INTERNATIONAL WORKSHOP

MULTILINGUAL RESOURCES, TECHNOLOGIES AND EVALUATION FOR CENTRAL AND EASTERN EUROPEAN LANGUAGES

held in conjunction with

The International Conference RANLP - 2009

PROCEEDINGS

Edited by

Cristina Vertan, Stelios Piperidis, Elena Paskaleva
and Milena Slavcheva

Borovets, Bulgaria

17 September 2009

International Workshop

MULTILINGUAL RESOURCES, TECHNOLOGIES AND EVALUATION FOR CENTRAL AND EASTERN EUROPEAN LANGUAGES

PROCEEDINGS

Borovets, Bulgaria
17 September 2009

ISBN 978-954-452-008-3

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Programme Committee

Tomaž Erjavec (Jozef Stefan Institute, Slovenia)

Maria Gavrilidou (ILSP, Greece)**;**

Walther von Hahn (University of Hamburg)

Svetla Koeva (Bulgarian Academy of Sciences)

Cvetana Krstev (University of Belgrad)

Steven Krauwer (University of Utrecht, the Netherlands)

Vladislav Kuboň (Charles University Prague)

Petya Osenova (University of Sofia, Bulgaria)

Elena Paskaleva (Bulgarian Academy of Sciences)

Stelios Piperidis (ILSP, Greece)

Adam Przepiórkowski (IPAN, Polish Academy of Sciences)

Milena Slavcheva (Bulgarian Academy of Sciences)

Marco Tadić (University of Zagreb, Croatia)

Dan Tufiş (Romanian Academy of Sciences)

Cristina Vertan (University of Hamburg)

Duško Vitas (University of Belgrade, Serbia)

Organising Committee

Elena Paskaleva (Bulgarian Academy of Sciences)

Stelios Piperidis (ILSP, Greece)

Milena Slavcheva (Bulgarian Academy of Sciences)

Cristina Vertan (University of Hamburg)

Foreword

The workshop on language processing for Central and Eastern European languages is organised this year for the 4th time in conjunction with the RANLP series of conferences. Looking at the titles of previous editions, one can see that they follow the development which NLP for those languages has faced from one edition of RANLP to the other.

Recent activities in the language technology community in Europe are concerned with the combination/ pipelining of already developed systems and use of very large language resources. This approach assumes that large language resources are available, that systems performances have been evaluated on such resources and that input and output are interoperable with other systems. European initiatives like CLARIN and FLAREET offer the frame for the development of a unified approach for languages all over Europe. For the first time methodologies, evaluation campaigns and roadmaps are planed for all European languages.

Language Processing is now seen as the main technology being able to give people access to information (no matter where it has been produced) in their native languages. Unfortunately, despite important developments, language resources for less popular languages, (especially Balkan and Slavic languages) are still far behind the achieved standard for major western European ones.

As most part of the current Language Technology applications rely on corpus-based methods, one major drawback in the development of language resources and tools for those languages is the lack of training and evaluation data, as well as reference systems for comparing results. Although well-known corpora like JRC-ACQUIS or OPUS are a significant step forward, they

- still do not cover all languages in the Balkan area,
- are collections of documents in specialised languages and therefore decrease the performance of systems trained on those data when testing on another domain.

In order to shorten this bottleneck, it is absolutely necessary to develop, promote and make available all data which can be used for training and evaluation. Additionally, it is important to know which systems have been developed for which applications, on which data have been tested, and what qualitative results have come out.

Therefore the workshop's topic focuses this year on *Multilingual resources, technologies and evaluation for Central and Eastern European languages.*

The selected papers for the current workshop proceedings focus on two issues: adaptation of tools for other languages and multilingual systems and language resources. The eight papers cover ten Central and Eastern European languages.

We would like to thank the authors for contributing to the workshop proceedings and the members of the scientific committee for their quality work. We are grateful to the organisers of RANLP 2009 for hosting this workshop as one of its satellite events. Especially we would like to thank Galia Angelova and Kiril Simov for their great support throughout the whole organisation period.

September 2009

Cristina Vertan, Milena Slavcheva, Stelios Piperidis and Elena Paskaleva

TABLE OF CONTENTS

Ludmila DIMITROVA, Violetta KOSESKA, Danuta ROSZKO and Roman ROSZKO
Bulgarian-Polish-Lithuanian Corpus – Current Development 1

Anca DINU and Liviu P. DINU
On the Behavior of Romanian Syllables Related to Minimum Effort Laws 9

Monica GAVRILA
SMT Experiments for Romanian and German Using JRC-ACQUIS 14

Georgi GEORGIEV, Preslav NAKOV, Petya OSENOVA and Kiril SIMOV
Easy Adaptation of English NLP Tools to Bulgarian 19

Cvetana KRSTEV, Ranka STANKOVIĆ, Duško VITAS and Svetla KOEVA
E-Connecting Balkan Languages 23

David MAREČEK and Natalia KLJUEVA
Converting Russian Treebank SynTagRus into Praguian PDT Style 30

Svetlin NAKOV, Elena PASKALEVA and Preslav NAKOV
A Knowledge-Rich Approach to Measuring the Similarity
between Bulgarian and Russian Words 36

Ivelina NIKOLOVA
New Issues and Solutions in Computer-aided Design
of MCTI and Distractors Selection for Bulgarian 44

Bulgarian-Polish-Lithuanian Corpus – Current Development

Ludmila Dimitrova
IMI-BAS
Acad. G. Bonchev St bl. 8
1113 Sofia, Bulgaria
ludmila@cc.bas.bg

Violetta Koseska
ISS-PAS
ul.Bartoszewicza 1B m.17
00-337 Warsaw
amaz@inetia.pl

Danuta Roszko
ISS-PAS
ul.Bartoszewicza 1B m.17
00-337 Warsaw
danuta.roszko@ispan.waw.pl

Roman Roszko
ISS-PAS
ul.Bartoszewicza 1B m.17
00-337 Warsaw
roman.roszko@ispan.waw.pl

Abstract

This paper discusses the building of the first Bulgarian–Polish–Lithuanian (for short, BG–PL–LT) experimental corpus. The BG–PL–LT corpus (currently under development only for research) contains more than 3 million words and comprises two corpora: parallel and comparable. The BG–PL–LT parallel corpus contains more than 1 million words. A small part of the parallel corpus comprises original texts in one of the three languages with translations in two others, and texts of official documents of the European Union available through the Internet. The texts (fiction) in other languages translated into Bulgarian, Polish, and Lithuanian form the main part of the parallel corpus. The comparable BG–PL–LT corpus includes: (1) texts in Bulgarian, Polish and Lithuanian with the text sizes being comparable across the three languages, mainly fiction, and (2) excerpts from E-media newspapers, distributed via Internet and with the same thematic content. Some of the texts have been annotated at paragraph level. This allows texts in all three languages and in pairs BG–PL, PL–LT, BG–LT, and *vice versa* to be aligned at paragraph level in order to produces aligned three- and bi-lingual corpora. The authors focused their attention on the morphosyntactic annotation of the parallel trilingual corpus, according to the Corpus Encoding Standard (CES). The tagsets for corpora annotation are briefly discussed from the point of view of possible unification in future. Some examples are presented.

Keywords

Bilingual and multilingual corpora, parallel and comparable corpora, corpus annotation, lexical database, bilingual dictionaries.

1. Introduction

Due to the recent development of information and communication technologies and the increased mobility of people around the globe, the number of electronic dictionaries has increased extraordinarily. This concerns, in particular, bilingual dictionaries, in which one of the languages is English. An Internet search shows that no electronic dictionaries exist at all for pairs of languages such as Bulgarian-Polish or Bulgarian-Lithuanian. Traditional printed paper dictionaries are either an antiquarian rarity (the most recent Bulgarian-Polish and Polish-Bulgarian dictionaries were published more than 20 years ago) or have never been published at all (Bulgarian-Lithuanian). It can not be expected however that all people know English to communicate with each other, especially if their native languages (Bulgarian and Polish) belong to the same language family. For the creation of a bilingual electronic or online dictionary for Bulgarian, Polish and Lithuanian an electronic corpus is necessary which will provide the material for lexical database, supporting the dictionary and its subsequent expansion and update. In the recent decades many multilingual corpora were created in the field of corpus linguistics, such as MULTEXT corpus [6], one of the largest EU projects in the domain of language technologies, the MULTEXT-East corpus (MTE for short, annotated parallel and comparable), an extension of the project MULTEXT for Central and Eastern European (CEE) languages [2], Hong Kong bilingual parallel English-Chinese corpus of legal and documentary texts [5], etc.

2. From Bilingual to Trilingual corpus

The MTE project has developed a multilingual corpus, in which three languages: Bulgarian, Czech and Slovene, belong to the Slavic group. The MTE model is being used in the design of the first Bulgarian-Polish corpus, currently under development in the framework of the joint research project “Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary” between Institute of Mathematics and Informatics—Bulgarian Academy of Sciences and Institute of Slavic Studies—Polish Academy of Sciences, coordinated by L. Dimitrova and V. Koseska. This bilingual corpus supports the lexical database (LDB) of the first experimental online Bulgarian-Polish dictionary [3].

2.1 Bulgarian-Polish corpus

The Bulgarian–Polish corpus consists of two parts: a parallel and a comparable corpus [4]. All texts in the corpus are texts published in and distributed over the Internet. Some texts in the ongoing version of the corpus are annotated at paragraph level. The **Bulgarian–Polish parallel corpus** includes two parallel sub-corpora:

1) a *pure* Bulgarian–Polish corpus consists of original texts in Polish – literary works by Polish writers and their translation in Bulgarian, and original texts in Bulgarian - short stories by Bulgarian writers and their translation in Polish.

2) a *translated* Bulgarian–Polish corpus consists of texts in Bulgarian and in Polish of brochures of the EC, documents of the EU and the EU-Parliament, published in Internet; Bulgarian and Polish translations of literary works in third language (mainly English).

The **Bulgarian–Polish comparable corpus** includes texts in Bulgarian and Polish: excerpts from newspapers and textual documents, shown in internet, excerpts from several original fiction, novels or short stories, with the text sizes being comparable across the two languages. Some of the Bulgarian texts are annotated at "paragraph" and "sentence" levels, according to CES [7].

2.2 Bulgarian–Polish–Lithuanian corpus

The first Bulgarian–Polish–Lithuanian (for short, BG–PL–LT) corpus (currently under development only for research) contains more than 3 million words and comprises two corpora: parallel and comparable. The **BG–PL–LT parallel corpus** contains more than 1 million words. A small part of the parallel corpus comprises original texts in one of the three languages with translations in two others, and texts of official documents of the European Union available through the Internet. The texts (fiction) in other languages translated into Bulgarian, Polish, and Lithuanian form the main part of the parallel corpus.

It turned out that it is extremely difficult to find electronic texts of translations from Bulgarian to Lithuanian or *vice versa* – the two languages are spoken by small nations in comparison to other languages of the EU and are distributed in remote areas of Europe. It can be assumed (provisionally of course) that the Polish language 'builds a bridge' between them: for the pairs of languages Bulgarian-Polish and Polish-Lithuanian one can find freely available translations on the Internet.

The comparable BG–PL–LT corpus includes: (1) texts in Bulgarian, Polish and Lithuanian with the text sizes being comparable across the three languages, mainly fiction, and (2) excerpts from E-media newspapers, distributed on the Internet and with the same thematic content.

Some of the texts have been annotated at paragraph level. This allows texts in all three languages and in pairs BG–PL, PL–LT, BG–LT, and *vice versa* to be aligned at paragraph level in order to produces aligned three- and bilingual corpora. "Alignment" means the process of relating pairs of words, phrases, sentences or paragraphs in texts in different languages which are translation equivalent. One may say that "alignment" is a type of annotation performed over parallel corpora. Excerpts of texts of the 3-languages parallel corpus, marked at paragraph level follow:

Bulgarian:

<p>Вместо отговор Гандалф гръмогласно подвикна на коня си:</p>

<p>- Напред, Сенкогрив! Трябва да бързаме. Няма време. Виж! Сигналните клади на Гондор горят, зоват за помощ. Войната е избухнала. Виж, огън бушува над Амон Дин, пламък покрива Ейленах, сигналът бърза на запад: Нардол, Ерелас, Мин-Римон, Каленхад и Халифириен на роханската граница.</p>

Polish:

<p>Zamiast odpowiedzieć hobbitowi, Gandalf krzyknął głośno do swego wierzchowca:</p>

<p>- Naprzód, Gryfie! Trzeba się spieszyć. Czas nagli. Patrz! W Gondorze zapalono wojenne sygnały, wzywają pomocy. Wojna już wybuchła. Patrz, płoną ogniska na Amon Din, na Eilenach, zapalają się coraz dalej na zachodzie! Rozbłyska Nardol, Erelas, Min-Rimmon, Kalenhad, a także Halifirien na granicy Rohanu.</p>

Lithuanian:

<p>Užuot atsakęs Gendalfas garsiai riktelėjo žirgui:</p>

<p>- Pirmyn, Žvaigždiki! Reikia skubėti. Laiko nebeliko. Žiūrėk! Jau dega Gondoro laužai, prašo pagalbos. Karo kibirkštis įžiebta. Matai, ant Amon Dino dega ugnis, liepsnoja ir Eilenachas, dar toliau vakaruose - Nardolas, Erelasas, Minas Rimonas, Kalenhadas ir Halifirienas prie Rohano sienos.</p>

//EN: For answer Gandalf cried aloud to his horse. 'On, Shadowfax! We must hasten. Time is short. See! The beacons of Gondor are alight, calling for aid. War is kindled. See, there is the fire on Amon Dîn, and flame on Eilenach; and there they go speeding west: Nardol, Erelas, Min-Rimmon, Calenhad, and the Halifirien on the borders of Rohan. (Part 3, Book 5 of *The Return of the King* of Tolkien's *The Lord of the Rings*)//

The BG-PL-LT corpus will be annotated according to the standards for morphosyntactic annotation of digital language resources. The main goal in collecting the trilingual corpus is the design and development of a BG–LT digital dictionary based on the BG-PL digital online dictionary.

The corpus will provide a sample of the vocabulary, which is to be included in an initial experimental versions of BG–LT digital dictionary.

We attempt to perform a comparison of the morphosyntactic characteristics of the words of parallel texts across the three languages from the point of view of a possible future unification.

3. Corpus annotation

Corpus annotation is the process of adding linguistic information in an electronic form to a text corpus [7], [8]. We would like to mention the following two most common types of corpus annotation: ***morphosyntactic annotation*** (also called *grammatical tagging* or *part of speech **(POS)** tagging*) and ***lemma annotation*** (where each word in the text is associated with the corresponding lemma). Lemma annotation is closely related to morphosyntactic annotation. Morphosyntactic annotation (POS tagging, where each word in the text is associated with its grammatical classification) is the task of labeling each word in a sequence of words with its appropriate part-of-speech. Words are often ambiguous with respect to their POS; for example, in Bulgarian the neuter singular forms of most adjectives serve double duty as adverbs, for example, BG: *внимателно* //EN: attentive/careful (neuter), attentively/carefully //:

(1) *внимателно* → POS specifications: adjective, Gender: neuter, Number: singular, Definiteness: no.

MTE MorphoSyntactic Descriptor (MSD) for this adjective is A--ns-n.

(2) *внимателно* → POS: adverb, Type: adjectival.

MTE MSD for this adverb is Ra.

The set of POS tags is called tagset. The size and choice of the tagsets vary across languages. The classical POS tagging system is based on a set of parts of speech including noun, adjective, numeral, pronoun, verb, participle, adverb, preposition, conjunction, interjection, particle, and often (depending on the language) article, etc. Of course, morphologically rich languages need more detailed tagsets that reflect to various inflectional categories.

The applications of the morphosyntactic annotation include lexicography, parsing, language models in speech recognition, disambiguation clues for ambiguous words (machine translation), information retrieval, spelling correction, etc.

4. Problems related to POS classification

The POS classification varies across different languages. Often there is more than one possible POS classification for a given language.

Here we would like to show that one cannot formally go about a direct use of the morphosyntactic annotation of a multilingual corpus. An in-depth contrastive study of specific phenomena in the respective languages is necessary. Next we will briefly review the POS classification of the *participle* (one of the important verbal forms) in the three languages, in comparison to another POS, the *adjective*.

4.1 Functions of the participle

The classification of a participle, not only as a verb form, is an important problem: the role of the participle varies significantly across languages, because its properties and functions are different. In contrast to English, for instance, where the participle are invariant, in the Slavic languages the forms of the participles are inflected and contain information about the aspect and tense of the verbal form. As is well-known the information about the aspect is important for the Slavic languages, but does not exist in English. Bulgarian, Polish and Lithuanian distinguish between the following functions of the *participle* form: predicative function, attributive function and adverbial function or semipredicative function, which are illustrated by the following examples:

(1) Examples of predicative function of the participle

BG: *украсен* // PL: *ozdobiony* // LT: *papuošta* [neuter], *papuoštas* [masculine] //EN: decorated//:

BG: *Коридорът е хубаво украсен*.

PL: *Korytarz jest ładnie ozdobiony*.

LT: *Koridorius gerai papuošta. / Koridorius gerai papuoštas.*

EN: *The corridor is beautifully decorated*.

(2) Examples of attributive function of the participle:

BG: *пишещ* // PL: *piszący* // LT: *rašantis //* EN: one who wrote //, in the sentences:

BG: *Пишещият тези писма старец беше осемдесетгодишен.*

PL: *Piszący te listy starzec był osiemdziesięciolatkiem.*

LT: *Rašantis tuos laiškus senelis buvo aštuoniasdešimtmetis.*

EN: *The old man who wrote these letters was eighty years old.*

(3) Examples of the semi-predicative function:

BG: *пишейки* // PL: *pisząc* // LT: *rašydamas* // EN: while writing //, in the sentences:

BG: *Пишейки, гледах през прозореца*.

PL: *Pisząc patrzyłem w okno*.

LT: *Rašydamas žiūrėjau per langą*.

EN: *While writing, I was looking out of the window*.

4.2 Participle and verb

It is important to emphasize that participles preserve some properties of the main form of the verb, such as voice, tense and aspect. In Bulgarian, Polish and Lithuanian there are active and passive participles:

a) Present active participle: BG: *говорещ* // PL: *mówiący* // LT: *kalbąs* / *kalbantis* // EN: *speaking* // (preserved active voice).

b) Present passive participle: BG: *любим*[1] //PL: *kochany* // LT: *mylimas* // EN: *beloved* // (preserved passive voice with information about present tense).

c) Past passive participle: BG: *написан* // PL: *napisany* //LT: *parašytas* // EN: *written* // (preserved passive voice with information about past tense and perfect aspect of the verbal form).

An interesting fact is that participles preserve the valency properties of the respective verbal form, for instance in Polish and Lithuanian:

PL: *Ten mężczyzna zajmuje się drobnym handlem. – Zajmujący sie drobnym handlem mężczyzna.* // LT: *Tas vyras užsiima mažmenine prekyba. – Mažmenine prekyba užsiimantis vyras*. // EN: *This man deals in retail. – A man dealing in retail.*

[1] Colloquial Bulgarian has lost this grammatical category. Such forms occur mostly in scientific writing, being literary loans from Russian or Church Slavonic. Because of their grammatical unproductiveness, they are classified as adjectives, corresponding to the Latin-derived adjectives in *-able*/*-ible* in English: (*не*)*допустим – (in)admissible, недосегаем – intangible, съвместим – compatible, etc.*

The phrase ‘deals in what? / dealing in what?’ requires the instrumental case in Polish and Lithuanian[2]. The valence of the Polish and Lithuanian participle is the same as the valence of the finite verb form.

A comparison of the three languages shows that in Bulgarian a subordinate clause in past perfect tense corresponds to a participle construction in Polish and Lithuanian:

BG: *След като си беше написал домашното, той започна да чете книга*. // PL: *Odrobiwszy lekcje zaczął czytać książkę*. // LT: *Paruošęs pamokas pradėjo skaityti knygą*. // EN: *Having written his homework, he started reading a book*.

Polish has a more modest stock of verbal forms with temporal meaning than Bulgarian or Lithuanian. In any case when the lexical means modifying the temporal meanings are taken into account, the participles, and verbal nouns, it is clear that Polish can express also the same temporal meanings.

4.3 Features of the adjective

Adjectives in Polish and Lithuanian can be declined for gender, number and case (in Bulgarian only for gender and number), but do not express a temporal or aspect relation on their own, unlike the participle. These arguments show that participles deserve a separate treatment from adjectives.

5. Towards development of annotated trilingual electronic resources

Morphosyntactic descriptions for Bulgarian have been developed in several projects, the first of which are for the purposes of corpora processing at the morpho-lexical level in MTE project of EC. The MTE consortium developed morphosyntactic specifications and word-form lexical lists (so called lexicons) covering at least the words appearing in the MTE corpus. For each of the six MTE languages, a lexical list containing at least 15,000 lemmata was developed for use with the morphological analyzer. Each lexicon entry includes information about the inflected-form, lemma, POS, and morphosyntactic specifications. A mapping from the morphosyntactic information contained in the lexicon to a set of corpus tags (used by the POS disambiguator) was also provided, according to the MULTEXT tagging model. The structure of the lexicon entry is the following:

word-form ‹TAB› **lemma** ‹TAB› **MSD** ‹TAB› **comments**

where **word-form** represents an inflected form of the lemma, characterised by a combination of feature values encoded by **MSD**-code (**MSD**: **M**orpho**S**yntactic **D**escription); the fourth (optional) column, comments, is currently ignored and may contain either comments or information processable by other tools. Here is an excerpt from the Bulgarian Lexicon:

обяснение	=	Ncns-n
обяснението	обяснение	Ncns-y
обяснения	обяснение	Ncnp-n
обясненията	обяснение	Ncnp-y

(обяснение ‘explanation’).

The **MSDs** are provided as strings, using a linear encoding; an efficient and compact way for the representation of the flat attribute-value matrices. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way: the character at position 0 encodes part-of-speech; each character at position 1, 2, …, *n*, encodes the value of one attribute (person, gender, number, etc.), using the one-character code; if an attribute does not apply, the corresponding position in the string contains the special marker “-” (hyphen). By convention, trailing hyphens are not included in the **MSD**s. Such specifications provide a simple and compact encoding, and are similar to feature-structure encoding used in unification-based grammar formalisms. When the word form is the very lemma, then the equal sign is written in the lemma field of the entry (“=”).

For Bulgarian the morphosyntactic descriptions were designed on the basis of the traditional POS classification according to the traditional Bulgarian grammar (Bulgarian Grammar 1993). Each word form is assigned a label encoding the major category (POS), type where applicable (e.g., proper *versus* common noun) and inflectional features. Punctuation is also included, as are abbreviations, numbers written in digits, and unidentified objects (residuals). A further non-standard category contains markers of degrees of comparison. Those are formed in Bulgarian with the particles *по* (comparative) and *най* (superlative), preposed to the adjective or adverb but separated from it by a hyphen (*лек* ‘light’, *по-лек* ‘lighter’, *най-лек* ‘lightest’; *леко* ‘easy’, *по-леко* ‘more easily’, *най-леко* ‘most easily’). These particles are annotated as separate words:

по → POS: Particle, Type: comparative, Formation: simple,
най → POS: Particle, Type: superlative, Formation: simple.

The morphosyntactic descriptions for Polish: the description of Polish by Saloni [15] serves as a basis for the morphosyntactic descriptions for Polish and has been adapted to a large degree to the MTE MSD format in [14].

The system of morphosyntactic tags developed for the Polish at the Institute of Computer Science, Polish Academy of Sciences (IPI PAN), is based on a sound methodological foundation comprising linguistic work by authors such as J.S. Bień, Z. Saloni, M.Świdziński. It is

[2] This does not apply to Bulgarian which lacks a case paradigm for nouns.

thanks to this foundation that the IPI PAN's tagset goes beyond the fossilised traditional framework dating back to Aristotle. On the other hand, the MTE tagset, which serves as a point of reference here, is based on the traditional subdivision into parts of speech (this is why, among others, pronouns have been singled out as a part of speech).

Consequently, the aim of our work is neither to revise the good and highly refined IPI PAN tagset nor to replace it with a new tagset for Polish. The issue in question is what kind of compromise should be sought when developing a joint tagset to be used for simultaneous description of the three languages in the BG-PL-LT parallel corpus. For some reasons the MTE tagset (developed previously for many languages) has been selected as the leading one for this corpus. Therefore, the aim of our work is to provide a theoretical study of various categories of Polish (and Lithuanian), to set priorities (e.g. morphological, semantic, syntactic) in identifying various meanings and to provide a classification of morphosyntactic phenomena which does not contradict the MTE standard and does not deviate too strongly from the IPI PAN tagset.

It cannot be excluded that due to the obvious difficulties in achieving consistency of the intertagset the BG-PL-LT corpus will use the IPI PAN tagset for Polish and its modification for Lithuanian. This solution would certainly necessitate a list of more or less close equivalents for the two tagsets: a tagset for Bulgarian on the one hand, and the IPI PAN tagset on the other (for Polish and an extended version for Lithuanian).

It is important to emphasise that only a coherent tagset for a parallel multilingual corpus 1) allows complete linguistic confrontation, 2) enables identification of linguistic facts, 3) enables a search based on pre-defined unambiguous morphosyntactic characteristics.

The morphosyntactic descriptions for Lithuanian: as a basis for morphosyntactic descriptions of Lithuanian serve the Academic grammar of the Lithuanian language [10] and the Functional grammar of Lithuanian [16]. A tool for morphosyntactic annotation for Lithuanian - *MorfoLema* - has been created by Vytautas Zinkevičius in Centre of Computational Linguistics of Vytautas Magnus University (Lithuania) [18]. The program *MorfoLema* can perform a morphosyntactic analysis and generate forms of Lithuanian words based on user's morphosyntactic characteristic. For disambiguation the *MorfoLema* uses „Two-level morphology" method of Kimmo Koskenniemi [9].

The next step of the development of a system for morphological annotation (Morfologinis anotatorius [20]) has been realised by Vidas Daudaravičius and Erika Rimkutė. Vidas Daudaravičius has created disambiguation tools for the *Morfologinis anotatorius*. More information about the *Morfologinis anotatorius* and used set of tags we can find on http://donelaitis.vdu.lt/main.php?id=4&nr=7_1 (in Lithuanian). (The names of tags are in Lithuanian, because the authors of the *Morfologinis anotatorius* didn't use English terms.) It is possible to perform online a morphosyntactic analysis through the web-page http://donelaitis.vdu.lt/main.php?id=4&nr=7_2. The results are visualized on the screen, and it is possible to receive the result as a file.

The tag list for Polish and Lithuanian, based on [11], [12], [13], [17], and used in the example below, follows:

subst - noun
sg – singular
pl – plurale
nom – nominative
gen – genitive
acc - accusative
loc - locative

m - masculine
f - feminine
-hum – nonhuman
-ani – nonanimate

nwok - nonvocal
adj - adjective
verb - verb
praes - present
nonpraet - nonpraeteritum
ter - 3rd person
bezosobnik - non person form of verb
perf - perfective
imperf - imperfective
particle - particle
prep – preposition

A comparison between experimental annotations of the following sentence "*The beacons of Gondor are alight, calling for aid.*[3]" of the parallel corpus was performed:

BG: Сигналните клади на Гондор горят, зоват за помощ.

PL: W Gondorze zapalono wojenne sygnały, wzywają pomocy.

LT: Jau dega Gondoro laužai, prašo pagalbos.

The annotation of the Bulgarian text is done with MTE MSDs, and ISSCO TAGGER [19] is used for disambiguation. For manual annotation of the Polish and Lithuanian text the above-mentioned descriptors are used, because these languages lack developed MTE language specifications. Establishing a 1-1-correspondence between the tags used and the MTE tagset does not present an insurmountable difficulty. The result follows:

Bulgarian (MTE annotation)

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok type=WORD>
<orth>Сигналните</orth>
<disamb><base>сигнален</base><ctag>AP</ctag></disamb>
<lex><base>сигнален</base><msd>A---p-
y</msd><ctag>AP</ctag></lex>
</tok>
<tok type=WORD>
<orth> клади </orth>
<disamb><base>клада</base><ctag>NCFP-N</ctag></disamb>
<lex><base>клада</base><msd>Ncfp-
n</msd><ctag>NCFPN</ctag></lex></tok>
<tok type=WORD>
<orth>на</orth>
<disamb><base>на</base><ctag>SP</ctag></disamb>
```

[3] Tolkien, J.R.R. The Lord of the Rings. Boston : Houghton Mifflin, 1994, p. 731.

```
<lex><base>на</base><msd>Qgs</msd><ctag>QG</ctag></lex>
<lex><base>на</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD>
<orth>Гондор</orth>
<disamb><base>Гондор</base><ctag>NPMS-
N</ctag></disamb>
<lex><base>Гондор</base><msd>Npms-n</msd><ctag>NPMS-
N</ctag></lex>
</tok>
<tok type=WORD >
<orth>горят</orth>
<disamb><base>горя</base><ctag>VMIP3P</ctag></disamb>
<lex><base> горя
</base><msd>Vmia3p</msd><ctag>VMIA3P</ctag></lex>
<lex><base> горя
</base><msd>Vmip3p</msd><ctag>VMIP3P</ctag></lex>
</tok>
<tok type=PUNCT >
<orth>,</orth>
<ctag>COMMA</ctag>
</tok>
<tok type=WORD >
<orth>зоват</orth>
<disamb><base>зова</base><ctag>VMIP3P</ctag></disamb>
<lex><base>зова</base><msd>Vmia3p</msd><ctag>VMIA3P</
ctag></lex>
<lex><base>зова</base><msd>Vmip3p</msd><ctag>VMIP3P</
ctag></lex>
</tok>
<tok type=WORD>
<orth>за</orth>
<disamb><base>за</base><ctag>SP</ctag></disamb>
<lex><base>за</base><msd>Sp</msd><ctag>SP</ctag></lex>
</tok>
<tok type=WORD>
<orth> помощ </orth>
<disamb><base> помощ </base><ctag>NCFS-
N</ctag></disamb>
<lex><base> помощ </base><msd>Ncfs-n</msd><ctag>NCFS-
N</ctag></lex>
</tok>
<tok type=PUNCT>
<orth>.</orth>
<ctag>PERIOD</ctag>
</tok>
</chunk>
</chunkList>
</cesAna>
```

Polish [11]

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok>
<orth>W</orth>
<lex><base>w</base><ctag>prep:loc:nwok</ctag></lex>
</tok>
<tok>
<orth>Gondorze</orth>
<lex><base>Gondora</base><ctag>subst:sg:loc:f</ctag></lex>
</tok>
<tok>
<orth>zapalono</orth>
<lex><base>zapalić</base><ctag>verb:bezosobnik:perf</ctag></l
ex>
</tok>
<tok>
<orth>wojenne</orth>
<lex><base>wojenny</base><ctag>adj:pl:acc:-hum</ctag></lex>
</tok>
<tok>
<orth>sygnały</orth>
<lex><base>sygnał</base><ctag>subst:pl:acc:-hum</ctag></lex>
</tok>
<ns/>
<tok>
<orth>,</orth>
<lex disamb="1"><base>,</base><ctag>interp</ctag></lex>
</tok>
<tok>
<orth>wzywają</orth>
<lex disamb="1"><base>wzywać</base><ctag>
verb:nonpraet:pl:ter:imperf</ctag></lex>
</tok>
<tok>
<orth>pomocy</orth>
<lex><base>pomoc</base><ctag>subst:sg:gen:f</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>
</chunk></chunkList></cesAna>
```

Lithuanian

```
<cesAna version="1.0" type="lex disamb">
<chunkList>
<chunk type="s">
<tok>
<orth>Jau</orth>
<lex><base>jau</base><ctag>particle</ctag></lex>
</tok>
<tok>
<orth>dega</orth>
<lex><base>degti</base><ctag> verb:praes.ter</ctag></lex>
</tok>
<tok>
<orth>Gondoro</orth>
<lex><base>Gondoras</base><ctag>subst:sg:gen:m</ctag></lex
>
</tok>
<tok>
<orth>laužai</orth>
<lex><base>laužas</base><ctag>subst:pl:nom:m</ctag></lex>
</tok>
<ns/>
<tok>
<orth>,</orth>
<lex disamb="1"><base>,</base><ctag>interp</ctag></lex>
</tok>
```

```
<tok>
<orth>prašo</orth>
<lex disamb="1"><base>prašyti</base><ctag>
verb:praes.ter</ctag></lex>
</tok>
<tok>
<orth>pagalbos</orth>
<lex><base>pagalba</base><ctag>subst:sg:gen:f</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>
</chunk>
</chunkList>
</cesAna>
```

6. Annotation of parallel corpus – problems and progress

A parallel corpus of two Slavic languages and one Baltic language is of great interest from the viewpoint of describing the similarities and differences of the formal means of these three languages. Bulgarian belongs to the South subgroup, Polish – to the West subgroup of the Slavic languages. Lithuanian belongs to the Eastern Baltic group. All three languages preserve the special features for each corresponding group.

A significant feature is the analytic character of Bulgarian, and the synthetic character of Lithuanian (with some analytic character, like word order in absolute constructions) and Polish. Bulgarian exhibits several linguistic innovations in comparison to the other Slavic languages (a rich system of verbal forms, a definite article), and has a grammatical structure closer to English, Modern Greek, or the Neo-Latin languages than Polish. The definite article in Bulgarian is postpositive, whereas in Lithuanian a similar function is served by qualitative adjectives and adjectival participial forms, both with pronominal declension. Bulgarian preserves some vestiges of case forms in the pronoun system. Polish and Lithuanian exhibit all features of synthetic languages (a very rich case paradigm for nouns). Although Lithuanian has lost the neuter gender of nouns, its case system is richer than the Polish one. Bulgarian and Lithuanian have a high number of verbal forms, but Polish has reduced most of the forms for past tense. Both Polish and Bulgarian have a strongly developed category of verbal aspect. In Lithuanian the verb can have more than one aspect depending on the usage of a base stem for present, past and future tense.

7. Conclusion

One of the main problems in human communication is the presence of a huge variety of written and spoken languages in the world. Finding ways to support the connection of people from different ethnical parts of the world is becoming more and more important. The advantage of processing a trilingual parallel corpus is to obtain context specific information about syntactic and semantic structures and usage of words in given language(s). The parallel BG–PL–LT corpus will enrich and uncover some unstudied features of the three languages. Furthermore, a trilingual corpus can find applications into the design and development of LDB of future bilingual dictionaries, for example, of a LDB supporting a BG–LT dictionary, based on a LDB that supports a BG–PL online dictionary.

Finally we note that the trilingual corpus can be used in education, in schools as well as universities; it will be useful to students, instructors, and linguists-researchers alike.

8. References

[1] Bulgarian Grammar. (1993). Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

[2] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., and Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: *Proceedings of COLING-ACL '98*. Montréal, Québec, Canada, pp. 315-319.

[3] Dimitrova, L., Panova, R., Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 36-47. ISBN 978-5-9900813-6-9.

[4] Dimitrova, L., V. Koseska-Toszewa. (2008). Some Problems in Multilingual Digital Dictionaries. In: *International Journal Études Cognitives*. 8, SOW, 237–254.

[5] May Fan, Xu Xunfeng. (2002). An evaluation of an online bilingual corpus for the self-learning of legal English. http://langbank.engl.polyu.edu.hk/corpus/bili_legal.html

[6] Ide, N., and Véronis, J. (1994). Multext (multilingual tools and corpora). In *COLING'94*, pages 90-96, Kyoto, Japan.

[7] Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference,*Granada, Spain, 463-70.

[8] Geoffrey Leech. (2004). Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm

[9] Kimmo Koskenniemi. (1983) Two-level morphology: a general computational model for word-form recognition and production. Publication No. 11. Helsinki: University of Helsinki, Department of General Linguistics.

[10] Lithuanian Grammar. (1997). Ed. Vytautas Ambrazas, Baltos lankos, Vilnius, pp.802.

[11] Piasecki, M. (2007). Polish Tagger TaKIPI: Rule Based Constru-ction and Optimisation. Task Quarterly. 11, p. 151-167

[12] Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN, Polonica, XXII-XXIII, p. 57-76 (In Polish)

[13] Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Warszawa: Akademicka Oficyna Wydawnicza EXIT (In Polish)

[14] Roszko, R. (2009). Morphosyntactic Specifications for Polish. Theoretical foundations. In: Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, 15-16 April 2009, Bratislava*. 140–150. ISBN 978-80-7399-745-8.

[15] Saloni, Z., W. Gruszczyński, M. Woliński, R.Wołosz (2007). Słownik gramatyczny języka polskiego, Wiedza Powszechna, Warszawa, CD + 177 s. (In Polish)

[16] Valeckienė, A. (1998). Funkcinė lietuvių kalbos gramatika, Mokslo ir enciklopedijų leidybos institutas, Vilnius, pp.415. (In Lithuanian)

[17] Woliński, M. (2003). *System znaczników morfosyntaktycznych w korpusie IPI PAN*, Polonica, XXII-XXIII, p. 39-55 (In Polish)

[18] Zinkevičius, V. (2000). Lemuoklis - morfologinei analizei. *Darbai ir dienos*, 24, Vytauto Didžiojo universitetas, p. 245-274 (In Lithuanian).

[19] ISSCO TAGGER: http://www.issco.unige.ch/staff/robert/tatoo/tagger.html#design

[20] Morfologinis anotatorius (tagger for Lithuanian): http://donelaitis.vdu.lt/main.php?id=4&nr=7_1

On the behavior of Romanian syllables related to minimum effort laws

Anca Dinu
University of Bucharest
Faculty of Foreign Languages
and Literature
Edgar Quinet 5-7
Bucharest, Romania,
anca_d_dinu@yahoo.com

Liviu P. Dinu
University of Bucharest
Faculty of Mathematics
and Computer Science
Academiei 14, 010014
Bucharest, Romania,
ldinu@funinf.cs.unibuc.ro

Abstract

The main goal of this paper is to investigate the behavior of Romanian syllables related to some classical minimum effort laws: the laws of Chebanow, Menzerath and Fenk. The results are compared with results of similar researches realized for different languages.

Keywords

syllable, minimum effort laws

1 Introduction

In the last decade, the building of language resources (LR) and theirs relevance to practically all fields of Information Society Technologies has been widely recognized. LRs cover basic software tools for their acquisition, preparation, collection, management, customisation and use and are used in many types of applications (from language services to e-learning and linguistic studies, etc.) The relevance of the evaluation for language technologies development is increasingly recognised. On the other hand, the lack of these resources for a given language makes the computational analyzes of that language almost impossible.

The lexical resources contain a lot of data base of linguistics resources like tree banks, morphemes, dictionaries, annotated corpora, etc. In the last years, one of the linguistics structures which regained the attention of the scientific community from Natural Language Processing area was the syllable (Kaplan and Kay 1994, Levelt and Indefrey 2001, Müller 2002, Dinu and Dinu 2005a,b). New and exciting researches regarding the formal, quantitative, or cognitive aspects of syllables arise, and new applications of syllables in various fields are proposed: speech recognition, automatical transcription of spoken language into written language, or language acquisition are just few of them.

A rigorous study of the structure and characteristics of the syllable is almost impossible without the help provided by a complete data base of the syllables in a given language. A syllable data base has not only a passive role of description, but an active role in applications as speech recognition. Also, the psycholinguistic investigation could greatly benefit from the existence of such a data base. These are some of the reasons which provided our motivation for investigating the behavior of Romanian syllable related to some cognitive laws, based on the corpus of Romanian syllable extracted from DOOM.

2 Motivation

The linguists refused to accord to the syllable the status of structural unity of the language, as opposed to the units as the phoneme and the morpheme. As a consequence, the mathematical models of the syllable failed to equal the complexity of the morpheme and phoneme mathematical models.

From the point of view of the language acquisition, the syllables are the first linguistical units learned during the acquisition process. Numerous studies showed that the children's first mental representation is syllabic in nature, the phonetic representation occurring only later.

Each language has its own way of grouping the sounds into syllables, as a result of its structure. The grouping of the syllables takes place depending on the innate psychic inclination of the group. If the vowels in a word are suppressed and only the consonants remain, the word form can be reconstructed with a high probability, when the syllabification of the word is known. This shows that from the existence of the consonant one can deduce the presence of the vowel, so one can determine the graphical form of the syllable an of the whole word. These aspects may have application in cryptography.

Numerous physiological experiments concerning the syllable are realized between the second part of the XIX-th century and the first part of the XX-th century. The experiments from 1899 made by Oussoff showed that the syllable does not always coincide with the respiratory act, because, during a single expiration, more then one syllable can be produced. In 1928 Stetson also showed that the syllable synchronizes with the movement of the thoracic muscles: each new movement of the muscles produces a new syllable (cf. Rosetti, 1963).

The psycholinguistic elements are situated inside the speech production area. Experiments revealed the presence of a library of articulatory pre-compiled routines, which is accessed during the speech production process. In 1994 these observations leaded to the so-called *mental syllabary.* The theory of Levelt and Wheeldon (1994) assumes the existence of this *mental syllabary*: for frequently used syllables there is a library of articulatory routines that is accessed during the process of speech production. The adjoining of such syllabic gesture generates the spoken word and greatly reduce the computational cost of articulatory programs.

These aspects determined us to study and analyze the syllable. In the following we will focus on the lexical (not phonological) aspects of the syllable.

3 Quantitative aspects of the syllable

Opposite to the lack of qualitative insight regarding the syllable, the quantitative, statistic nature of the syllable was intensely studied.

Determining the optimal values of the length of sentences and of the words depending on the certain groups of readers may prove to be very useful in practical application. By optimum value we understand the value for which the level of comprehensibility is the biggest for the class of readers. Knowing this value should be especially important for the teachers and for publishers who print text books. The main conclusion of (Elts and Mikk, 1996) is that, for a good understanding of a text, the length of sentences in the text must be around the average length of sentences. Some optimum values are presented in the next table:

The optimal length of the words (Bamberge, Vanecek, 1984-cf. Elts and Mikk, 1996) (the first row is the readers' level, the second row is the length of words in syllables, and the third row is the length of words in letters):

Reader's level								
4	5	6	7	8	9	10	11	12
1.62	1.68	1.72	1.80	1.88	1.91	1.99	2.08	2.11
6.16	6.39	6.54	6.84	7.15	7.26	7.57	7.91	8.02

Another experiment on 98 students which were given 48 texts, produced the following optimal values:

	Level 8	Level 10
Optimal words (in letters)	8.53	8.67
Optimal sentences (in letters)	71.5	76.0

3.1 On the data base of Romanian syllables

In order to properly investigate the cognitive aspects of the syllables (often embedded in *minimum effort laws*), it is necessary to have a data base of syllables. In [9] a such database of Romanian syllables is presented. We list here some of the main results of this study, with possible cognitive implications. Based on this database, in the next section we will investigate the behavior of Romanian syllables related to some cognitive laws.

Based on the DOOM dictionary, which contains $N_{words} = 74.276$ words, the following series of quantitative and descriptive results for the syllables of Romanian language was extracted ([9]):

1. it was identified $N_{Stype} = 6496$ (*type syllables*) in Romanian language. The total number of syllables (*token syllables*) is $N_{Stoken} = 273261$. So,the average length of a word measured in syllables is $Lwords_{syl} = N_{Stoken}/N_{words} = 273261/74276 = 3,678$.

2. The 74276 words are formed of $N_{letters} = 632702$ letters. So, the average length of a word measured in letters is $Lwords_{let} = N_{letters}/N_{words} = 632702/74276 = 8,518$.

3. In order to characterize the average length of a syllable measured in letters, two cases were investigated:
   (a) the average length of the *token syllables* measured in letters is: $Lsyl_{token} = N_{letters}/N_{Stoken} = 632706/273261 = 2,315$
   (b) The *type syllables* are formed of $N_{Tletters} = 24406$ letters. Thus, the average length of a *type syllable* measured in letters is $Lsyl_{type} = N_{Tletters}/N_{Stype} = 24406/6496 = 3,757$

4. The number of consonant-vowel structures which appear in the syllables is 56. Depending on the type-token rapport, the most frequent consonant-vowel structures are:
   (a) for the *type syllables*: *cvc* (22%), *ccvc* (14%), *cvcc* (10%).
   (b) for the *token-syllables*: *cv*(53%), *cvc* (17%), *v* (8%), *ccv* (6%), *vc* (4%), *cvv* (2%) and *cvcc* (2%).

   It is remarkable that these last 7 structures (i.e. 12% of the 56 structures) cover approximatively 95% of the total number of the existent syllables.

5. the most frequent 50 syllables (i.e. 0,7% of the syllables number N_{Stype}) have 137662 occurrences, i.e. 50,03% of N_{Stoken}.

6. the most frequent 200 syllables cover 76% of N_{Stoken}, the most frequent 400 cover 85% of N_{Stoken} and the most frequent 500 syllables (i.e. 7,7 % of N_{Stype}) cover 87% of N_Stoken. Over this number, the percentage of covering rises slowly.

7. the first 1200 syllables in there frequency order cover 95% of N_{Stoken}.

8. 2651 syllables of N_{Stype} occur only once (hapax legomena).

9. 5060 syllables (i.e. 78%) of N_{Stype} occur less then 10 times. These syllables represent 11960 syllables (4% of N_{Stoken}).

10. 158941 syllables (58% of N_{Stoken}) are formed of 2 letters; the syllables formed of 3 letters represent 27% of N_{Stoken}, those formed of 1 letter represent 9% of N_{Stoken} and those formed of 4 letters represent 6% of N_{Stoken}.

The upper results are similar to other results, from different languages. For Dutch (cf. Schiller et al., 1996), the first 500 *type syllables*, ordered after their frequency,($\approx$ 5% of the total number of *type syllables*), cover approximatively 85% of the total number of *token syllables*. For English, the result is similar, the first 500 syllables cover approximatively 80% of the total number of the *token syllables*. This results support the *mental syllabary* thesis.

4 The laws of Chebanow, Menzerath and Fenk for Romanian syllables

Several studies proposed laws of the *minimum effort type*: the famous Zipf's law, Menzerath's law which states that the bigger the number of syllables in a word, the lesser the number of phonemes composing these syllables. In cognitive economy terms, this means that *The more complex a linguistic construct, the smaller its constituents*. Fenk proposes another three forms of this law:

1. The bigger the length of a word, measured in phonemes, the lesser the length of its constituent syllables, measured in phonemes.

2. The bigger the average length of sentences, measured in syllables, the lesser the average length of syllables, measured in phonemes.

3. There is a negative correlation between the length of sentences, measured in words, and the length of the words, measured in syllables.

In this section we investigate the behavior of Romanian syllables related to the three above mentioned laws.

4.1 Chebanow's law

An intens studied problem in quantitative linguistics was the one regarding the existence of a correlation between the words' length (in syllables) and theirs occurrence's probability. In 1947, Chebanow investigated 127 Indo-European languages and he proposed a Poisson type law for the above problem.

For each particular language, he used a large number of texts to obtain the frequency of words. Denoting by $F(n)$ the frequency of a word having n syllables and by $i = \frac{\sum nF(n)}{\sum F(n)}$ the average length (measured in syllables) of the words, Chebanow proposed the following law between the average i and the probability of occurences $P(n)$ of the words having n syllables:

$$P(n) = \frac{(i-1)^{n-1}}{(n-1)!}e^{1-i}.$$

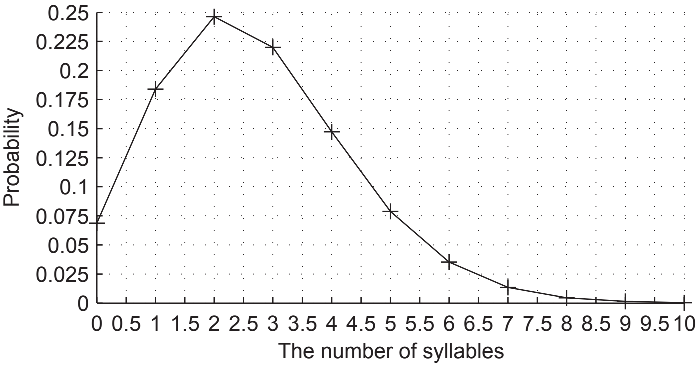


Fig. 1: *The Poisson distribution of length of words (parameter equal to 2.678)*

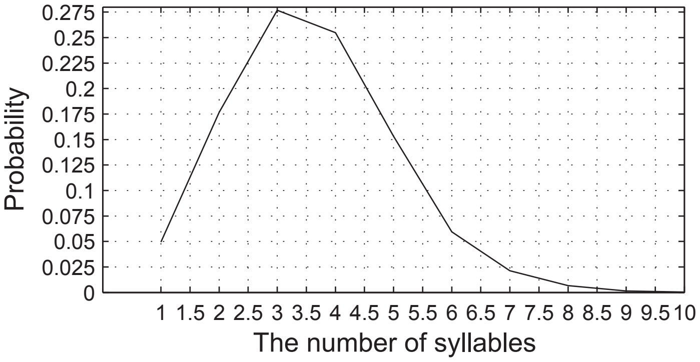


Fig. 2: *The probability distribution of the length of words*

We checked the Chebanow's law on the data base of Romanian syllables and we obtained a strong similarity between the Poisson's distribution (Fig.1) and the distribution of length (in syllables) of words (Fig. 2):

Remark 1 *It is important to see that the graphic from Fig. 2 must be translated with 1 to the left in order to overlap with Chebanow's law (probability $P(n)$ of the words of length n is the Poisson distribution with parameter $n-1$).*

Remark 2 *In the Fig. 1 we represented the following Poisson's distribution (the average length of word is 3.678, so we have to use the value 3.678-1=2.678, cf. Chebanow's law) :*

$$P(n) = \frac{2.678^n}{n!}e^{-2.678}.$$

4.1.1 Menzerath's law

We check the initial Menzerath's law, namely the one regarding a negative correlation between the length of a word in syllables and the lengths in phonemes of its constitutive syllables. The Fig. 3 shows that the law is satisfied.

4.1.2 Fenk's law

Fenk (1993) observed also that the bigger the length of a word, measured in phonemes, the lesser the length of

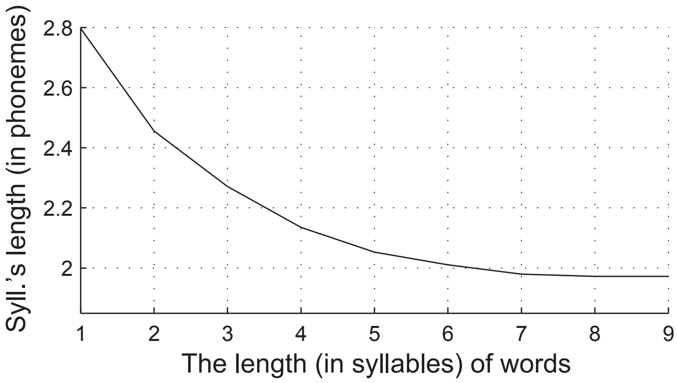


Fig. 3: *The Menzerath's law: The more syllables in a word, the smaller its syllables*

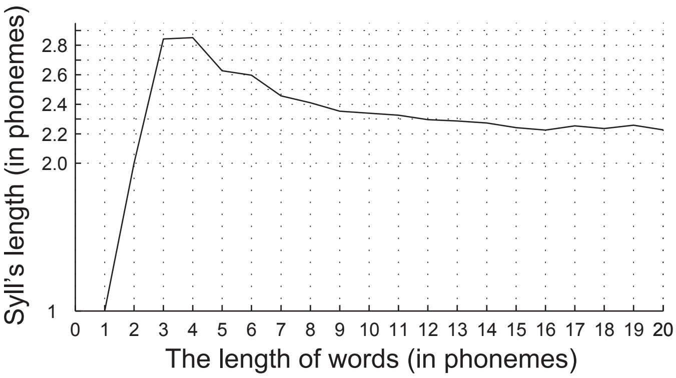


Fig. 4: *The Fenk's law: The more phonemes in a word, the lesser phonemes in its syllables*

its constituent syllables, measured in phonemes. We checked this correlation and the Fig. 4 confirms the first Fenk's law:

5 Conclusion and future works

In this paper we have presented some quantitative observations obtained from the analyze of a data base of Romanian syllables and we checked the behavior of the laws of Chebanow, Menzerath and Fenk for Romanian syllables. All of our results are similar to the results of other researches from different other natural languages (e.g. English, Dutch, Korean, cf. Schiller et. al 199 6, Choi 2000) . In some future work we hope to be able to present results obtained by analyzing a corpus of spoken Romanian language other then the one we used (DOOM) and compare them to the results in this paper.

Acknowledgements: Research supported by CNCSIS, PN2-Idei, Project 228 and University of Bucharest, project IdeiUnibuc 17.

References

[1] Alekseev, P.M. Graphemic and syllabic length of words in text and vocabulary. *Journal of Quantitative Linguistics*, 5, 1-2, 5-12, 1998.

[2] Altmann, G. Prolegomena to Menzerath's law. În *Glottometrika 2*, 1-10, ed. R. Grotjahn, Bochum, 1980.

[3] Altmann, G. Science and linguistics. În *Contributions to quantitative linguistics*, eds. R. Köhler, B. B. Rieger. Kluwer Academic Publishers, Netherlands, 1993.

[4] Chebanow,S.G. On conformity of language structures within the Indoeuropean family to poisson's law. *Comptes rendus de l'Academie de science de l'URSS*. 55(1947), S. 99-102

[5] Choi, S. W. Some statistical properties and Zipf's law in Korean text corpus. *Journal of Quantitative linguistics* 7, 1, 2000.

[6] Dinu, L.P. The alphabet of syllables with applications in the study of rime frequency. *Analele Univ. Bucureşti*, XLVI-1997, 39-44, 1997.

[7] A. Dinu, L.P. Dinu. On the Syllabic Similarities of Romance Languages. In *A. Gelbukh (Ed.): CICLing 2005. LNCS 3406*, 785-788, 2005a.

[8] L. P. Dinu, A. Dinu. A parallel approach to syllabification. In *A. Gelbukh (Ed.): CICLing 2005. LNCS 3406*, 83-87, 2005b.

[9] A. Dinu, L.P. Dinu. On the data base of Romanian syllables and some of its quantitative and cryptographic aspects. In Proceedings LREC 2006, Genoa, Italy, 1795-1798.

[10] *Dicţionarul ortografic, ortoepic şi morfologic al limbii române*. Ed. Academiei, Bucureşti, 1982.

[11] Elts, J., J. Mikk. Determination of optimal values of text. *Journal of quantitative linguistics* 3, 2, 1996.

[12] Fenk, A., G. Fenk-Oczlon. Menzerath's law and the constant flow of linguistic information. În *Contributions to quantitative linguistics*, eds. R. Köhler, B. B. Rieger. Kluwer Academic Publishers, Netherlands, 1993.

[13] Kaplan, R.M. and M. Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3), 331-379, 1994

[14] Levelt, W.J.M., L. Wheeldon. Do speakers have access to a mental syllabary? *Cognition* 50, 239-269, 1994.

[15] Levelt, W.J.M., P. Indefrey. The Speaking Mind/Brain: Where do spoken words come from. În *Image, Language, Brain*, eds. A. Marantz, Y. Miyashita, W. O'Neil, pp. 77-94. Cambridge, MA: MIT Press, 2001.

[16] Marcus, S., Ed. Nicolau, S. Stati. *Introduzione alla linguistica matematica*, Bologna, Patron, 1971.

[17] Markov, A.A. An example of statistical investigation in the text of Eugen Onyegin illustrating coupling of tests in chain. În *Proceedings of the Academy of Science of St. Petersburgh* VI Series, 7, 153-162, 1913.

[18] Menzerath, P. Die Architektonik des deutschen Wortschatzes. În *Phonetische Studien*, Heft 3. Bon: Ferd. Dümmlers Verlag, 1954.

[19] Müller, K. *Probabilistic Syllable Modeling Using Unsupervised and Supervised Learning Methods* PhD Thesis, Univ. of Stuttgart, Institute of Natural Language Processing, AIMS 2002, vol. 8, no.3, 2002

[20] Rosetti, A. *Introducere în fonetică*, Ed. Ştiinţifică, Bucureşti, 1963.

[21] Schiller, N., A. Meyer, H. Baayen. A Comparision of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics*, 3, 1, 8-28, 1996.

SMT Experiments for Romanian and German Using JRC-ACQUIS

Monica Gavrila
Hamburg University
Faculty of Mathematics, Informatics and Natural Sciences
Vogt-Kölln Str. 30, 20251, Hamburg, Germany
gavrila@informatik.uni-hamburg.de

Abstract

One of the LT[1]-applications that ensures the access to the information, in the user's mother tongue, is machine translation (MT). Unfortunately less spoken languages - a category in which the Balkan and Slavic languages can be included - have to overcome a major gap in language resources, reference-systems and tools. In its simplest form, statistical machine translation (SMT) is based only on the existence of a big parallel corpus and therefore it seems to be a solution for these languages. In this paper the performance of a Moses-based SMT system, for Romanian and German, is investigated using test data from two different domains - legislation (JRC-ACQUIS) and a manual of an electronic device. The obtained results are compared with the ones given by the Google on-line translation tool. An analysis of the obtained translation results gives an overview of the main challenges and sources of errors in translation, in these experimental settings.

Keywords

SMT, Romanian,German, Moses, Google in-line translation tool

1 Introduction

"Less interesting languages"[2] have to overcome a major gap in language resources, reference-systems and tools which ensure the development of an MT-system of higher quality. In its simplest form, statistical machine translation (SMT) is based only on the existence of a big parallel corpus, thereby it seems to be a solution for this kind of languages.

From the currently available corpora for the languages considered in the description of the workshop, JRC-ACQUIS is used for the experiments described in this paper. The languages addressed are Romanian and German. The size of bilingual subsets of JRC-ACQUIS differs strongly from language pair to language pair, e.g. for English-German the size of the corpus is over 1 million sentences, for German-Romanian is less than 350000 sentences. Compared to EUROPARL or to the "News Corpus" used in recent investigations in the EUROMATRIX project [1], bilingual subsets in JRC-ACQUIS have approximately six times less aligned sentences, for the language-pair considered.

In this paper the performance of a simplistic Moses-based SMT-system, when trained and tested on JRC-ACQUIS (version 2.2), is investigated. For one of the test-set, data from a small technical corpus is used. The obtained results are compared with the ones given by the Google SMT on-line translation tool. The outcome shows that for less resourced languages - in this case Romanian - the development of further parallel corpora on broader domains and the improvement of the existing resources seem to be unavoidable. In the case of JRC-ACQUIS, for Romanian, such a step has already been done with JRC-ACQUIS Version 3[3].

The paper is organized as follows: in section 2 the used corpora are presented; sections 3 and 4 describe the experiments performed and their results. The last section concludes the presented results.

2 Data Description

For the experiments described in this paper, German-Romanian was chosen as language pair. The tests were done for both directions of translation.

Romanian is a less-resourced language with a highly inflected morphology and high demand for translation after joining the European Union. Compared to widely spoken languages, few resources and tools were developed for Romanian. An overview of tools for Romanian was made in the CLARIN Project (http://www.clarin.eu). Bilingual resources including Romanian are not so many and with few exceptions (see [11], [10]) relate only to English-Romanian.

Few parallel corpora are available, in which one of the language is Romanian, that have a "*satisfactory*" size, and that do not consider, as the other language, only English, e.g. JRC-ACQUIS, OPUS[4].

One of the reasons for using in this paper JRC-ACQUIS is the fact that, to the author's knowledge, all MT experiments, where Romanian was considered,

[1] LT = Language Technology

[2] In this paper, "*less interesting languages*" means less spoken - as number of people - and politically uninteresting

[3] This last version was not used for the experiments, because, as stated on http://langtech.jrc.it/JRC-Acquis.html, at the moment, for this new version for Romanian, alignment information is not available.

[4] For more details on OPUS please see [9] and http://urd.let.rug.nl/tiedeman/OPUS/

are done using this corpus - see [4] and [2][5]. Although on-line or commercial translation tools for Romanian exist[6], they are all black-boxes.

2.1 Training Data

The training corpus is part of the JRC-ACQUIS (http://wt.jrc.it/lt/Acquis/ - last accessed on 18.04.09). Two types of alignments are available on the corpus homepage: Vanilla and HunAlign. The alignments realized with the Vanilla aligner[7] were used for the experiments presented here. Although not the best solution for MT, the alignment provided is done at paragraph-level. A *paragraph* can be a sentence, a sub-sentential phrase (e.g. noun phrase - NP), a phrase, or more sentences. This has an impact on the translation quality, as most of existing systems recommend sentence alignment.

In order to reduce the number of errors, only 1:1 paragraph alignments were considered for the experiments. This means that from 391972 links in 6558 documents, only 324448 links are used for the Language Model (LM). Due to the cleaning step of the SMT system, which limits the sentence length to 40 words[8], the number of 1:1 alignment links considered for the Language Model (LM) are reduced to 238172 links for the Translation Model (TM). This represents 61.38% of the initial corpus. More details on JRC-ACQUIS can be found in [8].

2.2 Test Data

The experiments were run on two different corpora: one is part of the JRC-ACQUIS corpus and the other is part of a technical manual of an electronic device.

897 sentences (299 from the beginning, 299 from the middle and 299 from the end) were removed from JRC-ACQUIS training data, in order to be used as test sets. Sentences were chosen from different parts of the corpus to ensure a relevant lexical, syntactic and semantic coverage. These 3 sets of 299 sentences represent **Test 1**, **Test 2** and **Test 3** of the experiments. As one of the goal of the experiments was to analyze the reaction of the evaluation scores to data-size, **Test 4** data-set contains all 897 sentences. In oder to see how the translation quality changes inside a corpus, several test-sets of the same size, from the same corpus, were chosen.

In order to evaluate the reaction of the SMT system to other input text type, the second test corpus was considered. It is extracted from a manual of an electronic device. It is sentence-aligned and the translation is manually verified. In the corpus dates, numbers and names were replaced by meta-words, e.g. numbers by NUM. Diacritics were not considered. From this corpus 300 sentences from the middle of the text were used as test data - **Test 5** in the experiments.

The detailed statistics on the data are presented in Table 1.

Corpus	No. of words	Vocabulary size	Average sentence length SL
	SL = German		
Training			
JRC-Acquis	3256047	69260	13.6
Test Data			
Test 1	5325	1067	17,8
Test 2	10286	1380	34,4
Test 3	5125	1241	17,23
Test 4	20763	2860	23.14
Test 5	4549	715	15.1
	SL = Romanian		
Training			
JRC-Acquis	3453584	48844	14.5
Test Data			
Test 1	5432	1198	18,16
Test 2	11488	1609	38,42
Test 3	5317	1298	17,7
Test 4	22237	3122	24.79
Test 5	4561	767	15.2

Table 1: *Corpora Statistics*

3 Experimental Settings

The SMT system used follows the description of the baseline system given for the EACL 2009 4th Workshop on SMT[9] and it is based on Moses[10] - see [5]. Wanting to see what results can be obtained by a very simple SMT, two parameters were changed: the tuning step is left out and the LM order is 3.

All test data-sets were translated with the Moses-based system and with the Google on-line translation tool[11]. In both cases, the same metrics were used for evaluation: BLEU and TER. For these experiments the use of other linguistic resources was avoided deliberately, in order to be able to evaluate the robustness of a pure SMT-System at domain change. When changing the domain it is expected that out-of-training-vocabulary words (*OOV-Words*) - especially in domain specific vocabulary - play a major role. In the following subsection this aspect is presented.

3.1 Out-of-training-vocabulary Words

The OOV-words were extracted, for both directions of translation, by comparing the training vocabulary and the test vocabulary for the source language (SL).

[5] The language-pair considered in this papers is Romanian-English. For the author was interesting to use the same corpus, as in previous work also Romanian-English experiments were run. These results were compared with the Romanian-German ones. The Romanian-English experiments are not part of the present paper.

[6] An overview of such systems can be found on www.euromatrix.net/euromatrix, last accessed on 17.06.2009.

[7] (http://nl.ijs.si/telri/Vanilla/ - last accessed 18.04.09.

[8] The sentence size limit is the one recommended for the EACL 2009 4th Workshop on SMT

[9] EACL 2009 Workshop on SMT: http://www.statmt.org/wmt09/index.html - last accessed on 18.04.09.

[10] http://www.statmt.org/moses/ (last accessed on 18.04.09)

[11] More on Google: http://translate.google.de/translate_t?hl=de# - last accessed on 08.05.09 See also [1]

As expected, the percentage of OOV-words, for the technical manual data-set, is higher - see Table 2. As seen in Section 4, the higher number of OOV-words leads to worse translation scores.

When manually analyzing the extracted words, it was noticed that, in the first corpus, due to segmentation and spelling errors and not-replacement of numbers, dates etc with meta-words, sometimes the extracted words are not correct, e.g. "*dreptulde*" (correct: "*dreptul de*" - English: "*the right of*"), or just symbols are extracted, e.g. "*2ev*", "*0155*", "****". After the removal of the wrong extracted words, the number of OOV-words for **Test 4** was reduced to almost 50% for Romanian-German and to 83% for German-Romanian. In Table 2 the number of OOV-words, after the removal procedure, are shown.

The words extracted for the second corpus were 99% right.

Corpus	No. of words	Percentage
	SL = German	
Test 1	47	4.4%
Test 2	37	2.68 %
Test 3	185	14.9 %
Test 4	267	9.3%
Test 5	280	39.16%
	SL = Romanian	
Test 1	17	1.41%
Test 2	20	1.24%
Test 3	82	6.36%
Test 4	130	4.16%
Test 5	279	36.327%

Table 2: *OOV-Words*

4 Experimental Results

In the experiments, due to the lack of multiple references, the comparison with only one reference translation is considered. The following metrics are used:

- BLEU (**bil**ingual **e**valuation **u**nderstudy) - The NIST/BLEU implementation, version 12 [12] is used. Although criticized, BLEU is mostly used in the last years for MT evaluation. It measures the number of n-grams, of different lengths, of the system output that appear in a set of references. More details about BLEU can be found in [6]. As for previous developed systems BLEU is one of the evaluation metrics, for comparison reasons, it is still important to calculate it.
- TER (**t**ranslation **e**rror **r**ate)[13] - It calculates the minimum number of edits needed to get from a obtained translation to the reference translations, normalized by the average length of the references. It considers insertions, deletions, substitutions of single words and an edit-operation which moves sequences of words. More information about TER one can find in [7].

The obtained results are shown in **Table 3**, **Table 4** and **Table 5**.

Score	Test 1	Test 2	Test 3	Test 4
	German - Romanian			
BLEU	**0.2955**	**0.4244**	**0.2884**	**0.3644**
TER	**0.6198**	**0.5905**	**0.6438**	**0.6112**
	Romanian - German			
BLEU	0.2953	**0.4411**	0.2939	**0.3726**
TER	0.6437	**0.5588**	0.6791	**0.6112**

Table 3: *Evaluation Results for the SMT System for the JRC-ACQUIS Test Data*

Score	Test 1	Test 2	Test 3	Test 4
	German - Romanian			
BLEU	0.2853	0.2809	0.274	0.2838
TER	0.6397	0.6707	0.6642	0.6612
	Romanian - German			
BLEU	**0.3277**	0.3301	**0.3208**	0.3332
TER	**0.5971**	0.6590	**0.6576**	0.6425

Table 4: *Evaluation Results for the Google On-line Translation System for the JRC-ACQUIS Test Data*

Score	SMT	Google
	German - Romanian	
BLEU	0.0192	**0.1041**
TER	0.9318	**0.836**
	Romanian - German	
BLEU	0.0223	**0.2242**
TER	0.9358	**0.7434**

Table 5: *Evaluation Results for the for the Manual Test Data - Test 5*

The BLEU scores from Table 3, 4 and 5 are graphically represented in Figure 1.

It is seen from Table 3, 4 and 5 that the BLEU and the TER scores are in all cases correlated.

The interpretation of the results is focused on three directions

1. variations of the evaluation metrics across sets of test data;
2. the comparison with the Google MT on-line tool;
3. manual evaluation.

A Moses-based system, that considers also Romanian and German, is described in [3]. Although not comparable, as the experimental settings are not the same, the BLEU scores reported in this paper are 0.2789 for Romanian-German and 0.2695 for German-Romanian.

[12] mteval_v12, as implemented on www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html - last accessed on 18.04.09

[13] TER as implemented on www.cs.umd.edu/ snover/tercom - last accessed on 18.04.09

4.1 Variation of the scores across different sets of test data

An interesting aspect of the evaluation is the variation of scores across sets of test data from the same corpus, using the same system. The corpus contains data in the time interval 1958-2006. Although terminology might have changed, both languages, Romanian and German, did not suffer major transformations, e.g. at syntactic level.

Several parameters can influence the results of automatic evaluation:

- **The creation of the test data**. As mentioned in Section 2, the test data was extracted from different parts of the aligned corpus. As there is no equal distribution of sentences per year included in the corpus, it might be possible that all sentences related to e.g. 1978 EU-Regulations are in the test data but not in the training data. OOV-words (see Section 3.1) and differences in lexical semantics among years can be in this case source for the variations of the scores.
- **Sentence limitation in the Moses translation model**. This was set to 40 words / sentence. Test data had no restrictions in this sense. The average sentence-length of the test data is higher for both translation-directions (see Table 1)
- **Variation of paragraph length in the alignment**. The 1:1 alignments vary strong in length, some of them are NPs, some of them are 1-verb sentences and some contain more than one sentence.
- **Verification of the test data**. The test data is not manually checked, so that only good and ”relevant” test paragraphs are used. In some test data-sets, paragraphs like ”*Article article_number*” are repeated several times. Sometimes, due to the automatic extraction of the test-sets, the reference translation is wrong (error of the alignment in JRC-ACQUIS).This reduces the BLEU score.
- **Rephrasing**. When manually analyzing part of the translations (see 4.3), it was noticed that some of the translations were correct from the human evaluation point of view, but they rephrased the reference translations. As BLEU calculation is based on n-grams, this leads to a decrease of the score.

4.2 Comparison with Google MT-System

The Google system is stable, i.e. the scores are close to each-other. The BLEU score varies between 0.274 and 0.2853 for German - Romanian (0.0113 score difference) and between 0.3208 and 0.3332 for Romanian - German (0.0124 score difference). The SMT system has the difference between the scores approximately ten times higher, e.g. the BLEU score difference for German - Romanian is 0.136, and for Romanian - German is 0.1472. In order to interpret the results a more detailed manual analysis of the translations is necessary.

For German-Romanian the Moses-based system has a higher BLEU (lower TER) score than the Google one. For Romanian-German on the test-sets of 299 sentence, in two cases out of three, Google has better scores. On the 897 test-set the scores of the Moses-based system are better.

However Google is not a reliable comparison as the system evolves dynamically, by contributions of users and there is no deep information about the architecture of the system. It is estimated that the training data is huge, comparable with the one used for the experiments reported in [1]. In favor of this argument is also the scores obtained for the electronic device corpus. The Google BLEU score is very similar to the one obtained in [1] when changing the domain. In conclusion, the availability of a larger training data set would increase the performance and robustness of a pure SMT-System. Also correcting the training and test data can lead to better results.

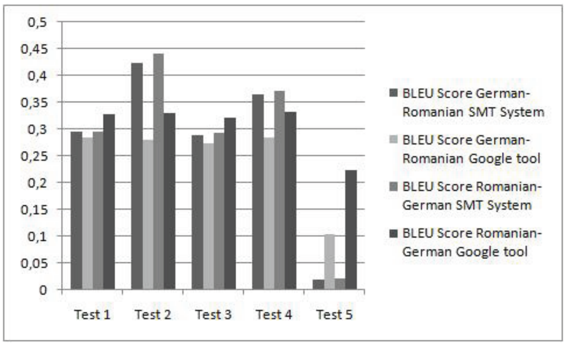


Fig. 1: *BLEU Score*

4.3 Manual Evaluation for German-Romanian

In order to extract the sources of errors, the translations of 100 paragraphs from **Test 4** data-set, obtained by the Moses-based SMT system, were manually analyzed. In order to have different paragraph-types, 50 were chosen from the beginning and 50 from the end. As the human evaluator has as mother tongue Romanian, the translation direction considered was German-Romanian.

If some paragraphs consist of only one word, it was observed that the last 50 paragraphs are longer: e.g. *paragraph 863* has 82 words and consists of one phrase and two sentences. There are 49 paragraphs shorter than 6 words.

The eight sources of translation errors are presented in **Table 6**. Some errors (e.g. OOV-words) presented in **Table 6** are due to the limited training data. Due to the German compounds and syntax, an important source of errors is the word alignment. These errors can be solved by adding more data or a bilingual dictionary.

In around 10% of the paragraphs, the translation was adequate and fluent, but it was the reference translation rephrased - e.g. passive voice translated

Error	Frequency	Explanation / Example
OOV-words	35 cases	Compounds or part of compounds ”*Forschungsfonds*” (”*Research fonds*”) Sometimes only half of the compound word is translated ”*anpassungsprotokoll*” (”*the Protocol adjusting...*”) translated as ”***protocolul*** *anpassungs**protokoll***” instead of ”*protocolul de adaptare*”
Punctuation		wrong position of ”*)*”
Prepositions	10%	wrong or word-to-word translation” ”*in das Abkommen*” (”*into the Agreement*”) translated as ”*din acord*” (”*from the agreement*”)
Agreement, case	12%	
Missing words	23 cases nouns, articles or prepositions	Missing definite article for genitive
Missing verb	14%	This is due to the German syntax Distance between the auxiliary and main verb Subordinate sentences
Extra words	less than 5%	
Word order	around 15%	
Wrong translation (semantics)	20 cases	

Table 6: *Manual Evaluation: Sources of Errors (% means percentage from the number of paragraphs; case means the appearance of the phenomenon (i.e. in one paragraph there can be more cases)*

as active voice - or it contained synonyms. This influences negatively the automatic evaluation

5 Conclusion

In this paper the performance of an SMT system based on Moses is investigated on test data from different domains for German - Romanian, in both directions. No additional linguistic tools were used. The article presents the comparison between the results of the Moses-based SMT system and the ones given by the Google on-line translation tool. The training corpus used is the JRC-ACQUIS. The test data are taken from the JRC-ACQUIS corpus and from a manual of an electronic device.

In the described experimental settings, in all cases for German-Romanian and in some cases for Romanian-German, the Moses-based SMT system, trained and tested on the same data type, scores better than the Google on-line tool. This, in spite of the fact that both languages are inflected, and that the corpus (JRC-ACQUIS) is small and includes errors.

In the other cases, with increased and better - i.e. sentence-aligned - training data, the Google performance can be reached with a Moses-based SMT-System. As it is not a black-box system, one has the possibility to control the workflow, and introduce in a targeted way linguistic components when available.

References

[1] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. Findings of the 2009 workshop on statistical machine translation. In *In Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March, 30-31 2009.

[2] D. Cristea. Romanian language technology and resources go to europe. Presentated at the FP7 Language Technology Informative Days, January, 20-11 2009. To be found at: ftp://ftp.cordis.europe.eu/pub/fp7/ict/docs/language-technologies/cristea_en.pdf - last accessed on 10.04.2009.

[3] C. Ignat. *Improving Statistical Alignment and Translation Using Highly Multilingual Corpora.* PhD thesis, INSA - LGeco- LICIA, Strasbourg, France, June, 16th 2009. It can be found on: http://sites.google.com/site/cameliaignat/home/phd-thesis - last accessed on 3.08.09.

[4] E. Irimia. Experimente de traducere automata bazată pe exemple pentru limbile engleza/romana. In *In Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române Workshop Proceedings*, pages 131–140, Iaşi, Romania, November 2008. Publisher: Ed. Univ. Alexandru Ioan Cuza, ISSN: 1843-911X.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June 2007.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine translation and evaluation*, pages 311 – 318, Philadelphia, Pennsylvania, 2002. Publisher: Association for Computational Linguistics Morristown, NJ, USA.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, 2006.

[8] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 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, May, 24-16 2006.

[9] J. Tiedemann and L. Ngygaard. The opus corpus - parallel and free. In *Proccedings of the 4th International Conference of Language Resources and Evaluation*, Lisbon, Portugal, May 26-28 2004.

[10] D. Tufiş, S. Koeva, T. Erjavec, M. Gavrilidou, and C. Krstev. Building language resources and translation models for machine translation focused on south slavic and balkan languages. In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pages 145–152, Dubrovnik, Croatia, September 25-28 2008. In Marko Tadi, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.).

[11] C. Vertan, W. von Hahn, and M. Gavrila. Designing a parole/simple german-english-romanian lexicon. In *In Language and Speech Infrastructure for Information Access in the Balkan Countries Workshop Proceedings - RANLP 2005*, Borovets, Bulgaria, September 2005.

Easy Adaptation of English NLP Tools to Bulgarian

Georgi Georgiev[*]
georgi.georgiev@ontotext.com

Preslav Nakov[†]
nakov@comp.nus.edu.sg

Petya Osenova[‡]
petya@bultreebank.org

Kiril Simov[‡]
kivs@bultreebank.org

Abstract

Nowadays, the dominating approach in natural language processing (NLP) is to acquire linguistic knowledge using machine learning methods. To demonstrate this approach for Bulgarian and to create a meaningful baseline, we automatically adapted to Bulgarian the OpenNLP tools, an open source machine learning NLP tool suite. The baseline of natural language processing components for Bulgarian is crucial since there is a lack of publicly available data based on systematic studies and toolsets of NLP components for well-defined comparisons. This intelligible baseline for Bulgarian uses the well-known framework of Maximum Entropy and features that have been found useful for other languages. As a first evaluation of a machine learning toolset of NLP components for Bulgarian, we demonstrate that the performance of OpenNLP's sentence splitter, tokenizer, part-of-speech tagger, shallow parser and parser can score near the state-of-the-art performance for other languages.

Keywords

part-of-speech tagging, parsing, chunking, tokenizing, sentence splitting

1 Introduction

Nowadays, the dominant approach in natural language processing (NLP) is to acquire linguistic knowledge using machine learning methods. Other approaches, e.g., relying on manual rules and constrains, have proven to be time-costly and error-prone. Still, using machine learning has a major limitation: it requires manually annotated corpora as training data, whose creation can be quite costly. Fortunately, for Bulgarian such a rich resource already exist (BulTreeBank[1]), which makes machine learning possible. In this paper, we stipulate that language adaption should be no harder than domain adaptation [2]. Similarly to the described in [2], we focus on the OpenNLP toolset[2] since it is open source and contains variety of platform-independent implementations of crucial NLP components (written in Java). Moreover, it is based on a single machine learning approach – maximum entropy (ME) [1]. In our experiments below, we focused on the following five basic components from the toolset: sentence detection, tokenization, POS tagging, chunking and parsing.

Maximum entropy models search for a distribution $p(x|y)$, that is consistent with the empirical observations about the feature representation $f(x, y)$, computed from the training examples $\mathcal{T} = \{x, y\}$ (e.g., a sentence x and its label y, see [6] for details). From all such distributions the one is chosen, that has the highest entropy [1]. It could be shown that the resulting distribution will have the following form:

$$p_w(y|x) \propto \exp(w \cdot f(x, y)). \tag{1}$$

The features in the OpenNLP framework combine diverse contextual information such as words around the endings of a sentence for the sentence splitter, or word, character n-grams and part-of-speech tag features alone and in combinations for the chunker. These components are based partially on the publications for chunking of Sha and Pereira [7] as well as the part-of-speech tagger and the parser described in the dissertation of Ratnaparkhi [6]. In our experiments for training and testing the described components, we use the training and testing sections of the BulTreeBank corpus [5, 8] without further modifications.

2 Converting the BulTreeBank XML to Penn Treebank-style Bracketing

Converting the BulTreeBank XML [8, 5] to Penn Treebank-style bracketing format is straightforward with some exceptions for which we define custom tools. Consider for example the following sentence

```
Всички на някакъв етап от живота си сме
изправени пред проблеми и предизвикателства .

'All at some point of life itself we face
 to problems and challenges.'

We all at some point of our lives face
problems and challenges.
```

[*] Ontotext AD, 135 Tsarigradsko Ch., Sofia 1784, Bulgaria

[†] Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417

[‡] Linguistic Modelling Laboratory, Institute for Parallel Processing, Bulgarian Academy of Sciences, 25A Acad. G. Bonchev St., 1113 Sofia, Bulgaria

[1] http://www.Bultreebank.org/

[2] http://OpenNLP.sourceforge.net/

that posseses the following simplified structure in a BulTreeBank XML format:

```
<S>
 <VPA>
  <VPS>
   <Pron>
    <w ana="Pce-op">Всички</w>
   </Pron>
   <PP>
    <Prep>
     <w ana="R">на</w>
    </Prep>
   </PP>
   <NPA>
    <NPA>
     <A>
      <w ana="Pfa--s-m">някакъв</w>
     </A>
     <N>
      <w ana="Ncmsi">етап</w>
     </N>
    </NPA>
    <PP>
     <Prep>
      <w ana="R">от</w>
     </Prep>
     <N>
      <w ana="Ncmsh">живота</w>
     </N>
     <Pron idref="id1">
      <w ana="Psxto">си</w>
     </Pron>
    </PP>
   </NPA>
   </PP>
   <VPC>
    <V>
     <w ana="Vxitf-r1p">сме</w>
    </V>
    <Participle>
     <w ana="Vppptcv--p-i">изправени</w>
    </Participle>
    <PP>
     <Prep>
      <w ana="R">пред</w>
     </Prep>
     <CoordP>
      <ConjArg>
       <N>
        <w ana="Ncmpi">проблеми</w>
       </N>
      </ConjArg>
      <Conj>
       <C>
        <w ana="Cp">и</w>
       </C>
      </Conj>
      <ConjArg>
       <N>
        <w ana="Ncnpi">предизвикателства</w>
       </N>
      </ConjArg>
     </CoordP>
    </PP>
   </VPC>
  </VPS>
 </VPA>
 <pt>.</pt>
</S>
```

This sentence is transformed to the following Penn Treebank-style bracketing structure:

```
((S
 (NP
  (PRP Всички)
 )
 (VP
  (PP
   (IN на)
   (NP
    (NP
     (PRP някакъв)
     (NN етап)
    )
    (PP
     (IN от)
     (NN живота)
     (PP$ си)
    )
   )
  )
  (VP
   (VB сме)
   (VB изправени)
   (PP
    (IN пред)
    (NP
     (NP
      (NNS проблеми)
     )
     (CC и)
     (NP
      (NNS предизвикателства)
     )
    )
   )
  )
 )
 (. .)
```

The process of transformation applies some simple rules for the coordinations in BulTreeBank for instance "CoordP" and "ConjArg" typically become "NP", and "Conj" becomes a "CC" phrase (see [5] for BulTreeBank syntactic phrase naming conventions and [4] for Penn Treebank). We also remove the outer verb phrases in BulTreeBank – in the above case the phrases "VPA" and "VPS" – as required by the Penn Treebank bracketing structure. We further defined specific rules for pronouns. For pronoun phrases defined as

```
<Pron><w ana="P...">....
```

in BulTreeBank the "ana" tag is very important [9]. First, in case the fourth position is filled by a "t" and the tag is "Ps*t*"[3], e.g., the tag starts with "Ps" this is a *possessiveform* and is part of the NP phrase in the transformation structure. Such expressions are:

[3] Here * stands for a character that is not important for the discussion.

хубавата ми кола

‘beautiful-the my-clitic car’

my beautiful car

and also:

майка ми

‘mother my-clitic’

my mother

Second, if the tag does not start with “Ps”, e.g., the “s” does not appear on second position in the tag, then the pronoun is part of the verb phrase, because it is a personal pronoun. is no “t” on position four, but there is “l” or “-” instead, we annotate this as an NP (see the example above).

We also reduce the BulTreeBank format [9] to a much smaller one with just 95 tags. In most cases, the tags collapsed to smaller ones by losing some of the surface morphology forms. For instance, in the example sentence the word “*си*” (”our own-clitic”) is annotated with the tag “Psxt” (pronoun, possessive, reflexive, short form), which is transformed to “PP$” (pronoun, possessive). The last word in the example sentence “*предизвикателства*” (challenges) has the tag “Ncnpi” (noun, common, neuter, plural, indefinite) and is collapsed to “NNS” (noun, plural). In some cases the tags are directly transformed, e.g., the word *от*” (from) with tag “R” (preposition) is transformed to the tag “IN” (preposition or subordinating conjunction).

3 Results

The results are shown in Table 1. As we can see, the sentence splitter achieved an F1 score of approximatively 92.54%. The false positives constituting the majority of errors constantly appear in complicated sentences rather than abbreviations of organization names. For instance, the tagger annotated the following chunk as a sentence:

Кой беше този човек?, би запитал той
Така че...

Who was that guy?, He would ask
So ...

However the actual sentence should be:

Кой беше този човек?, би запитал той

Who was that guy?, He would ask

Some errors appear in sentences annotated with direct speech, for instance the sentence tagger annotated the follows piece of text as a sentence:

Наведен напред, Той впери поглед в мрака
и рече: - Да бъде светлина.

Inclined forward, he took a look in the dark
and said: - Let there be light.

Inclined forward, he stared at the gloom
and said: - Let there be light.

However the sentence in BulTreeBank is:

Наведен напред, Той впери поглед в мрака
и рече:

Inclined forward, he took a look in the dark
and said:

Inclined forward, he stared at the gloom and said:

The tokenizer achieved F1=98.49%. The majority of the errors appear in sparse abbreviations. For instance, the abbreviation for kilograms, “*кг.*”, is frequently tokenized as “*кг*”. The abbreviation for vehicle “horse power”, “*к.с.*”, is tokenized as “*к.с*”. Some errors involve words that contain no space, e.g., “*предсказание*” (prediction), which are wrongly tokenized as “*казание*” (i.e., an old Bulgarian word that means “statement”). In a similar fashion, there are very rare errors in names of people and locations, e.g., the name “*Лазаръс*” (Lazarus) is tokenized as the Bulgarian name “*Лазар*” (Lazar), and the city “*Казанлък*” (Kazanlak) is tokenized as “*Казан*” (Kazan).

The part-of-speech tagger achieved F1=90.34% on the full morpho-syntactic tag set of BulTreeBank. We benefit from the tagger’s ability to include a tag dictionary that has been automatically generated from the training data. The dictionary enumerates a fixed set of possible tags for each word, which severely limits the number of decision options per word. The majority of the errors are not in a completely wrong morpho-syntactic tag; rather only part of the morphology forms are wrong. For instance, the word “*чорбаджиите*” is wrongly annotated as “noun, common, feminine, plural, definite”, rather than the correct “noun, common, masculine, plural, definite”.

wrong annotated
‘‘*чорбаджиите*’’ (gaffer) with POS=Ncfpd

correct annotation
‘‘*чорбаджиите*’’ with POS=Ncmpd

Another common error is the wrong annotation of proper nouns as common nouns. One example is the person name “*Странджата*:

wrong annotated
‘‘*Странджата*’’ (Strandjata) with POS=Ncfsd

correct annotation
Странджата’’ with POS=Npfsd

Another location example is the word “*Балканът*” (The Balkan mountain):

wrong annotated
’’*Балканът*’’ (Balkanut) with POS=Ncmsd

correct annotation
’’*Балканът*’’ with POS=Npmsd

Sentence splitter	Tokenizer	POS Tagger	Chunker	Parser
92.54	98.49	94.43	84.6	77.56

Table 1: ***The F1 scores (in %) of the OpenNLP components discussed in the study.***

We also collapsed the morpho-syntactic tagset to a much smaller one, containing 95 tags (see Section 2). The resulting POS tagger achieved F1=94.43%. We used this tagger in the following experiments on chunking and parsing.

The chunker achieved an F1 score of 84.6% on the BulTreeBank corpus. The result seems satisfactory since there is no adaptation of the chunker features to Bulgarian. It should be also noted that the result for English reported at the CoNLL-2000 competition was about 84-85%. However, these results must be treated with caution since there are some issues when using the ChunkLink[4] script, which was written especially for the English tagset and not for the collapsed BulTreeBank set (see Section 2).

The parser is evaluated with the standard metrics: Bracketing recall and Bracketing precision [3] on all sentences. The F1 score on the BulTreeBank with the collapsed tagset is 77.56%. We were unable to evaluate the parser with the full BulTreeBank morphosyntactic tagset since this approach will need direct re-coding of the parser implementation for some part-of-speech groups (e.g. nouns).

4 Conclusions

We have presented our experiments in adapting five important components of the OpenNLP toolset to Bulgarian: sentence detection, tokenization, POS tagging, chunking and parsing. We have further presented the first evaluation of variety of NLP components for Bulgarian within the same machine learning framework: maximum entropy. The resulting evaluation is crucial since there is a lack of publicly available data based on systematic studies and toolsets of NLP components for well-defined comparisons. Finally, we have defined a high baseline for Bulgarian using the well-known framework of Maximum Entropy using features that have been found useful for other languages. As a first evaluation of a machine learning toolset of NLP components for Bulgarian, we have demonstrated that the performance of OpenNLP's sentence splitter, tokenizer, part-of-speech tagger, shallow parser and parser can score near the state-of-the-art performance for other languages.

References

[1] A. Berger, S. D. Pietra, and V. D. Pietra. A maximum entropy approach to natural languageprocessing. *Computational Linguistics*, 22:39–71, 1996.

[2] E. Buyko, J. Wermter, M. Poprat, and U. Hahn. Automatically adapting an nlp core engine to the biology domain. In *Joint BioLINK-Bio-Ontologies Meeting*, pages 65–68, 2006.

[3] M. Collins. Three generative, lexicalised models for statistical parsing. In *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[4] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19, 1993.

[5] P. Osenova and K. Simov. Btb-tr05: Bultreebank stylebook. Technical report, BulTreeBank Project Technical Report, 2004.

[6] A. Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* PhD thesis, University of Pennsylvania, 1998.

[7] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 134–141, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[8] K. Simov. Btb-tr01: Bultreebank project overview. Technical report, BulTreeBank Project Technical Report, 2004.

[9] K. Simov, P. Osenova, and M. Slavcheva. Btb-tr03: Bultreebank morphosyntactic tagset. Technical report, BulTreeBank Project Technical Report, 2004.

[4] http://ilk.kub.nl/~sabine/chunklink

E-Connecting Balkan Languages

Cvetana Krstev
Faculty of Philology
University of Belgrade
cvetana@matf.bg.ac.rs

Ranka Stanković
Faculty of Mining and Geology
University of Belgrade
ranka@rgf.bg.ac.rs

Duško Vitas
Faculty of Mathematics
University of Belgrade
vitas@matf.bg.ac.rs

Svetla Koeva
Dep. of Computational Linguistics
Institute for Bulgarian
svetla@dcl.bas.bg

Abstract

In this paper we present a versatile language processing tool that can be successfully used for many Balkan languages. For its work, this tool relies on several sophisticated textual and lexical resources that have been developed for most Balkan languages. These resources are based on several *de facto* standards in natural language processing.

Keywords

Query expansion, e-dictionaries, wordnets, proper names, aligned texts

1. Introduction

The software tool WS4LR (shortened for WorkStation for Language Resources) has been developed by the Language Technology Group organized at the Faculty of Mathematics for several years now. Its first version was introduced in 2004 [8] and it dealt mainly with harmonizing various heterogeneous lexical resources. Subsequently, many new features have been added, particularly those that have helped in the production and exploration of aligned texts on the basis of the lexical resources incorporated [9]. The new tool WS4QE (shortened for Work Station for Query Expansion) was developed on the basis of WS4LR that enables the expansion of queries submitted to the Google search engine [10]. The integrated lexical resources enable modifications of users' queries for both monolingual and multi lingual searches.

When presenting WS4LR and WS4QE, we have always stressed that, although they have been mainly used for Serbian, they are by no means language dependent as long as compatible lexical resources exist for any two languages. Nevertheless, the full potential of these tools was until now used only for Serbian, and in bilingual context, for Serbian and English.

In this paper we will show that the tools WS4LR and WS4QE are truly independent both from Serbian, for which they were initially developed, and from English which seems to be in the background of many natural language processing tools. The main presupposition for the usage of these tools for other languages is the existence of textual and lexical resources developed in the same methodological framework. Since this prerequisite has been satisfied for Bulgarian, and, to some extent, for some other Balkan languages (Greek, Romanian, etc…), we will therefore show that WS4LR and WS4QE can be successfully used for these languages.

WS4LR, written in C#, is organized in modules which perform different functions. The core of the system WS4LR_Core comprises four .Net libraries, used by two components: the stand-alone windows application WS4LR.exe and the web service wsQueryExpand.asm. Web application WS4QE.asp manages user query request, than uses web service in order to expand user query, submits the expanded query to Google search engine and finally presents retrieved result.

2. Integrated Language Resources

In order to prove the usability of WS4LR and WS4QE for languages other than Serbian and English, we have used various resources, both textual and lexical. In the following sections we will briefly present these resources, what methodological framework was used for their development, and how they were integrated for their successful usage.

2.1 Textual Resources – Aligned Texts

The aligned texts as a special form of multilingual corpora were the focus of many projects in the past few decades. A systematic approach to the development of multilingual corpora was initiated within the Multext project, which subsequently included East-European languages through the Multext-East project [5]. In the meantime, many multilingual corpora have been compiled from large corpora, usually fully automatically prepared, which have a range comprised from texts in the limited technical domain [18] to more versatile literary corpora [5] that are often more modest in size but minutely prepared.

The main textual resource used to explore WS4LR is Jules Verne's novel *Around the World in Eighty Days*. This text was chosen for various reasons. First of all, the text is available in digital form for the majority of European languages, including Balkan languages. Regarding its content, it represents a suitable text for different types of analysis, especially in the domain of named entity recognition (geographical concepts and different measures). Besides this, it has already been used for some interesting research, e.g. multi-word tagging [13] and building models for machine translation [21]. Finally, from a practical point of view, its suitability stems from the fact that it presents a sample text for the French distribution of the Unitex system [15].

Versions of the novel in fifteen different languages have been acquired, but not all of these texts have yet been aligned. Among the already aligned texts are the French original and translations in English and four Balkan languages – Serbian, Bulgarian, Greek, and Romanian.

In the preparatory phase each translation was marked in accordance with the TEI-standard in XML, and the title (<head>), paragraph (<p>) and segments (<seg>) were included as units of a text logical layout. At the beginning of the alignment process, all segments coincided with sentences automatically tagged by Unitex. The XAlign system [1] was used for the alignment process. Starting from the French version, the goal of the alignment was to establish 1:1 relations with all other languages on the segmental level. In order to achieve this goal and after manually checking all aligned segments, some of them had to be divided into smaller units, and some were grouped into larger units. Thus, we arrived at the total of 4,409 segments in all texts. This way, the missing segments or the inconsistencies between the source text and its translations were identified in most of the cases. In the following example the English segment is given only for the sake of translation.

```
<tu id=" n2941">
  <seg lang="en">
      <s id="Verne80days.n2941">
        Between Omaha and the Pacific, the railway crosses a territory which is still infested by Indians and wild beasts, and a large tract which the Mormons, after they were driven from Illinois in 1845, began to colonise.</s></seg>
  <seg lang="fr">
      <s id="Verne80days.n2941">
        Entre Omaha et le Pacifique, le chemin de fer franchit une contrée encore fréquentée par les Indiens et les fauves, -- vaste étendue de territoire que les Mormons commencèrent à coloniser vers 1845, après qu'ils eurent été chassés de l'Illinois.</s></seg>
  <seg lang="sr">
      <s id="Verne80days. n2941">
        Između Omahe i Tihog okeana pruga prolazi kroz predeo u kome još ima Indijanaca i divljih zveri - prostranu zemlju koju su počeli naseljavati mormoni oko 1845. godine, kada su ih prognali iz države Ilinois.</s> </seg>
  <seg lang="bg">
      <s id="Verne80days. n2941">
        Между Омаха и Тихия океан железопътната линия прекосява район, все още населяван от индианци и диви зверове. Това е обширна територия, която мормоните са започнали да колонизират около 1845 г., след като са били прогонени от щата Илинойс.</s></seg>

  <seg lang="gr">
      <s id="Verne80days. n2941">
        Ανάμεσα στην Ομάχα και στον Ειρηνικό, το τρένο διασχίζει περιοχές όπου συχνάζουν ακόμα Ινδιάνοι και αγρίμια - τεράστια εδαφική έκταση την οποία αρχισαν να αποικίζουν οι μορμόνοι μετά το 1845, οπότε κυνηγήθηκαν από το Ιλινόις.</s></seg>
  <seg lang="ro">
      <s id=" Verne80days.n569">
        între Omaha şi Pacific drumul de fier trece printr-o regiune populatã încã de indieni şi fiare, - vastã întindere pe care mormonii au început s-o colonizeze pe la 1845 dupã ce au fost izgoniţi din Illinois.</s>
</tu>
```

2.2 Morphological Dictionaries in LADL Format

Morphological dictionaries are a necessary resource in various phases of the automatic analysis of a text. The tool WS4LR expects morphological dictionaries to be in the format known as DELAS/DELAF presented in [2] that was developed in LADL (*Laboratoire d'Automatique Documentaire et Linguistique*) under the guidance of Maurice Gross. The format of a DELAS-type dictionary basically consists of simple word lemmas accompanied with inflectional class codes which allow for the production of a DELAF-type dictionary, which consists of all inflectional forms with their grammatical information. In the Unitex environment, one finite-state transducer responsible for the generation of all inflectional forms of each DELAS lemma corresponds to each inflectional class code. The Serbian morphological dictionary of simple words contains 121,000 lemmas which yield the production of approximately 1,450,000 different lexical words. Close to 87,000 simple lemmas belong to the general lexica, while the remaining 34,000 lemmas represent various kinds of simple proper names [11]. The Bulgarian Grammar dictionary (DELAS dictionary) consists of 127,000 lemmas distributed as follows: app. 85,000 simple lemmas belong to the general lexis, app. 6,000 lemmas represent domain specific lexis and app. 36,000 lemmas are simple proper names. The corresponding DELAF dictionary consists of app. 1,260,000 entries [7].

2.3 Semantic Networks - Wordnet

Semantic networks, seen as one important node in the hierarchy of ontologies, are used more and more in various phases of the automatic analysis of texts. The tool WS4LR expects them to be in the form of wordnets, that is, nodes representing sets of synonymous words (synsets) which are linked by various semantic relations. The first built wordnet was an English wordnet, the so-called Princeton Wordnet (PWN), having today approximately 140,000 synsets. Due to its remarkable size and successful inclusion in various computer-based applications, it is considered as the de facto standard upon which wordnets for many other languages have been built. One successful application of this concept was achieved by the Balkanet project which was funded by the European Commission from (2001-2004). In the scope of this project, the development of wordnets for the following Balkan languages was initiated [20]: Bulgarian, Greek, Romanian, Serbian, and Turkish. The important feature of these wordnets is that they are all aligned with PWN via the Interlingual index (ILI) [22]. Namely, ILI consists of concepts, while wordnets represent the lexicalization of concepts in various languages and the way which they are connected.

The Serbian wordnet today consists of more than 15,000 synsets built by app. 25,000 literals. All of them are linked to PWN, except for 532 Balkan specific concepts that are connected with other Balkan languages, and 155 Serbian specific concepts that remain unconnected with other languages. The Bulgarian wordnet consists of more then 31,000 synsets built by more than 66,000 literals. The synsets are linked with the PWN as well; again there are 436 Balkan specific concepts shared with other Balkan languages and 182 Bulgarian language specific concepts. Both Serbian and Bulgarian wordnets, as well as wordnets for other Balkan languages, are represented in WS4LR using common XML schema.

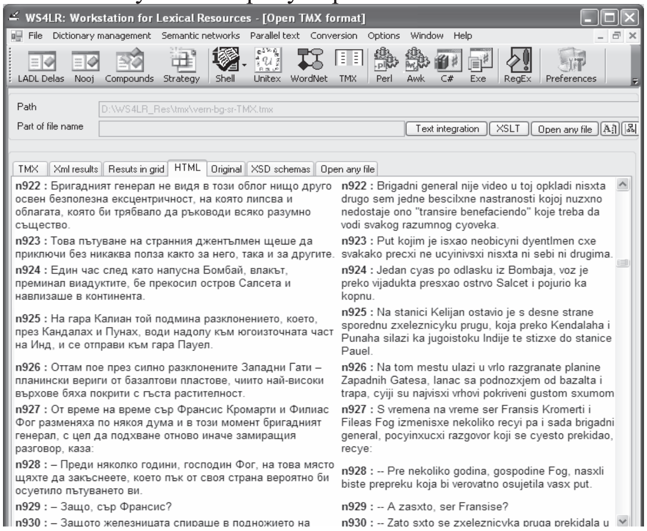
2.4 The Prolex Database

The *Prolex project* was initiated in the 1990's with the study of toponyms in French and had the aim of appropriately processing proper names in natural language applications [16]. This work was followed by the development of a Serbian version, which finally led to the design and construction of a relational multilingual dictionary of Proper Names, the Prolexbase, in the form of a relational database [19]. This model is based on two main concepts: the *pivot* (that represents the *conceptual proper name*) at a language independent level and the *prolexeme* (the projection of the pivot onto a particular language) which is a set of lemmas that includes the name, but also its aliases (variations in orthography, abbreviated forms, acronyms, etc.) and its derivatives. For instance, if a meronymy relation is established between the concepts "New York" and "the United States of America", then their Serbian Latin equivalents *Njujork* and *Sjedinjene Američke Države*, their Serbian Cyrillic equivalents *Њујорк* and *Сједињене Америчке Државе*, and their Bulgarian equivalents *Ню Йорк* and *Съединени американски щати* are all connected automatically.

3. Using WS4LR with Aligned Texts

The WS4LR module that works with aligned texts expects them to be in the Translation Memory eXchange (TMX) format[1]. It can transform texts previously aligned by XAlign into this format as well as into several other formats: textual, XML and tabular. This is particularly important since XAlign has been integrated into Unitex software starting from its version 2.1. In addition, the user can also produce various visualizations of aligned texts by applying appropriate XSLT transformations. Thus, the user can freely browse with such visualized texts. One such visualization is represented in Figure 1.

Browsing, however, is not a particularly successful form of text exploration. The WS4LR module for aligned texts offers the user the ability to posit different forms of queries that can be automatically expanded by using various bilingual lexical resources presented in the previous section. WS4LR offers the user the possibility to expand the query not only morphologically and semantically, but also to another language. If the first language is Serbian, the second language can be English, Bulgarian, or any other. A user can choose two working languages by adjusting the parameters in the "Preferences" menu of WS4LR. Besides this, WS4LR provides further possibilities for the user to control their query formulation, since in addition to expansion it also offers the query to be narrowed. Namely, a user can reject some of the automatically offered query expansions.



Path D:\WS4LR_Res\tmx\vern-bg-sr-TMX.tmx

Part of file name

n922 : Бригадният генерал не видя в този облог нищо друго освен безполезна ексцентричност, на която липсва и облагата, която би трябвало да ръководи всяко разумно същество.	**n922 :** Brigadni general nije video u toj opkladi nisxta drugo sem jedne bescilxne nastranosti kojoj nuzxno nedostaje ono "transire benefaciendo" koje treba da vodi svakog razumnog cyoveka.
n923 : Това пътуване на странния джентълмен щеше да приключи без никаква полза както за него, така и за другите.	**n923 :** Put kojim je isxao neobicyni dyentlmen cxe svakako precxi ne ucyinivsxi nisxta ni sebi ni drugima.
n924 : Един час след като напусна Бомбай, влакът, преминал виадуктите, бе прекосил остров Салсета и навлизаше в континента.	**n924 :** Jedan cyas po odlasku iz Bombaja, voz je preko vijadukta presxao ostrvo Salcet i pojurio ka kopnu.
n925 : На гара Калиан той подмина разклонението, което, през Кандалах и Пунах, води надолу към югоизточната част на Инд, и се отправи към гара Пауел.	**n925 :** Na stanici Kelijan ostavio je s desne strane sporednu zxeleznicyku prugu, koja preko Kendalaha i Punaha silazi ka jugoistoku Indije te stizxe do stanice Pauel.
n926 : Оттам пое през силно разклонените Западни Гати – планински вериги от базалтови пластове, чиито най-високи върхове бяха покрити с гъста растителност.	**n926 :** Na tom mestu ulazi u vrlo razgranate planine Zapadnih Gatesa, lanac sa podnozxjem od bazalta i trapa, cyiji su najvisxi vrhovi pokriveni gustom sxumom
n927 : От време на време сър Франсис Кромарти и Филиас Фог разменяха по някоя дума и в този момент бригадният генерал, с цел да подхване отново иначе замиращия разговор, каза:	**n927 :** S vremena na vreme ser Fransis Kromerti i Fileas Fog izmenisxe nekoliko recyi pa i sada brigadni general, pocyinxucxi razgovor koji se cyesto prekidao, recye:
n928 : – Преди няколко години, господин Фог, на това място щяхте да закъснеете, което пък от своя страна вероятно би осуетило пътуването ви.	**n928 :** -- Pre nekoliko godina, gospodine Fog, nasxli biste prepreku koja bi verovatno osujetila vasx put.
n929 : – Защо, сър Франсис?	**n929 :** -- A zasxto, ser Fransise?
n930 : – Защото железницата спираше в подножието на	**n930 :** -- Zato sxto se zxeleznicyka pruga prekidala u

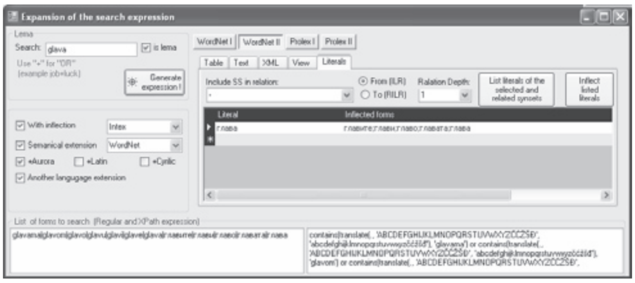
Figure 1. The HTML View of the Aligned Bulgarian-Serbian Text

User queries can be semantically expanded by the wordnets and by the Prolex database. WS4LR obtains the semantic expansion of a query by means of the wordnet of the first language (the Serbian wordnet – SWN, as is the case in our examples), selecting all synsets containing a given word and offering them to the user. This provides the user with an insight into all the concepts the keyword pertains to, through sets of synonyms used for these concepts. A user then gains the possibility to delete some of these synsets if he or she decides that they pertain to concepts which are not of interest at that particular moment. Also, the user can formulate a bilingual query by adding a second language to it. Namely, WS4LR can, for a given set of concepts, identify all corresponding concepts in the second language wordnet by using ILI. Thus, for an expanded Serbian query, one could obtain the corresponding expanded query in Bulgarian. The form used to bilingually expend a simple query *glava* "head" with the Bulgarian *глава* is presented in Figure 2. The semantic expansion is obtained by

[1] For details on TMX format see http://www.lisa.org/tmx/tmx.htm

checking the box “Semantic extension” in this form and by choosing the appropriate resource (Wordnet in this case), while the bilingual expansion is obtained by checking the box “Another language extension”.

In the same form, the user can choose to morphologically inflect all chosen keywords in both languages. If he or she wishes to do so the box “With inflection” should be checked. Morphological expansion is performed by Unitex modules that use morphological dictionaries of simple words as well as inflectional transducers. This options works only if a particular query keyword is listed in the morphological dictionary of the corresponding language. If this is not so, the aligned text will be searched only with the original keyword. As shown in Figure 2, the automatically added inflected forms of chosen keywords are presented in an editable form in which some of these inflected forms can be deleted or modified. For instance, the Serbian word *put* “path” has two plural forms: *putevi* and *puti*. The second one is restricted to poetical usage and a user can choose to delete it from the expended query if the working text is not of that kind.



☑ With inflection: Intex
☑ Semanical extension: WordNet
☑ +Aurora □ +Latin □ +Cyrilic
☑ Another language extension

Ralation Depth: 1

Literal	Inflected forms
глава	главите;глави;главо;главата;глава

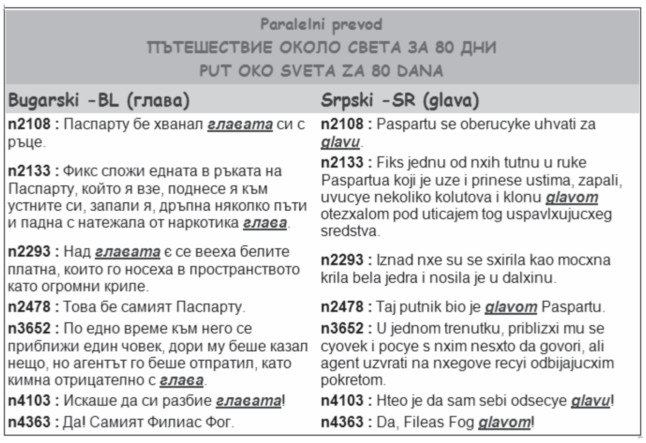
glavama|glavom|glavo|glavu|glavi|glave|glava|главите|глави|главо|главата|глава

contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'glavama') or contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'glavom') or contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ',

Figure 2. The original query keyword *glava* is shown in the upper left corner. The chosen query expansions are shown on the left side. The query expended by the Bulgarian wordnet is shown on the right side, together with the automatically obtained list of inflected forms that can be edited. The two fields at bottom show the final query set.

Finally, when a query is launched, the result is obtained with all the retrieved occurrences highlighted (see Figure 3).

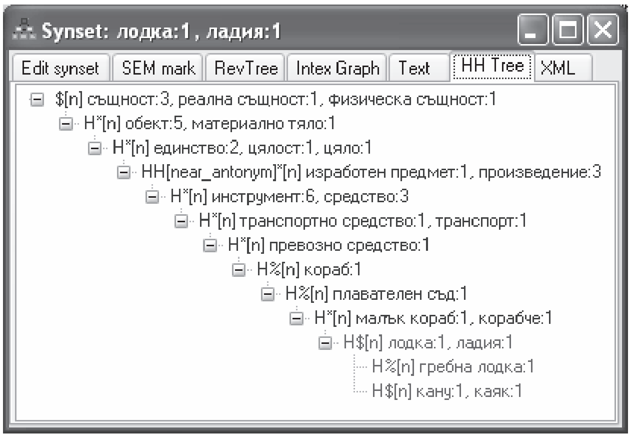
The query can be further semantically expanded by the choice of a particular semantic relation (e.g. hypernymy/hyponymy), in which case synsets pertaining to hypernyms/hyponyms of concepts from the initial group will also appear among the query set. This feature will be illustrated by a query which starts with the Serbian keyword *brodić* “small boat”. We would like to perform a bilingual search with a semantic expansion. The chosen Serbian keyword belongs to only one synset {brodica:1, brodić:1} whose corresponding Bulgarian synset is {лодка:1, ладия:1}. Figure 4 shows that these synsets are deep in the hypernymy/hyponymy hierarchy. In such a situation, expending query with hypernym synsets can be useful.



Paralelni prevod
ПЪТЕШЕСТВИЕ ОКОЛО СВЕТА ЗА 80 ДНИ
PUT OKO SVETA ZA 80 DANA

Bugarski -BL (глава)	Srpski -SR (glava)
n2108 : Паспарту бе хванал ***главата*** си с ръце.	**n2108 :** Paspartu se oberucyke uhvati za ***glavu***.
n2133 : Фикс сложи едната в ръката на Паспарту, който я взе, поднесе я към устните си, запали я, дръпна няколко пъти и падна с натежала от наркотика ***глава***.	**n2133 :** Fiks jednu od nxih tutnu u ruke Paspartua koji je uze i prinese ustima, zapali, uvucye nekoliko kolutova i klonu ***glavom*** otezxalom pod uticajem tog uspavlxujucxeg sredstva.
n2293 : Над ***главата*** є се вееха белите платна, които го носеха в пространството като огромни криле.	**n2293 :** Iznad nxe su se sxirila kao mocxna krila bela jedra i nosila je u dalxinu.
n2478 : Това бе самият Паспарту.	**n2478 :** Taj putnik bio je ***glavom*** Paspartu.
n3652 : По едно време към него се приближи един човек, дори му беше казал нещо, но агентът го беше отпратил, като кимна отрицателно с ***глава***.	**n3652 :** U jednom trenutku, priblizxi mu se cyovek i pocye s nxim nesxto da govori, ali agent uzvrati na nxegove recyi odbijajucxim pokretom.
n4103 : Искаше да си разбие ***главата***!	**n4103 :** Hteo je da sam sebi odsecye ***glavu***!
n4363 : Да! Самият Филиас Фог.	**n4363 :** Da, Fileas Fog ***glavom***!

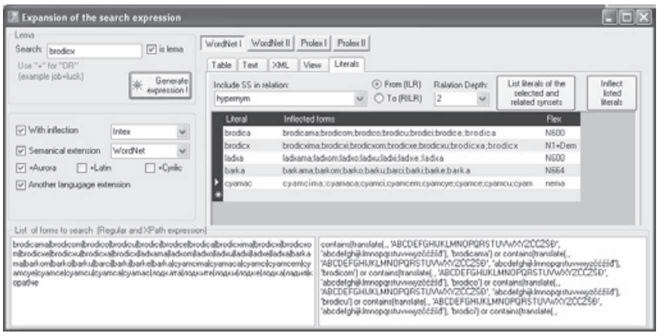
Figure 3. Some representative examples of aligned segments with the keywords *glava* and *глава* and their inflectional forms in HTML format.



- $[n] същност:3, реална същност:1, физическа същност:1
  - H*[n] обект:5, материално тяло:1
    - H*[n] единство:2, цялост:1, цяло:1
      - HH[near_antonym]*[n] изработен предмет:1, произведение:3
        - H*[n] инструмент:6, средство:3
          - H*[n] транспортно средство:1, транспорт:1
            - H*[n] превозно средство:1
              - H%[n] кораб:1
                - H%[n] плавателен съд:1
                  - H*[n] малък кораб:1, корабче:1
                    - H$[n] лодка:1, ладия:1
                      - H%[n] гребна лодка:1
                      - H$[n] кану:1, каяк:1

Figure 4. The hypernym/hyponym wordnet hierarchy of the Bulgarian synset {лодка:1, ладия:1}. The corresponding Serbian synset belongs to a similar tree.

Figure 5. shows the query expansion form in which the original query *brodić* is expanded, not only with a literal from its corresponding synset *brodica*, but also with the literals from synsets belonging to the hypernym branch of the length two, that are {barka:1, čamac:1, čun:1} “boat” and {lađa:1} “vessel”.



☑ With inflection: Intex
☑ Semanical extension: WordNet
☑ +Aurora □ +Latin □ +Cyrilic
☑ Another language extension

Include SS in relation: hypernym
Ralation Depth: 2

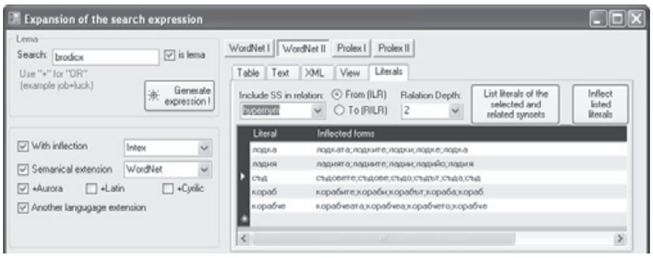
Literal	Inflected forms	Flex
brodica	brodicama;brodicom;brodico;brodicu;brodici;brodice;brodica	N600
brodicx	brodicxima;brodicxi;brodicxom;brodicxe;brodicxu;brodicxa;brodicx	N1+Dem
ladxa	ladxama;ladxom;ladxo;ladxu;ladxi;ladxe;ladxa	N600
barka	barkama;barkom;barko;barku;barci;barki;barke;barka	N664
cyamac	cyamcima;cyamaca;cyamci;cyamcem;cyamcye;cyamce;cyamcu;cyam	nema

brodicama|brodicom|brodico|brodicu|brodici|brodice|brodica|brodicxima|brodicxi|brodicxom|brodicxe|brodicxu|brodicxa|brodicx|ladxama|ladxom|ladxo|ladxu|ladxi|ladxe|ladxa|barkama|barkom|barko|barku|barci|barki|barke|barka|cyamcima|cyamaca|cyamci|cyamcem|cyamcye|cyamce|cyamcu|cyamca|cyamac|[illegible] opадие

contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'brodicama') or contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'brodicom') or contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'brodico') or contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'brodicu') or contains(translate(., 'ABCDEFGHIJKLMNOPQRSTUVWXYZČĆŽŠĐ', 'abcdefghijklmnopqrstuvwxyzčćžšđ'), 'brodici') or contains(translate(.,

Figure 5. In the query expansion form, the user can choose the type of semantic relation for the expansion and the length of the path with this relation he or she wishes to pursue.

Since in this case a bilingual search is initiated, the user can perform the same semantic expansion for the second language, presented in Figure 6. The two Bulgarian literals thus obtained are *плавателен съд* and *малък кораб* which are multi-word units. Since inflection of multi-word units for Bulgarian is not yet integrated in WS4LR, as will be explained in the final section, the user can choose to delete it from the final query set or to keep only the nouns *съд* and *кораб*, as we have done in our example search.



Expansion of the search expression

Lema

Search: brodicx ☑ is lema

Use "+" for "OR" (example job+luck)

Generate expression !

☑ With inflection: Intex

☑ Semanical extension: WordNet

☑ +Aurora ☐ +Latin ☐ +Cyrilic

☑ Another langugage extension

WordNet I | WordNet II | Prolex I | Prolex II

Table | Text | XML | View | Literals

Include SS in relation: hypernym ● From (ILR) ○ To (RILR)

Ralation Depth: 2

List literals of the selected and related synsets

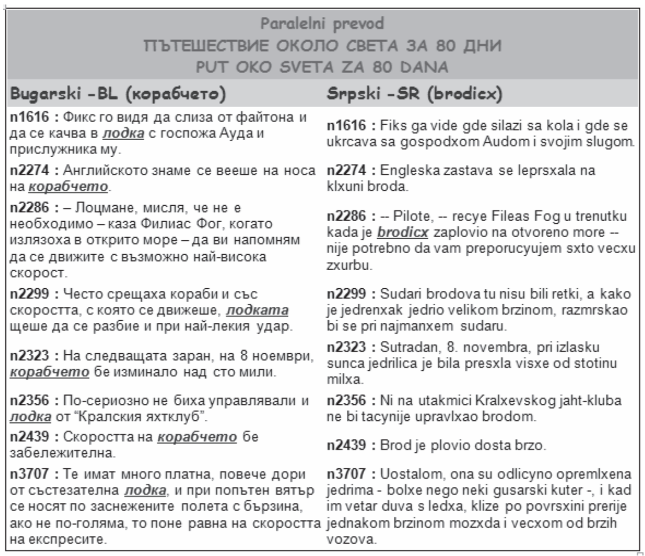
Inflect listed literals

Literal	Inflected forms
лодка	лодката;лодките;лодки;лодке;лодка
ладия	ладията;ладиите;ладии;ладийо;ладия
съд	съдовете;съдове;съдо;съдът;съда;съд
кораб	корабите;кораби;корабът;кораба;кораб
корабче	корабчеата;корабчеа;корабчето;корабче

Figure 6. The semantic expansion in the second language – Bulgarian – using hypernym relation

The results obtained by this query are very interesting and, show the potential this tool offers for various linguistic and literary researches on their own. This query retrieved 129 aligned segments, each of which contained at least one of the keywords from the produced query set in at least one of the languages. It comes as a surprise that only 8 of these segments contained query keywords in both languages. This is mainly due to the fact that adjectives *плавателен* and *малък* were omitted from the Bulgarian keywords thus broadening the query on the Bulgarian side too much. There were 5 segments with the keyword *съд*, with two occurrences of *плавателен съд* “vessel”; none of them corresponded to the Serbian wordnet equivalent *lađa*. There were also 90 occurrences of *кораб* among which there was not one *малък кораб*; in this case, however, the Serbian equivalent for *кораб* was almost unmistakably *brod*, as suggested by both wordnets.

Figure 7 shows some examples of a partial retrieval. The first (n1616) and third (n2286) segments in this sample occur due to the fact that the reference to a “boat” is missing in one of the languages. The other segments show that the Serbian *brod*, besides corresponding to the English *ship* and the Bulgarian *кораб*, is also a generic notion and should probably be added to the hypernym synset (segments n2274, n2356 and n2439). On the other hand, Serbian *jedrilica* and *jedrenjak* “sailing vessel” are translated in Bulgarian with the “sister” synsets *кораб* or *корабче* instead of using a more specific Bulgarian word *платноход* (segments n2299 and n2323). In the last example (n3707), in Bulgarian a rather arbitrary choice *лодка* is made for a more specific type of a vessel referred to in Serbian as *kuter* “cutter”.

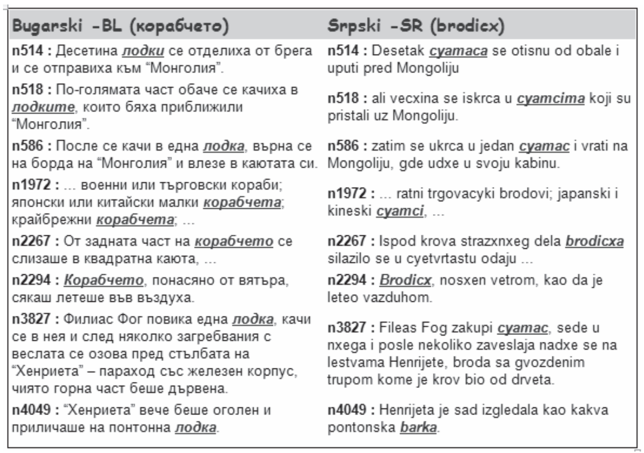


Paralelni prevod
ПЪТЕШЕСТВИЕ ОКОЛО СВЕТА ЗА 80 ДНИ
PUT OKO SVETA ZA 80 DANA

Bugarski -BL (корабчето)	Srpski -SR (brodicx)
n1616 : Фикс го видя да слиза от файтона и да се качва в ***лодка*** с госпожа Ауда и прислужника му.	**n1616 :** Fiks ga vide gde silazi sa kola i gde se ukrcava sa gospodxom Audom i svojim slugom.
n2274 : Английското знаме се вееше на носа на ***корабчето***.	**n2274 :** Engleska zastava se leprsxala na klxuni broda.
n2286 : – Лоцмане, мисля, че не е необходимо – каза Филиас Фог, когато излязоха в открито море – да ви напомням да се движите с възможно най-висока скорост.	**n2286 :** -- Pilote, -- recye Fileas Fog u trenutku kada je ***brodicx*** zaplovio na otvoreno more -- nije potrebno da vam preporucyujem sxto vecxu zxurbu.
n2299 : Често срещаха кораби и със скоростта, с която се движеше, ***лодката*** щеше да се разбие и при най-лекия удар.	**n2299 :** Sudari brodova tu nisu bili retki, a kako je jedrenxak jedrio velikom brzinom, razmrskao bi se pri najmanxem sudaru.
n2323 : На следващата заран, на 8 ноември, ***корабчето*** бе изминало над сто мили.	**n2323 :** Sutradan, 8. novembra, pri izlasku sunca jedrilica je bila presxla visxe od stotinu milxa.
n2356 : По-сериозно не биха управлявали и ***лодка*** от “Кралския яхтклуб”.	**n2356 :** Ni na utakmici Kralxevskog jaht-kluba ne bi tacynije upravlxao brodom.
n2439 : Скоростта на ***корабчето*** бе забележителна.	**n2439 :** Brod je plovio dosta brzo.
n3707 : Те имат много платна, повече дори от състезателна ***лодка***, и при попътен вятър се носят по заснежените полета с бързина, ако не по-голяма, то поне равна на скоростта на експресите.	**n3707 :** Uostalom, ona su odlicyno opremlxena jedrima - bolxe nego neki gusarski kuter -, i kad im vetar duva s ledxa, klize po povrsxini prerije jednakom brzinom mozxda i vecxom od brzih vozova.

Figure 7. A few examples of a partial retrieval

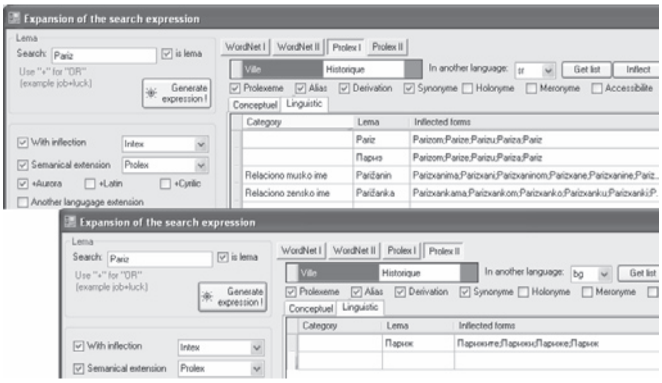
Figure 8 shows eight examples of the full retrieval. In one of these examples (n1972) for the Serbian *čamac* the near synonym in Bulgarian *корабчето* is used (as determined by wordnets). In two cases (n2267 and n2294) for the Serbian *brodić* the near hypernym *корабчето* is used, while in five cases (n514, n518, n586, n3827, n4049) for the Serbian *čamac* and *barka* the near hyponym *лодка* is used. This is not an unexpected result; rather it only proves that searching with the help of semantic networks, on the web for instance, can be useful, which is the ultimate goal of our experiments.



Bugarski -BL (корабчето)	Srpski -SR (brodicx)
n514 : Десетина ***лодки*** се отделиха от брега и се отправиха към “Монголия”.	**n514 :** Desetak ***cyamaca*** se otisnu od obale i uputi pred Mongoliju
n518 : По-голямата част обаче се качиха в ***лодките***, които бяха приближили “Монголия”.	**n518 :** ali vecxina se iskrca u ***cyamcima*** koji su pristali uz Mongoliju.
n586 : После се качи в една ***лодка***, върна се на борда на “Монголия” и влезе в каютата си.	**n586 :** zatim se ukrca u jedan ***cyamac*** i vrati na Mongoliju, gde udxe u svoju kabinu.
n1972 : ... военни или търговски кораби; японски или китайски малки ***корабчета***; крайбрежни ***корабчета***; ...	**n1972 :** ... ratni trgovacyki brodovi; japanski i kineski ***cyamci***, ...
n2267 : От задната част на ***корабчето*** се слизаше в квадратна каюта, ...	**n2267 :** Ispod krova strazxnxeg dela ***brodicxa*** silazilo se u cyetvrtastu odaju ...
n2294 : ***Корабчето***, понасяно от вятъра, сякаш летеше във въздуха.	**n2294 :** ***Brodicx***, nosxen vetrom, kao da je leteo vazduhom.
n3827 : Филиас Фог повика една ***лодка***, качи се в нея и след няколко загребвания с веслата се озова пред стълбата на “Хенриета” – параход със железен корпус, чиято горна част беше дървена.	**n3827 :** Fileas Fog zakupi ***cyamac***, sede u nxega i posle nekoliko zaveslaja nadxe se na lestvama Henrijete, broda sa gvozdenim trupom kome je krov bio od drveta.
n4049 : “Хенриета” вече беше оголен и приличаше на понтонна ***лодка***.	**n4049 :** Henrijeta je sad izgledala kao kakva pontonska ***barka***.

Figure 8. All occurrences of a full retrieval

When a search is performed not with common keywords but with proper nouns then a query expansion with Prolex database offers more possibilities. Semantic relations incorporated into this database are adapted to proper names. Here, the user can choose to expand his query both on the conceptual and the linguistic level. It can be seen in Figure 9 how a query launched with a pivot *Paris* is linguistically expanded into two languages. A morphological expansion can be chosen here as well and it is performed in the same way, using the same methods as for common words. In the given example, the query expansion for Serbian gives more results since the Prolex database for Bulgarian has only some sample entries.



Search: Pariz ☑ is lema
Use "+" for "OR" (example job+luck)
☑ With inflection: Intex
☑ Semanical extension: Prolex
☑ +Aurora □ +Latin □ +Cyrilic
□ Another language extension
Ville | Historique
In another language: sr
☑ Prolexeme ☑ Alias ☑ Derivation ☑ Synonyme □ Holonyme □ Meronyme □ Accessibilite

Category	Lema	Inflected forms
	Pariz	Parizom;Parize;Parizu;Pariza;Pariz
	Париз	Parizom;Parize;Parizu;Pariza;Pariz
Relaciono musko ime	Parižanin	Parizxanima;Parizxani;Parizxaninom;Parizxane;Parizxanine;Pariz...
Relaciono zensko ime	Parižanka	Parizxankama;Parizxankom;Parizxanko;Parizxanku;Parizxanki;P...

Search: Pariz ☑ is lema
Use "+" for "OR" (example job+luck)
☑ With inflection: Intex
☑ Semanical extension: Prolex
Ville | Historique
In another language: bg
☑ Prolexeme ☑ Alias ☑ Derivation ☑ Synonyme □ Holonyme □ Meronyme □

Category	Lema	Inflected forms
	Парис	Парисите;Парисы;Парисе;Парис

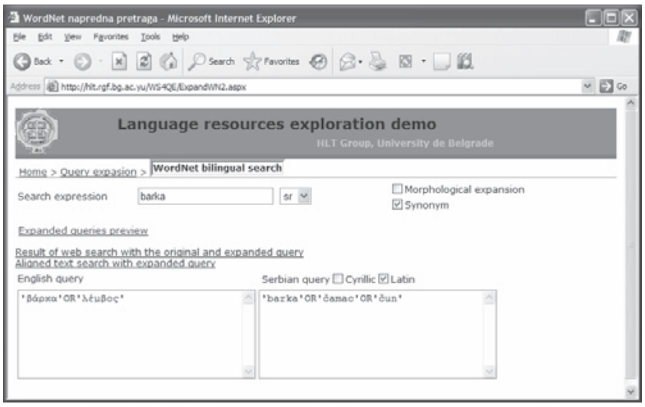
Figure 9. Prolex based semantic expansions

4. Additional Possibilities

We have illustrated the functions of WS4LR for working with aligned texts in the previous section by using the Serbian and Bulgarian pair. This can be successfully used for other Balkan languages as well. Wordnets have been being developed through the Balkanet project for Greek, Romanian and Turkish, which have enabled the experiments to have semantic query expansions for those languages as well. For Greek [12] and Romanian [3], morphological dictionaries in the LADL format have also been developed – however, these resources were not at our disposal so we have not been able to experiment with morphological expansion for these languages.

The possibility and the need for some of the functions developed within WS4LR to become available also on the web have led to the development of the WS4QE web application for lexical resources. This application is still under development, but some of its functions can already be used. Numerous user functions are envisaged for this tool, but the largest set is related to the expansion of queries submitted to the Google search engine, and they have already been implemented. In fact, they are very similar to those presented in the previous section. The only difference is that expanded queries are not applied to an aligned text but are rather forwarded to the search engine.

Figure 10 shows such a retrieval that starts with the Serbian keyword barka “boat” and is further expended by the Serbian synset {barka:1, čamac:1, čun:1} and Greek corresponding synset {βάρκα:0, λέμβος:0}. Figure 11 represents the first results retrieved with such an expanded query by Google.

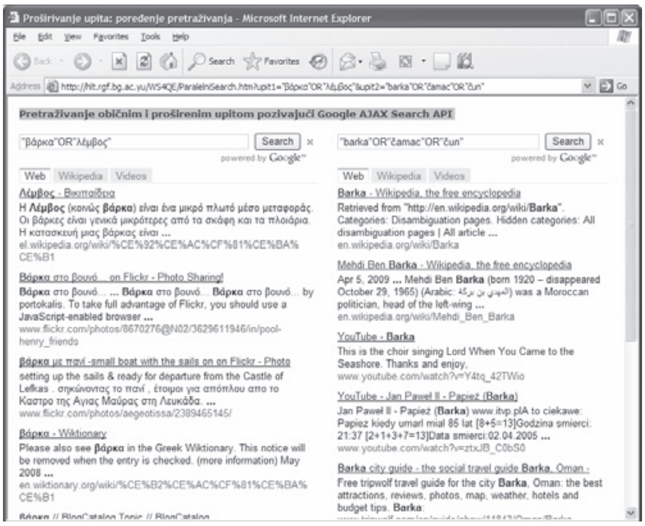

WordNet napredna pretraga - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
Go
Language resources exploration demo
HLT Group, University de Belgrade
Home > Query expasion > WordNet bilingual search
Expanded queries preview
Result of web search with the original and expanded query
Aligned text search with expanded query

Address: http://hlt.rgf.bg.ac.yu/WS4QE/ExpandWN2.aspx
Search expression: barka sr
□ Morphological expansion
☑ Synonym
English query: 'βάρκα'OR'λέμβος'
Serbian query □ Cyrillic ☑ Latin: 'barka'OR'čamac'OR'čun'

Figure 10. A bilingual query expansion with WS4QE – An example of Serbian and Greek

5. Further Work

Our main concern for our future work is the adequate processing of multi-word units. That is, we would like our tool to treat multi-word units in the same way as simple words and to inflect them correctly upon request. The first version of this approach was presented in [10]. Although this version gave promising results for Serbian, it was hardwired into the tool itself so that it was not easy to modify the Serbian module or to apply it to other languages. With a new approach that relies on the feature structure description of a particular language’s morphology [6] and widely uses XML technology, the portability to other languages will be much easier [17]. On a more practical level, our aim is enrich our lexical resources, first of all to enrich the Prolex database, as we plan to use it in a translation environment [14]. It is our wish to work in a future with a true aligned Balkan text – that is, a text originally written in a Balkan language and translated into other Balkan languages.



Address: http://hlt.rgf.bg.ac.yu/WS4QE/ParalelnSearch.htm?upit1="βάρκα"OR"λέμβος"&upit2="barka"OR"čamac"OR"čun"

Pretraživanje običnim i proširenim upitom pozivajući Google AJAX Search API

"βάρκα"OR"λέμβος"

Λέμβος - Βικιπαίδεια
Η **Λέμβος** (κοινώς **βάρκα**) είναι ένα μικρό πλωτό μέσο μεταφοράς. Οι βάρκες είναι γενικά μικρότερες από τα σκάφη και τα πλοιάρια. Η κατασκευή μιας βάρκας είναι ...
el.wikipedia.org/wiki/%CE%92%CE%AC%CF%81%CE%BA%CE%B1

Βάρκα στο βουνό... on Flickr - Photo Sharing!
Βάρκα στο βουνό... ... **Βάρκα** στο βουνό... **Βάρκα** στο βουνό... by portokalis. To take full advantage of Flickr, you should use a JavaScript-enabled browser ...
www.flickr.com/photos/8670276@N02/3629611946/in/pool-henry_friends

βάρκα με πανί -small boat with the sails on on Flickr - Photo
setting up the sails & ready for departure from the Castle of Lefkas . σηκώνοντας το πανί , έτοιμοι για απόπλου απο το Καστρο της Αγιας Μαύρας στη Λευκάδα. ...
www.flickr.com/photos/aegeotissa/2389465145/

βάρκα - Wiktionary
Please also see **βάρκα** in the Greek Wiktionary. This notice will be removed when the entry is checked. (more information) May 2008 ...
en.wiktionary.org/wiki/%CE%B2%CE%AC%CF%81%CE%BA%CE%B1

Βάρκα // BlogCatalog Topic // BlogCatalog

"barka"OR"čamac"OR"čun"

Barka - Wikipedia, the free encyclopedia
Retrieved from "http://en.wikipedia.org/wiki/**Barka**". Categories: Disambiguation pages. Hidden categories: All disambiguation pages | All article ...
en.wikipedia.org/wiki/Barka

Mehdi Ben Barka - Wikipedia, the free encyclopedia
Apr 5, 2009 ... Mehdi Ben **Barka** (born 1920 – disappeared October 29, 1965) (Arabic: المهدي بن بركة) was a Moroccan politician, head of the left-wing ...
en.wikipedia.org/wiki/Mehdi_Ben_Barka

YouTube - Barka
This is the choir singing Lord When You Came to the Seashore. Thanks and enjoy.
www.youtube.com/watch?v=Y4tq_42TWio

YouTube - Jan Paweł II - Papież (Barka)
Jan Paweł II - Papież (**Barka**) www.itvp.plA to ciekawe: Papiez kiedy umarl mial 85 lat [8+5=13]Godzina smierci: 21:37 [2+1+3+7=13]Data smierci:02.04.2005 ...
www.youtube.com/watch?v=ztxJB_C0bS0

Barka city guide - the social travel guide Barka, Oman -
Free tripwolf travel guide for the city **Barka**, Oman: the best attractions, reviews, photos, map, weather, hotels and budget tips. **Barka**:

Figure 11. Results of a query bilingually expanded by Wordnet

6. References

[1] P. Bonhomme, T. M. H. Nguyen, S. O'Rourke. XAlign: l'aligneur de Langue & Dialogue, http://www.loria.fr/equipes/led/outils/ALIGN/align.html, 2001.

[2] B. Courtois, M. Silberztein (eds.). *Dictionnaires électroniques du français.* Langue française. 87, Larousse, Paris, 1990.

[3] D.-M. Dimitriu. *Grammaires de flexion du roumain en format DELA,* Rapport interne 2005-02 de l'Institut Gaspard-Monge, CNRS, 2005.

[4] T. Erjavec and N. Ide. The MULTEXT-East Corpus. In *LREC'98*, Granada, pp. 971-974, 1998.

[5] A. Gelbukh, G. Sidorov, J.-A. Vera-Félix. A Bilingual Corpus of Novels Aligned at Paragraph Level. In proc. *FinTAL*-2006. *Lecture Notes in Artificial Intelligence*, no. 4139, Springer-Verlag, pp. 16–23, 2006.

[6] ISO 24610. *Language resource management – Feature Structures*, ISO/TC 37/SC 4, 2005.

[7] S. Koeva. M*odern language technologies – applications and perspectives,* in: Lows of/for language, Hejzal, Sofia, 2004, 111- 157, 2004.

[8] C. Krstev, et al. Combining Heterogeneous Lexical Resources, in Proc. of the Fourth International Conference LREC, Lisbon, Portugal, May 2004, vol. 4, pp. 1103-1106, 2004.

[9] C. Krstev, R. Stanković, D. Vitas, I. Obradović. *WS4LR: A Workstation for Lexical Resources*, Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 2006, pp. 1692-1697, 2006.

[10] C. Krstev, R. Stanković, D. Vitas, I. Obradović, The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines, in *Proceedings of the Sixth Interantional Conference on Language Resources and Evaluation* (LREC'08), Marrakech, Morocco, 28-30 May 2008, European Language Resources Association (ELRA), 2008.

[11] C. Krstev. *Processing of Serbian*, Faculty of Phylology, University of Belgrade, Belgrade, 2008.

[12] T. Kyriacopoulou. Les dictionnaires électroniques: Morphologie et syntaxe. Le cas du grec moderne, *Proceedings AILA 1990*, Chalcidique, 1990.

[13] E. Laporte, T. Nakamura, S. Voyatzi. A French Corpus Annotated for Multiword Nouns, in: *Towards a Shared Task for Multiword Expressions* (MWE 2008), in scope of the *Sixth Interantional Conference on Language Resources and Evaluation* (LREC'08), http://multiword.sourceforge.net/download/MWE2008-papers/8_Laporte.pdf, 2008.

[14] D. Maurel, D. Vitas, C. Krstev, S. Koeva. Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian, in *Bulag - Bulletin de Linguistique Appliquée et Générale*, Les langues slaves et le français : approches formelles dans les études contrastives, eds. A. Dziadkiewicz & I. Thomas, No. 32, pp. 55-72, Presses Universitaires de Franche Comté, Besançon, 2007.

[15] S. Paumier. *Unitex 2.1 User Manual,* http://www-igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf, 2008.

[16] O. Piton, D. Maurel. Beijing frowns and Washington takes notice: Computer Processing of Relations between Geographical Proper Names in Foreign Affairs, *Fourth International Workshop on Applications of Natural Language to Data Bases (NLDB'00),* Versailles, 28-30 juin (Actes p. 66-78), 2000.

[17] R. Stanković. Improvement of Queries using a Rule Based Procedure for Inflection of Compounds and Phrases. *Polibits (37) 2008, Special section: Natural Language Processing, Journal of Research and Development in Computer Science and Engineering, ed. Grigori Sidorov,* Centro Innovación y Desarrollo Tecnológico en Computo, Instituto Politécnico Nacional, Mexico, pp. 14-20, 2008.

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

[19] M. Tran, D. Maurel. Prolexbase : Un dictionnaire relationnel multilingue de noms propres, Traitement automatique des langues, Vol. 47-3, 2006.

[20] D. Tufiş (ed.). *Special Issue on BalkaNet Project*, Romanian Journal on Information Science and Technology. Bucureşti: Publishing house of the Romanian academy, Vol. 7, No.1-2, 2004.

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

[22] P. Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, 1998.

Converting Russian Treebank SynTagRus into Praguian PDT Style

David Mareček and Natalia Kljueva
Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00, Praha 1, Czech Republic
marecek@ufal.mff.cuni.cz *kljueva@ufal.mff.cuni.cz*

Abstract

In this paper, we report a work in progress on transforming syntactic structures from the SynTagRus corpus into tectogrammatical trees in the Prague Dependency Treebank (PDT) style. SynTagRus (Russian) and PDT (Czech) are both dependency treebanks sharing lots of common features and facing similar linguistic challenges due to the close relatedness of the two languages. While in PDT the tectogrammatical representation exists, sentences in SynTagRus are annotated on syntactic level only.

Keywords

Dependency treebank, tectogrammatical trees, dependency relations, parallel corpora

1 Introduction

Treebanking in Prague comprises not only the annotations of Czech. Besides the main project of Prague Dependency Treebank (PDT) [3], there are several other projects using the same schema for annotating other languages. We should mention the Prague Arabic Dependency Treebank (PADT) [4] and Prague English Dependency Treebank (PEDT) [1], which contains texts from Wall Street journal manually annotated in the PDT style. The Prague Czech-English Dependency Treebank (PCEDT) [2] was developed by translating PEDT into Czech and annotating it also on the Czech side.

Our goal is to convert the Russian corpus SynTagRus [7] into the PDT annotation scheme and build the tectogrammatical (deep-syntactic) layer for Russian. We also develop a small Russian-Czech parallel treebank so that we can compare the two closely-related languages and study structural similarities and differences, which could be useful for developing machine translation systems.

2 Description of the treebanks

2.1 Prague Dependency Treebank

Prague Dependency Treebank (version 2.0) [3] is a treebank of Czech, which consists of three interlinked annotation layers: the morphological layer, the analytical layer (describing the surface syntax) and the tectogrammatical layer (describing the deep syntax – transition between syntax and semantics). A highly simplified example of the annotation layers is in Figure 1. The theoretical background of PDT has its roots in the Prague School of Functional and Structural Linguistics, and especially in the language description framework called Functional Generative Description [9]. The following paragraphs summarize the main features of the three layers.

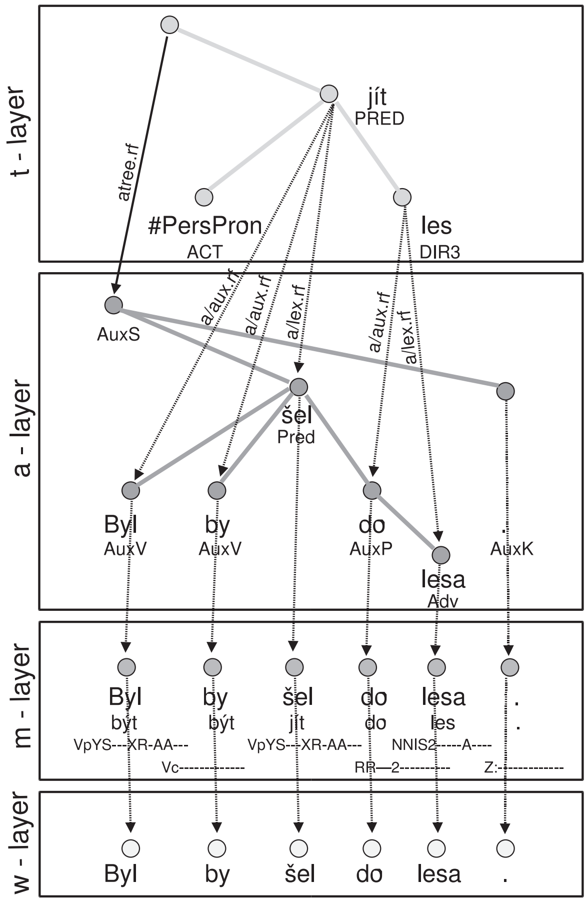


Figure 1: *PDT 2.0 annotation layers (and the layer interlinking) illustrated (in a simplified fashion) on the sentence "Byl by šel do lesa." ([He] would have gone into forest.)*

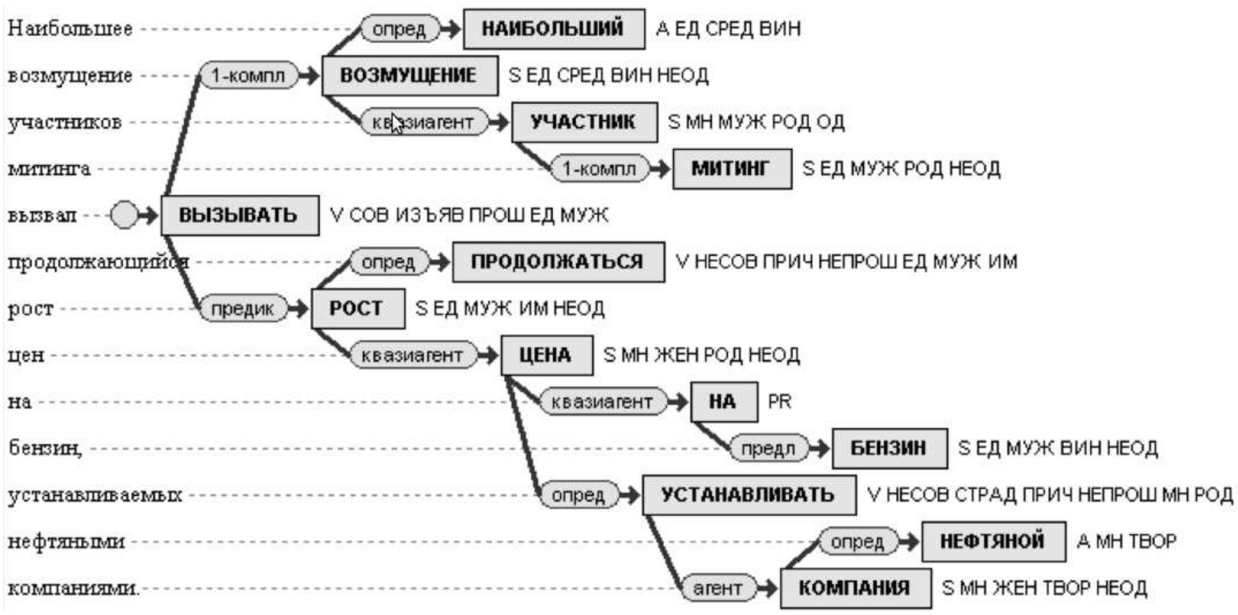


Figure 2: *A syntactically annotated sentence from the SynTagRus treebank. Lemmas in rectangles are followed by tags, syntactic relations are in ovals.*

At the morphological layer (m-layer), a sentence is divided into tokens (words, punctuation marks, and other symbols). Lemma and positional morphological tag are assigned to each token.

At the analytical layer (a-layer), a rooted dependency tree is being build for every sentence. Every token from the morphological layer becomes exactly one node in the analytical tree. Only one node – the "technical" root – is added. An analytical function (such as Subject, Object, Attribute) is assigned to each node, but in fact it captures the type of dependency relation between the given node and its parent node. However, there are also edges representing non-dependency relations (e.g. in coordination structures).

At the tectogrammatical layer (t-layer), each sentence is represented as a complex deep-syntactic dependency tree (tectogrammatical tree), in which only autosemantic words have nodes of their own. Functional words like prepositions, subordinating conjunctions, auxiliary verbs, and modal verbs are represented in the respective nodes in the form of their attributes. On the other hand, tectogrammatical trees contain nodes that have no counterparts in the surface shape of the sentences, for instance nodes corresponding to 'pro-dropped' subjects. Each node has its tectogrammatical lemma, functor (which determines the type of semantic relation between the node and its parent), semantic part of speech, grammatemes (semantically-oriented counterparts of morphological categories such as aspect, degree of comparison, modality, gender, iterativeness, negation, number, person, or tense).

The corpus contains 115,844 sentences (1,957,247 tokens including punctuation and other special characters) from newspapers and scientific articles. All of them are annotated on the m-layer, 75% on the a-layer and 45% on all three layers.

2.2 SynTagRus

SynTagRus is a syntactically annotated corpus of Russian based on the theory "Meaning-Text" [6]. In SynTagRus, sentences are represented as trees, in which words are nodes and edges between them are marked with the appropriate syntactic relation. Unlike in PDT, punctuation marks are not annotated in SynTagRus. They are included, but do not carry any labeling and they are not included in syntactic trees. An annotated sentence from SynTagRus is depicted in Figure 2.

Each word (node in a tree) has five attributes in the SynTagRus XML format:

- *id* – linear position of the word in the sentence,
- *dom* – id of its parent node,
- *lemma* – morphological lemma,
- *feat* – morphological tag.[1] Part of speech at the first position is followed by a sequence of respective features (e. g. number, gender, case, person, aspect, tense, ...),
- *link* – syntactic relation[1] between the node and its parent. It can be for example 'предик' (between a verb and its complement), '1-компл' (between a verb and its direct object), 'предл' (between a preposition and a noun), and many others.

The whole corpus contains 32,242 sentences and 461,297 tokens (excluding punctuation). Most of the texts are from journal articles and newspapers, but there are also texts belonging to the fiction genre.

[1] All morphological and syntactic features are described at `http://www.ruscorpora.ru/instruction-syntax.html`.

3 Adaptation of tectogrammatical layer for Russian

Here we discuss the ongoing process of constructing tectogrammatical representation on the basis of morphological information and syntactic relations. The conversion will be described in several steps.

3.1 Format conversion

Both PDT and SynTagRus are represented in XML based formats. In the case of PDT a special PML format was developed [8]. SynTagRus XML format was therefore transfered into PML, so that we can use the TectoMT[1] software framework [12] and TrEd[2] viewer.

As we can see from the corpora description, SynTagRus annotation covers all the features that are necessary to build morphological and analytical layer. The third – tectogrammatical layer will be derived from these two layers in the next steps.

3.2 Converting coordinations

Coordination relations do not belong among dependency relations. Their handling in SynTagRus is different from the PDT style. We will call the coordinated words (or clauses) *coordination members*, the word which governs all the coordination members will be *common parent* and the words depending on all the members will be *common dependents*.

In SynTagRus, according to the Meaning Text Theory [6], the first member of coordination is attached to the common parent. Common dependents are attached to the nearest member, often to the first one. Each other coordination member including conjunctions is attached to the previous member as it is depicted in Figure 3. The edges between coordination members are labeled by 'сочин' (composition relation) or 'соч-союзн' (composition-with-conjunction relation).

In our example, the verbs 'топали' (*stamped*), 'свистели' (*whistled*), and 'расходились' (*left*) are coordinated. They are head of the sentence (the first member is attached to the technical root 'SruA') and have one common dependent, the subject 'Собравшиеся' (*People*), which is attached to the first member 'топали'.

The same sentence but with the coordination handled in the PDT style is depicted in Figure 4. All members of coordination are attached here to the conjunction, the common dependent 'Собравшиеся' is attached also to the conjunction. Members of coordination are distinguished from common dependents with the special attribute '_co'.

The advantages and disadvantages of these two different handling of coordinations are discussed in more detail in [11]. Mel'čuk's approach needs less memory compared to PDT, because it needs no special attributes '_co' for marking coordinating members. It seems that it is also more suitable for annotators (missing '_co' attribute was very common and problematic error in PDT). On the other hand, Mel'čuk's theory can not reflect inner structure of coordination constructions (for example in *'Peter and Mary or Charlie and Suzanne'*) and does not allow different syntactic relations of coordinated words.

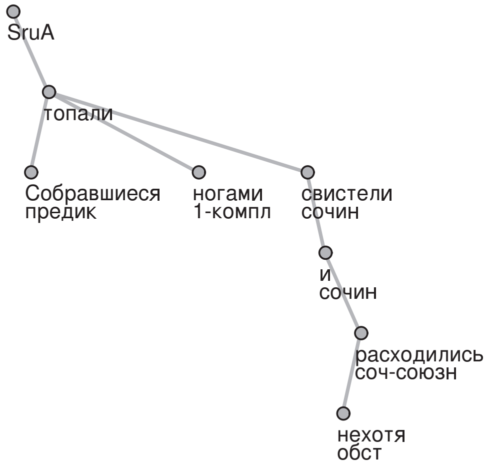


Figure 3: *Handling coordinations in SynTagRus, sentence 'Собравшиеся топали ногами, свистели и нехотя расходились.' (People stamped their feet, whistled and left unwillingly.)*

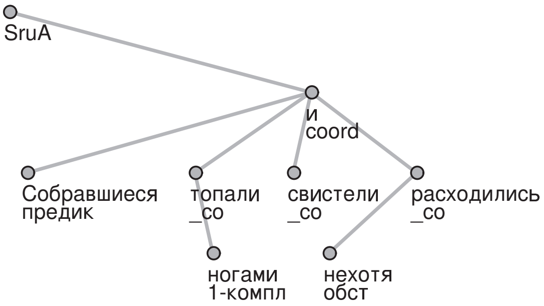


Figure 4: *Handling coordinations in the PDT style (the same sentence as in Figure 3).*

Several problems occurred in automatic conversion of coordinations into PDT style.

Firstly, it is not distinguished in SynTagRus whether a dependent of a member of coordination actually refers to the whole coordination or only to that one member. In our example, the words 'Собравшиеся' (*People*) and 'ногами' (*feet*) are attached both to the first coordination member 'топали' (*stamped*). While 'Собравшиеся' is a common dependent, the word 'ногами' depends on the first member only – on the word 'топали'. The authors of SynTagRus treebank decided not to distinguish them, because this is a notorious source of ambiguity in many cases, for example in *'old men and women'* vs. *'old men and women whose age is not specified'*. Nevertheless, the PDT representation requires this ambiguity to be resolved. The disambiguation can be partially facilitated by a couple of rules. For instance, a subject belonging to the coordinated verbs is almost certain the common subject if there is no other subject in the sentence. This is just the case of the word

[1] http://ufal.mff.cuni.cz/tectomt
[2] http://ufal.mff.cuni.cz/∼pajas/tred

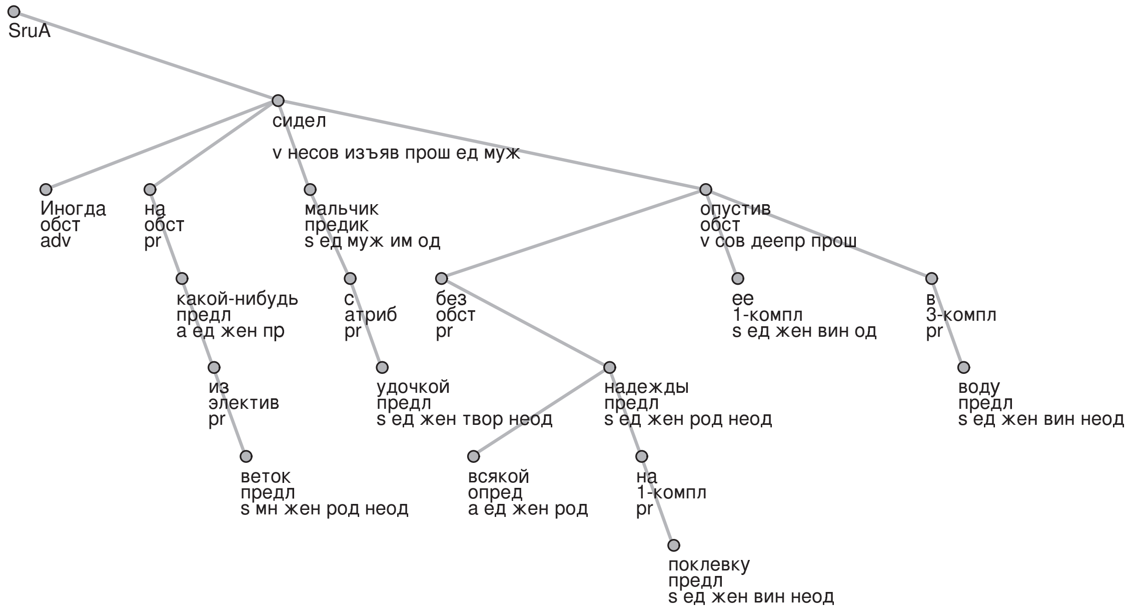


Figure 5: *Analytical representation of the Russian sentence 'Иногда на какой-нибудь из веток сидел мальчик с удочкой, без всякой надежды на поклевку опустив ее в воду. (Now and than a boy with a fishing rode was sitting on a branch, dropping it into the water without any hope to catch fish.)*

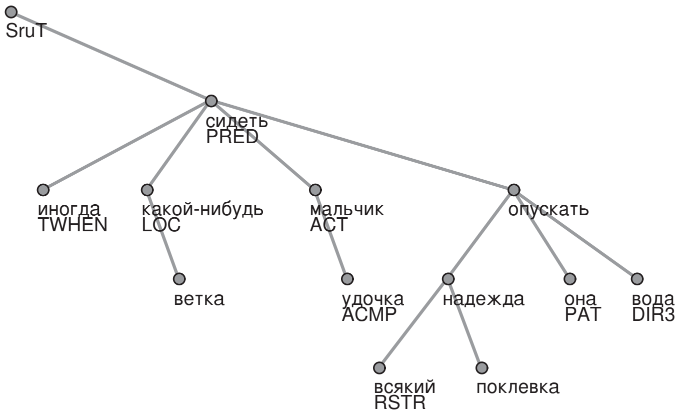


Figure 6: *Tectogrammatical representation of the sentence from the Figure 5, lemmas and functors are depicted.*

'Собравшиеся' (*People*). But by far not all cases can be solved.

Secondly, since punctuation marks are not included in the trees in SynTagRus, it is often the case that there is no node that could serve as the coordination head. In such situations, all coordination members are attached on their common parent instead on a conjunction. We also can not deal with common dependents in such structures, but this problem arises very rarely.

3.3 Function words

Function words (e. g. prepositions, subordinating conjunctions and auxiliary verbs) do not have their own nodes in the tectogrammatical trees. The conversion from analytical trees (in which every word is represented by one node) is done in several steps. Each function word is first marked and assigned to one of the content words. Afterwards the tectogrammatical tree is build using only content (non-function) words as nodes. The meaning of the function words is then expressed by functors and grammatemes (the attributes of respective content-word nodes).

An example of conversion from analytical tree into tectogrammatical tree is depicted in Figures 5 and 6.

For example, the prepositional phrase 'в воду' is represented by the node 'вода' in the tectogrammatical layer. The preposition 'в' is reflected in the functor 'DIR3', which means *to where*.

Some of the rules we use for assignment of function words to content words follow.

1. **prepositions** – A preposition is assigned to its child node (a noun), if the syntactic relation is 'предл' (prepositional).

2. **passive forms** – If there are two verbs which syntactic relation is ‘пасс-анал’ (analytical-passive) and the lemma of the parent verb is ‘быть’ (*to be*), the parent verb is assigned as a functional word to the child verb.

3. **future tense** – In Russian (as well as in Czech) future tense of imperfective verbs is expressed analytically as ‘to be’ + infinitive, e. g. ‘будут пользоваться’ (*will use*). Therefore, the rule is: If there are two verbs, their relation is ‘аналит’ (analytical), the lemma of the parent verb is ‘быть’ (*to be*), and the child verb is in infinitive form, the parent verb is assigned to the child.

4. **subordinated conjunctions** – Conjunctions ‘что’ (*that*), ‘чтобы’ (*so that*), or ‘потому что‘ (*because*) are assigned to their child nodes, if the syntactic relation between them is ’подч-союзн’ (subordinate clause with conjunction).

5. **modal verbs** – A verb which lemma is ‘хотеть’ (*want*), ‘мочь’ (*can*), ‘надо’ (*should*), or ‘должен’ (*must*) is assigned to its child node, if the child node is verb in infinitive form.

3.4 Elided ’to be’

In Czech, personal pronouns in subject positions are often dropped and have to be added (reconstructed) at the tectogrammatical layer. Analogically, we add special nodes into Russian tectogrammatical trees if the Russian verb ‘to be’ is dropped in the surface sentence shape, as it is for example in ‘Я студент’ (*I [am] a student*). This is currently approximated by the following simple heuristics: if there is a ‘предик’ (predicate) relation between two nodes and the parent node is not a verb, then generate a new node labeled with ‘#ToBe’ and attach both previously existing nodes below it (see Figure 7).

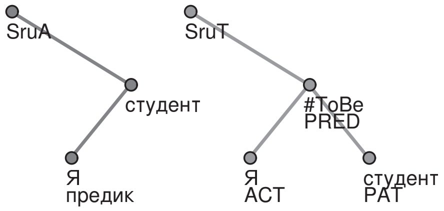


Figure 7: *Adding node ‘#ToBe’ into the tectogrammatical representation of the sentence ‘Я студент’ (I [am] a student).*

3.5 Assigning functors

Syntactic relations in SynTagRus bare not only syntactic information, but they go deeper towards semantic relations. Labels of semantic relations are called functors in the PDT terminology.

Yet, the classification of this relations within this two frameworks is very different, only a few of them can be mapped as one-to-one. It is for example apposition (syntactic relation ‘аппоз’ goes to the functor APPS), parenthetical relation (‘примыкат’ → PAR), and comparative relation (‘сравнит’ → CPR).

There are five functors for verb arguments in PDT: actor (ACT), patient (PAT), effect (EFF), addressee (ADDR), and origin (ORIG). We expected them to be closely-related to the completive syntactic relations within SynTagRus (1-компл, 2-компл, 3-компл, 4-компл, 5-компл). The apparent correspondence is between the functor PAT and syntactic relation ‘1-компл’ e. g. in ‘он читает письмо.PAT’ (*he is reading a letter.PAT*). The functor ACT (actor) is often the subject of the sentence and corresponds to ‘предик’.

Other relations however do not straightforwardly correspond to the PDT-style functors. In order to assign functors properly we need to know cognitive role of the word, but the argument relations in SynTagRus hardly give this information. Therefore we use several additional rules, for instance: If the relation between a verb and its child node is completive (?-компл) and the child node is a noun in dative case, we assign the functor ADDR (addressee) to it. Example: ‘Он дал ребенку.ADDR игрушку’ (*He gave to a child.ADDR a toy*).

Some other functors are assigned using lexical list. For example, the words ‘чтобы’ (*to*), ‘в интересах’ (*in order to*, ‘с целью’ (*with the aim of*) usually correspond to the functor AIM. A preposition ‘в’ (*in*, *to*) corresponds either to the functor LOC (where), if the noun is in locative case, or to the functor DIR3 (to where) for accusative case. A preposition ‘в’ followed by a noun representing a time, for example *Monday*, *January*, *yesterday*, *week*, corresponds to the functor TWHEN (when). A set of such temporal nouns is not too large to make a satisfactory list of them manually.

You can see an application of the described rules for functors assignment in Figure 6.

4 Small parallel treebank

We have built a small Russian-Czech parallel treebank. Luckily, there exist Czech translations for some of the prose texts included in SynTagRus. We have found one such book which contains Czech translation of one chunk in SynTagRus. We acquired 480 parallel sentences, so that we can compile a small parallel treebank. Whereas the Russian side is largely manually annotated (only the transfer form SynTagRus to tectogrammatics is automatic), the annotation on the Czech side is fully automatic. We use Morce tagger [10], McDonald maximum spanning tree dependency parser [5] and other mainly rule based scripts to generate the tectogrammatical layer. The corpus was compiled using TectoMT [12] framework, which includes all these tools. This parallel treebank, even if very small at the moment, can be once a valuable source of information in comparative language studies.

5 Conclusion and future work

We described the first steps of converting the Russian dependency treebank SynTagRus into the PDT style and developing tectogrammatical layer of Russian. We are on half of the way. We transformed the treebank into the PDT format, we changed the representation of coordination constructions, because their handling is very different in SynTagRus and in PDT. We hid the auxiliary words, that do not have their own nodes in the tectogrammatical layer, and the elided verbs 'to be' were added. We started with assigning functors (the deep-syntactic relations between tectogrammatical nodes).

In the future, we plan to continue with adding more (often more complex) rules for assigning functors. Other attributes as grammatemes are also going to be assigned to the tectogrammatical nodes.

As for the parsed parallel corpus, we also plan to experiment with aligning the tectogrammatical structures of the two languages on the node level.

Acknowledgments

The work on this project was supported by the grants GAUK 9994/2009, GAČR 201/09/H057, and GAAV ČR 1ET201120505.

We would like to thank Leonid L. Iomdin for providing the complete SynTagRus and for taking the time to answer our questions.

References

[1] S. Cinková, J. Toman, J. Hajič, K. Čermáková, V. Klimeš, L. Mladová, J. Šindlerová, K. Tomšů, and Z. Žabokrtský. Tectogrammatical Annotation of the Wall Street Journal. *Prague Bulletin of Mathematical Linguistics*, (92), 2009.

[2] J. Cuřín, M. Čmejrek, J. Havelka, J. Hajič, V. Kuboň, and Z. Žabokrtský. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, Catalog No.: LDC2004T25, 2004.

[3] J. Hajič, E. Hajičová, J. Panevová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, and M. Mikulová. Prague Dependency Treebank 2.0. Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

[4] J. Hajič, O. Smrž, and P. Pajas. Prague Arabic Dependency Treebank 1.0. Linguistics data Consortium, Catalog No,: LDC2004T23, 2004.

[5] R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*, pages 523–530, Vancouver, BC, Canada, 2005.

[6] I. Mel'čuk. *Dependency Syntax: Theory and Practice.* State University of New York Press, 1988.

[7] J. Nivre, I. Boguslavsky, and L. Iomdin. Parsing the SynTagRus Treebank. In *Proceedings of COLING08*, pages 641–648, 2008.

[8] P. Pajas and J. Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information*, Genoa, Italy, 2006.

[9] P. Sgall. *Generativní popis jazyka a česká deklinace.* Academia, Prague, Czech Republic, 1967.

[10] D. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

[11] J. Štěpánek. *Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu.* PhD thesis, Charles University in Prague, 2006.

[12] Z. Žabokrtský, J. Ptáček, and P. Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*, 2008.

A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words

Svetlin Nakov
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Boucher Blvd, Sofia, Bulgaria
nakov @ fmi.uni-sofia.bg

Elena Paskaleva
Linguistic Modeling Laboratory
Bulgarian Academy of Sciences
25A Acad. G. Bontchev St, Sofia, Bulgaria
hellen @ lml.bas.bg

Preslav Nakov
Department of Computer Science
National University of Singapore
13 Computing Drive, Singapore
nakov @ comp.nus.edu.sg

Abstract

We propose a novel knowledge-rich approach to measuring the similarity between a pair of words. The algorithm is tailored to Bulgarian and Russian and takes into account the orthographic and the phonetic correspondences between the two Slavic languages: it combines lemmatization, hand-crafted transformation rules, and weighted Levenshtein distance. The experimental results show an 11-pt interpolated average precision of 90.58%, which represents a sizeable improvement over two classic rivaling approaches.

Keywords

Orthographic similarity, phonetic similarity, cross-lingual transformation.

1. Introduction

We propose an algorithm that measures the extent to which a Bulgarian and a Russian word are perceived as similar by a person who is fluent in both languages. Leaving aside the full orthographical identity, we assume that words with different orthography can be also perceived as similar when they have the same or a similar stem and inflections, as in the Bulgarian word ***афектир**ахме* and the Russian ***аффектир**овались* (both meaning '*we were **affected***').

Bulgarian and Russian are closely related Slavonic languages with rich morphology, which motivates us to study the typical orthographical correspondences between their lexical entries (conditioned phonetically and morphologically), which we use to formulate and apply transformation rules for bringing a Russian word close to Bulgarian reading and vice versa. Our algorithm for measuring the similarity between Bulgarian and Russian words first reduces the Russian word to an intermediate Bulgarian-sounding form and then compares it orthographically to the Bulgarian word. The algorithm starts by transliterating the Russian word with the Bulgarian alphabet, and then transforms some typical Russian morphemes and word parts (*e.g.*, prefixes, suffixes, endings, *etc.*) to their Bulgarian counter-parts. Since both Bulgarian and Russian are highly-inflectional languages, lemmatization is used to convert the wordforms to their lemmata in order to reduce the differences at the morphological level. Finally, the orthographic similarity is measured using a modified Levenshtein distance with letter-specific substitution weights.

2. Method

The normalization of the Bulgarian and the Russian words into corresponding intermediate forms has phonetic and morphological motivation and is performed as a sequence of steps, which will be described below.

2.1. Transliteration from Cyrillic to Cyrillic

In a strict linguistic sense, *transcription* is the process of transition from sounds to letters, i.e., from speech to text; it is carried out generally in a monolingual context. In a bilingual context, the notion of *transliteration* is used to denote the transition of sounds and their letter correspondences in one language to letters in another language. The term *transliteration* is commonly used for the transition of letters when the two languages use different alphabets. In this paper, we deal with *transliteration* since we work with written texts.

The linguistic objective of our investigation is to introduce more formal criteria to the investigation of possible *cognates* between Russian and Bulgarian. By *cognates* we mean words with equal or close orthography denoting the same meaning; words with equal/close orthography but different meaning are *false cognates/friends*. For their further investigation in multilingual research, we need to define the exact expression of that identity/closeness by particular metrics and procedures.

For a pair of languages from different families, the source of cognates is borrowing between them or from a third language. Besides borrowing, an essential source of cognates in related languages is their common protolanguage. However, in the historical development of both languages, three factors lead to different grapheme shape for fully identical words: (1) language-specific phonetic laws and resulting changes, (2) settings of the spelling systems regulating the *sound-letter* transition, and (3) divergence in the grammatical systems and the grammatical formatives.

2.1.1 Full coincidence (equality) of letters

Both Russian and Bulgarian use the Cyrillic alphabet in their writing systems, but Russian uses two letters not present in Bulgarian: *ы* and э. Most other letters generally show a full coincidence with some exceptions to be listed in the following subsections. The list below presents the full identity of Cyrillic letters in both languages in the cognates: ***а**збука – **а**збука* , ***б**уква – **б**уква, **в**оля – **в**оля, гипс – гипс, **д**ух – **д**ух, **е**зда – **е**зда, **ж**ена – **ж**ена, **з**акон – **з**акон, **и**стина – **и**стина, **й**од – **й**од, **к**ипарис – **к**ипарис, **л**ак – **л**ак, **м**онета – **м**онета, **н**ож – **н**ож, **о**пора – **о**пора, **п**ост **–** **п**ост, **р**ека – **р**ека, **с**ом – **с**ом, **т**ом – **т**ом, **у**м – **у**м, **ф**акт – **ф**акт, **х**имия – **х**имия, **ц**арь – **ц**ар ,**ч**ай – **ч**ай, **ш**ум – **ш**ум, **щ**ит – **щ**ит, **ю**г – **ю**г, **я**хта – **я**хта*

As the above list shows, the full identity of the grapheme shape of cognates is manifested mainly when the transformed letter is in initial position.

2.1.2 Regular letter transitions

Replacing Russian letters that are missing in the Bulgarian alphabet. The transitions discussed here stem from historic differences in the phonetic and the spelling systems of the two languages. Bulgarian and Russian differ in their contemporary phonetic system mainly at the level of pronunciation; in the distinction of soft and hard consonants. The Russian-specific letters *ы* and *э* serve to denote the variant of a '*hard consonant+и/е*' while in Bulgarian all consonants preceding *и* and *е* are soft. This basic difference of the phonetic systems gives us the regular correspondence *ы-и* and *э-е* in all Russian-Bulgarian cognates containing these two letters, e.g., *рыба – риба*, *поэт- поет*.

Removing a Russian letter. Another regular phonetic difference between the two languages, which is also related to the opposition *soft/hard*, is the allowed softness of a consonant preceding another consonant (*пальто*) or in final position (*шесть*). Such phonetic combinations are not allowed in Bulgarian: see the corresponding *палто* and *шест*. This regularity allows us to remove all Russian *ь* in these positions in the initial stage of the process of cognate comparison.

Partial regularity of the letter transitions. In non-initial positions, other not so regular but repeated letter correspondences can be observed, e.g., *е-я* in *хлеб-хляб*, *е-ъ* in *серп-сърп*, *о-ъ* in *сон-сън, у-ъ* in *муж-мъж*, etc. The iterativity of such transitions is due to the specific development of the spelling systems in the two languages. One such example is the disappearance of some Old Slavic letters and their regular replacement with different letters in Russian and Bulgarian. The above-mentioned change *у-ъ* is due to the disappearance of an Old Slavic letter called 'big yus' and its regular replacement by different vowels in all contemporary Slavic languages. The transition is only partially regular since not all occurrences of the letter have the same etymological origin.

2.1.3 Tranformations of n-grams

The *sound-letter* transition legitimated by the spelling rules of the two languages is specific as well; its specificity is observed at the level of the grapheme composition of the full cognates, i.e., those that are borrowed from third languages or that are identical morphologically.

Transformations originating from spelling.

A fundamental difference between Russian and Bulgarian spellings is the treatment of *double consonants*. Russian allows them in every part of the word structure, while in Bulgarian they are only possible at the morpheme boundary. Thus, all words borrowed from third languages keep their double consonants in Russian, but lose them in Bulgarian, e.g., *процесс – процес*, *аффект – афект*, etc. In this way, a regular transition *ll-l* can be formulated for all double consonants with the following stipulation of grammatical origin.

In words of Slavic origin, consonant doubling occurs mainly at the morpheme boundary, but in Russian the phenomenon is more frequent since Russian spelling rules are more "phonetic". For example, they reflect the change *voiced-voiceless* for all prefixes ending with *з* and preceding the initial *с* of the next morpheme. Bulgarian spelling is more 'morphological' and conservative; it keeps the *з* in writing, although it is voiceless in pronunciation, e.g., *рассуждение – разсъждение*, *бессмертный – безсмъртен*, etc. This transformation of *hard-soft* consonants in the final prefix position is only valid for the couple *з-с*. Thus, the Bulgarian-Russian transition *зс-сс* can be formulated as regular for prefixes only and cannot be viewed as a universal for other parts of the word, e.g., *кавка**зс**кий – кавка**з**ки*.

Next, the following general question in treating double consonant correspondences arises: if we want to stay in the domain of uni- and bigram transformations, removing the second consonant in Russian can be ambiguous *поддержать – поддържам*, but *буддист – будист*, *вводить – въвеждам*, *раввин – равин.* The legal consonant doublings in Bulgarian can be only outlined in a larger context – a window of up to five letters, containing the prefix and the next consonant, as in *предд*, *надд*, *подд*, *изз*, *разз*, etc., where the second consonant should be preserved. Note that these exceptions from the rule are only valid for double *д*, *з* and *в* – final letter of prefixes, and for *н* – first letter of the affix *н*, e.g., *непременно – непременно*, but *аннотация – анотация*. .

Transformations of morphological origin.

In addition to the divergent development of phonetic and spelling systems, the two languages develop different grammatical systems, both at a systemic and at a morphemic level – different categories with different graphemic expressions. That divergence leads to different grapheme shapes for words that are lexically conceived as cognates, e.g., *жены – женатa*, and the difference is manifested in the ending part of the word, consisting of affixes, and ending and related to grammatical forms.

The transformations are made in two directions and for both languages. They can consist of removal of a letter sequence or its transformation.

1. Removing agglutinative morphemes.

Each of the two languages has one agglutinative mechanism of word formation (but for different parts of speech) – the reflexive morpheme *ся* and *сь* in Russian verb conjugation and the postpositioned article in Bulgarian in nominal inflections (for nouns and adjectives). The corresponding grammatical meanings are expressed in the twin language by other means (the article is totally missing in Russian and the reflexivity of verbs is expressed by a lexical element in Bulgarian – the particle *се*). Thus, removing these morphemes is the first step in the process of conversion to an intermediate form, e.g., *веселиться – веселить*, *квадратът – квадрат*. Note that the Russian agglutinative morpheme *ся/сь* at the end of the word are non-ambigous: all 212,000 wordforms with the ending *ся* in our Russian grammatical dictionary are reflexive verb forms. This is not the case with the Bulgarian article, where only removing the morpheme *ът* for masculin is non-ambiguous, while removing *та*, *ят* and other article morpheme can trim the stem, e.g., *жена-та*, but *квадрат-а*. We intentionally do not derive a transformation rule from the last correspondence.

Removing Bulgarian articles depends on the accepted conception about the place of lemmatization in the algorithm – should we set the orthographic similarity for all four members of the language pair – lemmata and wordforms – or should we measure the similarity at the lexical level only – the lemmata. In the latter case, no removal is necessary (see 1.3)

2. Transforming ending strings.

There is a big group of adjectives in the two languages derived from other parts of speech and formed with the suffix *н* and an adjectival ending, e.g., *шум – шумный*, *шум – шумен*. When the adjective is derived from a noun ending with *н*, we get a doubled *н* in the Russian lemma and in the Bulgarian wordforms, e.g., *гарнизон-гарнизонный* and *гарнизон – гарнизонни*. Another regular correspondence is manifested in the word derivation with the suffix *ск*. All these combinations of *н/нн/ск* and different adjectival endings give the correspondences shown in Table 1.

Russian Ending	**Bulgarian Ending**	**Examples**
-нный	*-нен*	*во**енный** → во**енен***
-ный	*-ен*	*веч**ный** – веч**ен***
-нний	*-нен*	*ра**нний** → ра**нен***
-ний	*-ен*	*вечер**ний** → вечер**ен***
-ский	*-ски*	*вражеск**ий** → вражеск**и***
-ый	*-и*	*стрелков**ый** – стрелков**и***
-нной	*-нен*	*сте**нной** – сте**нен***
-ной	*-ен*	*род**ной** – род**ен***
-ой	*-и*	*делов**ой** – делов**и***

Table 1: Transforming Russian adjectives to Bulgarian.

For verbs, there are some regularities in the correspondences of the endings of the Russian infinitive and the Bulgarian verb's main form in first person singular. Table 2 below shows some examples.

Russian Ending	**Bulgarian Ending**	**Examples**
-овать	*-ам*	*декорир**овать** → декорир**ам***
-ить, - ять	*-я*	*брод**ить** → брод**я*** *бле**ять** → бле**я***
-ать	*-ам*	*дав**ать** → дав**ам***
-уть	*-а*	*гасн**уть** – гасн**а***
-еть	*-ея*	*бел**еть** → бел**ея***

Table 2 – Transformation of Russian verbs to Bulgarian.

Concerning the transformation of endings, it is important to note that two linguistic problems are interrelated here: (1) the formal revelation of the morpheme boundary, and (2) the correct correspondence with the Bulgarian ending. The existing ambiguity in resolving these two problems requires serious statistical investigations before the rules can be formulated.

With ambiguity not taken into account, the proposed transformation rules for Russian word endings could sometimes generate the wrong Bulgarian wordform, e.g., *висеть* could become *висея*, while the correct Bulgarian form is *вися.* In order to limit the negative impact of that, we measure the similarity (1) *with* and (2) *without* applying rules for lemmatization; we then return the higher value of the two.

2.2. Lemmatization

Bulgarian and Russian are highly-inflectional languages, *i.e.*, they use variety of endings to express the different forms of the same word. When measuring orthographic similarity, endings could cause major problems since they can make two otherwise very similar words appear somewhat different. For example, the Bulgarian word *отправен**ата*** (*'the directed'*, a feminine adjective with a definite article) and the Russian word *отправленн**ому*** (*'the directed'*, a masculine adjective in dative case) exhibit only about 50% letter overlap, but, if we ignore the endings, the similarity between them becomes much bigger. Thus, if our algorithm could safely ignore word endings when comparing words, it might perform better.

If we could remove the ending, the similarity would be measured using the stem, which is the invariable part of the word. Unfortunately, both the ending as a letter sequence and the location of the morpheme boundary are quite ambiguous in both languages. Thus, we need to lemmatize the text, i.e., convert the word to its main form, the lemma. If every member of the pair of candidate cognates from L1 and L2 is represented by a wordform (WF) and its lemma (L), then we could compare: L1 with L2, WF1 with WF2, L1 with WF2 and WF1 with L2. Considering these four options, we can get a better estimation for the similarity not only between close wordforms like the Bulgarian *отправената* and the Russian *отправленному,* which look different orthographically, but have very close lemmata, but also between such very different words like the Bulgarian *къпейки* (‘*bathing*’, a gerund) and the Russian *копейки* (‘*copeck*’, plural feminine noun).

The lemmatization of the Bulgarian and the Russian words can be done using specialized dictionaries. In the present work, we will use two large grammatical dictionaries that contain words, their lemmata, and some grammatical information.

2.3. Transformation Weights

Let us now come back to the transliteration rules and to the next steps in our algorithm. There are orthographical correspondences between candidate cognates that are not as undisputable as the general rules, but are still observed in the development of the languages, at least for ones with a proven etymological basis. As was shown above, the regular correspondences between the languages can be due to phonetic and spelling reasons. Besides the unconditional letters transitions described above, not so regular ones occur in several cases, and their existence can be taken into account when constructing the weight scale for measuring similarity.

A general principle when building a weight scale is that the correspondences between letters denoting consonants and vowels (hereinafter ‘vowels’ and ‘consonants’ only) should be measured separately. The maximal ortographic distance between different letters is 1 (as for *а*-*ц*) and the maximal similarity has weight 0 (as for *а*-*а*). All weight values between 0 and 1 are assigned to letter correspondences that exist in a non-regular way in some cognates (the above-mentioned correspondence *у*-*ъ* was due to etymological reasons). Another general admission is that consonants and vowels with similar sequences of distinctive phonetic features (differing only in the place of articulation or in the presence/absence of voice, e.g., *б*-*в*, *б*-*п*) have lower weight distance. The same is valid for the pair of letters denoting a regular phonetic change, e.g., *reduction* (as in *а*-*ъ*, *о*-*у*) or *softening* of the preceding consonant (as in *у*-*ю*, *а*-*я*). Regular correspondences observed in a limited lexical sector (e.g., borrowed from Latin and Greek) such as *г*-*х* also have a lower distance.

Table 3 shows the letter transformation weights, which can be used to measure the orthographic similarity after the Bulgarian and Russian words have been transliterated to a subset of the Cyrillic alphabet.

The weights *w*(*a*, *b*) are used to transform the letter *a* into the letter *b* and vice versa. This weight function *w* is symmetric by definition, *i.e.*, *w*(*a*, *b*) = *w*(*b*, *a*). All other weights not given in Table 3 are equal to 1.

In order to write the Russian words in the modified Bulgarian alphabet used in Table 3, we make the following preliminary transformations for all Russian words:

э → е; ы → и; ь → (empty letter); ъ → (empty letter)

Table 3 shapes the match between letters and the sounds they denote in Bulgarian and Russian. It further correlates weights for letter transformation that have been phonetically justified.

а	*w*(*а*, *е*)=0.7; *w*(*а*, *и*)=0.8; *w*(*а*, *о*)=0.7; *w*(*а*, *у*)=0.6; *w*(*а*, *ъ*)=0.5; *w*(*а*, *ю*)=0.8; *w*(*а*, *я*)=0.5
б	*w*(*б*, *в*)=0.8; *w*(*б*, *п*)=0.6
в	*w*(*в*, *ф*)=0.6
г	*w*(*г*, *х*)=0.5
д	*w*(*д*, *т*)=0.6
е	*w*(*е*, *и*)=0.6; *w*(*е*, *о*)=0.7; *w*(*е*, *у*)=0.8; *w*(*е*, *ъ*)=0.5; *w*(*е*, *ю*)=0.8; *w*(*е*, *я*)=0.5
ж	*w*(*ж*, *з*)=0.8; *w*(*ж*, *ш*)=0.6
з	*w*(*з*, *с*)=0.5
и	*w*(*и*, *й*)=0.6; *w*(*и*, *о*)=0.8; *w*(*и*, *у*)=0.8; *w*(*и*, *ъ*)=0.8; *w*(*и*, *ю*)=0.7; *w*(*и*, *я*)=0.7
й	*w*(*й*, *ю*)=0.7; *w*(*й*, *я*)=0.7
к	*w*(*к*, *т*)=0.8; *w*(*к*, *х*)=0.6
м	*w*(*м*, *н*)=0.7
о	*w*(*о*, *у*)=0.6; *w*(*о*, *ъ*)=0.8; *w*(*о*, *ю*)=0.7; *w*(*о*, *я*)=0.8
п	*w*(*п*, *ф*)=0.8; *w*(*п*, *х*)=0.9
с	*w*(*с*, *ц*)=0.6; *w*(*с*, *ш*)=0.9
т	*w*(*т*, *ф*)=0.8; *w*(*т*, *х*)=0.9; *w*(*т*, *ц*)=0.9
у	*w*(*у*, *ъ*)=0.5; *w*(*у*, *ю*)=0.6; *w*(*у*, *я*)=0.8
ф	*w*(*ф*, *ц*)=0.8
х	*w*(*х*, *ш*)=0.9
ц	*w*(*ц*, *ч*)=0.8
ч	*w*(*ч*, *ш*)=0.9
ъ	*w*(*ъ*, *ю*)=0.8; *w*(*ъ*, *я*)=0.8
ю	*w*(*ю*, *я*)=0.8

Table 3– Letter substitution weights.

3. The MMEDR Algorithm

The MMEDR algorithm (*modified minimum edit distance ratio*) measures the orthographic similarity between a pair of Bulgarian and Russian words using some general phonetic and morphologically conditioned correspondences between the letters of the two languages in order to estimate the extent to which the two words would be perceived as similar by people fluent in both languages. It returns a value between 0 and 1, where values close to 1 express very high similarity, while 0 is returned for completely dissimilar words. The algorithm has been tailored for Bulgarian and Russian and thus is not directly applicable to other pairs of languages. However, the general approach can be easily adapted to other languages: all that has to be changed are the rules describing the phonetic and the morphological correspondences.

The MMEDR algorithm in steps:

1. Lemmatize the Bulgarian word.
2. Lemmatize the Russian word.
3. Transform the Russian word's ending.
4. Transliterate the Russian word.
5. Remove some double consonants in the Russian word.
6. Calculate the modified Levenshtein distance using suitable weights for letter substitutions.
7. Normalize and calculate the MMEDR value.

The algorithm first tries to rewrite the Russian word following Bulgarian letter constructions. As a result, both words are transformed into a special intermediate form and then are compared orthographically using Levenshtein distance with suitable weights for individual letter substitutions. The above general algorithm is run in eight variants with each of steps 1, 2 and 3 being included or excluded, and the largest of the eight resulting values is returned. A description of each step follows below.

3.1. Lemmatizing Bulgarian and Russian words

The Bulgarian word is lemmatized using a grammatical dictionary of Bulgarian as described in Section 1.3. If the dictionary contains no lemmata for the target word, the original word is returned; if it contains more than one lemma, we try using each of them in turn and we choose the one yielding the highest value in the MMEDR algorithm. The Russian word is lemmatized in the same way, using a grammatical dictionary of Russian.

3.2. Transforming the Russian Ending

At this step, we transform the endings of the Russian word according to Tables 1 and 2 and we remove the agglutinative suffix *ся*:

> *нный* → *нен*; *ный* → *ен*; *нний* → *нен*; *ний* → *ен*; *ий* → *и*; *ый* → *и*; *нной* → *нен; ной* → *-ен*; *ой* → *и*; *ский* - *ски*; *ься* → *ь; овать* → *ам; ить* → *я*; *ять* → *я*; *ать* → *ам*; *уть* → *а*; *еть* → *ея*

The substitutions rules are applied only if the left hand-side letter sequences are at the end of the word. Rules are applied in the given order; multiple rule applications are allowed. Note that we do not have rules for all possible endings in Russian, but only for the typical ones – object of transformation for adjectives and verbs.

Since all words have been already lemmatized in the previous step (if applied), verbs are assumed to be in infinitive and adjectives in singular masculine form. Adjective endings are transformed to their respective Bulgarian counter-parts, and reflexive verbs are turned into non-reflexive. Nouns are not considered since they generally have the same endings in the two languages (after having been lemmatized) and thus need no additional transformations.

Of course, there are many exceptions for the above rules, but our experiments show that using each of them has more positive than negative effect. Initially, we tried using few more additional rules, which were subsequently removed since they were found to be harmful.

3.3. Removing double consonants

According to 1.1.3, the following substitution rules are applied for the Russian word:

> *бб* → *б*; *жж* → *ж*; *кк* → *к*; *лл* → *л*; *мм* → *м*; *пп* → *п*; *рр* → *р*; *сс* → *с*; *тт* → *т*; *фф* → *ф*

3.4. Calculating the Modified Levenshtein Distance with Weights for Letter Substitution

Given two words, the Levenshtein distance [Levenshtein, 1965], also known as the *minimum edit distance* (MED), is defined as the minimum total number of single-letter substitutions, deletions and/or insertions necessary to convert the first word into the second one. We use a modification, which we call *modified minimum edit distance* (MMED), where the weights of all insertions and deletions are fixed to 1, and the weights for single-letter substitution are as given in Table 3.

3.5. Calculating MMEDR

At this step, we calculate MMEDR value by normalizing MMED – we divide it by the length of the longer word (the length is calculated after all transformations have been made in the previous steps). We use the following formula:

$$MMEDR(w_{bg}, w_{ru}) = 1 - \frac{MMED(w_{bg}, w_{ru})}{\max(|w_{bg}|, |w_{ru}|)}$$

3.6. Calculating the final result

The final result is given by the maximum of the obtained values for all eight variants of the MMEDR algorithm – with/without lemmatization of the Bulgarian word, with/without lemmatization of the Russian word, and with/without transformation of the Russian word ending. Note also, that lemmatization steps might result in calculating additional values for MMEDR – one for each possible lemma of the Russian/Bulgarian word.

3.7. Example

As we will see below, the proposed MMEDR algorithm yields significant improvements over classic orthographic similarity measures like LCSR (*longest common subsequence ratio*, defined as the longest common letter subsequence, normalized by the length of the longer word [Melamed, 1999]) and MEDR (*minimum edit distance* *ratio*, defined as the Levenshtein distance with all weights set to 1, normalized by the length of the longer word, also known as *normalized edit distance* /*NED*/ [Marzal & Vidal, 1993]). This is due to the above-described steps which turn the Russian word into a Bulgarian-sounding one and the application of letter substitution weights that reflect the closeness of the corresponding phonemes.

Let us consider for example the Bulgarian word *афектирахме* and the Russian word *аффектировались*. Using the classic Levenshtein distance, we obtain the following: MED(*афектирахме, аффектировались*) = 7. And after normalization: MEDR=1–(7/15) = 8/15 ≈ 53%. In contrast, with the MMEDR algorithm, we first lemmatize the two words, thus obtaining *афектирам* and *аффектировать* respectively. We then replace the double Russian consonant -*фф*- by -*ф*- and the Russian ending -*овать* by the first singular Bulgarian verb ending -*ам*. We thus obtain the intermediate forms *афектирам* and *афектирам*, which are identical, and MMEDR = 100%. Note that some pairs of words like *афектирахме* and *аффектировались* could be neither orthographically nor phonetically close but could be perceived as similar due to cross-lingual correspondences that are obvious to people speaking both languages.

Let us take another example – with the Bulgarian word *избягам* and the Russian word *отбегать* (both meaning '*to run out*'), which sound similarly. Using Levenshtein distance: MED(*избягам*,*отбегать*) = 5 and thus MEDR = 1 – (5/8) = 3/8 = 37.5%. In contrast, with the MMEDR algorithm, we first transform *отбегать* to its intermediate form *отбегам* and we then calculate MMED(*избягам, отбегам*) = 0.8 + 1 + 0.5 = 2.3 and MMEDR = 1 – (2.3/7) = 47/70 ≈ 67%, which is a much better reflection of the similarity between the two words.

Thus, we can conclude that, at least in the above two examples, the traditional MEDR does not work well for the highly inflectional Bulgarian and Russian. MEDR is based on the classic Levenshtein distance, which uses the same weight for all letter substitution, and thus cannot distinguish small phonetic changes like replacing *я* with *е* (two phonetically very close vowels) from more significant differences like replacing *я* with *г* (a vowel and a consonant that are quite different).

4. Experiments and Evaluation

We performed several experiments in order to assess the accuracy of the proposed MMEDR algorithm for measuring the similarity between Bulgarian and Russian words in a literary text.

4.1. Textual resources

We used the Russian novel *The Lord of the World* (*Властелин мира*) by Alexander Belyayev [Belayayev, 1940a] and its Bulgarian translation by Assen Trayanov [Belayayev, 1940b] as our test data. We extracted the first 200 different Bulgarian words and the first 200 different Russian words that occur in the novel, and we measured the similarity between them.

#	Bulgarian word	Russian word	MMEDR	Sim	Precision	Recall
1	беляев	беляев	1.0000	Yes	100.00%	0.68%
2	на	на	1.0000	Yes	100.00%	1.37%
3	глава	глава	1.0000	Yes	100.00%	2.05%
4	кандидат	кандидат	1.0000	Yes	100.00%	2.74%
5	за	за	1.0000	Yes	100.00%	3.42%
6	наполеон	наполеоны	1.0000	Yes	100.00%	4.11%
7	не	не	1.0000	Yes	100.00%	4.79%
8	ми	нас	1.0000	No	87.50%	4.79%
9	ми	мой	1.0000	Yes	88.89%	5.48%
10	ми	мы	1.0000	Yes	90.00%	6.16%
...	...	...	...	...	...	...
93	четвъртият	четвертым	0.9375	Yes	94.57%	59.59 %
94	оставят	остается	0.9286	Yes	94.62%	60.27 %
...	...	...	...	...	...	...
39998	са	в	0.0000	No	0.37%	100%
39999	са	к	0.0000	No	0.37%	100%
40000	боядисвали	к	0.0000	No	0.37%	100%

Table 4 – Results of the MMEDR algorithm.

4.2. Grammatical Resources

We used two monolingual dictionaries for lemmatization:

- **A grammatical dictionary of Bulgarian,** created at the Linguistic Modeling Department, Institute for Parallel Processing, Bulgarian Academy of Sciences [Paskaleva, 2002]. This electronic dictionary contained 963,339 wordforms and 73,113 lemmata. Each dictionary entry consisted of a wordform, a corresponding lemma, and some morphological and grammatical information.
- **A grammatical dictionary of Russian,** created at the Institute of Russian language, Russian Academy of Sciences, based on the Grammatical Dictionary of A. Zaliznyak [Zaliznyak, 1977]. The dictionary consisted of 1,390,613 wordforms and 66,101 lemmata. Each dictionary entry consisted of a wordform, a corresponding lemma, and some morphological and grammatical information.

4.3. Experimental Setup

We measured the similarity between all 200x200=40,000 Bulgarian-Russian pairs of words. Among them, 163 pairs were annotated as very similar by a linguist who was fluent in Russian and a native speaker of Bulgarian; the remaining 39,837 were considered unrelated.

We used the MMEDR algorithm to rank the 40,000 pairs of words in decreasing order according to the calculated similarity values. Ideally, the 163 pairs designated by the linguist would be ranked at the top. We can determine how well the ranking produced by our algorithm does using standard measures from information retrieval, *e.g. 11-point interpolated average precision* [Manning et al., 2008].

We compared the MMEDR algorithm with two classic orthographic similarity measures: LCSR and MEDR. Unfortunately, we could not directly compare our results to those in other work, since there were no previous publications measuring orthographic or phonetic similarity between words in Bulgarian and Russian.

4.4. Results

Table 4 shows part of the ranking produced by the MMEDR algorithm. The table shows an excerpt of the ranked pairs of words along with their similarity calculated by the MMEDR algorithm, the corresponding human annotation for similarity (the column "Sim"), as well as precision and recall calculated for all rows from the beginning to the current row.

Table 5 shows the 11-*pt interpolated average precision* for LCSR, MEDR and MMEDR. We can see that MMEDR outperforms the other two similarity measures by a large margin: 18-22% absolute difference.

Algorithm	**11-pt interpolated average precision**
LCSR	69.06%
MEDR	72.30%
MMEDR	**90.58%**

Table 5 – Comparison of the similarity measuring algorithms.

5. Discussion

As Tables 4 and 5 show, the MMEDR algorithm works quite well. Still, there is a lot of room for improvement:

- Bulgarian and Russian inflectional morphologies are quite complex, with many exceptions that are not captured by our rules. This is probably a limitation of the general approach rather than a deficiency of the particular rules used: if we are to capture all exceptions, we would need to manually specify them all, which would require a lot of extra manual work.
- The transformation rules between Bulgarian and Russian are sometimes imprecise as well, *e.g.*, for very short words or for words of foreign origin.
- While linguistically motivated, the letter-for-letter substitution weights we used are *ad hoc*, and could be improved. First, while we used symmetric letter substitution weight in Table 3, asymmetric weights might work better, *e.g.* the Bulgarian prefixes *раз*- and *из*- are spelled as *рас*- and *ис*- in Russian when followed by a voiceless consonant. Thus, the substiution weight for *з* → *с* should probably be higher than for *с* → *з*. We could further extend the rules to take into account the local context, *e.g.*, changing *раз*- to *рас*- could have a different weight than changing -*з*- то -*с*- in general.
- Another potential problem comes from us using only one linguist for the annotation, which might have yielded biased judgments. To assess the impact of the potential subjectivity, we would need judgments by at least one additional linguist.

6. Related Work

Many algorithms have been proposed in the literature for measuring the orthographic and the phonetic similarity between pairs of words from different languages.

The simplest ones considered as orthographically close words with identical prefixes [Simard & al., 1992].

Much more popular have been orthographic similarity measures based on normalized versions of the Levenshtein distance [Levenshtein, 1965], the longest common subsequence [Melamed, 1999], and the Dice coefficient [Brew and McKelvie, 1996].

Somewhat less common have been phonetic similarity measures, which compare sounds instead of letter sequences. Such an approach has been proposed for the first time by [Russel, 1918]. Guy [1994] described an algorithm for cognate identification in bilingual word lists based on statistics of common sound correspondences. Algorithms that learn the typical sound correspondences between two languages automatically have also been proposed: [Kondrak, 2000], [Kondrak, 2003] and [Kondrak & Dorr, 2004].

Instead of applying similarity measures for symbolic strings on the words directly, some researchers have first performed transformations that reflect the typical cross-lingual orthographic and phonetic correspondences between the target languages. This is especially important for language pairs where some letters in the source language are systematically substituted by other letters in the target language. The idea can be extended further with substitutions of whole syllables, prefixes and suffixes. For example, Koehn & Knight [2002] proposed manually constructed transformation rules from German to English (*e.g.*, the letters *k* and *z* are changed to *c*; and the ending -*tät* is changed to -*ty*) in order to expand lists of automatically extracted cognates.

Finally, orthographic measures like LCSR and MEDR have gradually evolved over the years, enriched by machine learning techniques that automatically identify templates for cross-lingual orthographic and phonetic correspondences. For example, Tiedemann [1999] learned spelling transformations from English to Swedish, while Mulloni & Pekar [2006] and Mitkov & al. [2007] learned transformation templates, which represent substitutions of letters sequences in one language with letter sequences in another language.

7. Conclusions and Future Work

We have described and tested a novel algorithm for measuring the similarity between pairs of words based on transformation rules between Bulgarian and Russian. The algorithm has shown very high precision and could be used to identify possible candidates for cognates or false friends in text corpora. It can also be used in machine translation systems working on related languages where it could help overcome the incompleteness of translation dictionaries used in the system.

There are many ways in which we could improve the proposed algorithm. For example, we could adapt the algorithms described in [Mitkov et al., 2007] and [Bergsma & Kondrak, 2007] to Bulgarian and Russian and try to learn cross-lingual transformation rules for morphemes and other sub-word sequences automatically. We could then try to combine MMEDR with such rules.

Acknowledgments

The presented research is supported by the project FP7-REGPOT-2007-1 SISTER.

8. References

[Belyayev, 1940a] Belyaev A. "Lord of the World" (1940), in Russian, publisher "Onyx 21 Century", 2005, ISBN 5-329-01356-9, http://lib.ru/RUFANT/BELAEW/lordwrld.txt

[Belyayev, 1940b] Belyaev A. "Lord of the World" (1940), translation from Russian to Bulgarian by A. Trayanov, publisher "National Youth", 1977, http://www.chitanka.info/lib/text/2130

[Bergsma & Kondrak, 2007] Bergsma S., Kondrak G. "Alignment-Based Discriminative String Similarity". *Proceedings of the 45th Annual Meeting of the ACL*, pages 656–663, Prague, Czech Republic, 2007

[Brew and McKelvie, 1996] Brew C. and McKelvie D. "Word-Pair Extraction for Lexicography". *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Turkey, 1996

[Guy, 1994] Guy J. "An Algorithm for Identifying Cognates in Bilingual Wordlists and Its Applicability to Machine Translation", *Journal of Quantitative Linguistics*, Volume 1 (1), pages 35-42, 1994

[Koehn & Knight, 2002] Koehn P., Knight K. "Learning a Translation Lexicon from Monolingual Corpora". In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9-16, Philadelphia, PA, 2002.

[Kondrak and Dorr, 2004] Kondrak G., Dorr B. "Identification of Confusable Drug Names: A New Approach and Evaluation Methodology". *Proceedings of COLING 2004*, pages 952–958, Geneva, Switzerland, 2004

[Kondrak, 2000] Kondrak G. "A New Algorithm for the Alignment of Phonetic Sequences". *Proceedings of NAACL/ANLP 2000: 1st conference of the North American Chapter of the Association for Computational Linguistics and 6th Conference on Applied Natural Language Processing*, pages 288-295, Seattle, WA, USA, 2000

[Kondrak, 2003] Kondrak G. "Identifying Complex Sound Correspondences in Bilingual Wordlists". *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003)*, pages 432-443, Mexico City, Mexico, 2003

[Levenshtein, 1965] Levenshtein V. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". *Doklady Akademii Nauk SSSR*, Volume 163 (4), pages 845-848, Moscow, Russia, 1965

[Manning et al., 2008] Manning C., Prabhakar R. and Schütze H. "Introduction to Information Retrieval". *Cambridge University Press*, ISBN 0521865719, New York, USA, 2008

[Marzal & Vidal, 1993] Marzal A., Vidal E. "Computation of Normalized Edit Distance and Applications". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 15, Issue 9, pages 926-932, USA, 1993

[Melamed, 1999] Melamed D. "Bitext Maps and Alignment via Pattern Recognition". *Computational Linguistics*, Volume 25 (1), pages 107-130, ISSN:0891-2017, 1999

[Mitkov et al., 2007] Mitkov R., Pekar V., Blagoev D. and Mulloni A. "Methods for Extracting and Classifying Pairs of Cognates and False Friends". *Machine Translation*, Volume 21, Issue 1, pages 29-53, Springer Netherlands, 2007

[Mulloni & Pekar, 2006] Mulloni A. and Pekar V. "Automatic Detection of Orthographic Cues for Cognate Recognition". *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 2387–2390, Genoa, Italy, 2006.

[Paskaleva, 2002] Paskaleva E. “Processing Bulgarian and Russian Resources in Unified Format”. *Proceedings of the 8th International Scientific Symposium MAPRIAL*, pages 185-194, Veliko Tarnovo, Bulgaria, 2002.

[Russel, 1918] Russel R. "U.S. Patent 1,261,167", Pittsburgh, PA, USA, 1918

[Simard et al., 1992] Simard M., Foster G., Isabelle P. "Using Cognates to Align Sentences in Bilingual Corpora". *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal, Canada, 1992

[Tiedemann, 1999] Tiedemann J. "Automatic Construction of Weighted String Similarity Measures". *SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 213-219, College Park, MD, USA, 1999

[Zaliznyak, 1977] Zaliznyak A. "Grammatical Dictionary of the Russian Language", publisher "Russian Language", Moscow, Russia, 1977

New Issues and Solutions in Computer-aided Design of MCTI and Distractors Selection for Bulgarian

Ivelina Nikolova
Institute for Parallel Processing, Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str.
1113 Sofia
iva@lml.bas.bg

Abstract

We describe a methodology for improving the generation of multiple-choice test items through the usage of language technologies. We apply common natural language processing techniques, like constituency parsing and automatic term extraction together with additional morpho-syntactic rules on raw instructional material in order to determine its key terms. These key terms are then used for the creation of fill-in-the blank test items and the selection of distractors. Our work aims at proving the availability and compatibility of language resources and technologies for Bulgarian, as well as at assessing the readiness for implementation of these techniques in real-world applications.

Keywords

information extraction, natural language processing application in e-learning

1 Introduction

Multiple-choice tests (MCT) are a common tool to assess learners achievements. They are widely proven to be efficient. During the last years MCT gained even more popularity due to the growth of the e-learning programmes. In these programmes, which are offered by universities and other educational institutions, multiple-choice questions appear to be the most frequently used evaluation tool. Multiple-choice is a form of assessment in which respondents are asked to select the best possible answer(s) out of a list of choices. We refer to the questions as *stems*, the best option as *correct answer* and the rest of the given choices as *distractors*. The demand for great quantities of such tests and the availability of already advanced learning technologies gave rise to a new research area dealing with the generation of multiple-choice test items (MCTI) and the suggestion of distractors from raw text.

The manual preparation of MCT is a time and effort consuming task. Teaching experts who prepare the tests have much broader knowledge in their field in general, compared to the specific content which is explicitly included in the particular instructional material. They have to tune the tests carefully to the knowledge of the test takers. Hence one of the most difficult subtasks during the creation of test items is to decide whether a question does really have its answer in the taught material. With an automatic extraction of test items from the instructional material, this problem is easily solved and the time for test designing is significantly reduced. An automatic extraction allows the test designers to oversee large instructional materials in a new manner, giving them a content overview and helping them to take faster decisions about the topics to be included in a test and concrete questions which could be given to the learners.

The generation of multiple-choice questions with the help of natural language processing (NLP) technologies is an active research area in which different tools for text processing are used in order to transform the facts from the instructional materials to questions for students assessment. The items produced in this way are often used in Computer Assisted Language Learning (CALL), for vocabulary [2], grammar [3, 4, 1] or language proficiency testing [11, 5], as well as in comprehension testing in specific subject areas in the native language [6]. Our aim is to produce multiple-choice test items for testing learners achievements especially in the second area - learners comprehension of specified instructional material.

We present the design of a workbench for test designers employing language technologies for generation of MCTI (stem, correct answer and distractors), which are to be wrapped as learning objects (LO) and can be loaded in an e-learning environment. The task is divided into three subtasks: *automatic keyterm extraction*; *sentence extraction and stem transformation* and *distractors selection*. In particular, we discuss our contributions to an improved methodology for keyterm and distractors selection and stem transformation.

The remainder of the article is organised as follows: Section 2 describes the state-of-the-art; Section 3 reveals the motivation of the author; Section 4 outlines the overall architecture of the workbench; Section 5 presents a detailed view of the text processing phases; Section 6 presents a discussion on tests done with the system and Section 7 gives a conclusion and issues for future work.

2 Related Work

One of the first works on our topic was presented by [3]. Fairon implemented a corpus search for finding sentences or short parts of text that match initially preselected linguistic patterns. Later, [5] proposed a word sense disambiguation method for locating sentences in which designated words carry specific senses, and applied a collocation-based method for selecting distractors that are necessary for multiple-choice cloze items. Our work differs from these approaches as far as we detect relevant terms automatically, henceforth called *keyterms*. Furthermore, for distractors selection we employ morpho-syntactic information.

Authors working on vocabulary testing [2] use definitions or examples given for the focal term in WordNet in order to produce a non-interrogative stem. [6] also employ WordNet, but only as a tool for distractors selection. Their approach is domain independent; furthermore, the authors report a 6-10 times speed-up in comparison with a manual test elicitation. Similarly, [11] uses a thesaurus in order to find distractors for stem, generated by replacing the verb of the chosen sentence with a blank. [4] apply standard classification methods in order to decide the position in the gap in the generation of fill-in-the-blank (FIB) test items. Other researchers who are actively working in the area include [1], who are focusing on the different types of question models with application mainly in the language learning. In our approach we extract sentences which contain the central terms for the given material in Bulgarian and produce FIB type of questions out of them. Along with that, we also suggest the correct answer and distractors.

3 Motivation

The fact that we are not familiar with any related work for learning materials in Bulgarian (except for previous work of the author [8]), together with the presence of sophisticated language technologies for Bulgarian, which allow for complex text analysis strongly inspired us to work out the practical potential of our ideas. Moreover, the growing interest in the field, which is due to its significant practical importance, was a motivating factor to concretise our aims and more precisely to apply the developed technology for e-learning purposes.

4 Workbench Outline

The system is designed in a way that it accepts instructional material from the test designer in form of raw text and produces draft learning objects - MCTI of FIB type with their correct answer and possible distractors.

Our approach is based on the assumption that the learner knowledge is tested over the terms, central to the learning materials. As shown in Fig. 1, once the text is submitted, a list of generated FIB questions, concerning keyterms from the instructional material, is presented to the test designer. At this moment, she can modify all MCTI components and then save or export them as a learning object or as a plain text document.

The list of FIB stems serves as a cross-reference to the whole text and facilitates for the test-designer in summarising the learning topics.

5 Data Processing

This section describes the processing of the data from the user input to the output of the draft learning objects. Well-established language technologies, like parsing and automatic term extraction are employed. Additionally, linguistic assumptions are taken into consideration. An overview of the data processing chain is shown in Fig. 2.

The input of the test designer is plain text instructional material, which has to be parsed in order to extract lexico-syntactic features from the text. Due to the importance of parsing as a basic source of information used later on for the test items generation, we have picked a statistical parser which reports state-of-the-art results and has been tuned to work with Bulgarian - the Berkeley parser [9, 10]. The parser was trained on BulTreeBank[1]. Parsing texts from the same domain as the training corpus gave highly satisfactory results.

After reformatting the parsed text, we extract from it all nouns and noun phrase structures as well as names. From the tools offering fast structure querying for our purposes the most appropriate turned out to be the CLaRK system[2]. As it is based on Xpath expression querying it is fully configurable. In contrast with the NP extractor *Morena* we used earlier, CLaRK allows for the manual specification of sequences of constituents.

In order to overcome the language inflection, the extracted morpho-syntactic structures are stemmed using BulStem [7] and organised in an internal representation format, where each stem[3] maps to all NPs having the same stem. Here is an example of a partial record for the stem закон (law):

*[stem value="*закон*" occurences=51*
type="N" isKeyterm=false]
*[instance value="*закон*"]*
*[instance value="*законите*"]*
*[instance value="*закона*"]*
....

In this case, the stem is закон (law). It has a total of 51 occurences in the document. Some of them are *закон* (law), *законите* (the laws), закона (the law) and it is type *noun*. Other NP types are *np-A-N* - NP composed of adjective and noun, *np-N-PP* - NP composed of noun and prepositional phrase, *NE-loc* - name of the type location, *NE-org* - name of the type organisation, *NE-Pers* - name of the type person, *NE-other* - name of the type other). For each wordform (phrase) corresponding to the stem, only one instance is generated. If the wordform appears at least twice, only the counter *occurences* is incremented, but no new instance is created. The attribute *isKeyterm* is initially set to *false* for all stems. After the keyterm threshold is set, it is turned to *true* for the terms which belong to the keyterms list. This representation is the starting

[1] A HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank), `http://bultreebank.org/`.

[2] CLaRK - an XML Based System for Corpora Development, `http://bultreebank.org/clark/index.html`.

[3] A stem is the common prefix of several wordforms.

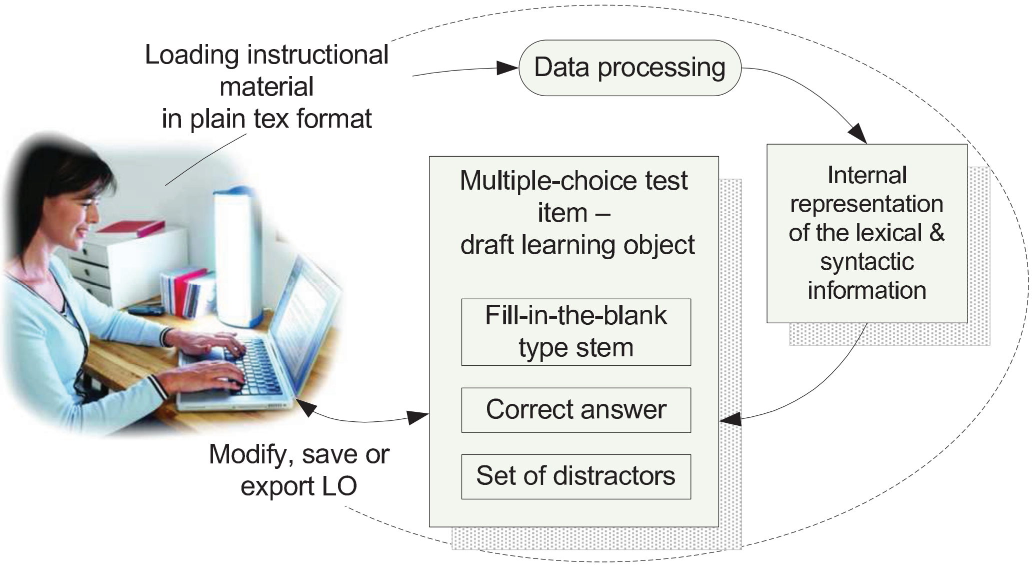

Loading instructional material in plain tex format
Data processing
Internal representation of the lexical & syntactic information
Multiple-choice test item – draft learning object
Fill-in-the-blank type stem
Correct answer
Set of distractors
Modify, save or export LO

Fig. 1: *Workbench supporting the development of multiple-choice test items.*

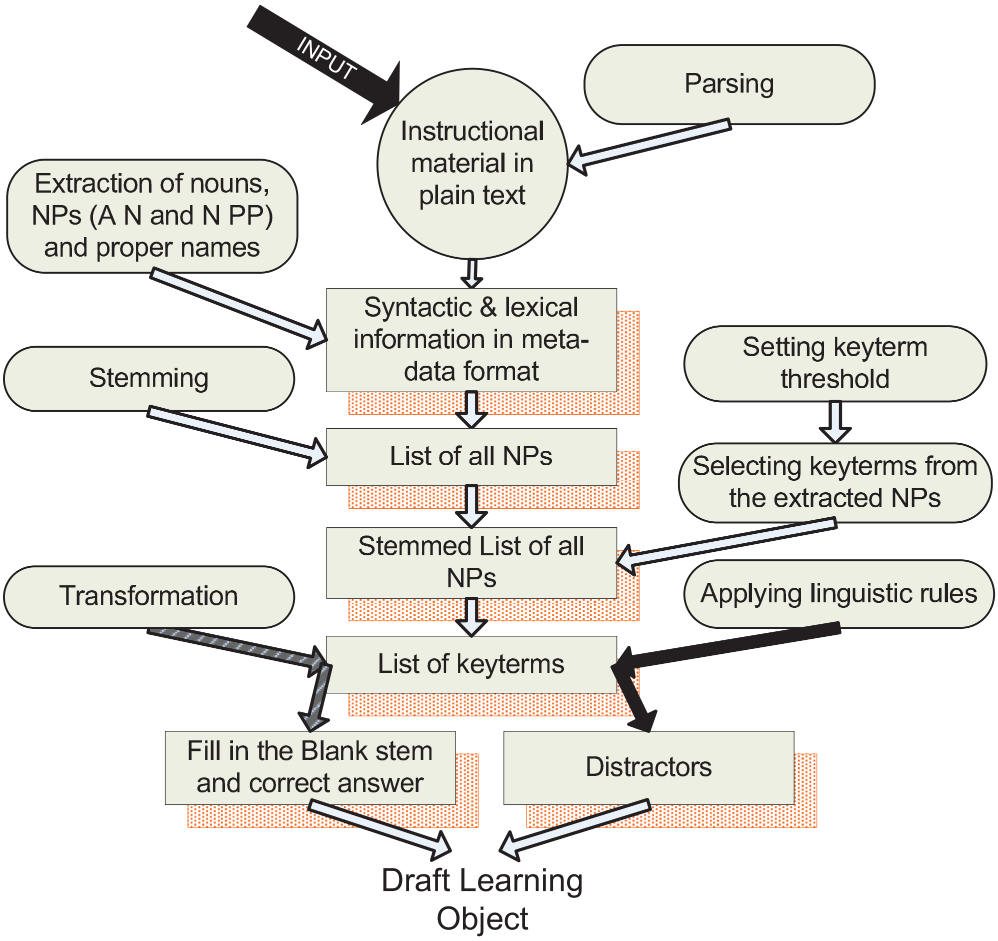

INPUT
Instructional material in plain text
Parsing
Extraction of nouns, NPs (A N and N PP) and proper names
Syntactic & lexical information in meta-data format
Stemming
List of all NPs
Setting keyterm threshold
Selecting keyterms from the extracted NPs
Stemmed List of all NPs
Transformation
Applying linguistic rules
List of keyterms
Fill in the Blank stem and correct answer
Distractors
Draft Learning Object

Fig. 2: *Data processing.*

point for any further processing in order to generate the test item stem and distractors.

The *occurences* attribute of the stem field is used later on for calculating the threshold for important terms. The instances are used in order to expand the stem and query the text for sentences containing the exact keyterms. In the extracted sentences, we replace the keyterm with a blank and offer it as an FIB item together with the keyterm as its correct answer. Then, applying some linguistic rules on the keyterm yields a set of distractor suggestions.

5.1 Keyterms Extraction

In order to extract the terms which are central for the instructional material and to filter out the less important ones we establish some requirements a keyterm should adhere to:

– keyterms are nouns and noun phrases from the text, which frequency is higher than a set threshold;

– keyterms are the noun phrases, which contain a keyterm with frequency higher than the set threshold;

– all names are keyterms.

The first step in this respect is to extract all potential keyterms. In previous research, we have concentrated on extracting nouns, noun phrases of the type $np - A - N$ and names, but now, in order to extend the list of valuable keyterms, we have inserted an additional type of noun phrases: $np - N - PP$. In domains like Law, where the specific terms tend to be longer, exactly this structure greatly helps in detecting keyterms. After all potential keyterms are extracted they are stemmed and the two lists of terms – the stemmed and the original one – are arranged in the internal representation shown above.

As we have determined in previous research and is reported also by other authors [6], in instructional materials the keyterms are often repeated in order to make the learner remember them. That is why simple term frequency is a better measure than TF-IDF, which tends to lower the score of the most often used words. We store the frequencies of our terms (f_{ti}) in the *occurences* attributes for each stem. We sort these frequencies and calculate the number of words having equal frequency ($r_{f_{ti}}$). Then we set the threshold as follows:

$$threshold = max f_{ti} - 2, \{r_{f_{ti}} \geq f_{ti}\} \qquad (1)$$

We have established this procedure for threshold setting empirically, by observing manually prepared test items and analysing the keyterms used in questions and answers.

Once the threshold is set, we initialise the list of keyterms by adding to it all nouns or noun phrases[4] that have frequency higher than the *threshold*. In the second step, we add to the keyterm list all noun phrases that contain a keyterm, without regarding their occurrence frequency. In a third step, we add all names as keyterms. For example: Along with the stem of the noun право (Right/Law) all these noun phrase stems will be added to the list of keyterms:

право на жалби (right of complaint)
право на живот (right of living)
право на законодател инициатив (right of legislation initiative)
международ прав (International Law) etc.

All NPs containing stop words are removed. In our case stop words most often appear to be personal and possessive pronouns.

The belonging of each stem to the keyterm list is implemented by turning the value of the attribute *isKeyterm* to *true* in the internal representation shown above.

5.2 Stem Extraction

We aim to produce a MCTI, which is of FIB type, and along with it, to suggest a correct answer and possible distractors. Seen from this perspective, our task may be thought of as vocabulary testing where optional answers are available. Taking into account the constraints we have put on the extraction of keyterms, we decided to relax the syntactic restrictions about the position of the keyterm, resp. the blank. As only requirement in this respect, we set the extraction of well-formed sentences. In terms of the grammar we use, these are sentences wrapped in a VPS constituent[5].

We extract all sentences from the text which contain at least one of the keyterms. For each of the keyterms in a sentence, we check whether it is a part of a longer keyterm:

– if it is not, we replace it with blank and save the so-produced stem;

– if it is contained in a longer keyterm, then we replace with blank the longest keyterm it is a part of and then save the stem.

Consider the following example. A sentence containing a keyterm *право* (law) is:

Външната политика на Република България се осъществява в съответ-ствие с принципите и нормите на международното *право*.

(The foreign policy of Republic of Bulgaria is realised in conformity with the International Law.)

The longest sequence of words containing the keyterm *право* (right/Law) and being a keyterm is *международното право* (International Law). That is why the system catches *международното право* and replaces it with a blank and produces the stem:

Външната политика на Република България се осъществява в съответ-ствие с принципите и нормите на ....

Respectively in the following sentence:

Чужденците и чуждестранните юридически лица не могат да придобиват право на собственост върху земя освен при наследяване по закон.

[4] All string comparisons are done in stemmed fashion.

[5] VPS -head-subject verb phrase for full definitions - HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank), BulTreeBank Project Technical Report 05. 2004, http://bultreebank.org/TechRep/BTB-TR05.pdf

(The foreigners and foreign legal bodies cannot acquire land property rights except for the case of inheritance by law.)

the longest sequence containing the term *право* and being a keyterm is the phrase "*право на собственост върху земя*" (land property right). Hence the system will replace this keyterm with blank and will produce the following FIB stem (instead of considering for a keyterm *право* only).

Чужденците и чуждестранните юридически лица не могат да придобиват .... освен при наследяване по закон.

5.3 Suggestion of Distractors

In well designed multiple-choice questions, the distractors are semantically close to the correct answer, as well as to each other in a sense. On the basis of our previous work and observations over manually prepared tests we suppose that distractors are close to each other if they look alike, too. Very often in tests for beginners, the distractors are noun phrases which contain the same noun as the one in the correct answer but with a different modifier or the other way round - the same modifier, but different noun. Our approach was based mainly on this assumption so far and now we want to extend the idea to the following:
– distractors of NPs of the type $np - N - PP$ are only NPs of the same type and the same head noun;
– distractors of NPs of the type $np - A - N$ are only NPs of the same type, which contain the same noun and different modifier or contain the same modifier and different noun;
– distractors of nouns are all NPs containing the given noun;
– distractors of names are names of the same type (for ex. for a keyterm which is a name of the type Org all names of the type Org are distractors).

The distractors are matched with the keyterms in a stemmed fashion too. Later on they are expanded to their full form and they are offered to the test-designer. Given the previously shown examples we may produce the following distractors:

Stem: Външната политика на Република България се осъществява в съответ-ствие с принципите и нормите на международното *право*. (The foreign policy of Republic of Bulgaria is realised in conformity with the International Law.)

Keyterm/correct answer: *международното право* (the International Law)

Type: np-A-N

Possible distractors: вътрешното право (the Domestic Law); избирателно право (franchise)

Consider the case when the keyword is a part of np-N-PP phrase:

Stem: Чужденците и чуждестранните юридически лица не могат да придобиват право на собственост върху земя освен при наследяване по закон. (The foreigners and foreign legal bodies cannot acquire land property rights except in case of inheritance by law.)

Keyterm/correct answer: *право на собственост върху земя* (land property right)

Type: np-N-PP

Possible distractors: *право на адвокат-ска защита* (right of advocate defense), *право на жалби* (right of complaint), *право на живот* (right of life), *право на законодателна инициатива* (right of legislation initiative), *право на лична свобода* (right of personal freedom), *право на ползване* (right to use), *право на строеж* (right of construction), *право на труд* (right to work).

When the keyterms are names, all names of the same type are distractors to each other. For example when processing the Bulgarian Constitution the names of the type NE-Loc are only *София* (Sofia) and *България* (Bulgaria) and they are treated always as options to each other. As Sofia is a city and Bulgaria a country they hardly assimilate each other and unfortunately we can not cope with this issue yet as we do not rely on any external resources which could deliver us this additional information. Consider the following problematic example where a question complying with this rule would look like.

Stem: Столицата на Република Българиа е град София. (The capital of Republic of Bulgaria is Sofia.)

Keyterm/correct answer: *София* (Sofia)

Type: NE-Loc

Possible distractors: *България* (Bulgaria)

We want to make clear that this is not the general figure but only a specific case. Our observation is that with the additional rule for distractors selection of names, the performance of the system significantly increases. Of course the recognition of the names is a matter of good parsing, but as the parser model is trained on BulTreeBank, which contains annotation for named entities, we rely on comparatively high rate of recognition at least in the domains in which the parser was trained.

6 Testing and Evaluation

6.1 Assessment of Results Obtained from the Bulgarian Constitution

As a first try of evaluation of our system, we run it over an extracted part of the Bulgarian Constitution and asked several experts to evaluate the quality of the resulting MCTI and the features of the system. The Law domain is characterised with its relatively long terms and also the high frequency of the terms and importance of most of the sentences. The linguistic patterns we chose for extracting keyterms and distractors turned out to suit very well the keyterms in the

Law domain. This is apparent from the results shown in Fig 3. 75% of the input sentences were extracted as MCTI. The high number of selected FIB stems means that a high percentage of sentences in the text contain keyterms. This is explicable with the nature of the laws where the redundant information is reduced to a minimum.

The average number of suggested distractors is between 2 and 3. The number of distractors for the examined texts varies between 0 and 7 and often all of the distractors are good suggestions. For about one third of the resulting MCTI the system could not offer distractors, which is due to the fact that the distractors are selected only from the submitted instructional materials and no external sources are used.

The criteria for evaluating the results were the following:
Quality of the question (1-3):
1 - the question is not proper for testing learners on this material; 2 - the question is unclear; 3 - the question is a well formed sentence, concerning terms which are central for the instructional material.
Quality of the answer (1-3):
1 - the answer is not central for the instructional material; 2 - the answer is central for the instructional material but more specific or general than the desired answer; 3 - the answer is a central for the instructional material and concrete enough.

Fig. 3 shows the trends in average scores given by the experts when assessing the quality of the stems and correct answers (keyterms) of the selected questions (shown on the X-axis), according to the criteria given above (shown on the Y-axis). Both the stems and answers received most often the highest mark (3). Half of the evaluators have given the highest score (3) to all of the questions and the other half have given different marks maximum to the half of the sentences. This means that keyterms are correctly chosen by the algorithm and they have high importance for the material. Given this fact, we can explain the high score given to the stems by the fact that in Law the sentences structure is very compact. Almost each sentence represents a separate rule and they are often independent from each other. In this respect the legal texts differ significantly from texts in humanities like Geography and History where references are often used and terms are not that strictly defined. The high scores given for question quality could be also explained with the fact that the questions are directly extracted from the text, and thus their grammatical well-formness is preserved.

6.2 Discussion

Our aim with this work was to explore the field of automatic generation of MCTI and to prove the availability and compatibility of the language resources and technologies for Bulgarian as well as to assess the readiness for the implementation of these techniques in real-world applications. We were attracted by the high level of the work done in this area for English and we wanted to check whether it is possible to build a working prototype using some existing tools and to make inferences about the directions in which language technologies (LT) development for Bulgarian should take. Given the fact that the state-of-the-art in LT for English and for Bulgarian is incomparable, we wanted to point out concrete steps which must be taken in consideration to help the development of the next generation LT for Bulgarian.

From this experiment we can clearly point out several decisive factors, whose improvement will lead to more satisfactory results and overall progress in the area. First of all, as a fundamental basis of all further processing, improvements in parsing will result in more correct extraction of target morpho-syntactic structures. In the experiment we noticed that for documents from the same domains as the ones in the training corpus the parsing performs with very high precision which is comparable to the state-of-the-art results declared by the parser-developers, but for other documents, however, the precision drops dramatically. This is due to the fact that the parser is fully statistical and does not accept any additional POS input with the parsing string as some other parsers do. Improved syntactic analysis would mean more correct keyterm extraction and better distractors selection.

Our work so far, although employing several different language processing techniques is strongly dependent on the parsing results and limited to the lexemes available in the instructional material. A complementary resource like a Thesaurus of any kind would give us the options to go beyond the limits of the processed text and will extend the capabilities of our system. Dictionary of synonyms/antonyms, dictionaries of names in Bulgarian will also be of great help for defining better possible distractors and go one step further and form a question-like stem instead of a FIB one. For this purpose in future we intend to integrate BalkaNet [6] as a component in the described system.

The lack of additional resources for conceptual processing in Bulgarian is tangible. A terminological dictionary would set a common terminological frame for the analysed materials and would facilitate the keyterm and distractors selection; dictionaries of names would be of great help in defining better possible distractors as well, variety of annotated corpora in different domains would improve the parser performance.

When talking about resources we must mention as well the quality of the input resources. From the processing algorithm it is clear that some kinds of texts are hard to analyse. For example, tabular data just transformed to plain text format will would not constitute good sentences. Mathematical or chemical formulae will hardly fit in any of the patterns adapted for other domains. The input used for similar systems should be carefully adjusted for the specific needs.

Stemming seems to be satisfactory enough. There is no need for applying lemmatisation on extracted terms. In the observed samples, we have not found examples of overstemming or understemming which would be better solved by lemmatisation. We explain this with the fact that after stemming we work mostly with phrases and then inflexional ambiguity is much lower which makes the technique for transforming the wordforms to a single one (stem/lemma) less significant.

[6] Multilingual lexical database comprising of individual WordNets for the Balkan languages

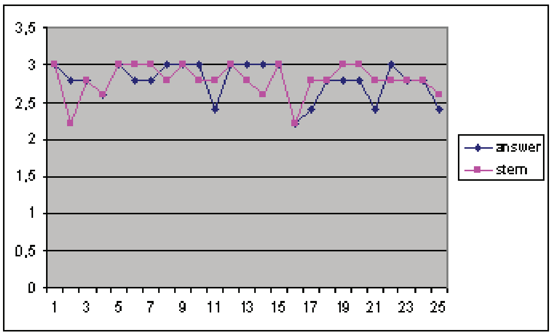


Fig. 3: *Average scores given to stems and answers.*

The chosen morpho-syntactic categories prove to be efficient and catch most of the terminology available in the instructional materials. We are especially satisfied with the addition of the noun phrases of the type np-N-PP which tend to match keyterm phrases in domains with comparatively longer terms like Law. We notice that even more categories coul be added (like the ones satisfying the regular expression $A^{+}N$). In comparison with previously reported work we noted that the new approach in distractor suggestion gains significant improvement from filtering useless distractors. Our expectation is that in a large-scale evaluation, the distractors, which are names, would contribute significantly to the overall efficiency of the system.

Under a direct comparison, the results we obtain for Bulgarian are not as good as those obtained for English, but this discrepancy can be explained by the presence of much more sophisticated language technologies for English. The presence of such tools and resources for Bulgarian will help us to gain conceptual knowledge about the target terms, to build more semantically-grounded distractors and to better filter significant from insignificant terms and sentences. Due to the limited resources available for Bulgarian, the capabilities of our system are also limited. However, we have implemented the main idea of the automatic MCTI generation and have shown what can be done with some of the existing language resources for Bulgarian as well as we have also scatched the gaps that need to be addressed in the future.

7 Conclusion and Future Work

Our aim with this work was to explore the field of automatic MCTI generation and to prove the availability and compatibility of the language resources and technologies for Bulgarian as well as to assess the readiness for the implementation of these techniques in real-world applications.

Our ideas for future development are related to experiments with a larger variety of question types and better distractors selection by involving dependency parsing and more external resources. We are working on improvement of the user interface as it is a main issue concerning the test designers' efficiency and will allow a real-time evaluation. Deeper evaluation, including classical test theory and error analysis in order to improve the output is also one of our future goals.

References

[1] I. Aldabe, M. L. De Lacalle, and M. Maritxalar. Automatic acquisition of didactic resources: generating test-based questions. In I. F. de Castro, editor, *Proceeding of SINTICE 07*, pages 105–111, 2007.

[2] J. C. Brown, G. A. Frishkoff, and M. Eskenazi. Automatic question generation for vocabulary assessment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[3] C. Fairon. A web-based system for automatic language skill assessment: Evaling. In *Proceedings of Computer Mediated Language Assessment and Evaluation in Natural Language Processing Workshop*, pages 62–67, 1999.

[4] A. Hoshino and N. Hiroshi. A real-time multiple-choice question generation for language testing: A preliminary study. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 17–20, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[5] C.-L. Liu, C.-H. Wang, Z.-M. Gao, and S.-M. Huang. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 1–8, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[6] R. Mitkov, L. A. Ha, and N. Karamis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12.:177–194, 2006.

[7] P. Nakov. Bulstem: Design and evaluation of inflectional stemmer for bulgarian. In *Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics*, Thessaloniki, Greece, 2003.

[8] I. Nikolova. Language technologies for instructional resources in bulgarian. In K. Balogh, editor, *Proceedings of 13th Student Session at ESSLLI 2008*, pages 135–142, 2008.

[9] S. Petrov, L. Barrett, R. Thibaux, and D. Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July 2006. Association for Computational Linguistics.

[10] S. Petrov and D. Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April 2007. Association for Computational Linguistics.

[11] E. Sumita, F. Sugaya, and S. Yamamoto. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 61–68, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.