In the article the authors present the experimental Polish-Lithuanian corpus (ECorpPL-LT) formed for the idea of Polish-Lithuanian theoretical contrastive studies, a Polish-Lithuanian electronic dictionary, and as help for a sworn translator. The semantic annotation being brought into ECorpPL-LT is extremely useful in Polish-Lithuanian contrastive studies, and also proves helpful in translation work.

**Keywords:** corpora, parallel and comparable corpora, annotation, Polish, Lithuanian.

**Introduction**

The Internet provides vast text resources both for Polish and Lithuanian:

- The National Corpus of Polish (Narodowy Korpus Języka Polskiego, http://nkjp.pl/),
- The IPI PAN Corpus (Korpus Instytutu Podstaw Informatyki PAN, http://korpus.pl/),
- The PWN Corpus of Polish Language (Korpus Języka Polskiego Wydawnictwa Naukowego PWN, http://korpus.pwn.pl/);

Nevertheless, there are no corpora to amalgamate the resources of the both languages mentioned in the article’s headline. Obviously, there are small and big corpora vastly available on-line, including both Polish and Lithuanian texts. However, they do not meet basic corporal requirements, and that is why they do not make it possible to successfully conduct Polish-Lithuanian contrastive studies and construct a Polish-Lithuanian dictionary of standard value. The corpora are as follows:
Danuta Roszko & Roman Roszko

- ParaSol corpus — http://parasol.unibe.ch/
- Opus — http://opus.lingfil.uu.se/
- EMEA — http://opus.lingfil.uu.se/EMEA.php

The usefulness of the above-mentioned corpora and those similar to them is limited. First, the corpora resources are mainly based on specific texts (e.g. medical or union legislation). Second, the volume of universally available Polish-Lithuanian parallel corpora is insufficient for advanced linguistic studies.

1. Reasons for the experimental Polish-Lithuanian corpus coming into existence

In the years 2008–2011, a joint Polish-Bulgarian team (alphabetically) composed of Ludmila Dimitrova, Violetta Koseska-Toszewa, Danuta Roszko and Roman Roszko worked on the experimental Bulgarian-Polish-Lithuanian corpus (refer to Dimitrova, Koseska, Roszko, D. & Roszko, R. 2009a–b, 2010, 2011). The authors of herein had great hopes with the (BG-PL-LT) corpus. They counted on a tool that not only could streamline, but also provide a high factual standard for the Polish-Lithuanian contrastive studies conducted by them. Over the course of time, it turned out that it was not possible to create a Bulgarian-Polish-Lithuanian corpus whose resources would give a start to: (1) conduct advanced Polish-Lithuanian studies and (2) create dictionaries containing the contemporary lexis and terminology. One of the reasons for this was (except for translations from third languages) lack of common mutual translations (e.g. from Polish at the same time to Bulgarian and Lithuanian, etc). Some Polish novels (e.g. *In Desert and Wilderness* by Henryk Sienkiewicz, *Ashes* by Stefan Żeromski, to mention a few) were translated only into the Bulgarian language, whereas Lithuanian literature works were mostly translated into Polish and rarely into Bulgarian. Similarly, Bulgarian literature works were mostly translated into Polish and rarely into Lithuanian. For these very reasons, the texts common only for the two languages could not be included in the trilingual corpora. During the research into the experimental Bulgarian-Polish-Lithuanian corpus, it was found that in principle, except for an extensive union legislation base, there are no other texts coming into life on a large scale for these three languages. The economic cooperation resulting from the territorial closeness of Poland and Lithuania generates hundreds-thousands of mutually translated documents: bilateral agreements, terms of tenders, lists of tasks, business plans, joint European projects, resolutions, court conclusions and recommendations, correspondence etc. Moreover, the world markets division results in the fact that products hitting Poland and Lithuania are different from those hitting Bulgaria. A further consequence of that division is a small number of texts, of identical or similar content, common for Polish, Lithuanian and Bulgarian (common translations, mainly from English, being full of the contemporary terminology of many everyday life areas).

It should be emphasized that the experimental Bulgarian-Polish-Lithuanian corpus is not only a parallel corpus, but also a comparable one. Whereas the parallel subcorpus resources developed thanks to translations from third languages, the
comparable subcorpus resources did not have this possibility. The reason for this state of things is the location of the countries. Poland and Lithuania bordering with each other are a part of Central Europe, whilst Bulgaria is a Balkan state, recognized by Poles and Lithuanians as a holiday destination, gaining profits from tourism. Preliminary research into the Polish, Lithuanian and Bulgarian press and web sites in the years 2009–2011 confirmed the practical lack of common threads to be of interest to all the three nationalities. Any events important to Poland and Lithuania had no coverage in the Bulgarian press. And conversely, events widely analysed in the Bulgarian press went unnoticed by the Polish and Lithuanian media. Of course, in a jumble of information there were found texts concerning the same problems, however they concerned events of global character and usually came from the same sources, reported by mainstream news agencies. Nevertheless, if we limit our interests to hot issues common for Poland and Lithuania, it will turn out that there is a large number of texts complying with the conditions for comparable corpora. These are mostly texts referring to Polish-Lithuanian issues. There can be mentioned here independently-published Polish and Lithuanian articles, commentaries and reports concerning the same events and problems, sometimes having a different interpretation of facts and including the spelling of names and surnames, education, school-leaving examinations, textbooks, complaints, devastation of monuments, tablets, plaques, signboards with names of towns, establishing the joint Polish-Lithuanian committees, the Polish authorities’ visits to Lithuania and Puńsk, the Lithuanians’ governmental visits to Poland, and the like.

Facing the above-mentioned facts, the conclusion has arisen spontaneously. The emergence of a Polish-Lithuanian corpus is inevitable.

2. Creation of the experimental Polish-Lithuanian corpus /EcorpPL-LT/
2.1. First stage of the experimental Polish-Lithuanian corpus coming into existence
For nearly 20 years, the authors have been dealing with professional translation. The structures of translations and their conversions into other languages collected for years in the electronic form did not create any organised structure to give instant access to resources, lexis, terminology, etc to use when needed. The important aspect of the job of a translator is not to translate the same things repeatedly, but to keep the same terms and acronyms always identically conveyed in translation. Moreover, the authors deal with the Polish-Lithuanian linguistic contrast. They compare the both languages, describe the ways of formalization of particular semantic categories for Polish and Lithuanian. Therefore, the creation of a Polish-Lithuanian corpus has just been a matter of time.

The authors’ own translations were the first texts to be included in the corpus. At present, the volume of the corpus resources exceeds 6 million words. The resources do not include all sort of translations. Schematic records/documents are presented by one or two copies (e.g. a vehicle registration book, consignment note, identity card, birth certificate, death certificate, some agreements and other documents in the form imposed by the union legislation). The following documents are amongst the resources based on the authors’ own translations, such as the Civil Code of the Lithuanian Republic, particular acts of law (e.g. personal
income tax act) and directives, European (partner) projects, lists of medicines, activities, abridged and unabridged copies of corporate/business activity register, typical forms to use at tax offices, customs houses, police stations, social insurance institutions, insurance companies etc. Also, medical documentation (e.g. epicrises), court documentation (e.g. conclusions, resolutions, sentences, correspondence etc.), business and technical documentation (e.g. bilateral agreements, bid conditions, pleadings, certificates, specifications, operational and maintenance instructions, warranties, technical requirements, regulations, plans, brochures, commissioning documentation etc.

The significant majority of the records/documents collected in the first stage of ECorpPL-LT coming into existence are only for internal use. On the basis of them, the bases of Polish and Lithuanian equivalents of specialist terminology are being formed. To achieve this, there are used such programmes as ApSIC Xbench (http://www.apsic.com/en/products_xbench.html) and Terminotix LogiTermPro (http://www.terminotix.com/index.asp?name=LogiTerm_Pro&content=item&brand=2&item=12&lang=en).

2.2. Second stage of the experimental Polish-Lithuanian corpus coming into existence
At first, research work on ECorpPL-LT was narrowed to the texts the translation of which was being done by the authors of the article. In the course of time, a decision was made to form another sector of the corpus to include works universally available on the Internet, and also belles-lettres. The other sector of ECorpPL-LT is planned to be helpful for the Polish-Lithuanian contrastive studies and along with the main sector for creating a Polish-Lithuanian electronic dictionary.

ECorpPL-LT described herein has all features characteristic for parallel corpora. This is completely understandable, since such a purpose motivated the creators of the corpus. In a later period of time, facing some changes taking place in Poland and Lithuania, they decided to undertake an extra task, namely, to form a comparable corpus. The idea of the comparable subcorpus, different from the parallel one, consists in the inclusion (in the resources) of the texts which are neither mutual translations nor translations from other languages. There is a certain rule of the texts selection in this seeming chaos. That is, in Poland and Lithuania, there are coming into existence, irrespective of each other, the texts which have, however, the following in common: a topic, a similar size/content and date of edition. An example of this kind of texts is a report of the visit from Bronisław Komorowski, the President of the Republic of Poland, to a Polish school in Soleczniki:

- version one (of the report) by a Polish journalist, released in a Polish newspaper in the Polish language, and
- version two (of the report) by a Lithuanian journalist, released in a Lithuanian newspaper in the Lithuanian language.

3. Structure of the experimental Polish-Lithuanian corpus
ECorpPL-LT is a corpus created for research purposes. It is a typical bilingual corpus, whose resources are divided into two subcorpora:
A parallel,
B comparable.

There are two sectors distinguished within subcorpus A. Sector A1 are texts being the authors’ own translations, and sector A2 are texts representing different styles and kinds (incl. belles-lettres), not being the authors’ own translations.

The subcorpus A volume amounts to about 8 million words (sector A1 — over 6 million, sector A2 — below 2 million). The said numbers regard the resources in total for both the languages. Whereas, the volume of subcorpus B amounts to about 200 000 words.

3.1. EcorpPL-LT. Subcorpus A
3.1.1. EcorpPL-LT. Subcorpus A. Sector 1
An overview of ECorpPL-LT sector 1 is presented above in point 2.1. The texts have been aligned (at first, for this purpose a commonly available program, TextAlign by Andrew Manson, was used). Recently, however, because of the limited possibilities of the program the researchers have switched to other programs, commercial ones this time – Nova Text Aligner and Terminotix AlignFactory.

The goals for which sector 1 was created did not demand carrying out the morphosyntactic annotation. The resources were loaded to the Athel ParaConc program (http://www.athel.com/para.html) and additionally, there were created bases in the programs: ApSIC Xbench (http://www.apsic.com/en/products_xbench.html) and Terminotix LogiTerm Pro (http://www.terminotix.com/index.asp?name=LogiTerm_Pro&content=item&brand=2&item=12&lang=en), with the aim of determining the pairs of Polish-Lithuanian terms.

3.1.2. EcorpPL-LT. Subcorpus A. Sector 2
General rules leading to the creation of sector 2 are presented above in point 2.2. Here, first of all, there should be demonstrated the features to diversify sector 1 and sector 2. The texts being the authors’ own translations are in sector 1. Therefore, the principle of balanced character of the resources cannot be observed there. The selection of texts results from the character of tasks carried out. Of course, as stated above in point 2.1, schematic texts have not been copied in sector 1. In sector 2, care for appropriate internal balancing of texts has been taken. Diverse materials, representing a wide thematic range are being included in the resources. The resources of sector 2 comprise literary (representing different styles and kinds) as well as technical, medical, legal, judicial texts, materials connected with new technologies and civilisational achievements. In connection with the principles of internal balancing of the resources, commonly available resources of the union legislation have been withdrawn from uncontrolled inclusion. The union resources have been limited to a few essential ones in view of the resources of lexical acts, e.g.

• Commission Regulation (EU) No 715/2010 of 10 August 2010 amending Council Regulation (EC) No 2223/96 as regards adaptations following the revision of the statistical classification of economic activities NACE Revision 2 and the statistical classification of products by activity (CPA) in national accounts (nearly 140,000 unique terms in both languages)
  
  – pol. Woda w postaci naturalnej; usługi związane z uzdatnianiem i dostarczaniem wody,
  
  – lit. Gamtinis vanduo; vandens valymo ir tiekimo paslaugos,
  
  ‘Natural water; water treatment and supply services’.


  – pol. Działalność związana z produkcją filmów, nagrań wideo i programów telewizyjnych,
  
  – lit. Filmy cinema, vaizdo filmy ir televizijos programy gamyba,
  
  ‘Motion picture, video and television programme production, sound recording and music publishing activities’.

Similarly, synchronized medical texts, commonly available on the Internet have been withdrawn and limited to:

• Lists of medicines, e.g.

  – pol. Zleca lekarz neurolog, świadczący usługi trzeciego stopnia, jeśli leczenie choroby Parkinsona (kod G20 według ICD-10) produktami leczniczymi, zawierającymi analogi generyczne, i produktem leczniczym Rasagilinum, jest niedostatecznie skuteczne,
  
  – lit. Skiria gydytojas neurologas, teikiantis tretinio lygio paslaugas, for her Parkinsono ligos (kodas G 20 pagal TLK-10) gydymas vaistiniais preparatais, turinčiais generinius analogus, ir vaistiniu of preparation Rasagilinum yra nepakankamai efektyvus,
  
  ‘A neurologist, providing the third degree services, recommends if curing the Parkinson’s disease (G20 code according to ICD-10) with medicinal products, containing generic analogues, and with the medicinal product Rasagilinum is insufficiently effective’.
Experimental Polish-Lithuanian Corpus...

- International Statistical Classification of Diseases and Related Health Problems ICD-10, e.g.:
  - pol. Demielinizacyjne choroby centralnego układu nerwowego: stwardnienie rozsiane, inne demielinizacyjne choroby centralnego układu nerwowego,
  - lit. Demielinizuojančios centrinės nervų sistemos ligos: išsėtinė sklerozė, kitos demielinizuojančios centrinės nervų sistemos ligos,
  ‘Demyelinating diseases of the central nervous system: Multiple sclerosis, Other demyelinating diseases of the central nervous system’.


A considerable part of sector 2 is belles-lettres. An effort was made to complete mutual Polish-Lithuanian translations after the second world war, e.g. works by A. Kuklys, R. Černiauskas, J. Šikšnelis, E. Bialołęcka, S. Lem, W. Gombrowicz and others. Translations of world literature have also been included, e.g. works by P. Coelho, J. K. Rowling and others. Some of the works are presented as a whole, others as a representative part. At present, sector 2 comprises 36 pieces of belles-lettres. Further 60 pieces, including dramas and prose, are in preparation.

Apart from the literary works mentioned in sector 2, there are also technical texts, operating manuals, travel brochures, all sorts of guides etc.

According to the principles of ECorpPL-LT, the resources of sector 2 are to be aligned. To provide this, the above-mentioned programs Nova Text Aligner and Terminotix AlignFactory are used. Next, morphosyntactical annotation with the help of the programs Morpheus (http://sgjp.pl/morfeusz/) for the Polish language, and Anotatorius (http://donelaitis.vdu.lt/main.php?id=4&nr=7_2) for the Lithuanian language is carried out. At present, the resources of sector 2 have been loaded into the program Athel ParaConc (http://www.athel.com/para.html).

3.2. EcorpPL-LT. Subcorpus B

Subcorpus B is a typical comparable corpus. At the present stage of the development of this part of ECorpPL-LT, the resources have been stored in the electronic version and appropriately arranged within the directories reflecting the thematic tree. In each directory, besides two appropriate files (Polish and Lithuanian) there are informative files to hold data on the source, author of the text, date of publishing and basic keywords, comp. the example of metadata in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Example of texts included in subcorpus B (along with translations into English as well as metadata and keywords)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
</tbody>
</table>
| 2 | **Puńsk: dwujęzyczne tablce zamazane**  <br>**Puńsk: bilingual signboards vandalized**<br><br>**Autor: pl, bzato; Źródło: PAP, TVP Info**<br>*18:23*<br>*22.08.2011*<br><br>BNS  <br>*atnaujinta 14.43*  <br>*22.08.2011*<br>**Updated 14.43**<br><br>**In Lithuanian regions in Poland — excesses aimed at Lithuanian signboards updated 14.43**<br><br>**W gminie Puńsk (Podlaskie) nieznani sprawcy oblali białą i czerwoną farbą 14 tablic z litewskimi nazwami. Miejscowość, gdzie dopuszczono się wandalizmu, jest skupiskiem mniejszości narodowej naszych wschodnich sąsiadów.**  
*Lenkijos Punsko savivaldybėje, kur gausu lietuvių gyventojų, pirmdienio naktį raudonos ir baltos spalvos dažais užtepti lietuviški miestelių ir kaimų pavadinimai, išpaįšyti lenkų nacionalistų organizacijos ženklai, praneše miestelio vadovas. Lenkijoje gyvenantys lietuviai taip pat pasipiktino sprendimu nutraukti lietuvišku laidu transliacijas Białstogės televizijoje.*<br><br>**In the administrative commune of Puńsk (of the Podlasie region) unknown perpetrators vandalized 14 signboards containing Lithuanian names with white and red paint. The town, where the vandalism was committed, is a national minority community of our eastern neighbours.**  
*“In the Polish autonomy of Puńsk, where Lithuanians live in large numbers, on Monday night, Lithuanian-language signboards containing the names of small towns and villages were vandalized with white and red paint, and Polish nationalist organizations emblems were painted on them”, the head of a small town informed. The Lithuanians living in Poland were also scandalized by a decision to cease broadcasting a Lithuanian program from a television studio in Białystok.*
Skrajni nacjonalisci dali znać o sobie w Puńsku.
Wandale działali w nocy lub nad ranem. Nieznani sprawcy zniszczyli 14 tablic i jeden pomnik, na którym namentalowano farbą znak nacjonalistycznej organizacji Falanga — poinformował rzecznik podlaskiej policji Andrzej Baranowski. (…)

Extreme nationalists attracted public attention in Puńsk.
The vandals acted at night or before dawn. “Unknown perpetrators vandalized 14 signs and one monument on which the Falanga nationalist organization emblem was painted” — the Podlasie region police spokesman, Andrzej Baranowski, informed. (…)

Polska, województwo podlaskie, Puńsk, Sejnai; mniejszość litewska w Polsce; Falanga; tablice miejscowości, pomnik, zniszczone; wystąpienia antylitewskie; wójt gminy puńskiej, Witold Liszkowski

Lenkija, Palenkės vaivadija, Punskas, Seinai; lietuvių tautinė mažuma Lenkijoje; Falanga; kaimų pavadinimų ženklai, paminklas, sunaikinti; antilietuviškos įšpuoliai; Punsko valsčiaus viršaitis, Vytautas Liškauskas; LR konsulas Seinuose, Liudvikas Mišašius
Danuta Roszko & Roman Roszko

Poland, Podlachia province, Puńsk, Sejny; Lithuanian minority in Poland; Falanga; town signboards, memorial, vandalized; anti-Lithuanian rallies; Puńsk commune leader, Witold Liszkowski

| Poland, Podlachia province, Puńsk, Sejny; Lithuanian minority in Poland; Falanga; town signboards, memorial, vandalized; anti-Lithuanian rallies; Puńsk commune leader, Witold Liszkowski | Poland, Podlachia province, Puńsk, Sejny; Lithuanian minority in Poland Falanga; town signboards, memorial, vandalized; anti-Lithuanian rallies; Puńsk commune leader, Witold Liszkowski; Lithuanian consul in Sejny, Liudvikas Milašius |

Table 1 explanatory notes: Source data have been given in line 1. These are the web portals of the Polish TVP INFO TV and of the Lithuanian information portal Delfi.lt. The articles titles have been put in line 2. Line 3 provides the information source and the online publication date (also the date of last updating in the Lithuanian version). Line 4 provides the headlines distinguished by the publishers. Line 5 contains the initial fragments of both the articles. Line 6 gives keywords for the given texts. The deletion means that in the given text there is no information on the topic, but in the text contrasted with the given one such information is included.

4. Semantic annotation of subcorpus A sector 2
The semantic annotation is to be supplied for the parallel corpus (subcorpus A sector 2). It is a new kind of annotation, so far not met in corpus linguistics.

4.1. Morphosyntactic annotation as opposed to semantic annotation
All parallel corpora presently coming into existence have the morphosyntactic annotation. It is so because the morphosyntactic annotation is an indispensable corpus element and at the same time an indicator of the corpus quality. It is also obvious that the annotation facilitates the corpus exploration and makes searching more effective. That is why it is possible to ask such corpora a question the answer to which would be all cases of using any adjective in the plural genitive form coming at the beginning of a sentence. It is also possible to search for all uses of any derivative form of a given verb (e.g. participial, personal, verbal noun form etc). However, the rule should be followed here that the lemma for these forms must be the same, e.g. in the form of the infinitive. Yet, when verbal nouns, participles and personal forms of a verb have different lemmas, then finding all regularly created forms for a particular verb is not possible. Moreover, as far as the corpora based on the morphosyntactic annotation are concerned, it is not possible to ask a question in the syntax of which the argument structure would be any meaning. For example, in traditional corpora, it is not possible to give an order to search for the forms expressing the meanings of quantificative universality or imperceptive modality. It results from the applied morphosyntactic annotation the idea of which consists in
ascribing purely formal parameters to every form, i.e. morphological and chosen syntactic parameters connected with collocation, e.g.

PL. dom ‘home’: [lemma: home], MSD: noun+, masculine+, singular+, nominative+ (optional: Animal−, Human−, depreciativeness−, common+, countable+ etc.)

Let us consider some simple sentences:

LT Jonas jau kažką nupirkо.
John (has) already bought something.

LT Jonas dar kas nors nupirks ir einame.
John will still buy something and we are going.

In the Polish version of the two sentences [1–2] the identical form of coś ‘something’ occurs. However, the meanings connected with using this form in these sentences are not identical. In sentence [1] we say that the thing which John has bought exists — that is the thing (which is mentioned in the sentence) was chosen and bought by John. However, in sentence [2] we assume that potentially the thing which John will buy exists, and the thing can be every item being at the moment on offer at the store. As we can see, formally the Polish coś can have at least two meanings: real existentiality and habitual universality (as for the definition of these terms, comp. (Roszko, R., 2004), as for the superior terms of quantification, uniqueness, existentiality and universality comp. (Koseska-Toszewa, Gargov, 1990). Two Lithuanian formal equivalents certify for the fact that there are different meanings ascribed to the Polish form coś: in sentence [1] the Polish coś corresponds with the Lithuanian kažką, in sentence [2] the Polish coś corresponds with the Lithuanian kaškas. Other examples are provided by ECorpPL-LT:

LT Prietemoje kaškas blausiai švytėjo.
There was something unclearly glistening at dusk.

[4] PL Myślisz, że jest coś, co się oprze antymaterii?
LT Manai, kad yra kas nors, kas gali atsispirti antimaterijai?
Do you think there is something that will resist the antimatter?

In the pairs of sentences [1] and [3], and [2] and [4] identical meanings are conveyed, and this is ascribed respectively to the Polish form coś and the Lithuanian form kažkas (the meaning of real existentiality), and to the Polish form coś and Lithuanian form kas nors (the meaning of habitual universality).
Below, some examples of the first efforts to ascribe the semantic annotation to Polish sentences [1–2]. As for now, not all meanings get specified in the description. In this specific example the attention is focused on individual meanings of range quantification.

[1a] PL Jan już coś kupił.

Jan — uniqueness of the entity

już kupił — uniqueness of the past event

cóż — real existentiality of the entity

[2a] PL Jeszcze Jan coś kupi i idziemy.

Jan — uniqueness of the entity

jeszcze kupi — uniqueness of the future event

jeszcze kupi — uniqueness of the future event

i —

idziemy — uniqueness of the future event

Below, there are given particular elements of semantic annotation so far specified in the experimental part of sector 2 in subcorpus A:

Modality:
Possibility:
Imperceptiveness (testimonial modality):
I1 — neutral degree
I2 — enhanced degree
Hypothetical (modality):
H1 — degree of the lowest probability
H2–H5 — particular degrees of growing probability
H6 — degree of the highest probability

Range quantification:
Uniqueness of the entity / state / event
Uniqueness of the set of entities / states / events
Existentiality presupposing the individuality of the entity (-ies) / state (-s) / event (-s)
Real existentiality of the entity (-ies) / state (-s) / event (-s)
Habitual universality of the entity (-ies) / state (-s) / event (-s)
Real universality of the entity (-ies) / state (-s) / event (-s)

An example of a sentence with the semantic annotation grasping the meanings of possibility modality of hypothetical nature (comp. 3.1.2 Roszko, D. 2013, in this volume). As regards annotation, also comp. (Koseska-Toszewa 2013, in this volume). The annotation presented here in the above examples would be welcome not only by the corpora users, but also by the computer specialists working on the
experimental Polish-Lithuanian corpus.

Constant development of both the subcorpora is on table. It makes possible to include new texts and the full semantic annotation of the part of the corpus available online (that is sector 2 in subcorpus A). The corpus inclusion in general online resources will require applying a new software to organise the resources.

Summary

The experimental Polish-Lithuanian corpus is the first extended bilingual Polish-Lithuanian corpus whose resources have been divided into two subcorpora: parallel and comparable. The parallel subcorpus (A) is widely applied in contrastive studies carried out at the Institute of Slavic Studies of the Polish Academy of Sciences by the Corpus Linguistics and Semantics Team. Moreover, on the basis of the parallel subcorpus (A) a Polish-Lithuanian electronic dictionary and a Polish-Lithuanian terminological dictionary are coming into existence. The recipients of the parallel subcorpus (A) available online in the near future are supposed to be not only linguists, but also IT specialists, literary scholars, librarians, teachers, translators, specialists for linguistic information machine processing, programmers participating in creating automatic translation systems. Also, irrespectively of the education and the job being done, Poles studying Lithuanian (e.g. students) and Lithuanians studying Polish.

The semantic annotation planned for the parallel corpus (A) is bringing a new value into corpus linguistics. It reflects the content plan in isolation from the formal
side of both the languages. The semantic annotation is considered to have a big influence on the development of the machine translation.

The resources of the comparable subcorpus (B) are definitely more modest in comparison with the parallel subcorpus (A). However, the materials stored in the comparable subcorpus (B) reflect mutual Polish-Lithuanian relations, a little bit differing views about the world, history, nature etc demonstrated by Poles and Lithuanians. Therefore, making the subcorpus B available online is supposed to be of interest to wide circles of recipients, such as historians, ethnographers, folklorists, political scientists, sociologists, anthropologists, culturologists, researchers of the linguistic image of the world. The long shared history of Lithuania and Poland, the common border, the issues of the Polish minority in Lithuania and those of the Lithuanians living in Poland, also the issues of Polish schools in Lithuania and those of Lithuanian schools in Poland are among some problems to look at from Polish and Lithuanian perspective. This fact can result in people who shape up the foreign policy of Poland and the national minorities internal policy getting interested in the subcorpus B resources. There is no doubt that the Polish-Lithuanian comparable corpus (B) can be a valuable source of reliable information for linguists, history teachers, translators, students of different branches of humanities and social sciences and those searching the knowledge about the world, art etc.

References


Corpora and web resources

Anotatorius
ApSIC Xbench
Corpus of the Contemporary Lithuanian Language — Tekstynas. (Dabartinės lietuvių kalbos tekstynas)
EMEA
IPI PAN Corpus — Institute of Computer Science, Polish Academy of Sciences’ Corpus (Korpus Instytutu Podstaw Informatyki PAN)
KDE 4
MonoConc
Morpheus
National Corpus of Polish (Narodowy Korpus Języka Polskiego)
Opus
(http://opus.lingfil.uu.se/). 30.09.2012
ParaConc
ParaSol corpus
Polish Scientific Publishing House PWN Corpus of the Polish Language (Korpus Języka Polskiego Wydawnictwa Naukowego PWN)
Terminotix AlignFactory
Terminotix LogiTerm Pro