586	0.3327	-	0.1873	0.3521
tr	0.0566	0.1995	**0.2522**	0.1620	0.2396	0.2215	0.1928	0.1851	-	0.1886
Mean to	0.1235	0.3108	0.3905	0.3104	0.3707	0.3413	0.3242	0.3300	0.1838	-

Table 4: BLEU scores on the test set for all the language pairs. Highest scores translating *to* a language are given in **bold face**, while highest scores translating *from* a language are given in *italics*.

Also, as noted by Virpioja et al. (2007), in morphologically rich languages words include more information on average, and one mistake in a suffix is enough to mark the whole word as incorrect, although it may not prevent understanding. Another factor is that the texts are probably all translated *from* English and not from each other, which could provide a less literal translation, effectively making any translation not aligned with English more of a paraphrase.

While membership of the *Sprachbund* does not seem to have any relationship with the translation quality between typologically related languages, further work might look at how morphological or syntactic information might be included, for example in factored translation models (Koehn and Hoang, 2007), or even look at creating rule-based systems between these languages, which have been shown to outperform phrase-based SMT between related languages (Tyers and Nordfalk, 2009). However, the lack of free morphological and syntactic analysers for the Balkan languages (with the exception of Romanian and Bulgarian) makes this more difficult.

We hope that the release of this corpus can provide a basis for other multilingual projects for the Balkan languages, one could, for instance, envisage a pan-Balkan aligned tagged corpora or treebanks based on this text.

Acknowledgements

This article is inspired by Philipp Koehn's article on EuroParl, any similarities are intentional. We are thankful to Miloš Stolić for his help in validating alignments for the Slavic languages and English. This work has also received the support of the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01.

6. References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. *Proceedings of EACL-2006*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. *EACL 2009 Fourth Workshop on Statistical Machine Translation*, pages 1–28.

Petr Homola and Vladislav Kuboň. 2004. A Translation Model For Languages of Acceding Countries. *Proceedings of the Conference of the European Association of Machine Translation 2004*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and*

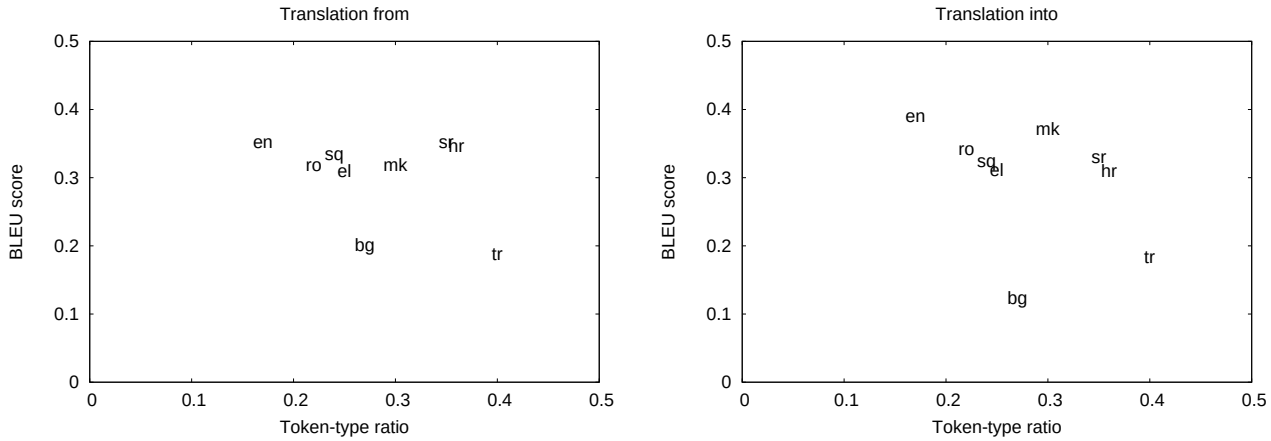

Figure 2: Plots showing token-type ratio versus mean BLEU score for translating into and from a given language. As a trend the token-type ratio increases, the BLEU score decreases. In both cases, Bulgarian is an outlier due to bad training data, or simply less training data, and Macedonian, Serbian and Croatian are outliers due to high translation quality between very closely-related languages regardless of token-type ratio.

Computational Natural Language Learning (EMNLP-CoNLL), pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *ACL 2007, demonstration session*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005*.

Gorka Labaka, Nicholas Stroppa, Andy Way, and Kepa Sarasola. 2007. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal*.

Jouko Lindstedt. 2000. Linguistic Balkanization: Contact-induced change by mutual reinforcement. *Languages in Contact*, 28:231–246.

Federico Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: An open source toolkit for handling large scale language models. *Proceedings of Interspeech 2008*, pages 1618–1621.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Elena Paskaleva. 2007. Balkan South-East Corpora Aligned to English. *Proceedings of the Workshop on Common Natural Language Processing Paradigm for Balkan Languages, RANLP 2007*, pages 35–42.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy, June.

Francis M. Tyers and Jacob Nordfalk. 2009. Shallow-transfer rule-based machine translation for Swedish to Danish. *Proceedings of FREERBMT2009, the First Workshop on Free/Open-Source Rule-based Machine Translation*, pages 28–33.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. *Proceedings of RANLP 2005*, pages 590–596.

Jaakko J. Virpioja, Mathias Creutz Väyrynen, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, September.