SLIEKKAS | KALT

Senosios **LIE**tuvių **K**albos **K**orpus**AS** | **K**orpus **A**lt**L**i**T**auisch

Jolanta Gelumbeckaitė
Institute of Empirical Linguistics
gelumbeckaite@em.uni-frankfurt.de

Aim of SLIEKKAS

- ➢ to create a qualitative multilevel electronic retrieval engine for multilateral linguistic research of Old Lithuanian,
- ➢ to allow for reliable results of diachronic Lithuanian language studies,
- ➢ to enable the implementation of the two biggest desiderata of Baltic linguistics, the **Old Lithuanian grammar** and the **historic dictionary of Lithuanian**.

What is SLIEKKAS doing?

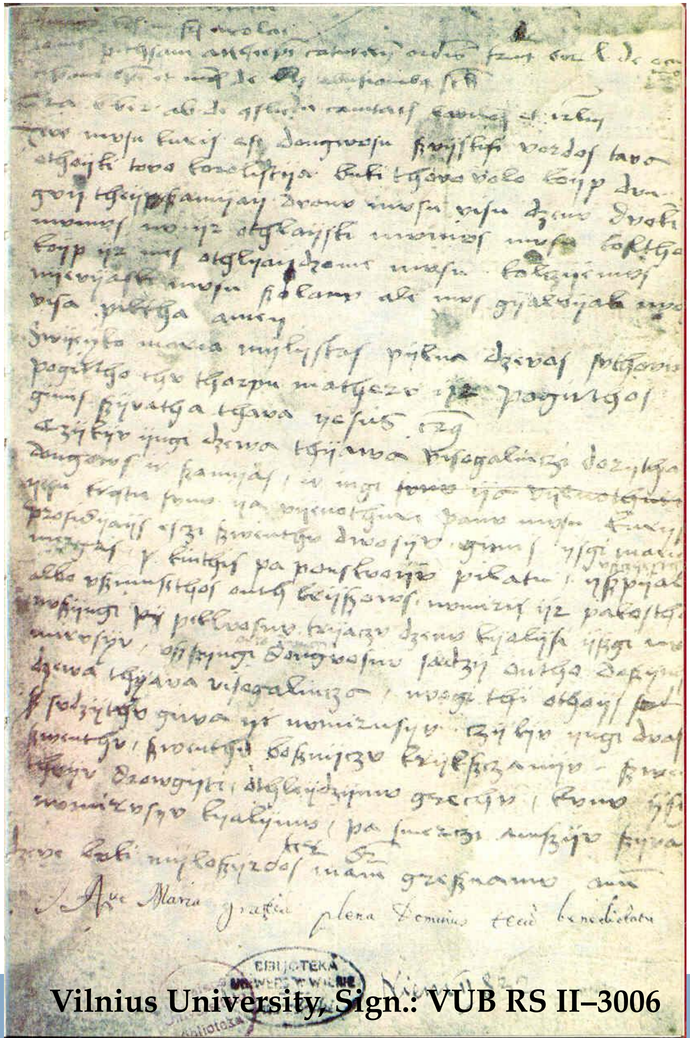
- ✓ Joining the current process of contemporary historical corpora, and not copying them or reinventing the wheel.
- ✓ Consolidating already existing databases of Old Lithuanian and using already existing resources optimally.
- ✓ Developing a concept of the Old Lithuanian reference corpus as well as its scientific and technical base.

Arrangement of the Corpus

- Digitisation according to the principles of the diplomatic literal transcription and conversion to the XML format;
- Multi-layer-stand-off annotation:
  - Header-Informations,
  - Structural, palaeographic, and textological annotations,
  - Lemmatising of word forms (according to the language period), glossing in Standard Lithuanian, and translating into Modern Lithuanian,
  - POS-Tagging. Parts of speech and grammatical forms (morphology and in certain cases syntactic morphology);
- Alignment of the Lithuanian texts with the digital facsimiles (autograph, print) and with such parallel corpora as with the annotated Latin, German, Polish etc. source texts.

Old Lithuanian (1500–1800), ca. 10 million words

The oldest known Lithuanian text:
Pater noster, Ave Maria and *Credo*
in the *Tractatus sacerdotalis* (Straßburg: Martin Flach, 1503) by Nicolaus de Blony

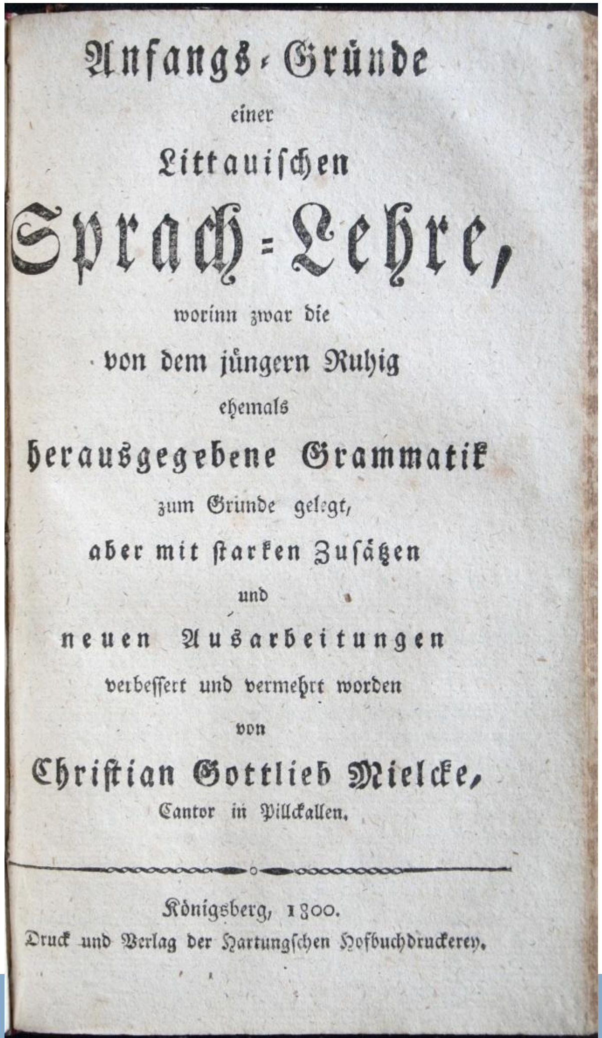


[illegible]
[illegible]
[illegible]
[illegible]
Tewe muſu kuriſ aſſ danguweſu ſzwiſkiſ vardaſ tavo
athaiki tavo karaliſtya buki thavo vala kaip dan-
guy thaypſſamyay dwonu muſu wiſu dzenu duoki
mumuſ nu yr athlayſki mumuſ muſu kaltha
kaip yr meſ athlaydzem muſu kaltzimuſ
yr newedi muſu ſokuſimu ale muſ gialbeki nu
viſa pikta amen
Swiekia maria myliſtaſ pilna dziewaſ ſu tavim
pagirtaſ tu tharpu mathezr yr pagirtaſ
gimiſ ſzywatha thawa yeſuſ. czy
Aſs tikiu ingi dziewa thiawa viſagalincza dazytha
dunguwſ ir ſzamyaſ ir ingi [illegible]
yeſu kriſtu ſunu ya wienatigi pono muſu kury
praſidyaſſ iſz ſzwenthe dwaſyo gimuſ iſz marie
mergaſ y kinthyſ pa ponſtwayo pilato iſz [illegible]
albo viſmuſſthoſ [illegible] nomirzuſ iſz pakoſtha
[illegible] pi pilkwoſnu [illegible] dzenu kyaliſſ iſzgi [illegible]
[illegible] dangwaſnu ſedzy dutha [illegible]
dzewa thyava viſagalinczo [illegible] thi athaiſ [illegible]
ſudzytho gyva yr nomirzuſyu [illegible] dwaſ
ſzwentha ſzwenthe baſnyczu [illegible]
[illegible] atlaydzymu grecziu [illegible]
nomirzuſyu kyalyms pa ſmerczi [illegible]
[illegible] buki nyſkoſyzdoſ mam grekſnamu amen

Ave Maria gratia plena Dominus tecū benedicta tu

Vilnius University, Sign.: VUB RS II–3006

Lithuanian Grammar by
Christian Gottlieb Mielcke (1732–1807)
Anfangs=Gründe einer Littauischen Sprach=Lehre (Königsberg, 1800)

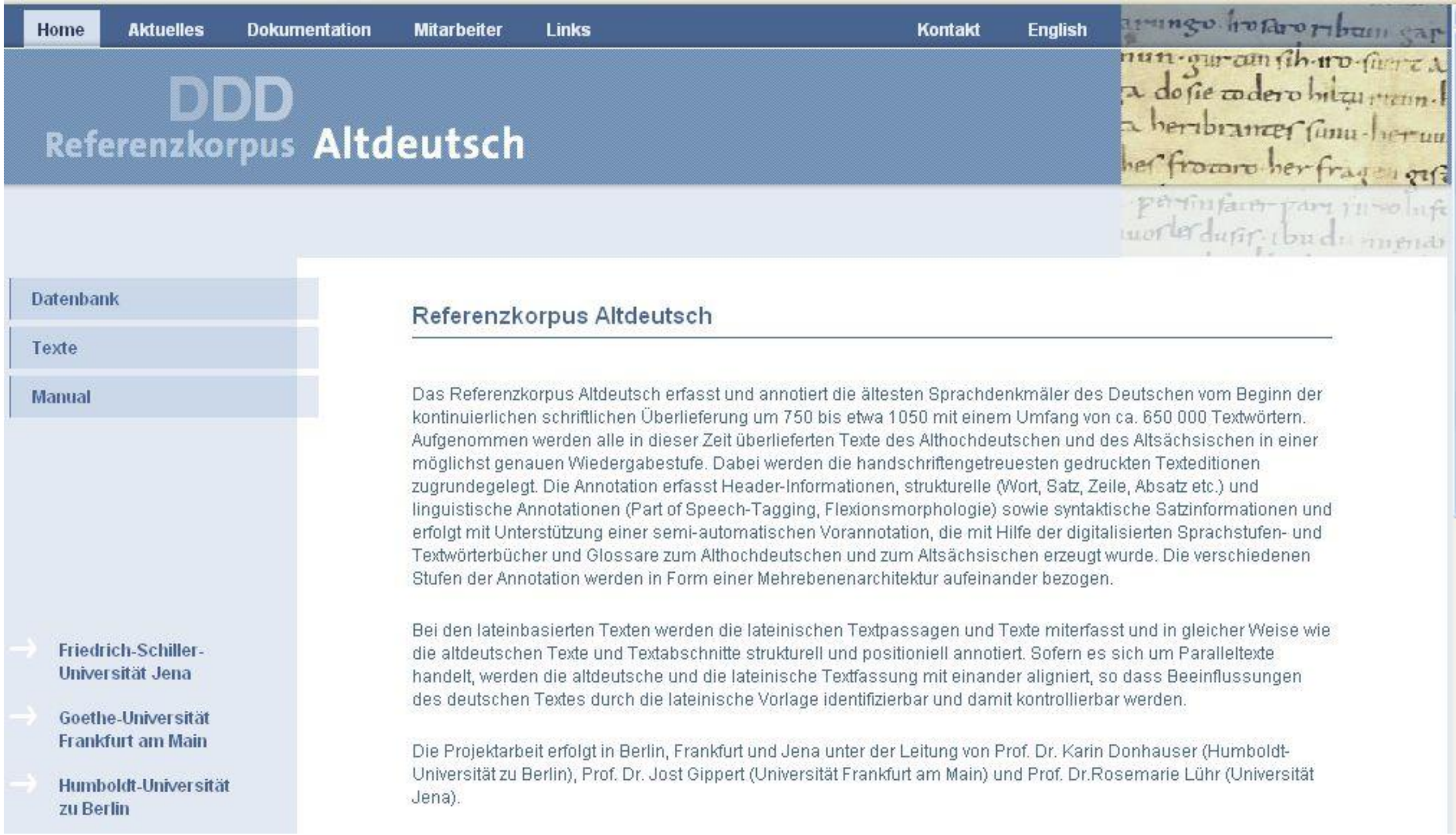


Anfangs=Gründe
einer
Littauiſchen
Sprach=Lehre,
worinn zwar die
von dem jüngern Ruhig
ehemals
herausgegebene Grammatik
zum Grunde gelegt,
aber mit ſtarken Zuſätzen
und
neuen Ausarbeitungen
verbeſſert und vermehrt worden
von
Chriſtian Gottlieb Mielcke,
Cantor in Pillkallen.

Königsberg, 1800.
Druck und Verlag der Hartungſchen Hofbuchdruckerey.

Referenzkorpus Altdeutsch

Scientifically as well as technologically SLIEKKAS leans onto the Old German Reference Corpus *DDD – Referenzkorpus Altdeutsch (750–1050)*


Home
Aktuelles
Dokumentation
Mitarbeiter
Links
Kontakt
English
DDD
Referenzkorpus Altdeutsch
Datenbank
Texte
Manual
Friedrich-Schiller-Universität Jena
Goethe-Universität Frankfurt am Main
Humboldt-Universität zu Berlin

Referenzkorpus Altdeutsch

Das Referenzkorpus Altdeutsch erfasst und annotiert die ältesten Sprachdenkmäler des Deutschen vom Beginn der kontinuierlichen schriftlichen Überlieferung um 750 bis etwa 1050 mit einem Umfang von ca. 650 000 Textwörtern. Aufgenommen werden alle in dieser Zeit überlieferten Texte des Althochdeutschen und des Altsächsischen in einer möglichst genauen Wiedergabestufe. Dabei werden die handschriftengetreuesten gedruckten Texteditionen zugrundegelegt. Die Annotation erfasst Header-Informationen, strukturelle (Wort, Satz, Zeile, Absatz etc.) und linguistische Annotationen (Part of Speech-Tagging, Flexionsmorphologie) sowie syntaktische Satzinformationen und erfolgt mit Unterstützung einer semi-automatischen Vorannotation, die mit Hilfe der digitalisierten Sprachstufen- und Textwörterbücher und Glossare zum Althochdeutschen und zum Altsächsischen erzeugt wurde. Die verschiedenen Stufen der Annotation werden in Form einer Mehrebenenarchitektur aufeinander bezogen.

Bei den lateinbasierten Texten werden die lateinischen Textpassagen und Texte miterfasst und in gleicher Weise wie die altdeutschen Texte und Textabschnitte strukturell und positioniell annotiert. Sofern es sich um Paralleltexte handelt, werden die altdeutsche und die lateinische Textfassung mit einander aligniert, so dass Beeinflussungen des deutschen Textes durch die lateinische Vorlage identifizierbar und damit kontrollierbar werden.

Die Projektarbeit erfolgt in Berlin, Frankfurt und Jena unter der Leitung von Prof. Dr. Karin Donhauser (Humboldt-Universität zu Berlin), Prof. Dr. Jost Gippert (Universität Frankfurt am Main) und Prof. Dr.Rosemarie Lühr (Universität Jena).

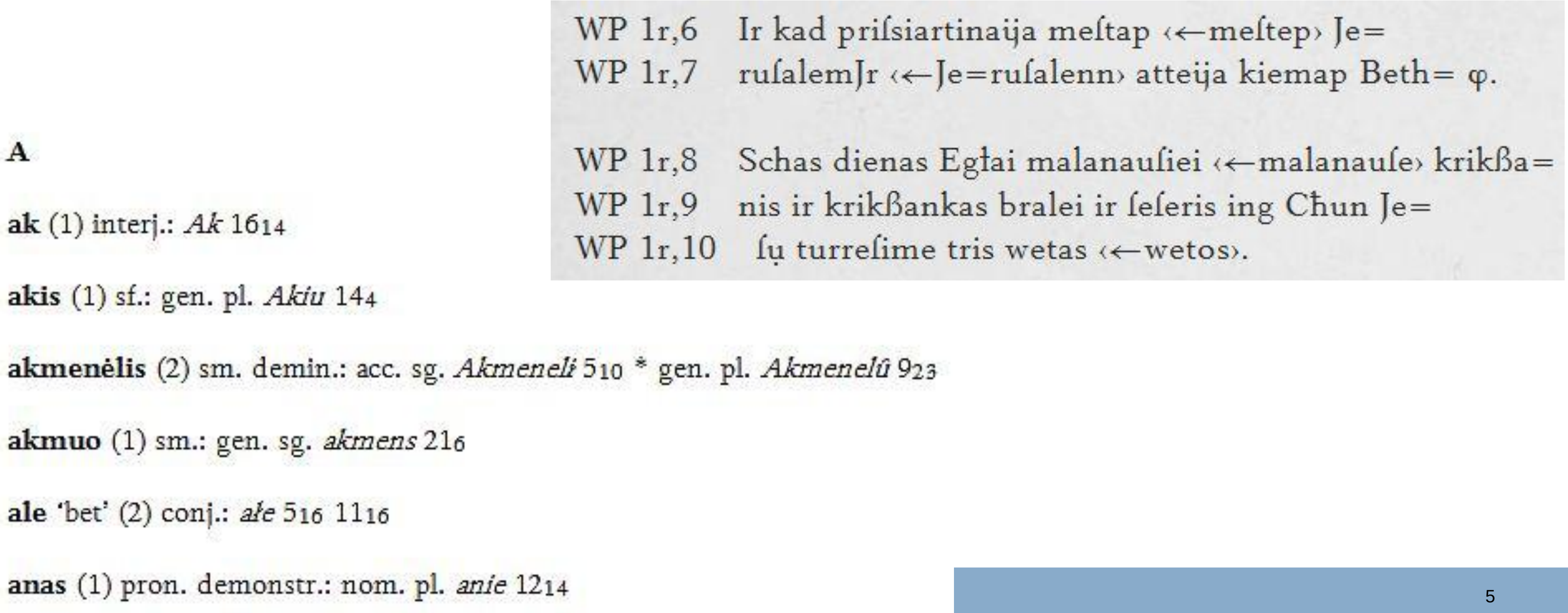
Senieji raštai

www.lki.lt/seniejirastai

Institute of Lithuanian Language, Vilnius. The most comprehensive and constantly increasing database of Old Lithuanian Writings. The digital versions of the texts are prepared according to the principles of the diplomatic literal transcription. Out of 72 texts (1573–1816, over 3 m. text words) 40 texts along with their KWIC concordances are currently accessible to external users online and for downloading (*Word-doc* format).

Deficiencies:

- no special search engine,
- no clear solution for the lemmatising and the grammatical tagging.



WP 1r,6 Ir kad priſsiartinaịja meſtap ‹←meſtep› Je=
WP 1r,7 ruſalemJr ‹←Je=ruſalenn› atteịja kiemap Beth= φ.

WP 1r,8 Schas dienas Egłai malanauſiei ‹←malanauſe› krikßa=
WP 1r,9 nis ir krikßankas bralei ir ſeſeris ing Cħun Je=
WP 1r,10 ſụ turreſime tris wetas ‹←wetos›.

A

ak (1) interj.: *Ak* 16_{14}

akis (1) sf.: gen. pl. *Akiu* 14_{4}

akmenėlis (2) sm. demin.: acc. sg. *Akmeneli* 5_{10} * gen. pl. *Akmenelû* 9_{23}

akmuo (1) sm.: gen. sg. *akmens* 21_{6}

ale ‘bet’ (2) conj.: *ale* 5_{16} 11_{16}

anas (1) pron. demonstr.: nom. pl. *anie* 12_{14}

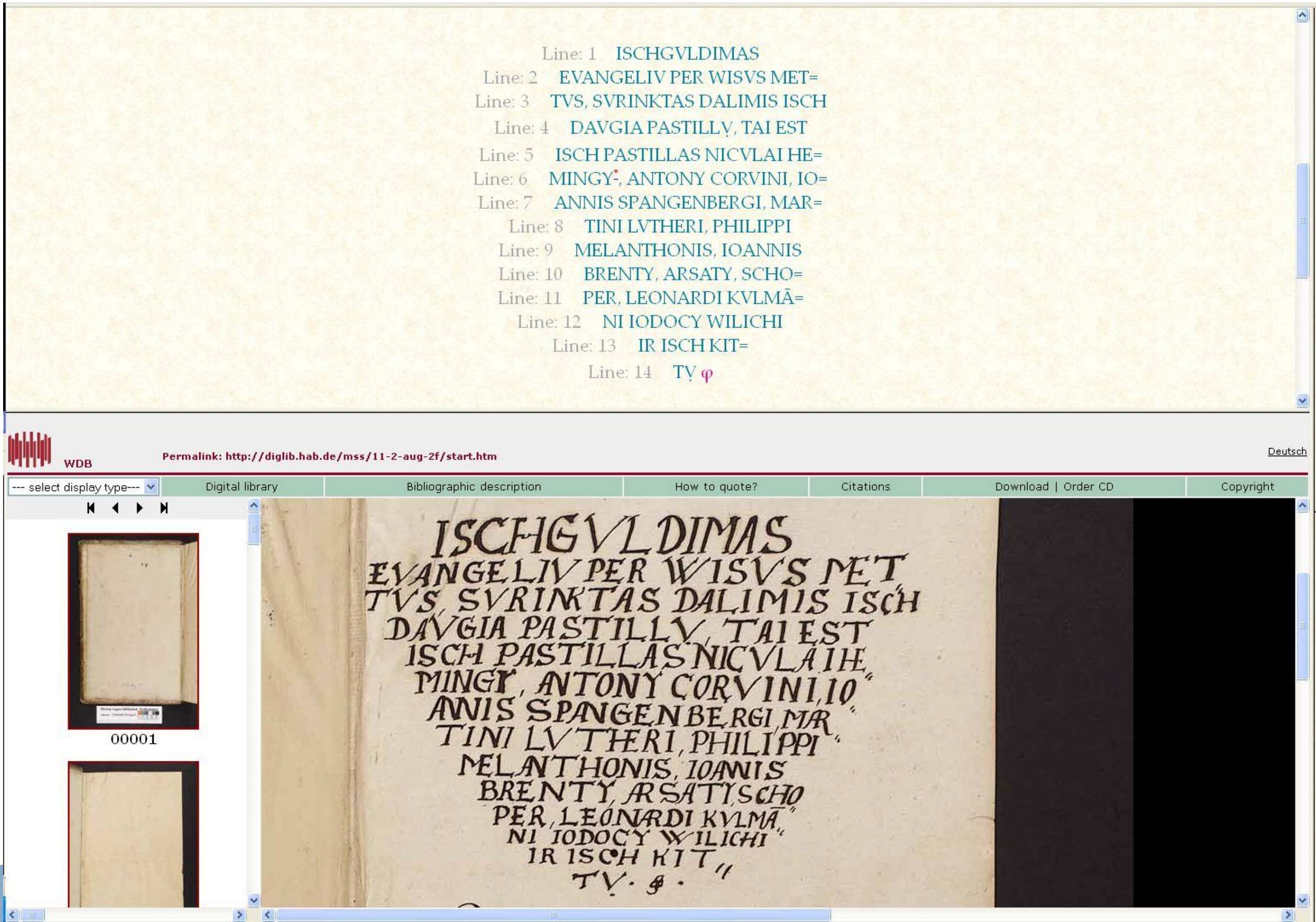
The Lithuanian material consists of ca. 200.000 words.
Every character (resp. every word) is aligned with the index databank. This makes it possible to register all parallel positions of the words in the corpus as well as with the facsimile of the autograph (folio to folio so far).

Single corpora as well as the reference corpus can be searched by using special database query forms.

The texts are annotated merely with regard to the lemmatisation.

Deficiency:

- no grammatical annotations.

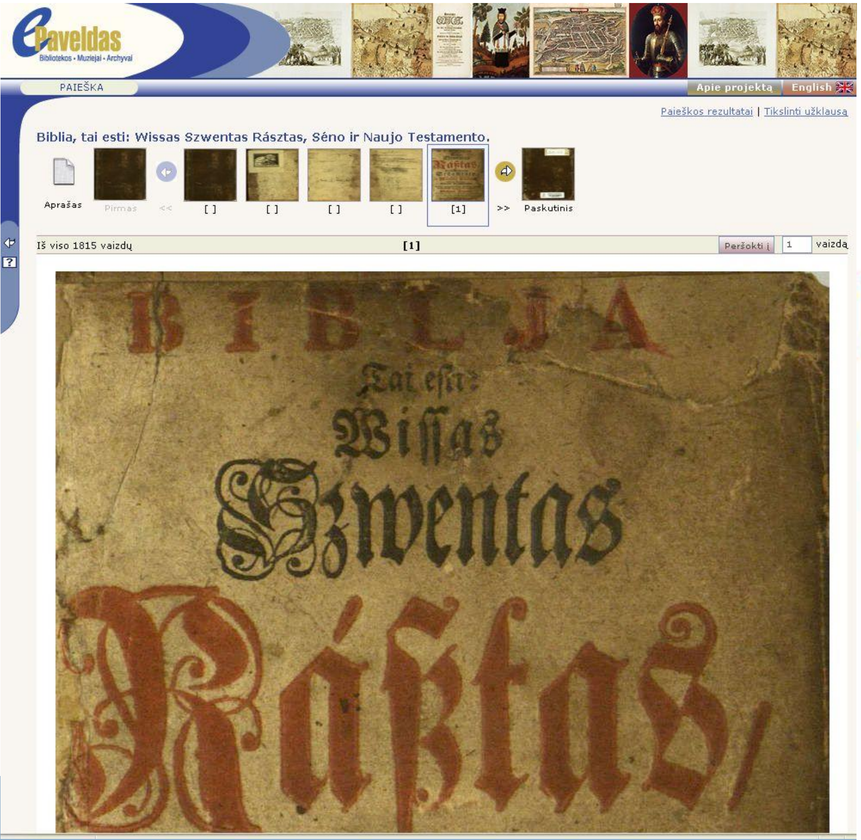


Line: 1 ISCHGVLDIMAS
Line: 2 EVANGELIV PER WISVS MET=
Line: 3 TVS, SVRINKTAS DALIMIS ISCH
Line: 4 DAVGIA PASTILLỴ, TAI EST
Line: 5 ISCH PASTILLAS NICVLAI HE=
Line: 6 MINGY*-, ANTONY CORVINI, IO=
Line: 7 ANNIS SPANGENBERGI, MAR=
Line: 8 TINI LVTHERI, PHILIPPI
Line: 9 MELANTHONIS, IOANNIS
Line: 10 BRENTY, ARSATY, SCHO=
Line: 11 PER, LEONARDI KVLMĀ=
Line: 12 NI IODOCY WILICHI
Line: 13 IR ISCH KIT=
Line: 14 TỴ φ

Permalink: http://diglib.hab.de/mss/11-2-aug-2f/start.htm

ePaveldas
www.epaveldas.lt

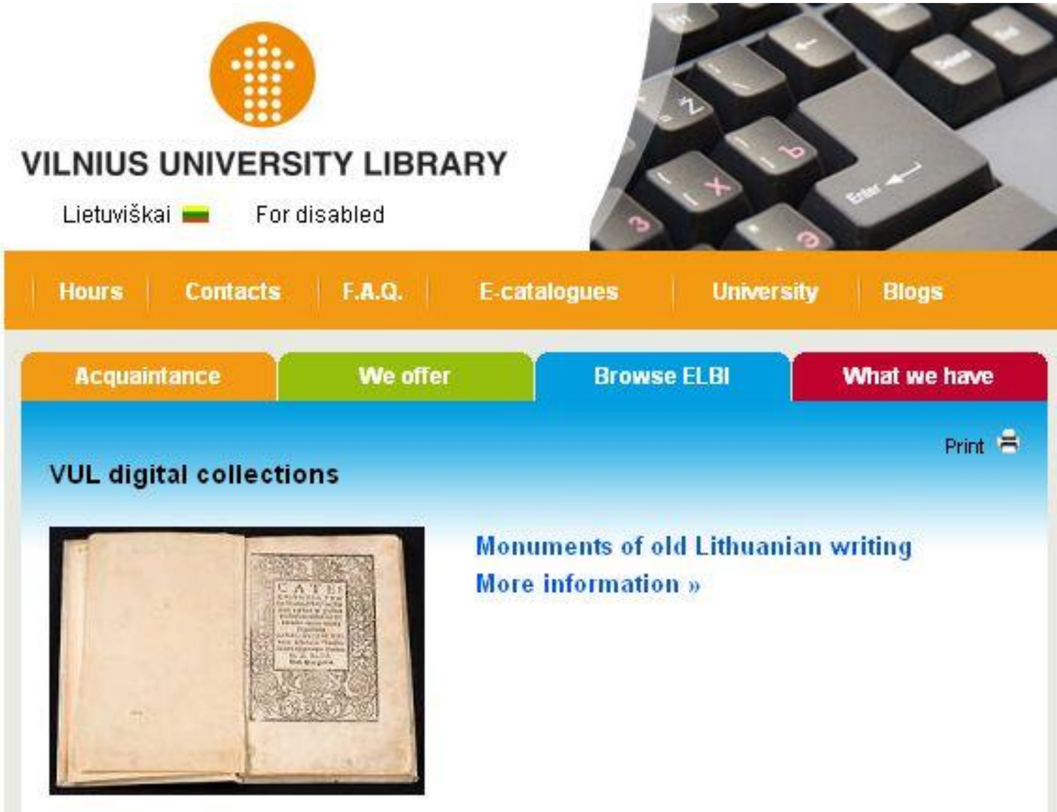
The databank of the virtual cultural heritage system contains 314 digitised facsimiles of handwritten and printed Lithuanian texts from the period 1547–1863.



Biblia, tai esti: Wissas Szwentas Rásztas, Séno ir Naujo Testamento.

Vilnius University Library
www.mb.vu.lt/sk-kolekcijos

VUL digital colletions comprise among others unique copies of Old Lithuanian grammars, dictionaries, and catechisms.



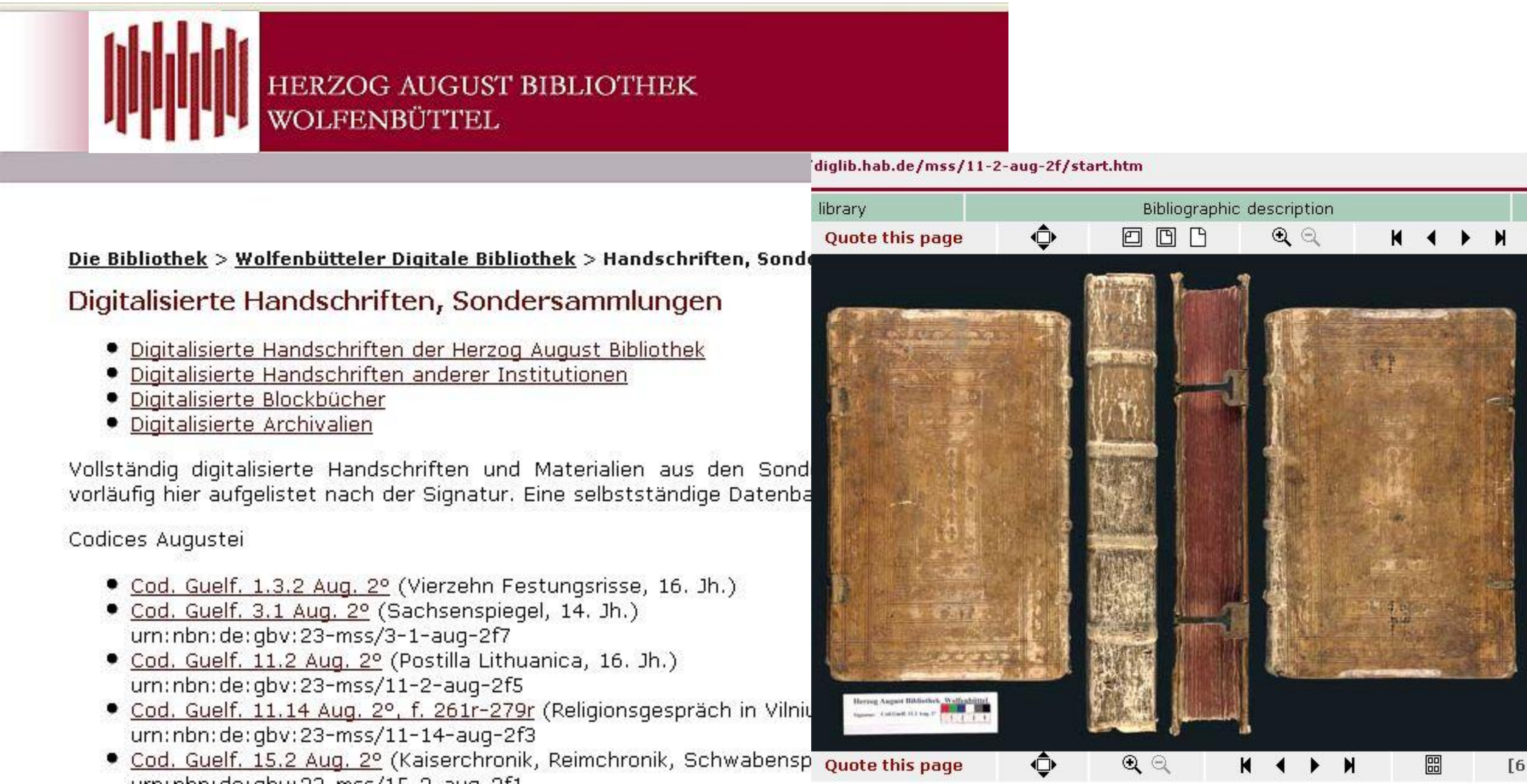
VUL digital collections

Monuments of old Lithuanian writing
More information »

Herzog August Bibliothek

www.hab.de

Digital facsimiles in the databases of such single libraries as the Duke August Library in Wolfenbüttel. This library both preserves the original and places at the disposal a digital facsimile of the oldest Lithuanian Codex from 1573.



HERZOG AUGUST BIBLIOTHEK
WOLFENBÜTTEL

Die Bibliothek > **Wolfenbütteler Digitale Bibliothek** > **Handschriften, Sond**

Digitalisierte Handschriften, Sondersammlungen

- Digitalisierte Handschriften der Herzog August Bibliothek
- Digitalisierte Handschriften anderer Institutionen
- Digitalisierte Blockbücher
- Digitalisierte Archivalien

Vollständig digitalisierte Handschriften und Materialien aus den Sond
vorläufig hier aufgelistet nach der Signatur. Eine selbstständige Datenba

Codices Augustei

- Cod. Guelf. 1.3.2 Aug. 2° (Vierzehn Festungsrisse, 16. Jh.)
- Cod. Guelf. 3.1 Aug. 2° (Sachsenspiegel, 14. Jh.)
  urn:nbn:de:gbv:23-mss/3-1-aug-2f7
- Cod. Guelf. 11.2 Aug. 2° (Postilla Lithuanica, 16. Jh.)
  urn:nbn:de:gbv:23-mss/11-2-aug-2f5
- Cod. Guelf. 11.14 Aug. 2°, f. 261r-279r (Religionsgespräch in Vilni
  urn:nbn:de:gbv:23-mss/11-14-aug-2f3
- Cod. Guelf. 15.2 Aug. 2° (Kaiserchronik, Reimchronik, Schwabensp
  urn:nbn:de:gbv:23-mss/15-2-aug-2f1

/diglib.hab.de/mss/11-2-aug-2f/start.htm

library
Bibliographic description
Quote this page
Herzog August Bibliothek Wolfenbüttel
Quote this page
[6

Selected texts for the beginning (ca. 200 000 words)

1. 3 prayers (*Pater noster, Ave Maria, Credo*), ca. 1520, 168 text words.
2. Martynas Mažvydas, *Catechismusa prasty žadei* (*MžK* 1547), 7 000 text words.
3. Martynas Mažvydas, *Giesme S. Ambrasziejaus* (*MžGA* 1549), 2 418 text words.
4. Wolfenbüttel Postilla (*WP* 1573), 130 559 text words.
5. Baltramiejus Vilentas, *Enchiridion* (*VE* 1579), 10 199 text words.
6. Mikalojus Daukša, *Kathechismas* (*DK* 1595), 13 702 text words.
7. Heinrich Johann Lysius, *Mažas Katgismas* (*LyK* 1719), 6 136 text words.
8. Gabrielis Engelis, *Mažas Katgismas* (*EnK* 1722), ca. 10 000 text words.
9. Kristijonas Donelaitis, *Metai* (ca. 1765–1775), 17 291 text words.

Grammatical annotation

- *communis opinio* – any grammatical tagging of a text is a “perilous activity, because the text thereby loses [its] integrity” and becomes contaminated through various views of different linguistic schools (John M. Sinclair, *Current Issues in Corpus Linguistics*, 2004);
- no in any way grammatically annotated contemporary Lithuanian text-corpus;
- no standards for tagging of the Lithuanian texts;
- Lithuanian corpora as merely raw text accumulations;
- The Contemporary Lithuanian Corpus (donelaitis.vdu.lt), over 100 million words.

Kompiuterinės lingvistikos centras

Dabartinės lietuvių kalbos
TEKSTYNAS

Teksto suvokimą lemia mūsų lūkesčiai

Grammatical annotation of SLIEKKAS

- predominantly restricted to morphology,
- POS-Tagging and the morphological description of single word forms,
- hierarchical: starting with the grammatical class of the lemma and going to individual unchangeable, and changeable morphological categories of the lemma as well as of the individual word form attested.

– The difficulties begin in the establishing of the parts of speech already ("There are ten parts of speech, and they are all troublesome", Mark Twain, *The Awful German language*). The Lithuanian grammar traditionally distinguishes 11 parts of speech, since the beginning of the 20th century setting apart interjections and onomatopoeica. The POS of the SLIEKKAS consits although of 10 more or less "troublesome" grammatical classes:

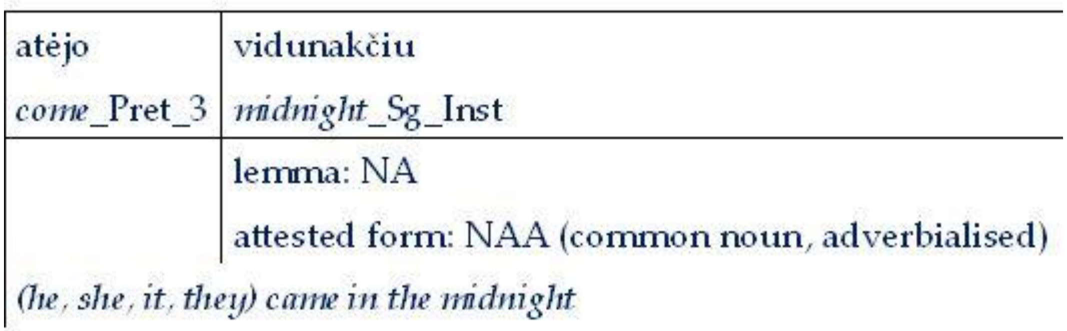
Parts of speech

1. **ADJ** – adjective,
2. **ADV** – adverb,
3. **AP** – adposition,
4. **CARD** – cardinal number, **ADJO** ordinal number,
5. **ITJ** – interjection + onomatopoeic,
6. **KO** – conjunction,
7. **N** –noun: **NA** common noun, **NT** proper noun,
8. **P** – pronoun: **PD** demonstrative, **PI** indefinite, **PK** interrogative, **PPER** personal,
9. **PTK** – particle,
10. **V** – verb: **VA** auxiliary, **VV** main.

POS

The grammatical class is indicated according to the lemma and to the word form attested. This distinction enables us to indicate:

- changes in the grammatical classes (nominalisation of the adjectives, adjectivisation of the participles, adverbialisation of the nouns, adjectives or participles, and turning of some nouns into adpositions),
- ambivalent issues as adverbialisation of the oblique case forms:



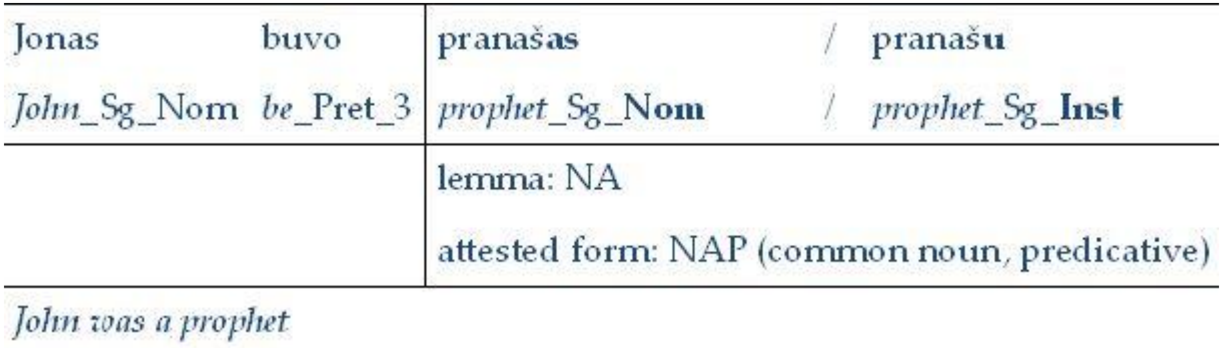
atėjo	vidunakčiu
*come*_Pret_3	*midnight*_Sg_Inst
	lemma: NA
	attested form: NAA (common noun, adverbialised)
(he, she, it, they) came in the midnight	

- distinction of prepositions and postpositions within the class of the adposition,
- a clear definition of such non-finite forms of the verb as infinitives, participles, semiparticiples, gerunds, gerundives, supines, and adverbialised infinitives.

Morphology and syntactic morphology (1)

- analytic (compound) tense forms, consisting of a copula and a participial (present or past, active or passive) predicative:
  - distinguishing between an auxiliary and a main verb already in the definition of the grammatical class of the lemma,
  - between predicative, attributive or adverbial usage of the participles in the characterisation of the word forms attested.

- compound nominal predicates, consisting of a copula and of a nominal predicative:

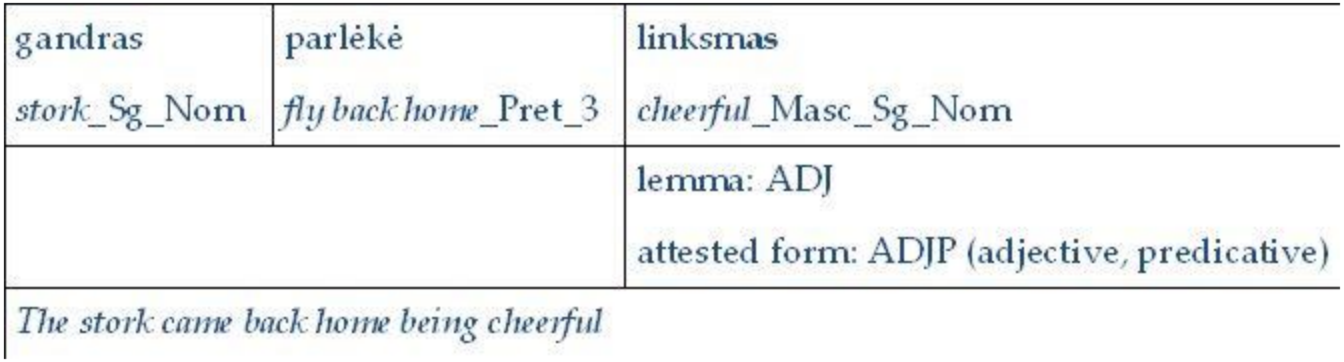


Jonas	buvo	pranaš**as**	/	pranaš**u**
*John*_Sg_Nom	*be*_Pret_3	*prophet*_Sg_**Nom**	/	*prophet*_Sg_**Inst**
		lemma: NA attested form: NAP (common noun, predicative)		

John was a prophet

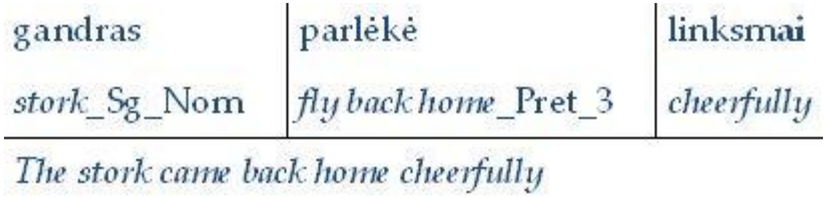
Morphology and syntactic morphology (2)

- secondary predicatives, expressed by nouns, adjectives, pronouns, numerals or participles:



gandras *stork*_Sg_Nom	parlėkė *fly back home*_Pret_3	linksmas *cheerful*_Masc_Sg_Nom
		lemma: ADJ attested form: ADJP (adjective, predicative)
The stork came back home being cheerful		

✓ cf. the adverb *linksmai* in this position:



gandras *stork*_Sg_Nom	parlėkė *fly back home*_Pret_3	linksmai *cheerfully*
The stork came back home cheerfully		

Morphological annotation

The further morphological annotation is also hierarchical. It consists of 3 layers, which include the information about:

1) the unchangeable morphological categories of the lemma (like gender and flexional class),

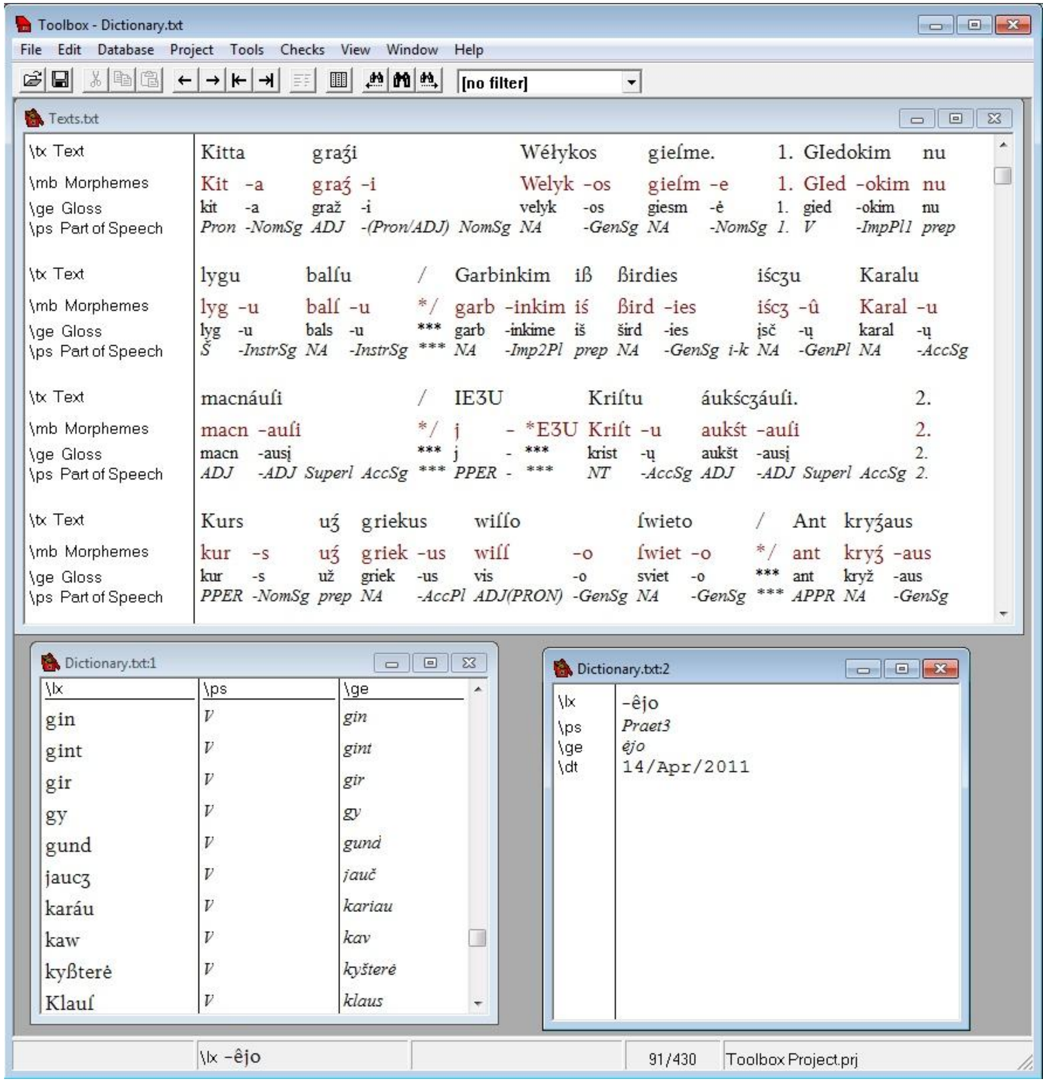
2) the unchangeable morphological categories of the word form attested: they can differ from these of the lemma,

3) the flexional morphological categories or the word form attested.

Data analysis tools (Toolbox)

There are no specialised diachronic dictionaries of Old Lithuanian, which we could apply for the (semi)automatisation of the annotation process. Therefore we have to either perform the morphological annotation manually or to choose such data analysis tools like the Toolbox, which enables semiautomatisations of interlinear annotations.

Through recognising and defining the problems we search for the best solutions. Going through the 'earthworm' limb by limb.



```
\tx Text             Kitta         graʒi                       Wélykos      gieſme.        1. Gledokim     nu
\mb Morphemes        Kit  -a       graʒ -i                     Welyk -os    gieſm -e       1. Gled -okim   nu
\ge Gloss            kit  -a       graž -i                     velyk -os    giesm -ė       1. gied -okim   nu
\ps Part of Speech   Pron -NomSg   ADJ  -(Pron/ADJ) NomSg      NA    -GenSg NA    -NomSg   1. V    -ImpPl1 prep

\tx Text             lygu        balſu          /    Garbinkim      iß    ßirdies        iśczu       Karalu
\mb Morphemes        lyg -u      balſ -u        */   garb -inkim    iś    ßird -ies      iścʒ -û     Karal -u
\ge Gloss            lyg -u      bals -u        ***  garb -inkime   iš    šird -ies      įsč  -ų     karal -ų
\ps Part of Speech   Š   -InstrSg NA  -InstrSg  ***  NA   -Imp2Pl   prep  NA   -GenSg i-k NA  -GenPl  NA    -AccSg

\tx Text             macnáuſi                   /    IE3U          Kriſtu        áukścʒáuſi.                 2.
\mb Morphemes        macn -auſi                 */   j    - *E3U   Kriſt -u      aukśt -auſi                 2.
\ge Gloss            macn -ausį                 ***  j    - ***    krist -ų      aukšt -ausį                 2.
\ps Part of Speech   ADJ  -ADJ Superl AccSg     ***  PPER - ***    NT    -AccSg  ADJ   -ADJ Superl AccSg     2.

\tx Text             Kurs         uź    griekus      wiſſo                 ſwieto        /    Ant   kryźaus
\mb Morphemes        kur  -s      uź    griek -us    wiſſ         -o       ſwiet -o      */   ant   kryź -aus
\ge Gloss            kur  -s      už    griek -us    vis          -o       sviet -o      ***  ant   kryž -aus
\ps Part of Speech   PPER -NomSg  prep  NA    -AccPl ADJ(PRON)    -GenSg   NA    -GenSg  ***  APPR  NA   -GenSg
```

\lx	\ps	\ge
gin	*V*	*gin*
gint	*V*	*gint*
gir	*V*	*gir*
gy	*V*	*gy*
gund	*V*	*gund*
jaucʒ	*V*	*jauč*
karáu	*V*	*kariau*
kaw	*V*	*kav*
kyßterė	*V*	*kyšterė*
Klauſ	*V*	*klaus*

\lx -êjo
\ps *Praet3*
\ge *ėjo*
\dt 14/Apr/2011

\lx -êjo | 91/430 | Toolbox Project.prj

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

www.goethe-universitaet.de

Thank you for your attention!