DANUTA ROSZKO

Institute of Slavic Studies, Polish Academy of Science, Warsaw
danuta.roszko@ispan.waw.pl

**EXPERIMENTAL CORPUS OF THE LITHUANIAN LOCAL
DIALECT OF PUŃSK IN POLAND.
EXAMPLES OF THE LEXICAL AND SEMANTIC ANNOTATION**

Abstract

In the article the author describes the experimental corpus of the Lithuanian local dialect of Puńsk in Poland (ECorp-of-Punsk). It is the first corpus of this type for the Lithuanian local dialect. The corpus consists of three subcorpora. The first one (referred to as fundamental) contains utterances given by Lithuanians in the local dialect, the second one — utterances given by Lithuanians in Polish, the third one — aligned Polish-dialectal texts. The texts recorded in the years 1986–2012 have been included in the Ecorp-of-Punsk resources.

Keywords: corpora, annotation, Lithuanian local dialect of Puńsk in Poland, experimental dialectal corpus.

Introduction

The development of corpus linguistics has been gaining momentum in the recent years. After a period of intensive work on monolingual corpora (the so-called national corpora created for standardized languages) and multilingual parallel ones (mainly in comparison with the English language) the time has come for forming the dialectal corpora. These, however, on account of the narrowed circle of potential recipients (mainly dialectologists) and incomparably large amounts of labour, as for now are not commonly formed. It cannot, however, be ruled out that as today in large numbers monolingual and multilingual corpora are coming into existence as in the future dialectal corpora will be developed. These are some examples of dialectal corpora: Catalan Corpus Oral Dialectal, Estonian Dialect Corpus, FRED — Freiburg Corpus of English Dialects, Helsinki Dialect Corpus, Nordic Dialect Corpus, Russian National Corpus (Dialectal corpus), YADAC — Dialectal Arabic Corpus etc. (see Corpora and Web Resources)

As far as dialectal corpora are concerned, the basic question is a limited access to materials. It is known that one of the features of local dialects is that they

don't have their own written version. Therefore, the first step to form a dialectal corpus is a recording of utterances within a given local dialect. It is a long-time task, and in many cases it requires a few years of work in the field. It is important to select informants on the grounds of generation, sex, education. You should also bear in mind that dialectal texts to be recorded should represent as broad lexical spectrum as possible. Converting these audio recordings to the text form is the next stage of work on dialectal corpora (e.g. to TXT files). An inherent problem at this stage of work is the form of record (phonetic or of transliterational). After converting the audio texts to text files, the way of annotation (morphological-syntactic, lemmatization) and metadata (the annotation containing the information on informants as well as the place and the date) should be established. Not always the morphosyntactic features of a local dialect and general language correspond with each other. Therefore, there is a need to define new morphosyntactic units for a local dialect.

1. The first stage of ECorp-of-Puńsk coming into existence

In the late 80-ties of the 20th century, an accidental recording of a conversation between Puńsk Lithuanians initiated a number of dialectological expeditions to Puńsk and its environs (north-east end of Poland, right by the border with the Lithuanian Republic) with the aim of recording the utterances given by people of Lithuanian origin.

In the years 1986–1992, short-term dialectological expeditions were run only in holiday months. The size of the recording equipment (which required plugging in) and the necessity to put the microphone in the direct proximity of the people talking might have influenced the subject matter and the way of giving utterances by respondents. Puńsk Lithuanians, knowing that they are being recorded, consciously avoided characteristic dialectal features and replaced them with literary equivalents.

After 1992, part of recording was made on video cassettes (with VHS-C and Hi8 cameras). The camera placed so as not to catch the locals' eye (with its recording function on) did not arouse any suspicions that anything was being recorded. In the 90-ties of the 20th century a frequency of trips to Puńsk and its environs rose. The expeditions were run not only in the vacation spring-summer months, but also in autumn-winter months. The advantage of the spring and autumn expeditions was not only a better opportunity to start a conversation with the locals (who have then less land work), but also a better quality of recording. In cold days the windows of their homes are usually closed, which considerably deadens sounds coming from the outside.

At the turn of centuries, different digital recorders (so-called dictaphones) came into use. Small sizes and relatively long time of the incessant recording are among the assets of the devices. Sometimes, the mentioned assets of dictaphones were exploited — e.g. at a shop counter where being left on made it possible for the recorded material not to get burdened with the possible influence of the researcher on the way of constructing the utterances by the respondent, also — on the content of the utterances.

The quality of recording dating from the years 1986–2006 is not one of the best. A high level of noise and different interferences are characteristic of them.

Moreover, the majority of the so-called optical miniDiscs lost their data (at present the disks read as empty). Fortunately, the minidiscs on account of their high cost never constituted a basic data carrier.

It should be emphasized that initially the best quality was provided by video recordings. Recently, handy dictaphones were completely withdrawn from being used and professional sound recorders and semi-professional video cameras came into use.

2. The second (main) stage of ECorp-of-Punsk coming into existence

For a long period of time the material was only being collected. However, an attempt of its systematic listing was never taken. It was not until the beginning of 2010 that a decision was made to convert the collected sound materials to text files. Then it turned out that the quality of recording on some cassette tapes was low (high noise level), and some miniDiscs completely lost their data. However, no problems with files coming from electronic recorders were found, despite the fact (on account of the easiness of copying the data) the files were frequently copied and put in different archival files. Loss of part of the recordings (originating from miniDiscs) and, from the present-day point of view, a poor quality of the first recordings on cassette tapes can bring the researcher to frustration. Fortunately, a considerable part of the (dialectal) recordings survived on video cassettes (VHS-C, Hi-8 and miniDV). The people who were accompanying the author on her dialectal expeditions recorded (independently of her) part of the conversations with a video camera, and the recording material is still kept by them.

The task of listing the recordings undertaken by the author turned out to be time-consuming. Polish and Lithuanian companies providing the service of listing of recordings did not show interest, whereas some of the Puńsk inhabitants willing to list the recordings unintentionally brought changes, which rather reflected their personal approach than that of the recorded respondents. The material listed in that way would require some detailed correction. After all, the author undertook the task. On account of a limited amount of time available for the author to spend on the ECorp-of-Punsk research, she decided to give up chronologically listing the utterances for the sake of a representative selection of the texts to be listed. Therefore, the author pays close attention to the proper relationships between the utterances given by Puńsk particular generation groups and between the years of making the recordings. Thanks to it, at every stage of the research the ECorp-of-Punsk linguistic material represents an almost thirty-year period of changes taking place in the local dialect of Puńsk. The author takes great care to make sure that part of the corpus resources originates from the same informants, which considerably raises the aspect of credibility of the changes taking place in the local dialect of Puńsk.

2.1. The dialectal material record problem

For the needs of ECorp-of-Punsk a simplified record (transliteration) has been used. It was well-known from D. Krištopaitė's works (1998, 1999) and earlier W. Smoczyński's studies (1984a,b, 1986a,b). Resignation from phonetic transcription resulted from a) the fact that the phonetic and phonological aspect of the

dialect having been sufficiently described, b) purposes motivating the creation of the corpus (semantic studies and morpho-syntactic description of the dialect), and c) the corpus form available to wide circles of researchers. In practice, the record of dialectal texts is based on the rules known from the orthographic record used for the standard Lithuanian. Only in places where the dialect and the standard language differ, the elements indicating this dissimilarity were introduced. For instance, a distinction between the phonemes [l] and [l'] is not applied in the orthographical record for Lithuanian because their distribution is unambiguous. The hard phoneme [l] appears before back vowels (e.g. [l]*aukas* 'field'), phoneme [l'] — before front vowels [l']*ekti* 'fly' and back vowels [l']*iaudis* 'nation'; 'the people', which, however, is indicated by the character *i* after *l* (= *liuadis*). Exceptions to the presented rule are possible in the dialect — the hard phoneme [l] may occur also before front vowels, for example *už-[l]-ėké* 'arrived'; 'came' (therefore in transliteration, the character *l*: *užlėké* was used toward the literary *užlėké* [už^(ʰ)l'ėk'ė]).

2.2. The text record format and the annotation.

In ECorp-of-Punsk all of the texts have been recorded in the standardized format. The standard of UTF8 coding and the format of the TXT record have been implemented.

ECorp-of-Punsk is annotated on the word level. A lemma has been ascribed to each lexeme, e.g.

medzy: word="medzy" lemma="medzis" (the noun tree)
dzirbo: word="dzirbo" lemma="dzirbc" (the verb work)

The ECorp-of-Punsk resources annotation is under compilation. On account of limited possibilities and time, there was a decision to use an annotator designed for Lithuanian, i.e. Anotatorius (http://donelaitis.vdu.lt/main.php?id=4&nr=7_1) — for the corpus resources annotation. Due to the differences between the standard language and the local dialect, such kind of solution is not a target. As part of experiment, the automatic annotation of a significant part of the resources by using the above-mentioned programme Anotatorius was carried out. Next, there were adjustments made by hand. There were some changes in the record introduced in order to keep the recognition accuracy of dialectal texts maximally high, for example, the dialectal *c* was changed consistently for the literary equivalent *t* and the dialectal *dz* for *d*. Thanks to this change, the correct annotation was acquired for the lexemes, which in the dialectal record would be unrecognized by the programme Anotatorius, comp:

The lexeme recorded in the slang version:
 <word="dzirbo" lemma="dzirbo" type="nežinomas"/>
 where "nežinomas" = "unknown"

The lexeme recorded according to the standards of Lithuanian:
 <ambiguous>
 <word="dirbo" lemma="dirbti(-a,-o)" type="vksm., teig., nesngr., tiesiog. n., būt. k. l., vns., 3 asm."/>

```
<word="dirbo" lemma="dirbti(-a,-o)" type="vksm., teig., nesngr., tiesiog. n.,
būt. k. l., dgs., 3 asm."/>
</ambiguous>
where "vksm., teig., nesngr., tiesiog. n., būt. k. l., vns., 3 asm." = "verb,
positive form, irreflexive, indicative, single past tense, singular, third person",
"vksm., teig., nesngr., tiesiog. n., būt. k. l., dgs., 3 asm."/> = "verb, positive
form, irreflexive, indicative, single past tense, plural, third person"
```

Having conducted the process of annotation automatically, the adjustment by hand is indispensable. You need to restore the lexeme dialectal form of lexeme and check the correctness of the lemma attributed to it. In case of the ambiguous form, the appropriate meaning is expected to be indicated, e.g:

```
<word="dzirbo" lemma="dzirbc" type="vksm., teig., nesngr., tiesiog. n., būt. k.
l., dgs., 3 asm."/>
where "vksm., teig., nesngr., tiesiog. n., būt. k. l., dgs., 3 asm."/> = "verb,
positive form, irreflexive, indicative, single past tense, plural, third person."
```

An example of the annotation of a dialectal sentence is presented below: *Aš tai sakiau, tį nieko neraikalaokit* ‘I said it so that you would demand nothing’:

```
<p>
<word="Aš" lemma="aš" type="Iv., vns., V."/>
<space/>
<word="tai" lemma="tus" type="Iv., neįvardž., bev. g."/>
<space/>
<word="sakiau" lemma="sakyc" type="vksm., teig., nesngr., tiesiog. n., būt. k.
l., vns., 1 asm."/>
<sep=","/>
<space/>
<word="tį" lemma="tį" type="prv., teig., nelygin. l."/>
<space/>
<word="nieko" lemma="niekas" type="dkt., vyr. g., vns., K."/>
<space/>
<word="neraikalaokit" lemma="nereikalauk" type="vksm., neig., nesngr., liep.
n., dgs., 2 asm."/>
<sep="."/>
<p/>
```

2.2.1. During the corpus resources automatic annotation carried out in the Anotatorius program, there were certain accuracies noticed between the percentage of the recognised text and the generation (young, middle, old) and the year of the utterances’ recording. As for the recordings of the late eighties of the past century, the utterances given by the old and middle generations are usually in an inconsiderable percentage recognised by the Anotatorius program. The majority of the forms are provided with the annotation ‘unknown’. As for the recordings

coming from XXI century, only the utterances given by the old generation do not comply with the process of annotation in the Anotatorius programme, which was to anticipate. The Lithuanian national minority inhabits the Polish-Lithuanian border regions, in the eighties of the 20th century — in the immediate vicinity of the USSR, and later — of the Republic of Lithuania. Until in Poland and in the neighbouring states political and economic transformations took place, the areas inhabited by the Lithuanian population were at the very end of Poland, entirely cut away from Lithuania (then the Lithuanian SRR) by the tightly guarded border. The Lithuanian national minorities in Poland were not usually in everyday contact with Lithuanians living abroad. Similarly, contacts with other inhabitants of Poland were not common. If a Puńsk Lithuanian left his place to study, he often came back to Puńsk after getting a university degree. Hardly anyone arrived in Puńsk or its environs from other areas of Poland. Simply, that was because of Puńsk being situated just at the border of Poland and the USSR, where no trade or tourist routes existed. A considerable distance from the centre of Poland as well as the fact that travelling to Puńsk one passed attractive tourist regions (e.g. Mazuria) resulted in the Lithuanians of Puńsk living in isolation. It was not until the political changes in the Republic of Poland and the USSR, the border opening for the east and the west, the accession of the Republic of Poland and the Republic of Lithuania to the EU (the Schengen area), new economic conditions, cultural changes and the accelerating technical revolution that the lifestyle of the Lithuanians of Puńsk changed and the unification of the local dialect and the standard Lithuanian language took place.

The material collected in ECorp-of-Punsk depicts the decadent period of functioning of this local dialect. The interferences revealed in the corpus between the dialectal system and the Polish language on the one hand and the literary Lithuanian language on the other show that the dialectal elements are being replaced with the Lithuanian general and linguistic versions (mainly with regard to morphology, phonetics, lexis). Polonisms and calques of the Polish language also appear in the local dialect.

2.3. MonoConc — the program supporting the Ecorp-of-Punsk resources

After the proper adjusting and conducting the lemmatization and annotation of the text, the standardized material in respect to the coding (UTF8) and record format (TXT) was imported to the MonoConc program (<http://www.athel.com/mono.html>). MonoConc is a simple program providing minimum requirements for such kind of programs. Amongst the available functions, it is possible to distinguish: searching using the annotation data, rich statistical characteristics and the concordance automatic finding. The metadata cannot be included in the function of searching, however they are visible in the reply obtained.

2.4. ECorp-of-Punsk statistical data

In January 2012, the ECorp-of-Punsk volume amounted to 1,300,043 of signs, which corresponds with about 225,000 words, including 16,279 lemmas and 68,183 unique forms. The data given here refers to the basic pillar of the corpus resources —

utterances given by the Lithuanians of Puńsk using the local dialect (comp. below Subcorpus A, point 3.1.).

3. Structure of the experimental corpus of the Lithuanian local dialect of Puńsk in Poland

E Corp-of-Punsk has a complex structure. It is not a typical monolingual corpus. The material collected allowed to extend the structure and form a few subcorpora:

- A a monolingual subcorpus of the utterances given by Lithuanians in the local dialect of Puńsk (the main core of the corpus.),
- B a monolingual subcorpus of the utterances given by Lithuanians in Polish,
- C a bilingual Polish-Lithuanian parallel subcorpus.

3.1 Subcorpus A contains utterances given by the Lithuanians of Puńsk (residents of Puńsk and its environs) in the local dialect (Lithuanian). The problems with the structure of the corpus described above in points 2.–2.3 are just connected with subcorpus A.

Table 1 demonstrates model utterances of the years 2007–2009 given by the three generations’ representatives.

Table 1. Subcorpus A. Examples of utterances given by three generations’ representatives.

Item 1. Informants: a 70-year-old man (farmer, completed 4 classes of elementary school) [M70], a 70-year-old woman (completed 4 classes) [W70], a 9-year-old child, a recording of 2007 [C9]	
Example of subcorpus A	English Translation
[M70] Aš tai sakiau, <i>tį</i> nieko <i>neraikalaukit</i> , <i>laimė</i> — <i>ciej</i> vaikai gyvi liko ir ... <i>ale</i> anoj pusė <i>tį biskį</i> iš <i>bagotų</i> , tai <i>ciej nenorėj dovanoc</i> .	[M70] I said this, demand nothing from there, luckily those children remained alive and ... but that party a bit from the rich, it was them not to want to forgive.
[K70] <i>Tį kap biskį jiem</i> išmokėjo.	[W70] They were somewhat paid to.
[M70] Ale <i>an</i> pamokos, <i>ba</i> jis <i>prisgėris</i> kiek sykių... Žinokit, kad va <i>mūs toj</i> ...	[M70] But it as a lesson, because whenever he is drunk... you know, where this our ... is.
[K70] Jis girtas <i>važavo</i> .	[W70] He was driving being drunk.

[M70] ... <i>toj</i> kur <i>dartės</i> Sigitai gyvena tokiu <i>pakranti pakialėj važau kap is tį</i> išsivertė ... tai jau vieni, <i>tuom</i> šonu, kad jis <i>važavo kairi pusi</i> , o <i>in ciek in jį</i> <i>pavercimas</i> , tai <i>kap ca sėdėj</i> , jau <i>vietom</i> ratai <i>nesiekj</i> , ir <i>išvažau</i> gerai <i>kadaisi da</i> o ...	[M70] ... there where now the Sigitasa family lives, along that tilted roadside, he was driving by the road, when he lost control there... some people that he was driving this side of the road, that he was driving on the left, in this direction — on his side a (visible) track of the overturn, the way he was sitting here, the wheels didn't touch the ground, and once he still drove out well, and...
[K70] O <i>tadu</i> skubinosi, <i>ciej vaikucai šoni</i> savo šonu ėjo.	[W70] And that time he was in a hurry, those babes were walking along the roadside on their side.
[D9] Mes ėjom <i>žolų da tadu</i> .	[C9] We were still walking on the grass then.

Item 2. Informants:

a 45-year-old-woman (teacher, after studies in Poland) [W45],
a 46-year-old woman (teacher, after studies in Poland and Lithuania) [W46],
a 15-year old girl (middle school pupil) [G15], a recording of 2009

[K15] [...] <i>nu</i> tai Vaitakiemio uždarė.	[G15] [...] so in Wojtokie mie they closed (the school)
[K45] Ir Navinykuose, <i>kap</i> JV pradėjo dirbt. Taigi jis dešimt metų gal virš <i>važinėj</i> in Navinykus, <i>da</i> vis jis turėdavo pusė etato, <i>Punski</i> dirbo. <i>Nu</i> tai jis už <i>manį dzviem</i> metais jaunesnis. Tai devyniasde- ... apie šimtas vaikų buvo Navinykų mokykloj, o dabar gal trisdešimt <i>likj</i> .	[W45] And in Nowiniki, when JV started working. He is likely to have kept going to Nowiniki for more than ten years, still he had a part-time job, he worked in Puńsk. Well, he is two years older than me. There were nine... about one hundred children at school in Nowiniki, and now perhaps thirty of them have remained.
[K15] Ir daugiausia iš visų kaimo mokyklų yra navinykuose dabar. Pristavonyse tai gal penki, šeši.	[G15] And the most (children) from all the country schools are now in Nowiniki. In Przystawańce there are perhaps five, six [pupils].
[K45] Nežinau, devyni <i>tį</i> buvo ...	[W45] I don't know, [probably] there were nine there ...
[K15] <i>Ale</i> mažai labai. Koki vienas antroj klasėj, du tračoj klasėj ir <i>tep va</i> .	[G15] Yet, very few. Somehow, there was one in the second class, two in the third, and the like.
[K46] <i>Dartės</i> labai mažai mokinių. Išvis visose mokyklose mažai.	[W46] There are very few pupils now. Generally speaking, there are few in all the schools.

<p>[K45] Taigi <i>kap</i> pati pradėjau dirbti, <i>kap</i> Birutė <i>išvažavo</i> <i>in Kanadą</i> ... paėmiau po jos, ar aš <i>dzvidešimt</i> aštuonias ar trisdešimt turėjau valandų, buvo popietinių daug <i>tį</i> kiek nori [...] Dėl auklėjimo ... <i>nu</i> tai darbo daugiau. [...] <i>Ciej</i> vaikai nenori ir ... visi patogūs, <i>cik šokc, dainuoc</i>, o daugiau nieko [...] Kad mokslinių <i>būralių</i> kokių <i>būt</i> ... kad niekas niekur neprasimuša.</p>	<p>[W45] When I myself started working, when Birute went away to Canada ... I took over from her, I had twenty or thirty hours, there were many afternoon classes there, as many as you want [...] as for the class teacher's duties... there is more work then [...] These children don't want, and ... everyone is comfort-seeking, only to dance, to sing, and nothing else [...] if only there were any special interests groups...but no one shows any interest.</p>
---	---

The words in italics in table 1 do not follow the standards of Lithuanian. Among the indicated forms there are lexemes (a) not known to the literary language, e.g. the dialectal *bagotas* 'rich' (comp. the Lithuanian *turtingas*) (b) differing only in pronunciation, e.g. the dialectal *išvažavo* 'she went away' (comp. the Lithuanian *išvažiavo*) (c) having a different inflection, for example *nenorėj* 'they did not want' (comp. the Lithuanian *nenorėjo*). Proportionally, the most dialectal elements are noted in utterances given by the old generation (comp. Table 1, item 1). There are definitely fewer dialectal elements in utterances given by the middle generation (comp. Table 1, item 2). The fewest dialectal elements are displayed in utterances given by the young generation, comp. the informants' utterances [C9] and [G15] in table 1. However, you should take into account that in utterances given by the youngest representatives of the young generation dialectal elements are distinct. The number of these features undergoes a significant reduction along with the school education going on, comp. the informant's utterances [G15] in Table 1, item 2.

At the present stage of studies on subcorpus A, we can say that we are dealing with a balanced corpus. The texts evenly represent the utterances given by the three generations within thirty years. As for the dialectal material metadata, the following is taken into account: the year and the place of the recording as well as the informant's age, education, sex and the place of residence.

In case of the corpus being published online, the resources' translation into Polish is considered. Translation of the subcorpus A resources into Polish can affect greater interest not only in the local dialect, but Lithuanians themselves — the residents of the commune of Puńsk. The subcorpus A potential recipients (along with the translation of the resources into Polish) can be: sociologists, ethnologists, historians, culturologists, researchers of the linguistic image of the world and even politicians dealing with the problems of the national minorities in Poland.

3.1.1. Lexical annotation

E Corp-of-Puńsk presented here is not a purpose-in-itself. Based on its resources, a monograph of the local dialect of Puńsk is being compiled. Therefore, an additional annotation, for which the working name of lexical annotation was taken, has been carried out in subcorpus A. The purpose of implementing this annotation

was to distinguish all forms included in subcorpus A on the basis of their origin. Therefore, the following indicators have been singled out:

- LIT — form consistent with the literary form
- GERM — germanism
- SLAV — slavism
- GWAR — dialectal innovation or archaism
- Gwar — dialectal form morphologically consistent with the literary form, however, with distinct phonetic dialectal features.

In Table 2, an example of the lexical annotation has been presented for the sentence: *Ale anoj pusé tĭ biskĭ iš bagotŭ, tai cieĭ nenorěj dovanoc.* ‘But that party a bit from the rich, it was them not to want to forgive.’

Table 2. Example of the lexical annotation

Item	Wordform	Lexical Annotation
1	ale	SLAV
2	anoj	GWAR
3	pusé	LIT
4	tĭ	GWAR
5	biskĭ	GERM
6	iš	LIT
7	bagotŭ	SLAV
8	tai	LIT
9	cieĭ	GWAR
10	nenorěj	GWAR
11	dovanoc	gwar

3.1.2. Semantic annotation

Annotation is an indispensable element of each corpus. Almost each corpus is morphosyntactically annotated. Along with the development of corpus linguistics there are expectations with reference to corpora themselves. One of the expectations is semantic annotation which contains important vital characteristics describing the present meaning of a given lexeme on the semantic level of the sentence. More about semantic annotation, comp. the articles included in this volume (Koseska-Toszewa, 2013; Roszko, D. & Roszko, R., 2013).

In ECorpus-of-Punsk, the semantic annotation elements were implemented in regard to exponents of the semantic categories of hypothetical nature and exponents of imperceptivity. According to the divisions established in *Bulgarian-Polish Contrastive Grammar*, within particular categories the following parameters are distinguished: M, H1, H2, H3, H4, H5, H6, I1, I2. The letter M means modality,

H — hypothetical nature, I — imperceptivity, numbers from 1 to 6 indicate a degree of probability. As far as hypothetical nature, 6 degrees of probability are established, where H1 means the size probability close to "0" (false), and H6 — close to "1" (true). As far as imperceptivity, 2 degrees of probability are established, where I1 — neuter size, and I2 — enhanced size. Below, an example of a dialectal text fragment, for which the semantic annotation of lexemes bringing the meaning of modality was conducted.

Kiba atjojis tas ponas su sūnum. Tas pamatis tu mergu ir insimylėjis.

‘Probably this man has arrived with his son. This son (in turn) saw this girl and fell in love.’

Kiba [M:H4] **atjojis** [M:H4] tas ponas su sūnum. Tas **pamatis** [M:H4] tu mergu ir **insimylėjis** [M:H4].

The form *kiba* is a lexical exponent of hypothetical nature, to which degree 4 of probability is ascribed. The form *atjojis* is a present perfect form without the copula, which in this sentence becomes a morphological exponent of hypothetical nature, cooperating with the lexical exponent. A probability degree ascribed to the present perfect form is dependent on the proper value of the lexical exponent *kiba*. In the next sentence lexical exponents do not appear, but present perfect forms without the copula (*pamatis*, *insimylėjis*) appear as a morphological exponent of hypothetical nature. Degree 4 of probability is also being ascribed to these forms. Generally speaking, perfect forms reflect a degree of probability initially expressed with the lexical exponent. You can find more on this, comp. (Roszko, D., 2013).

3.2. Subcorpus B contains utterances given by Lithuanians (residents of Puńsk and its environs) in Polish. Certainly, it is a brand new thing in corpus linguistics, which should influence the extension of the circle of potential recipients of ECorpus-of-Punsk to include dialectologists studying Polish local dialects of Podlasie and the Suwałki region.

Table 3. Subcorpus B. A fragment of an utterance given by a Puńsk Lithuanian in Polish directed to tourists from central Poland.

Informant: 60-year-old man, farmer, elementary education, resident of Puńsk (his farmland in close vicinity of Puńsk), once a week goes shopping to Suwałki, stayed in Germany. A recording of 2010	
Example of subcorpus B	English Translation
Tutej z naszej strony to nie było żadnych patroli, a tu z Litvy strony ¹ , nie, tutaj były był patrol, tu wszystko było przywiezione tam te azjaci. Oni byli tak nastawieni	Here, from our side there were no patrols, and here from the Lithuania's side, no, here were, was a patrol, everything here was brought, there those Asians. They were oriented that way.
[nadjeżdża samochód]	[there appears a car]

Tak byli nastaviony, że za granico to sami vrogovie tutej mieszkajo. Zaras tutaj taki był nauczycielem², dyrektorem i sekretarzem gminnym był, jak kiedyś. To były siedemdziesiąte jakieś drugi rok, jakiś tak o mniej viencej. To oni przy tu taki młody lasanek, a tu taka przestrzeń, tutej jeszcze po polskiej stronie, nie?, a jusz tutej już granica. Teraz oni przyjechali, usiedli, przywieźli żony, dzieci, tu usiedli tutej, pijo(m) sobie po trochu, zagryzajo, pijo. Idzie akurat ten patrol, żołniesz. U nas v tych czasach żołniesz to był v takim poszanowaniu, bo on bronił ojczyzny. To, to jak vszedł do gospody, czy tam gdzieś jego podwieść, czy, czy, to jemu i jeść dali i pić dali, vszystkiego³, bo on się liczył, służy dla ojczyzny. Teraz on nic nie myśląc móvi „khadzi siuda to vypijom”, a ten móvi „khadzi ty siuda”. Ten nic nie myśląc v jednym renku tależ z zagrycho, a v drugim butelkie i pszez granice. Jak tylko pszeszedł pszez granice, ten krzyknoł „stop” i „ruki vverkh”. Ten myśli, że żartuje i idzie idzie do niego dalej. To ten od razu automatycznie automat na plecach jak miał, to, móvi, tak automatycznie ściognoł, załadował i móvi „ruki vverkh kak nie to streliaju”. To zaczęli tam żony krzyczeć, vszyscy, że rzuć vszystko, podnieś rence, bo zastrzeli. Tam nie ma. I, kurcze, potem przed automatem i vpieriod. Tu vszyscy płakać, krzyczeć, a on vpieriod. Tylko tyle, że był sekretarzem, to on na tych sesjach, zebraniach różnych tam z naszymi tam vopistami był, tam z vojskiem granicznym, to podrzymali jego tutej Litvini podrzymali do nocy. Tamte przyjechali v nocy i oddali, bo oni tutej vspółpracovali, vszystko jedno, i v tych⁴ czasach. Oto jagby teras tako zgrajo by zajechali, to by my pszyjechali do domu za jakieś trzy miesioce jak kiedyś.

They were so oriented that here abroad only enemies live. Right away here such was a teacher, a director, and a commune secretary, like before. It was in the seventies, something seventy two, more or less. So they here at such a young grove, and here such a space, here still on the Polish side, right?, and just close here is the border. Now they arrived, got seated, with their wives, children, here they got seated, started drinking and snacking, and drinking. Just this patrol coming by, a soldier. Those days the soldier was so respected, because he defended the homeland. When he used to enter the inn, was approached to be given a lift to somewhere, or was given something to eat and drink, everything, because he was respected, served the homeland. Now not thinking such says "get here and drink", says "you get here". Not thinking, a plate with snacks in one hand and a bottle in the other, such crosses the border. Once he crossed the border, the soldier shouted "stop" and "hands up". That one thinks that he is joking and keeps approaching him. This one at once automatically the gun machine from his back, and says, so automatically moved it down, loaded and says "hands up or I will shoot". Then the wives started to shout, everyone [to that one] to throw everything, raise hands, or he will shoot dead. No two ways. And, oh gosh, then towards the machine gun that one keeps approaching. Here everyone crying, screaming, and he approaching. But he was a secretary, he used to be at those sessions, different meetings there with our border soldiers with the border army, so they only detained him here, Lithuanians detained him till night. Those arrived at night and took him, because they cooperated here, all the same, so in these days. Now, if they

	arrived as such a pack, we would arrive home in some three months, like before.
--	---

Like in case of subcorpus A, also here the transliteration based on the Polish spelling has been applied. Only in certain phonetic contexts the norms of the Polish spelling are disturbed in order to portray phonetic phenomena typical of Lithuanians speaking Polish. The subcorpus B resources, after the text proper adjustment and preparation in regard to coding (UTF8) and record format (TXT), were imported to the above-mentioned MonoConc program (<http://www.athel.com/mono.html>), comp. above 2.3.

3.3. Subcorpus C is a typical parallel corpus (here bilingual). The materials included in it are utterances given by Lithuanians (the residents of Puńsk and its environs) in the local dialect (or in Lithuanian), and the translations of the utterances into Polish. The Lithuanians of Puńsk are the authors of the translations. A considerable part of the subcorpus C resources comes from the Internet. A definitely small part of them are the texts recorded during the meetings (mostly official ones) Poles participated in.

Table 4. Subcorpus C. An example of a text.

Lithuanian version	Polish version	English Translation
Seniausi žmogaus gyvenimo pėdsakai šiame krašte siekia 10.000 metus prieš m.e. (paleolito saulėlydį). Aptikta juos Valinčiuose, Ožkinuose, Vaiponioje ir Šlynakiemyje. Ankstyvaisiais viduramžiais šios žemės sudaro Jotvos dalį. Jotvingių pėdsakus rodo gyvenviečių tipai (atviros gyvenvietės) ir piliakalniai.	Najstarsze ślady bytności człowieka na ziemi puńskiej sięgają 10 000 lat p.n.e. (schyłek paleolitu). Odnaleziono je w miejscowościach Wołyńce, Oszkinie, Wojponie i Szlinokiemie. We wczesnym średniowieczu ziemie te stanowiły część Jacewicz. Na ślady jacewiczkie wskazuje typ osadnictwa (osady otwarte) i góry zamkowe.	The oldest remains of human presence on the land of Puńsk date back to 10,000 BC (the end of the palaeolith). They are to be found in the town of Wołynce, Oszkinie, Wojponie and Szlinokiemie. In the early Middle Ages these areas constituted part of Jacwicz. The type of settlements (open settlements) and castle hills are evidence of the history of Jacwicz.

¹The word sequence in the phrase *z Litwy strony* is typical of the local dialect and the Lithuanian language, comp. *Lt iš Lietuvos pusės*.

²The instrumental case used here can be a result of the local dialect, which can mean that — in the given case — this person was not a teacher by profession, but for some time worked as a teacher.

³A significant influence of the dialectal use of the so-called partitive genitive.

⁴With the meaning of *w tamtych czasach*. The use of the Polish pronoun *ten* results from the dialectal use of *tas* to express the meaning of definiteness.

Įdomiausias yra 9 km nuo Punska į šiaurę nutolęs Eglinės piliakalnis.	Najciekawsza góra zamkowa znajduje się w miejscowości Jegliniec, oddalonej o 9 km od Puńska na północ.	The most interesting castle hill is in the town of Jegliniec, which is about 9 km northward of Puńsk.
Jotvingius XIII a. pradžioje nukariavo kryžiuočiai. Tačiau šio krašto nepajėgta intensyviai apgyvendinti, todėl iki XV a. čia ošė miškai. Tik XV a. pradžioje pradėta naujai šiam krašte kurtis. Naujieji krašto šeimininkai buvo lietuviai, kilę nuo Merkinės ir Punios. Jie čia tyvuliuojantį ežerą ir pavadino Punia, nuo kurio ir gyvenvietė gavo vardą.	Jaćwingów na początku XIII w. podbili krzyżacy, ale ich ziem nie byli w stanie zaludnić, dlatego też do XV w. porastała je puszcza. Dopiero na początku XV w. pojawiają się tu nowi osadnicy. Nowymi gospodarzami ziemi puńskiej stali się Litwini wywodzący się z nad Merecza i Puni. To oni miejscowe jezioro nazwali Punia, od którego później nazwę przejęła także osada.	At the beginning of the 13th century, the Jaćwingi people were conquered by the Teutonic Knights, who were not able to populate their lands, which, therefore, remained covered by a forest till the 15th century. It was not until the beginning of the 15th century that new settlers started to appear here. The Lithuanians coming from Merecz and Puni became the new hosts of the land of Puńsk. It was them to call the local lake Punia, from which later also the settlement took its name.
Girininkas Stanislovas Zalivskis 1597 metais pastatė Punske bažnyčią, ir čia buvo įsteigta parapija. Sekmadieniais ir švenčių dienomis Punske vykdavo turgūs.	Leśniczy Stanisław Zaliwski w r. 1597 wznosił w Puńsku kościół, erygowano tu również nową parafię. W niedziele i święta odbywały się targi.	Forester, Stanislaw Zaliwski raised a church in Puńsk in 1597, also a new parish was founded here. On Sundays and holidays, fairs were held.
[...]	[...]	[...]
Metadata: http://punkskas.pl/pkv2-pl.htm	Metadata: http://punkskas.pl/?page_id=18	

Table 4 demonstrates the initial fragments of the texts included in the subcorpus 3 resources. The paragraphs are grafically distinguished. In Table 5, a file fragment is presented in the TMX format, being a result of alignment on the level of sentences placed in Table 4. At the early stage of research on alignment the TextAlign program (by Andrew Manson) was used. Currently, the Terminotix and Nova companies commercial programs are used for this purpose.

Table 5. Subcorpus C. The initial fragment of a TMX file containing the aligned texts placed in Table 4.

```

<?xml version="1.0" encoding="UTF-8" ?>
<tmx version="1.4">
<header adminlang="Lithuanian" creationdate="20090416T140039Z" creationtool
="TextAlign" creationtoolversion="1.0.0.0" datatype="PlainText" segtype="sent
ence" srclang="Lithuanian" o-tmf="TextAlign TMX"></header>
<body>
<tu tuid="0000000001">
<tuv xml:lang="Lithuanian">
<seg>Seniausi žmogaus gyvenimo pėdsakai šiame krašte siekia 10.000 metus prieš
m. e. (paleolito saulėlydį).</seg>
</tuv>
<tuv xml:lang="Polish">
<seg>Najstarsze ślady bytności człowieka na ziemi puńskiej sięgają 10 000 lat
p. n. e. (schyłek paleolitu).</seg>
</tuv>
</tu>
<tu tuid="0000000002">
<tuv xml:lang="Lithuanian">
<seg>Aptikta juos Valinčiuose, Ožkiniuose, Vaiponioje ir Šlynakiemyje.</seg>
</tuv>
<tuv xml:lang="Polish">
<seg>Odnaleziono je w miejscowościach Wołyńce, Oszkinie, Wojponie i Szlino-
kiemie.</seg>
</tuv>
</tu>
...

```

After being aligned, the resources were loaded to the ParaConc program (<http://www.athel.com/para.html>). The ParaConc tool makes the option of simultaneous searching and asking questions available for both the languages. On account of little value of these texts to linguistic studies (the editor's interference in the written text and the efforts of the Lithuanians of Puńsk to use the standardized Polish and Lithuanian during formal meetings) the texts included in subcorpus C were not subjected to the process of lemmatization or annotation.

Summary

The dialectal material collected for nearly 30 years was partly listed during the last two years, provided with annotation and loaded to the programs organising the resources (MonoConc and ParaConc). A basic pillar of the corpus is subcorpus A containing the utterances of the Lithuanians of Puńsk using the local dialect. The two other subcorpora came into existence as secondary. It turned out that besides the utterances of the Lithuanians of Puńsk in the local dialect there are plenty of utterances of these Lithuanians in Polish included in the resources. Taking into

account the fact that it is not entirely correct Polish, there was a decision to include also this material in the corpus as an additional pillar marked as subcorpus B. As concluded, the material collected in subcorpus B can be useful for researchers of the Polish language on Podlasie and the Suwałki region, and for linguists dealing with the problems of interference. The recordings also include utterances given by Lithuanians in the local dialect (in Lithuanian), with simultaneous translation into Polish (e.g. at formal meetings where Poles participate). So, these texts were also included, moreover, they have been supplemented with bilingual materials coming from the local publishing companies and websites run by Puńsk Lithuanians.

The resources (subcorpus A) collected in ECorp-of-Punsk are extremely useful, since they reflect the changes lasting nearly thirty years in the local dialect. The dialect evolution was largely forced by external processes, such as the change of the political system of the Republic of Poland at the turn of the eighties and nineties of the past century, the regaining of independence by Lithuania, the accession of Poland and Lithuania to the European Union, the border opening for the east and the west (the Schengen area), moreover new economic conditions, cultural changes and the accelerating technical revolution. The changes recorded in ECorp-of-Punsk confirm the thesis that the local dialect is disappearing, is becoming similar to the standard Lithuanian language.

References

- Koseska-Toszeńska, V. (2013). About Certain Semantic Annotation in Parallel Corpora, *Cognitive Studies / Études Cognitives*, 13, p. 67–78 (this volume). DOI: 10.11649/cs.2013.004
- Krištopaitė, D. (Ed.) (1998). *Nuo Punsko iki Seiny. Iš Juozo Vainos rinkimų. II. Punsko "Aušros" leidykla*, pp. 282.
- Krištopaitė, D. (Ed.) (1999). *Nuo Punsko iki Seiny. Iš Juozo Vainos tautosakos rinkimų. I. Punsko "Aušros" leidykla*, pp. 317.
- Roszko, D. (2013 / in print). *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej*, SOW, Warszawa..
- Roszko, D. & Roszko, R. (2013). Experimental Polish-Lithuanian corpus with elements of the semantic annotation, *Cognitive Studies / Études Cognitives*, 13, p. 97–111 (this volume). DOI: 10.11649/cs.2013.006
- Smoczyński, W. (1984a). Szkic morfologiczny litewskiej gwary puńskiej, *Acta Baltico-Slavica*, 16, p. 235–261.
- Smoczyński, W. (1984b). Zapożyczenia słowiańskie w litewskiej gwarze puńskiej. In: *Studia nad gwarami Białostockimi. Morfologia i słownictwo [Prace Białostockiego Towarzystwa Naukowego Nr 27]*, Warszawa, p. 179–222.
- Smoczyński, W. (1986a). System fonologiczny litewskiej gwary puńskiej, *Acta Baltico-Slavica*, 17: p. 369–385.
- Smoczyński, W. (1986b). Zapożyczenia niemieckie w gwarze litewskiej okolic Puńska na Suwalszczyźnie. In: *Zeszyty Naukowe Uniwersytetu Jagiellońskiego, Prace Językoznawcze, Zeszyt 82*, Warszawa–Kraków, p. 35–45.

Corpora and web resources

Anotatorius

(http://donelaitis.vdu.lt/main.php?id=4&nr=7_1). 30.09.2012

Catalan Corpus Oral Dialectal

(<http://www.uv.es/foncat/cat/Treballs/10.Clua-Lloret.pdf>). 30.09.2012

Corpus Gesproken Nederlands

(<http://lands.let.kun.nl/cgn/ehome.htm>). 30.09.2012

Estonian Dialect Corpus

(<http://www.murre.ut.ee/estonian-dialect-corpus/>). 30.09.2012

FRED — Freiburg Corpus of English Dialects

(<http://www.helsinki.fi/varieng/CoRD/corpora/FRED/index.html>). 30.09.2012

Helsinki Dialect Corpus

(http://blogs.helsinki.fi/hes-eng/files/2011/03/HES_Vol2_Peitsara_Vasko.pdf). 30.09.2012

MonoConc

(<http://www.athel.com/mono.html>). 30.09.2012

Nordic Dialect Corpus

(<http://www.tekstlab.uio.no/nota/scandiasyn/index.html>). 30.09.2012

NoTa Corpus — Norwegian speech corpus — Oslo part

(<http://www.tekstlab.uio.no/nota/oslo/>). 30.09.2012

ParaConc

(<http://www.athel.com/para.html>). 30.09.2012

Russian National Corpus (Dialectal corpus)

(<http://www.ruscorpora.ru/en/corpora-structure.html>). 30.09.2012

Scottish Corpus of Text and Speech

(<http://www.scottishcorpus.ac.uk/>). 30.09.2012

Spoken Japanese Dialect Corpus (GSR-JD)

(<http://research.nii.ac.jp/src/eng/list/detail.html#GSR-JD>). 30.09.2012

Swedia 2000

(<http://swedia.ling.gu.se/>). 30.09.2012

YADAC — Dialectal Arabic Corpus

(http://www.lrec-conf.org/proceedings/lrec2012/pdf/663_Paper.pdf).
30.09.2012