EURAC research

LULCL II 2008

Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics

Bozen-Bolzano, 13th-14th November 2008

The initiative is co-financed by the Autonomous Region Trentino – South Tyrol,
Department III - Linguistic Minorities and European Integration,
Office for Linguistic Minorities

Bestellungen bei:

Europäische Akademie Bozen
Drususallee 1
39100 Bozen – Italien
Tel. +39 0471 055033
Fax +39 0471 055099
E-mail: press@eurac.edu

Ordinare Libro:

Accademia Europea Bolzano
Viale Druso, 1
39100 Bolzano – Italia
Tel. +39 0471 055033
Fax +39 0471 055099
E-mail: press@eurac.edu

Verantwortlicher Direktor: Stephan Ortner
Herausgeber: Verena Lyding
Redaktion / Koordination: Verena Lyding
Revision: Peter Farbridge
Druckvorstufe: Typoplus, BZ
Druck: ESPERIA srl

Direttore responsabile: Stephan Ortner
Curatore: Verena Lyding
Redazione / Coordinazione: Verena Lyding
Revisione: Peter Farbridge
Prestampa: Typoplus, BZ
Stampa: ESPERIA srl

An online version of the proceedings is available at:
http://www.eurac.edu/Org/LanguageLaw/Multilingualism/index

Last modified: August 2009

ISBN 978-88-88906-52-2

LULCL II 2008

Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II)

"Combining efforts to foster computational support of minority languages"

Bozen-Bolzano, 13th-14th November 2008

Verena Lyding (Ed.)

Index

The Latgalian Component in the Latvian National Corpus

Aleksey Andronov and Everita Andronova

The paper deals with the Latgalian written language used by a part of Latvians living mostly in the eastern Latvia. There is an initiative to start compiling a Latvian National Corpus, which is supposed to be a long-term activity to cover all printed Latvian texts and a wide scope of the speech. Since there are two standardised varieties of Latvian: the Latvian literary language and the Latgalian written language (with a tradition of more than 250 years), Latgalian should be represented in the Latvian National Corpus. The authors describe the first experience with compiling a corpus of Latgalian texts published in Soviet Russia (1917–1937) and detect some possible issues of compiling a corpus of Modern Latgalian, which can be delimited by the National Awakening and the reestablishment of the Republic of Latvia in 1991. Although Modern Latgalian is characterised by restricted usage and the lack of some text types, there are rich Latgalian data from other time periods and regions, and this would make it possible to develop some specialised corpora in future. Various linguistic resources and tools should be developed for Latgalian in order to raise its prestige.

1. The Latvian Corpus: the State-of-the-Art

Nowadays language corpora are a clear prerequisite for a comprehensive study of a language and its very existence in a global high-tech society ("linguistic corpora are intended to be the basis for the analysis and description of the structure and use of languages and for various applications" [Kennedy 1998: 60]). In these terms, the Latvian language could be considered as a lesser-used language, because there is still a lack of corpus resources and the community of linguists is still not very enthusiastic about using modern technologies in their everyday research. There is a certain gap observed between language resource developers and users. There is a rather long-term tradition of collecting Latvian texts in electronic form, dating back to the beginning of the 90s (Milčonoka et al. 2004; Grūzītis et al. 2004). Today the main language

resource developers are the Institute of Mathematics and Computer Science at the University of Latvia (henceforth IMCS, UL), the National Library of Latvia, the IT company 'Tilde', as well as some academic institutions. The corpus activities were started earlier this century—a pilot morphological annotation has been performed (Levāne et al. 2000) and some studies of a parallel English-Latvian corpus have been carried out (Skadiņa 2005; Milčonoka 2001). Apart from this, in 2003 a diachronic Corpus of Early Written Latvian was launched, and its development is in progress (Andronova 2007). In 2005, a design of the Latvian corpus was developed by the IMCS, UL (Koncepcija 2005). In 2007, a one-million-token balanced corpus of Modern Latvian was compiled according to the guidelines set in this design. The compilation was supported by the State Language Agency and done at the IMCS, UL[1]. In 2009, the size of the corpus will be extended by another 2.5 million tokens from balanced and representative Latvian texts. There are some activities carried out towards an unbalanced large corpus from texts available on the Web (Džeriņš et al. 2007). There are a number of experimental language tools developed at the IMCS, although they are still not available for off-the-shelf use; there are plans to provide a graphical corpus interface for a semi-automatic morphological and syntactic analysis within the SEMTI-Kamols project[2].

The Latvian State Language Commission, established in 2002, aims to study "the situation of Latvian as the country's state language and to draft recommendations on how to strengthen its status and develop it further" (The State Language Commission). The year 2008 was marked by several initiatives supported by the State Language Commission: in April, a wider researcher community was introduced to the concept of a language corpus, and for the first time the idea of a Latvian National Corpus came up. Later the initiative was taken by the National Library of Latvia, which is the leader of the National Digital Library project[3] and inspired the Agreement of Intention between the main language resource developers and holders in Latvia, both academic and industry partners. In November, an international workshop was held in Riga, organised by the Latvian State Language Commission; the IMCS, UL; and the National Library to get acquainted with the practice of the Czech National Corpus and to set some further tasks. Unfortunately, the initiative is now slightly slowed down.

The Latvian National Corpus is supposed to be a long-term activity to cover all printed Latvian texts and a wide scope of the speech (cf. Vasiļjevs 2008). It will be a

1 It is now available from www.korpuss.lv via the Manatee platform http://www.textforge.cz/products

2 www.semti-kamols.lv

3 www.lnb.lv/lv/digitala-biblioteka

complex system of separate sub-corpora, both synchronic and diachronic, monolingual and multilingual, developed by different partners such as the University of Latvia, the Institute of the Latvian Language, the Institute of Mathematics and Computer Science, the National Library, and so forth.

The National Corpus is expected to represent the Latvian language in full. According to the State Language Law, there are two standardised varieties of Latvian: the Latvian literary language and the Latgalian written language. The law states: "The official language in the Republic of Latvia is the Latvian language (§3.1). The State shall ensure the maintenance, protection and development of the Latgalian written language as a historical variant of the Latvian language (§3.4)" (VVL 1999). Thus, the Latvian National Corpus cannot be considered complete without the Latgalian component, which is the topic of the present report.

2. What is Latgalian?

Latvia is historically divided into several ethnographic regions; the eastmost of them is Latgale, former Polish Livonia. Latvian has been divided into three dialects: the Central dialect, the Tamian dialect and the High Latvian Dialect (for more detailed information see Balode et al. [2001]).

The Latgalian written language is a standardised variety of the language used by a part of Latvians, living mostly in eastern Latvia (Latgale). Due to the history of the region the native population of Latgale differs from other Latvians, not only in language, but also in ethnography, cultural life and religion (Latgalians are mostly Roman Catholics, while other Latvians are mostly Protestants). For almost three centuries (1629–1917), Latgale was separated from the rest of Latvia. In the 17th century its territory came under the rule of the Polish–Lithuanian Commonwealth, and it was known as Polish Inflantia or Polish Livonia. In 1772 it was incorporated into Vitebsk Province of the Russian Empire. Therefore, the Latgalian language has been exposed to influence from Polish and East Slavonic (Russian and Belarussian).

There is no common agreement on the linguistic status of the language spoken in Latgale: it is considered either one of the three main dialects of the Latvian language or a separate Baltic language on equal terms with Latvian and Lithuanian (Brejdak 2006: 195). Some linguists try to achieve for Latgalian the status of regional language in Latgale. Further in the text it will be referred to as just Latgalian, and its standardised variety as Standard Latgalian.

The linguistic distinction between Standard Latgalian and Standard Latvian is large enough to complicate mutual understanding. The differences are mainly found in the phonological system, as well as in the vocabulary, but certain important deviations exist also in morphology and syntax. See the text of prayer 'Pater noster' in Latvian and in Latgalian in Table 1.

Latvian	Latgalian
Mūsu Tēvs debesīs! Svētīts lai top Tavs vārds. Lai nāk Tava valstība. Tavs prāts lai notiek kā debesīs, tā arī virs zemes. Mūsu dienišķo maizi dod mums šodien. Un piedod mums mūsu parādus, Kā arī mēs piedodam saviem parādniekiem. Un neieved mūs kārdināšanā. Bet atpestī mūs no ļauna. [Jo Tev pieder valstība, spēks un gods mūžīgi mūžos.] Āmen.	Tāvs myusu, kas esi debesīs, svieteits lai tūp Tovs vuords, lai atīt Tova vaļsteiba, Tova vaļa lai nūteik kai debesīs, tai ari viers zemis. Myusu dīniškū maizi dūd mums šudiņ un atlaid mums myusu poruodus, kai ari mes atlaižam sovim poruodnīkim, un naīved myusu kārdynuošonā, bet atpestej myus nu ļauna. Amen.

Table 1: Example of 'Pater noster' in Latvian and Latgalian

Latgalian has a well-established written tradition dating back to 1753 (cf. Leikuma 2008); it has experienced the ban of publishing books in Latin script during the process of Russification as a part of Russia's anti-Polish policy (1865–1904). Probably more than 750 books have been published till now (cf. Seiļ 1936). There are several linguistic descriptions of Latgalian (practical grammars and dictionaries) reflecting deliberate work on developing a literary norm. A precise statistic evaluation is difficult, but according to the Research Institute of Latgale some 150–200,000 people speak Latgalian in their everyday life.[4] It is used not only at home, but also has a notable place in public life, cultural events, local authorities' work, and Catholic church services. The amount of the linguistic and social linguistic problems to be commented on in a comprehensive description of Latgalian corresponds to the frame of a language, not a dialect.

Among the 'alternative' languages of the Baltic States (compared to Võro in Estonia and Samogitian in Lithuania), Latgalian is the most prominent and fully-fledged. If we look at the *Unesco Digital Atlas of the World's Languages in Danger of Disappearing*, Latgalian is marked as unsafe (UNESCO 2008).

4 Retrieved May 15, 2009, from http://dau.lv/ld/latgale(english).html

3. Why is the Modern Latgalian Corpus Necessary?

In spite of considerable usage of Latgalian in fiction and mass media (including radio broadcasting and the Internet)[5], the government pays no special attention to it, and it lacks linguistic research, thus making the language endangered in Latvia today. There are several courses on the Latgalian written language and its history at universities (in Rezekne Higher Education Institution, Daugavpils University and the University of Latvia), but the practical language is not taught at schools.

Several linguistic resources and tools should be developed for Latgalian in order to raise its prestige and to ensure its development. A standard dictionary and grammar, schoolbooks and readers, a spell-checker and a morphological analyser, together with a linguistic corpus are necessary.

A modern language corpus would serve as a basis for other resources. The modern language period began together with the National Awakening and the reestablishment of the Republic of Latvia in 1991, which gave a new impulse to the rebirth of Latgalian after its being almost neglected during the years of the Soviet rule. In June 1990, a public non-profitable organisation, The Latgalian Culture Centre[6] was established, and its publishing house is the main publisher of books of different genres in Latgalian today.

4. The first experience with compiling the Corpus of Latgalian texts published in Soviet Russia (1917–1937)

At the end of the 19th century, there was an organised movement to get free land in the Russian Empire, especially in areas of Siberia. Thousands of Latgalian people moved to Russia. A pioneer initiative has been started at St. Petersburg State University in close cooperation with the National Library of Russia to compile a corpus of the Latgalian texts published in Soviet Russia (Andronov et al. 2008). Concerning this corpus, all the texts (100 books and 11 periodicals) published during the period 1917–1937 are representative to generalise the language of that period as a whole, as we do

5 There is a blogger in a daily newspaper of Latvia writing in Latgalian (http://www.diena.lv/lat/tautas_balss/blog/saprge), 'Latgales Radio' (which has existed since 2006) broadcasts mostly in Latgalian (http://www.lr.lv/).

6 http://www.lkcizdevnieciba.lv/

not have any other sources to be included in the corpus. The issue of representativeness here might be associated with the argument concerning the case of diachronic corpora where "it can only be based on the body of preserved texts and the authenticity of those included in the corpus. However, the linking up of representativeness of diachronic corpora to the body of preserved texts means that the corpora reflect, in fact, the skewed stylistic, genre and other proportions in the body of texts rather than the characteristics of the real language of the time" (Kučera 2007: 1).

The corpus in process will be a static corpus, including full texts of newspapers, fiction (mostly translated from Russian and Latvian, but also original pieces), schoolbooks and social and political brochures. One of the data collection challenges in this case is the lack of some seven sources caused by the repressive policy of national minorities in Soviet Russia in 1937, when books in Latgalian were forbidden and destroyed. Therefore, one of the main tasks of this corpus is to provide researchers with unique data little explored till now. This, of course, requires a systematic and profound search of sources in the largest libraries and archives, and luckily there is still a chance to find lost books (Andronova et al. 2008). The task is to scan approximately 8000 pages and to ensure that facsimiles of the sources are made available on the Web via the server of the National Library of Russia[7]. At the moment 21 sources have been scanned and OCR has been carried out.

The first observations of the data reveal a great amount of linguistic variants in these sources. These are not only spelling versions (found both in the same text and in texts published by different authors, by different publishing houses), but also morphological versions which can be explained by the influence of the native spoken vernacular (since the settlers were from different parts of Latgale which have their own peculiarities) and syntactic versions influenced by the source language and Russian as a close contact language. For instance, different calqued constructions are observed: in Latgalian *jaunotne draudzejas ar komunarim un* ***jem nu jim lobu pimaru*** 'the youth make friends with Communards and **takes them as a good example**' there is a calqued construction from Russian *berët s nich primer*. This gives us an interesting picture of the language processes which were taking place in written Latgalian in Soviet Russia.

The corpus is supposed to provide several versions of the same text: an original form, a normalised orthography (removing imperfect spellings explained by the gradual adaptation of appropriate graphic means and lack of necessary letters in the typographies), a text with morphological annotation, and a lemma translation into Russian.

7 http://www.nlr.ru/coll/onl/fonds_onl/latgalsk.htm

As there are no language processing tools for Latgalian, there are two ways to provide a morphological annotation of the text: either manually or by using a morphological analyser for Modern Latvian with some implemented transponation rules, which may be applied for a certain amount of the Latgalian lexicon. Although in this corpus there is obviously a rather high number of lemmas that are influenced by Russian. As for Latvian, there exists a lexicon-based morphological analyser developed at the Institute of Mathematics and Computer Science (IMCS) (Paikens 2007). If we want to make use of this analysis, we may add a specific Latgalian lexicon to the Latvian one.

Compilers of the Corpus of Latgalian texts published in Soviet Russia believe that in the future it will further foster the comparative studies of the varieties of Latgalian used in Latvia and Russia respectively.

5. Problems of the Corpus of Modern Standard Latgalian

There are common issues in corpus design and compilation that should be discussed before any activities are undertaken.

Today, the usage of Latgalian is restricted to a few spheres of social life. It is quite common in oral conversation, but its written form is less popular. The Corpus of Modern Standard Latgalian (CMSLg), a written synchronic corpus, will serve to strengthen the image and status of Standard Latgalian.

This restricted usage and lack of some text types and genres (Biber 1993: 244–245) affect the size, representativeness and balance of the CMSLg. To start with, some 2–5 million running words can be processed in the corpus, although estimating the size is problematic before one has compiled a complete list of sources and studied their availability (issues of authorship, etc.) and quality (see Table 2 below). Thus, composing a comprehensive bibliography of Modern Latgalian publications is a prerequisite, which can be a topic for a separate project. The main part of a corpus of modern language usually consists of texts from periodicals, but CMSLg is quite different in this respect because there are only few periodicals in Latgale publishing more or less sporadic articles in Latgalian ('Katōļu dzeive', 'Latgales Laiks', 'Vietējā Latgales Avīze', 'Rēzeknes Vēstis', 'Vaduguns'). There is an on-line newspaper, 'LaKuGa'[8], edited by the Latgalian Students Centre, which is also a good source for the corpus. Here we can find readers' commentaries, which are usually in a colloquial form; this will make

8 Retrieved May 15, 2009, from http://www.lakuga.lv/lg/

the data of the corpus more varied. Seemingly, fiction (mainly original) will be the main source of data. An important publishing house is the Cultural Center of Latgale in Rēzekne, which prints fiction and poetry books in Latgalian as well as academic and popular studies in the cultural history of Latgale; a collection of scholarly articles in humanities, 'Acta Latgalica', is published annually by the Research Institute of Latgale in Daugavpils. In addition, we should not ignore the significant role of the Catholic Church in maintaining the Latgalian language both in printed religious texts and in public worship. Modern Latgalian lacks or has a very small amount of medical, juridical, business and technical texts.

Data acquisition and processing in the CMSLg can be solved on the same grounds as in the Latvian part of the National Corpus (Koncepcija 2005: 13, 75–88), although text selection and sampling procedures might differ. One should pay special attention to the input data quality. Many Latgalian texts are created just by mere phonetic transponations from Latvian according to sound correspondence rules, which gives an inadequate impression of the authentic lexicon, morphology and syntax. For instance, see Table 2 below:

Latvian	**so called 'Latgalian'**	**Latgalian**
Arī šorīt pamodos ļoti agri. Rūpēja ikdienas darbi. Kūtī brēca aitas, bubināja nedzirdītais kumeļš, māva neslauktā govs. Nelika mierā tās pašas domas, kuras mocīja jau vairākas nedēļas. Vai atradīs mana Anna cerēto laimi svešumā? Un kā tālāk dzīvot pašam?	Ari šūreit pamūdūs ļūti agri. Ryupēja ikdīnas dorbi. Kūtī brēce aitas, bubynōja nadzirdeitais kumeļš, mōve naslauktō gūvs. Nalyka mīrā tōs pošas dūmas, kuras mūceja jau vairōkas nedeļas. Voi atradeis muna Anna carātū laimi svešumā? Un kai tōļōk dzeivōt pošam?	I šūreit pasamūdu cīši agri. Pruotā stuovēja kasdīnys dorbi. Klāvā viekše vuškys, bubinēja nadzirdeitais kumeļs, bļuove naslauktuo gūvs. Nadeve mīra tuos pat dūmys, kuruos mūceja jau nazcik nedeļu. Voi atrass muna Ane īdūmuotū laimi svešumā? I kai tuoļuok dzeivuot pošam?

Table 2: Example of Latvian transponations into Latgalian[9]

An approximate translation would be as follows: *This morning I woke up early again. I was thinking about today's chores. In the byre sheep bleated, the horse, still unattended, neighed and the cow, still unmilked, mooed. The same thoughts that had bothered me already for some weeks, were again coming to my mind. Will my Anna find the happiness she hoped to get, there, in a foreign country? And how am I supposed to live further?*

Here we may see that instead of original Latgalian words (e.g. *i* 'again'; *cīši* 'very'; *pruotā stuovēja* 'the mind was occupied with', *kasdīnys* 'everyday') phonetically latgalianised forms of Latvian lexemes are used (cf. *ari* 'again'; *ļūti* 'very'; *ryupēja* 'concerned';

9 Many thanks to Prof. Lidija Leikuma, University of Latvia, who composed the text samples.

ikdīnas 'everyday'). This transponation also concerns the morphology, for example, the original Latgalian reflexive verb form *pasamūdu* 'woke up' is replaced by the transponation of the Latvian form, where the reflexive marker in the prefixed verbs is placed at the end, not after the prefix as in Latgalian (*pamūdū*s vs. *pa*s*amūdu*), and so forth.

Obviously, there is a question how to deal with such texts, that is, whether they can serve a source of the CMSLg or should be ignored.

Despite the publication of several practical grammars and a few dictionaries in the 20th century and the work of special commissions elaborating the literary norm, there is no generally accepted orthography, and a considerable variation is observed in the morphology and lexicon (not to mention the pronunciation, which is not yet even touched by the literary standard). The problem of mixing odd elements coming from the tradition and those promoted by the linguistic authorities should be solved to ensure the automatic processing of the corpus. An intelligent search engine is necessary to identify the spelling variants (cf. recent orthography rules—LPN 2008).

To sum up, there are two general problems complicating the development of the CMSLg: the objective peculiarities of a minor language and the lack of linguistic research of Latgalian. One should emphasise that developing a corpus will stimulate the research and language progress, contributing to the creation of a fully-fledged Latgalian literary language.

The IMCS together with partners at Rezekne Higher Education Institution are planning to start a compilation of corpus of Modern Latgalian.

6. Possible Types of Latgalian Corpora in the Future

Apart from the corpus of Modern Latgalian, which should be our first task to compile, we may consider the compilation of several possible specialised corpora in order to provide further resources and promote a deeper analysis of all aspects of Latgalian. The main emphasis here is placed on the monolingual corpora, as there are not many Latgalian original works translated into Latvian and vice versa. There are some activities that have been observed of the work of Latgalian authors being translated into Russian and vice versa. On the other hand, we cannot exclude the possibility of compiling a parallel Latgalian–Latvian or Latgalian–Russian corpus (or even other language pairs) in the future.

6. 1 Geographical varieties of Latgalian

On one hand, we should start with the modern Latgalian language spoken and written in Latvia. On the other hand, there is a pretty large Latgalian community that settled in Europe and the United States after the Second World War. While in Soviet Latvia Latgalian was used only as a colloquial language at home, a number of books and articles were printed in Germany by Vladislavs Locis' Press in 1945–1984. This might serve as a basis for the specialised corpus of the different varieties of Latgalian.

6. 2 Dialectal varieties of Latgalian

One should not exclude highly valuable data collected in Latvia during expeditions organised by the academic institutions after the Second World War: linguists, historians, folklorists and ethnologists have recorded Latgalian songs, narratives, and so forth, for more than 50 years. The data collection is still going on. These data are scattered all around Latvia, and the information about these collections and their distribution and characteristics is rather vague. Here, a question of a level of co-operation might rise. Hopefully, this might be partly solved within the CLARIN-Latvia framework, in which almost all Latvia's research institutions expressed their will to participate. Another question concerns the technical possibilities to digitalise these data from the old tape recorders. The usage of metadata is important to ensure the reusability of this valuable information.

Apart from this, there are still Latgalians living in Russia, and serious fieldwork is in progress there. There are some research projects carried out by the University of Latvia and St. Petersburg State University to investigate linguistic, sociolinguistic, folklore and culture issues of Siberian Latgalians (cf. the Estonian-Latvian joint conference 'Compatriots in Siberia' held in Tartu 2008, where a number of papers on various topics were presented). One of the latest results is a Latgalian-Latvian-Russian phrase book (Andronovs et al. 2008). There are a number of recordings that have been collected during expeditions to Siberia organised in 2004–2009[10], and which in the future may serve as a solid ground for the spoken sources of the modern counterpart of corpus of the Latgalians in Russia. This, in turn, raises the question of the transcription principles to be used.

10 This research has been financed by the Russian Foundation for Humanities (project no. 07–04-00208a) and University of Latvia (project no. 2007/ZP-38).

6. 3 Diachronic Corpus of Latgalian

Nonetheless, the written tradition stretching more than 250 years back provides a good foundation for developing a diachronic corpus in the future. The first printed Latgalian book which has survived till our days, 'Evangelia toto anno' (1753), is already included in the Corpus of Early Written Latvian[11] in order to give a complete picture of the texts from the 16–18th century. 'Evangelia toto anno' laid the foundation for a second written tradition of Latvian. While the orthography of the first Latvian printed sources was based on the German orthography, the Latgalian prints were using Polish orthography. The first period of written Latgalian ended in 1865, when the ban against printing Lithuanian and Latgalian books in Latin script came into force.

6. 4 Learner Corpus of Latgalian

One might consider the development of a Learner's Corpus. Considering that there are regular winter schools organised in Siberia and regular summer schools in Latgale ('Vosoruošona' and 'Atzolys'), there is a possibility to make a collection of essays written by learners with different backgrounds (national, educational, etc.). If Latgalian is taught as a facultative course in some Latgalian schools, their data is also very valuable source.

7. Conclusions

The restricted usage and lack of some linguistical text types and genres of Latgalian affect the size, representativeness and balance of the corpus of Modern Latgalian. One of the main issues to be dealt with is the input data quality, as there is a risk of 'noisy' texts, which are mere transponations from Latvian, but not Latgalian texts.

There are enough sources that can eventually turn into a number of sub-corpora (diachronic, regional, learner's, etc.). Different corpora will serve a basis for a profound linguistic analysis and will promote the further development of the language processing tools. This will counteract the present state, where Latgalian as a lesser-used language also has fewer resources.

Last, but not the least, the Latgalian corpora will be integrated into the Latvian National Corpus.

11 www.korpuss.lv/senie

References

Andronov, A. / Andronova, E. (2008). "Latgalian in Soviet Russia: A Pilot Model of the Linguistic Corpus of Printed Texts", in *Proceedings of the 21st Conference on Baltic Studies 'Baltic Crossroads: Examining Cultural, Social, and Historical Diversity'.* Indiana University, Bloomington, Indiana. Retrieved May 15, 2009, from http://depts.washington.edu/aabs/documents/confabstr.pdf

Andronova, E. (2007). "The Corpus of Early Written Latvian: current state and future tasks" in *Proceedings of Corpus Linguistics 2007*. Birmingham, UK. Retrieved May 15, 2009, from http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf

Andronova, E. / Andronovs, A. / Leikuma, L. (2008). "«Mozī draugi Sibirī» (Tomska, 1918)—pirmā ābece Sibīrijas latgaliešiem un Krievijas latgaliešu valodas korpuss" in *J. Endzelīna 135. dzimšanas dienas atceres starptautiskās zinātniskās konferences 'No skaņas un burta līdz tekstam un korpusam tēžu krājums'*, Rīga: LU LVI, 3–6.

Andronovs, A. / Leikuma, L. (2008). *Latgalīšu-latvīšu-krīvu sarunu vuordineica*. Krasnojarskys nūvoda regionaluo sabīdriskuo organizaceja «Latgalīšu kulturys centrs» / Latvīšu volūdys apgivis vaļsts agentura. Ačynskys / Reiga.

Balode, L. / Holvoet, A. (2001). "The Latvian language and its dialects" in Dahl, Ö. / Koptjevskaja-Tamm, M. (eds.) (2001). *Circum-Baltic Languages.* Volume I: Past and Present. Amsterdam / Philadelphia: John Benjamins Publishing Company, 3–40.

Biber, D. (1993). "Representativeness in Corpus Design", *Literary and Linguistic Computing*, 8 (4), 243–257.

Brejdak, A. B. (2006). "Latgal'skij jazyk" in Toporov, V. N. / Zav'jalova, M. V. / Kibrik, A. A. et al. (eds.) (2006). *Jazyki mira: Baltijskie jazyki*. Moskva: Academia, 193–213.

Džeriņš, J. / Džonsons, K. (2007). "Harvesting National Language Text Corpora from the Web" in *Proceedings of the 3rd Baltic Conference on Human Language Technologies*. Kaunas, 87–94.

Grūzītis, N. / Auziņa, I. / Bērziņa-Reinsone, S. / Levāne-Petrova, K. / Milčonoka, E. / Nešpore, G. / Spektors, A. (2004). "Demonstration of resources and applications at the Artificial Intelligence Laboratory, IMCS, UL" in *Proceedings of the first Baltic conference 'Human Language Technologies—the Baltic Perspective'*. Riga, 38–42.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London / New York: Longman.

Koncepcija (2005). *Latviešu valodas korpusa koncepcija*. Unpublished. Rīga: Latvijas Universitātes Matemātikas un informātikas institūts. Retrieved May 15, 2009, from http://www.korpuss.lv/

Kučera, K. (2007). "Mapping the Time Continuum: A Major Raison D'être for Diachronic Corpora" in *Proceedings of Corpus Linguistics 2007*. Birmingham, UK. Retrieved May 15, 2009, from http://ucrel.lancs.ac.uk/publications/CL2007/paper/27_Paper.pdf

Leikuma, L. (2008). "The beginnings of written Latgalian" in Ross, K. / Vanags, P. (eds.) (2008). *Common Roots of the Latvian and Estonian Literary Languages*. Frankfurt am Main / Berlin / Bern / Bruxelles / New York / Oxford / Wien: Peter Lang, 211–233.

Levāne, K. / Spektors, A. (2000). "Morphemic Analysis and Morphological Tagging of Latvian Corpus" in *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, May 31 - June 2, 2000. V. 2, 1095–1098.

LPN (2008). *Latgaliešu pareizrakstības noteikumi*. Tieslietu ministrijas Valsts valodas centrs. Rīga / Rēzekne.

Milčonoka, E. (2001). "Some observations about English-Latvian Translation Equivalents in a new Bible for Europe: A Study Based on the EU legislation and its translation" in *Proceedings of COMPLEX2001 6th Conference on Computational Lexicography and Corpus Research.* Birmingham, 175–187.

Milčonoka, E. / Grūzītis, N. / Spektors, A. (2004). "Natural language processing at the Institute of mathematics and computer science: 10 years later" in *Proceedings of the first Baltic conference 'Human Language Technologies—the Baltic Perspective'*. Riga, 6–11.

Paikens, P. (2007). "Lexicon-Based Morphological Analysis of Latvian Language" in *Proceedings of the 3rd Baltic Conference on Human Language Technologies*. Kaunas, 235–240.

Seiļ, V. (1936). *Grāmatas Latgales latviešiem*. Riga: Valtera un Rapas akc. sab. apgāds.

Skadiņa, I. (2005). "Studies of English-Latvian Legal texts for Machine Translation" in Barnbrook, G. / Danielsson, P. / Mahlberg M. (eds.) (2005). *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Continuum, 188–195.

The State Language Commission. Retrieved May 15, 2009, from http://www.president.lv/pk/content/?cat_id=8

Unesco Digital Atlas of the World's Languages in Danger of Disappearing. Work in progress, beta 0.2 version October 8, 2008. Retrieved May 15, 2009, from http://www.unesco.org/culture/ich/atlas/.

Vasiļjevs, A. (2008). "Kā veidosim Latviešu valodas nacionālo korpusu?" Speech given at CLARIN project and the National Corpus workshop on November 3, 2008. Retrieved May 15, 2009, from http://www.clarin.lv/materiali/clarin-vasiljevs.ppt.

VVL (2000). "Valsts valodas likums", *Latvijas Vēstnesis* 428/433 (1888/1893), December 21, 1999. (The English translation is available at: http://isec.gov.lv/normdok/oflanglaw.htm—retrieved May 15, 2009.)