Bilingual Corpus – Digital Repository for Preservation of Language Heritage

Ludmila Dimitrova, Radovan Garabík

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
Ľudovit Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
ludmila@cc.bas.bg,garabik@kassiopeia.juls.savba.sk

Abstract. The article briefly reviews bilingual Slovak-Bulgarian/Bulgarian-Slovak parallel and aligned corpus. The corpus is collected and developed as results of the collaboration in the frameworks of the joint research project between Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, and Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. The multilingual corpora are large repositories of language data with an important role in preserving and supporting the world's cultural heritage, because the natural language is an outstanding part of the human cultural values and collective memory, and a bridge between cultures. This bilingual corpus will be widely applicable to the contrastive studies of the both Slavic languages, will also be useful resource for language engineering research and development, especially in machine translation.

Keywords: Natural Language, Bilingual Corpus, Parallel Corpus, Aligned Corpus, Annotation

1 Introduction

The multilingual digital resources are valuable multilingual dataset for language engineering research and development; they contribute to preserving and supporting the multilingual and multicultural world heritage, of which language is an outstanding part.

For many of the pair languages, so called low and medium density languages, there are no multilingual or bilingual resources easily available for scientific community. The parallel sentence-aligned Bulgarian-Slovak/Slovak-Bulgarian corpus is currently developed in the framework of the joint research project between IMI–BAS and ĽŠIL–SAS, coordinated by L. Dimitrova and R. Garabík. The corpus is widely applicable to the contrastive and terminology studies of the both Slavic languages and will also be useful digital resource for machine translation research, for lexical and terminology databases and bilingual dictionaries development.

2 Standards and Models for Corpora Encoding

In linguistics, a corpus or text corpus is a large and structured set (repository) of texts, stored and proceed electronically. A corpus may contain texts in a single language (monolingual corpus) or texts in multiple languages (multilingual corpus). Multilingual corpora, as great repositories of digital natural language data, are very useful for preservation and support of language cultural heritage.

A multilingual corpus that consists of parallel texts (a text placed alongside its translation or translations) is called a parallel corpus. An aligned parallel corpus is a corpus containing relations between corresponding chunks of text of multiple languages. An alignment is the process of relating pairs of words, phrases, sentences or paragraphs in the texts in different languages which are translation equivalent. Commonly, parallel corpora are aligned at the sentence level – the alignment aims to produce a set of corresponding sentences (original and its translation(s)). (One of the most well-known examples of parallel text alignment is inscribed on the famous Rosetta stone.) The result of the alignment of two parallel texts is a merged document, called usually bi-text, composed of both source- and target-language versions of a given text that retains the original sentence order. The software tools, generating bi-texts, are called alignment tools, or bi-text tools, which automatically align the original and translated versions of the same text. The tools generally match these two texts sentence by sentence.

In order to make the corpora more useful for linguistic research, they are often subjected to a process known as annotation: corpus annotation is the process of adding linguistic or structural information to a text of the corpus. An example of a corpus annotation is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) in the form of labels – tags – is added to the words of the corpus. Another example of a linguistic annotation is indicating the lemma – the base form of each word. The structural annotation of a corpus indicates the different structural levels of a corpus or text, for example, parts and chapters of novels, sections of newspapers, articles of reference works, etc. The most common division in this structural hierarchy is the paragraph. The structural annotation allows the texts in the two languages (Bulgarian/Slovak and vice versa) to be aligned at the corresponding level in order to produce aligned bilingual corpora.

2.1 Bulgarian Parallel Corpora

Bulgarian digital multilingual resources were developed for the first time under multilingual research project MULTEXT-East (*Multilingual Text Tools and Corpora for Central and Eastern European Languages*, MTE for short), (1995-1997). The MTE project is a continuation of MULTEXT project (*Multilingual Text Tools and Corpora*), [8], that produced the language resources for six western European languages (Dutch, English, French, German, Italian, and Spanish) and a freely available set of extensible, coherent, and language independent tools for natural language processing (NLP).

MTE project developed significant language resources for six Central and Eastern European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as for English [1]. Three of these languages belong to the Slavic language group: Bulgarian, Czech, and Slovene. The MTE digital lexical resources include a dataset of language specific resources and multilingual MTE corpus, produced as a well-structured and lemmatized CES-corpus [8, 9]. The MTE project has succeeded in providing foundational resources for work in language engineering in Bulgarian, for morphological, grammatical, semantic or other research, or as the basis for development of new applications in NLP.

Recently created Bulgarian-Polish [4] and Bulgarian-Polish-Lithuanian [5] parallel corpora follow the MTE model for the development of parallel corpora.

2.2 Slovak Parallel Corpora

The first bilingual parallel corpus has been the Slovak-Russian parallel corpus, produced as a common project of ĽŠIL and Saint-Petersburg State University [7], accessible through a web interface. It was later followed by the parallel Slovak-French corpus [12], using very similar structure and interface. Both of these corpora are lemmatized and POS tagged.

The Slovak-Czech parallel corpus contains mostly translations of fiction between Czech and Slovak (in both directions), with a small part of independent translations from other languages (75 % English, the rest German, Polish, Italian, French and Ancient Greek) into both Czech and Slovak. Both Czech and Slovak parts of the corpus are morphologically analyzed and lemmatized with the ***morče*** software, the Czech part uses the tagset used in the Czech National Corpus.

The Slovak-English corpus consists of original English language fiction and their Slovak translations. The texts were obtained from various sources; the bulk of the Slovak translations were already collected in the Slovak National Corpus, some of them were scanned, OCRed and then proofread. A small amount of OCR-induced errors remained in the texts, but presumably, it will not have significant effects on the overall corpus quality. The Slovak texts are automatically morphologically annotated with the same tagset as in the Slovak National Corpus. The English texts are lemmatized and POS-tagged, using the TreeTagger software [11].

All these parallel corpora use the Hunalign software for the alignment [13].

3 Description and Current Content of Bilingual Bulgarian-Slovak/Slovak-Bulgarian Corpus

The parallel and aligned Bulgarian-Slovak/Slovak-Bulgarian corpus is currently developed as language resource for the research and development of machine and human translation systems, language analysis, automatic term extraction, semantic analysis, supervised and unsupervised NLP tools training, incl. for machine translation. All collected texts in this corpus are texts published in and distributed over the

Internet, so copyright issues for the texts are not a concern. The recent version of the corpus is available via a web interface at http://korpus.sk:8090/.

3.1 Specific Features of Bulgarian and Slovak Languages

These two Slavic languages exhibit some specific features, occurring repeatedly in several categories. First, there are different orthography traditions – the corpora are based on written languages and the orthography forms an inseparable part of language analysis. Another significant feature is the analytic character of Bulgarian, and the synthetic character of Slovak. Old-Bulgarian had an elaborate case system but in the process of evolution from a synthetic (inflectional language) to an analytic (flectional) language Bulgarian has lost most of the traditional old Slavic case system. Bulgarian case forms were replaced with combinations of different prepositions with a common case form. Bulgarian has a grammatical structure closer to English or the Neo-Latin languages than other Slavic languages. Bulgarian indicates also some innovations such as a rich system of verbal forms, and a definite article that is morphological indicator of the grammatical category determination (definiteness) – one of the most important grammatical characteristics of the modern Bulgarian language, whereas the other Slavic languages lack the definiteness attributes altogether.

The differences between Bulgarian and Slovak morphology specification were analyzed: these differences can be caused either by inherent language dissimilarities, or to different way of analyzing morphology in traditional grammars. All the parts of speech category for Bulgarian and Slovak, with emphasis on the differences of the MTE morphology tagset [10], were analyzed in detail [3].

3.2 Structure of the Corpus

The corpus currently contains translations of fiction in both languages, either from Slovak into Bulgarian or from Bulgarian into Slovak. The main part of parallel corpus contains texts in other languages translated into both Bulgarian and Slovak. The corpus consists of two subcorpora: **direct** and **translated** [2].

The **direct** Bulgarian–Slovak parallel subcorpus consists of original texts in Bulgarian, such as novels and short stories by Bulgarian writers and their translation in Slovak, and original texts in Slovak, such as literary works by Slovak writers and their translation in Bulgarian.

The **translated** Bulgarian–Slovak parallel subcorpus consists of Bulgarian and Slovak translations of literary works in third language.

3.3 Bulgarian-Slovak/Slovak-Bulgarian Corpus – Parallel and Aligned

The Bulgarian–Slovak/Slovak-Bulgarian corpus contains parallel texts, aligned at the sentence level. To align the text on sentence level we used the Hunalign software. The program foresees the use of a corresponding bilingual dictionary to ensure a higher accuracy of the alignment; however, no such dictionary has been available for the use with the corpus.

The corpus contains 376 200 words in parallel texts, aligned at the paragraph level and at the sentence level. The set of aligned texts includes two Bulgarian novels: Dimitar Dimov's ***Doomed Souls*** (*Осъдени души*) and Pavel Vezhinov's ***The Barrier*** (*Бариерата*) and their Slovak translations, the novel of Slovak writer Klára Jarunková ***The silent wolf's brother*** (*Brat mlčanlivého vlka*) and its Bulgarian translation, the Slovak and Bulgarian translations of Jaroslav Hašek's ***The Good Soldier Švejk***.

The following tables present some examples of aligned sentences (without a dictionary) of the parallel and aligned corpus:

Table 1. Slovak original and Bulgarian translation of Jarunková's *The silent wolf's brother*

To sa teda rozhodne múdrejšie dá stráviť čas so Strážom a Bojom . Така че много по - мъдро е да прекараш времето си със Страж и Бой .
Keď som bežal od potoka , náročky som bral zákrutu poza chatu . Като се връщах тичешком от потока , взех нарочно завоя зад хижата .
Pri ľadovni obyčajne ležia , aby sa skryli pred slnkom , pretože bernardíny najväčšmi na svete nenávidia teplo . Обикновено те лежат край ледника , крият се от слънцето , защото от всичко на света бернардините най - много мразят топлината .
V lete úplne schudnú a sú celé utrápené od strachu , že už nikdy neuvidia sneh . През лятото съвсем отслабват и се измъчват от страх , че никога вече няма да видят сняг.

Table 2. Bulgarian original and Slovak translation of Vezhinov's *The Barrier*

Тя сякаш изчезна в някакъв пад , после отново се появи . Akoby sa bola odrazu pohrúžila do hlbokej vody a potom sa opäť vynorila .
– Аз съм била на един ваш концерт . – Bola som raz na vašom koncerte .
И знаете ли какво ми направи най - силно впечатление ? . . . A viete , čo ma najväčšmi zaujalo ? . . .
Вашата безупречна , изящна логика . Vaša bezchybná , skvelá logika .

Table 3. Slovak and Bulgarian translations of J. Hašek's novel *The Good Soldier Švejk*

„ Má ich tam byť dvanásť , " povedal Švejk , keď si upil. - Трябваше да са дванадесет - рече Швейк , като отпи.
„ Prečo myslíte dvanásť ? " opýtal sa Bretschneider. - Защо пък дванадесет ? - запита Бретшнайдер.
„ Aby to išlo do počtu , do tucta , lepšie sa to ráta a na tucty je to vždy lacnejšie , " odpovedal Švejk. - За по - лесно , като са дузина , по - лесно се броят , пък и на дузини всичко е по - евтино - отговори Швейк.

4 Applications of Bilingual Corpus

The corpora are the main base of knowledge in corpus linguistics. The digital bilingual corpora are widely applicable to the contrastive studies of Slavic languages [6]. The structural annotation allows the texts in the two languages (Bulgarian/Slovak and vice versa) to be aligned at the corresponding level in order to produce aligned bilingual corpora. Currently, the corpus is automatically aligned at the sentence level without the help of a bilingual dictionary.

The aligned at the sentence level parallel corpora give more correct approach – in contrastive studies, we are not comparing "word" with "word", we compare word-forms in a broader context, which allows us to obtain the word's meaning. Such corpora are prerequisite for contrastive studies or other linguistics research, and can also be used for searching/extracting of linguistic information. There one could find and extract many examples of a word's usage because the corpus provides samples of the word's meaning and usage in a wide context.

The main area of application of the corpora is language translation. The new developments are mainly in the direction of computer means to support the translators in their activities. The bilingual aligned corpora are valuable resources for many NLP applications: in systems for machine-aided human translation, for machine translation research; and can be also used as a translation database and language learning materials for training of translators – human and programming tools.

The web-presented bilingual aligned corpora are oriented both to human and machine users and are available for a wide area of applications: corpora and frequency lists derived from them are useful for language teaching. Recently, the aligned corpora serve as a basis for development of new applications in multilingual digital libraries.

In addition, the aligned corpora are the best resource for the development of bi- and multilingual lexical or terminological databases, different kinds of digital dictionaries, and other special type of lists, namely concordances.

A concordance is a list of occurrences of a given (specified) word or phrase used in given large text, a book or a corpus, together with their immediate context.

Concordances have many applications in contrastive studies: they are used for comparison of different uses of the same word (in a different context) and for creating indexes and lists of words; in a keyword analysis and analysis of the frequency of words; to locate and analyze phrases and idioms in a given text; to find the translation of the essential elements of text, such as terms (in multilingual texts).

A dialogue box of the request for searching in Bulgarian–Slovak/Slovak-Bulgarian corpus is presented at the **Fig.1**:

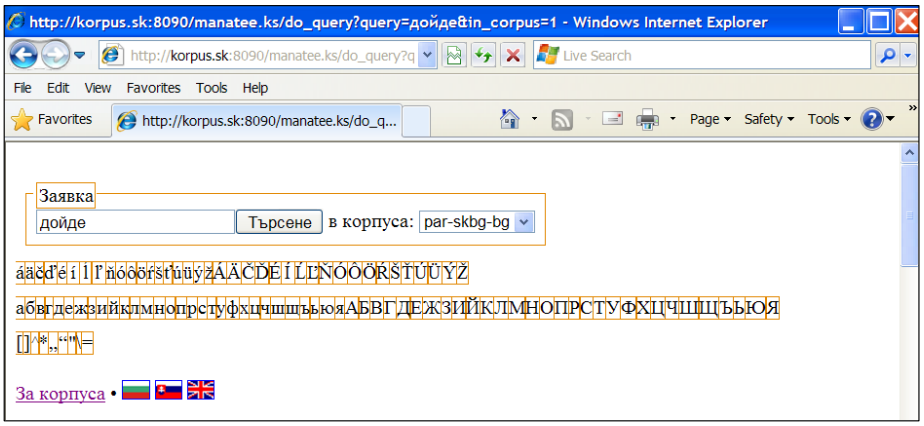
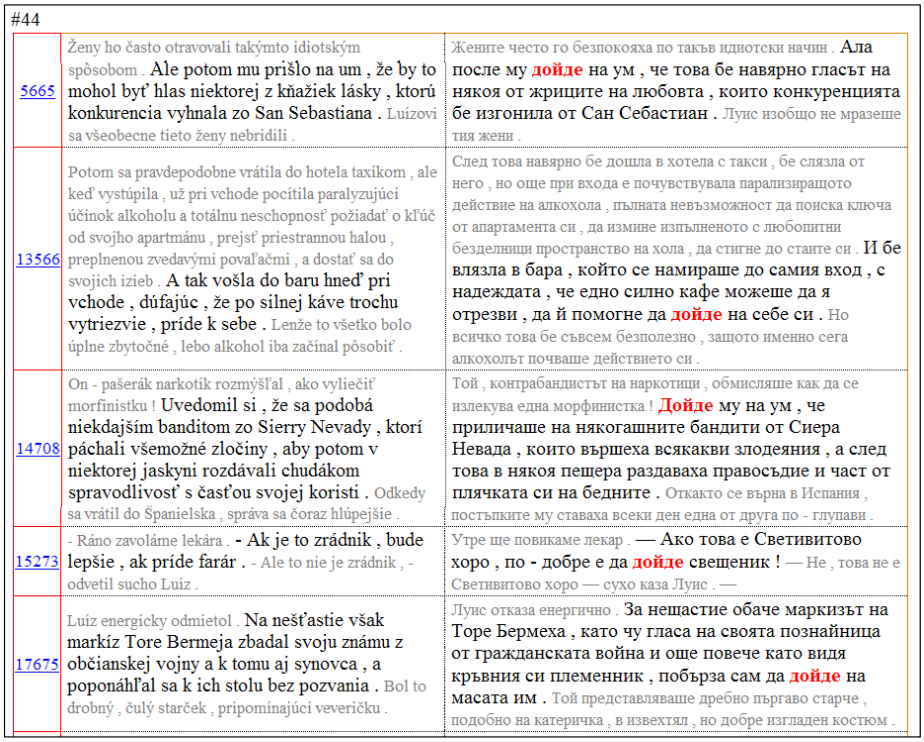


Fig. 1. Web search interface – a dialogue box in Bulgarian: searching the wordform ***дойде***

The result of the request execution (some of the concordances of the wordform ***дойде*** of Bulgarian verb ***дойда*** /come, arrive/) follow at the **Fig. 2**:



#44

5665	Ženy ho často otravovali takýmto idiotským spôsobom . Ale potom mu prišlo na um , že by to mohol byť hlas niektorej z kňažiek lásky , ktorú konkurencia vyhnala zo San Sebastiana . Luízovi sa všeobecne tieto ženy nebridili .	Жените често го безпокояха по такъв идиотски начин . Ала после му **дойде** на ум , че това бе навярно гласът на някоя от жриците на любовта , които конкуренцията бе изгонила от Сан Себастиан . Луис изобщо не мразеше тия жени .
13566	Potom sa pravdepodobne vrátila do hotela taxíkom , ale keď vystúpila , už pri vchode pocítila paralyzujúci účinok alkoholu a totálnu neschopnosť požiadať o kľúč od svojho apartmánu , prejsť priestrannou halou , preplnenou zvedavými povaľačmi , a dostať sa do svojich izieb . A tak vošla do baru hneď pri vchode , dúfajúc , že po silnej káve trochu vytriezvie , príde k sebe . Lenže to všetko bolo úplne zbytočné , lebo alkohol iba začínal pôsobiť .	След това навярно бе дошла в хотела с такси , бе слязла от него , но още при входа е почувствувала парализиращото действие на алкохола , пълната невъзможност да поиска ключа от апартамента си , да измине изпълненото с любопитни безделници пространство на хола , да стигне до стаите си . И бе влязла в бара , който се намираше до самия вход , с надеждата , че едно силно кафе можеше да я отрезви , да й помогне да **дойде** на себе си . Но всичко това бе съвсем безполезно , защото именно сега алкохолът почваше действието си .
14708	On - pašerák narkotík rozmýšľal , ako vyliečiť morfinistku ! Uvedomil si , že sa podobá niekdajším banditom zo Sierry Nevady , ktorí páchali všemožné zločiny , aby potom v niektorej jaskyni rozdávali chudákom spravodlivosť s časťou svojej koristi . Odkedy sa vrátil do Španielska , správa sa čoraz hlúpejšie .	Той , контрабандистът на наркотици , обмисляше как да се излекува една морфинистка ! **Дойде** му на ум , че приличаше на някогашните бандити от Сиера Невада , които вършеха всякакви злодеяния , а след това в някоя пещера раздаваха правосъдие и част от плячката си на бедните . Откакто се върна в Испания , постъпките му ставаха всеки ден една от друга по - глупави .
15273	- Ráno zavoláme lekára . - Ak je to zrádnik , bude lepšie , ak príde farár . - Ale to nie je zrádnik , - odvetil sucho Luiz .	Утре ще повикаме лекар . — Ако това е Светивитово хоро , по - добре е да **дойде** свещеник ! — Не , това не е Светивитово хоро — сухо каза Луис . —
17675	Luíz energicky odmietol . Na nešťastie však markíz Tore Bermeja zbadal svoju známu z občianskej vojny a k tomu aj synovca , a poponáhľal sa k ich stolu bez pozvania . Bol to drobný , čulý starček , pripomínajúci veveričku .	Луис отказа енергично . За нещастие обаче маркизът на Торе Бермеха , като чу гласа на своята познайница от гражданската война и още повече като видя кръвния си племенник , побърза сам да **дойде** на масата им . Той представляваше дребно пъргаво старче , подобно на катеричка , в извехтял , но добре изгладен костюм .

Fig. 2. Concordances of the Bulgarian wordform ***дойде*** in the corpus

Fig. 3. Concordances of the Slovak noun ***človek*** /man, person/ in the corpus

5 Future Work

Future work will focus on the interface: to ensure an efficient visualization of these massive and humanities relevant data. A creation of a web page for presentation of the parallel bilingual digital resources, with appropriate trilingual interface in Bulgarian, Slovak and English for easy access to the corpus, is envisaged.

In order to achieve reasonable quality of the corpus, the alignment should be as precise as possible. The project aims to create a small experimental bilingual Bulgarian-Slovak/Slovak-Bulgarian dictionary (several thousand words) suitable for automatic alignment of the bilingual corpus.

We will also continue the enlargement and the enrichment of the first Bulgarian-Slovak parallel and aligned corpus (in volume and by additional annotated information).

6 Conclusion

Parallel corpora are the most effective means for the creation of bi- and multilingual dictionaries and contrastive grammars. This is of great importance not only for language confrontation, but also for the typology of the studied languages. One has to remember that parallel corpora comprise direct material for the evaluation of translations and their analysis will bring out the improvement of the quality of both traditional, human translation, and machine translation. Besides, texts extracted from parallel or aligned corpora prove the necessity of evaluating translations: it is common that in translation words get omitted or word meanings get changed.

The parallel and aligned corpora are successfully used as language materials for the training of translators, as well as in education – for language learning in schools and universities. That is why online free-use parallel texts are also useful educational resource.

References

1. Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., and Tufis, D.: Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: COLING-ACL '98. Montréal, Québec, Canada, pp. 315–319 (1998)
2. Dimitrova, L., Garabík, R.: Bulgarian-Slovak Parallel Corpus. In: 6th International Conference NLP, Multilinguality. SLOVKO 2011, Modra, Slovakia, 20–21 October 2011, pp. 44–50 (2011)
3. Dimitrova, L., Garabík, R., Majchráková, D.: Comparing Bulgarian and Slovak Multext-East morphology tagset. In: Organisation and Development of Digital Lexical Resources. MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009, pp. 38–46 (2009)
4. Dimitrova, L., Koseska, V.: Bulgarian-Polish Corpus. J. Cognitive Studies/Études Cognitives. Vol. 9, SOW, Warsaw, pp. 133–141 (2009)
5. Dimitrova, L., Koseska, V., Roszko, D., Roszko, R.: Bulgarian-Polish-Lithuanian Corpus– Current Development. In: International Workshop "Multilingual resources, technologies and evaluation for Central and Eastern European languages" in conjunction with International Conference Recent Advance in NPL'2009. Borovec, Bulgaria, 17 September 2009. INCOMA Ltd., Bulgaria, pp. 1–8 (2009)
6. Garabík, R., Dimitrova, L., Koseska–Toszewa, V.: Web-presentation of bilingual corpora (Slovak-Bulgarian and Bulgarian-Polish). In: J. Cognitive Studies/Études Cognitives. Vol. 11, SOW, Warsaw, pp. 227–239 (2011).
7. Garabík, R. and В. П. Захаров.: Параллельный русско-словацкий корпус. In: Труды международной конференции Корпусная лингвистика, pp. 81–87, Санкт-Петербург, Издательство С.-Петербургского университета (2006)
8. Ide, N., Bonhomme, P., and Romary, L.: XCES: An XML based Encoding Standard for Linguistic Corpora. In: 2nd International Language Resources and Evaluation Conference. Paris: ELRA, pp. 825–830 (2000)
9. Ide, N., Veronis, J.: Multext (multilingual text tools and corpora). In: COLING'94. Kyoto, Japan, pp. 90–96 (1994)
10. MTE, 2004: MULTEXT-East Morphosyntactic Specifications – version 3, edition 10th May 2004 (2004)

11. Schmid, H. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, Studies in Computational Linguistics, pp. 154–164. UCL Press, London, GB. (1997)
12. Vasilišinová, D. and Garabík, R. Parallel French-Slovak Corpus. In *Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007*. Tribun, Brno, pp. 261–266. (2007)
13. Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pp. 590–596. (2005)