Argument structure in Slavic diachrony: perspectives for corpus-based research

Roland Meyer
Institut für Slavistik
Universität Regensburg

Polska akademia nauk - Instytut podstaw informatyki
Spotkanie DAAD-PPP/PAN, 19. kwietnia 2006

Plan

- Argument structure in diachrony
- Regensburg Diachronic Corpus
- Other diachronic corpora of Slavic languages
- First results
- Perspectives and Conclusion

Argument Structure in diachrony (1)

- Synchronic fact: differences in argument realisation among Slavic languages
- subject realization and diathesis
  - pro-drop [Russian vs. others]
  - subject expletives [Czech vs. others]
  - pass. participle/reflexive + accusative object [Polish and Ukrainian vs. others]
  - reflexive verbs and impersonals [messy]

pro-drop (1) [cf. Lindseth 1998, Lindseth & Franks 1994]

- pronominal subject ±emphatic

(1) a. *ja ne ponimaju* [R, possibly non-emphatic]
b. *já nerozumím* [CZ, oblig. emphatic]
'I don't understand.'

- oblig. referential reading in 3.pl.

(2) a. *vo Francii (oni) edjat ulitok* [R, possibly non-referential]
b. *we Francji oni jedzą ślimaki* [PL, oblig. referential]
'In France they eat snails.'

- bound variable readings

(3) a. *Ivan dumaet, čto on polučit pjaterku.* [R, possibly coreferential]
b. *Jan myslí, že on dostane jedničku.* [CZ, oblig. noncoreferential]
'John thinks that he will get an A.'

pro-drop (2)

[cf. Lindseth 1998]

- other opinions:
  - R is canonically pro-drop. No relevant differences [Müller 1989, Kosta 1990]; differences in use of anaphoric pronouns [Koktová 1992]
- Why is Russian different?
  - note: lacks verbal agreement marking in past tense

(4) a. *(ja) čital* [R - no person marking]

b. *czytałem, četl jsem, čitala sam …* [PL, CZ, BCS - person marking]

'I was reading.'

  - existence of a consistent person agreement paradigm licenses referential null subjects

pro-drop (3)

[cf. Lindseth 1998]

- Sources in diachrony

(5) a. *ne dalъ jesi kozlęte* [Ostromir Gospel 1056-7, R Church Sl.]
not given aux-2sg goat
'you did not give a goat'

b. *gdě že ne xodili ni dědi naši* [Laurentius chronicle 1377, Old R]
where MP not went nor grandfathers our
'where our grandfathers didn't go'

c. *ja dalъ rukoju svojeju* [Mstislav's charter 1130, Old R]
I gave my-ins hand-ins
'I gave with my own hand'

- aux lost early in 3rd, gradually in 1st/2nd p.
- use of non-emphatic pronouns spreads widely

pro-drop (4)

- Hypothesis: defective agreement paradigm is the reason for loss of pro-drop
- But: data unclean, vast variation, diglossia

(6) *sego ty že esi xotelъ* [Laurentius chronicle 1377, Old R]
this you MP aux-2sg wanted
'This is what you wanted.'

- gradual drift or change in parameter setting?
- **Constant Rate Effect** [Kroch 1999]

"The rate of change in different surface contexts reflecting a single underlying parameter is the same."

pro-drop: agenda

- frequency distributions of
  - agreeing auxiliaries
  - overt pronouns
  - agreeing auxiliaries with/without overt subject
  - general measures of Church Slavonic influence
- hypotheses
  - amount of 1st/2nd p. auxiliaries compared to overt 1st/2nd p. pronouns remains constant as the former decline and the latter increase
  - 3rd p. auxiliaries dropped earlier

subject expletives (1)

- Czech has developed expletive-like elements

(7) a. *vono se tam nepracuje* [non-agreeing; emphatic]
it refl there not-works
'no work is being done there'

b. *vona ta myšlenka má něco do sebe* [agreeing]
it-fem this-fem thought-fem has something to it
'this thought has something to it'

c. *vono je možný, že tam nepřijede* [non-agreeing]
it is possible that there not-comes
'it is possible that he will not arrive there'

- cf. also Russian èto

(8) èto počemu že vy mne daete stol'ko deneg? [Bulgakov]
that why MP you me-dat give so-much money
'Why are you giving me so much money?'

subject expletives (2)

- historical development [cf. Trávníček 1962, Lindseth 1998]
  - interjection *ano, aj* - outside the sentence

(8) *ano, těžař čaká drahého óvoce* [Olomouc Gospel 14th cent.]
see farmer waits precious fruit

  - agreeing forms; integrated into the sentence

(9) *a oni Polaci volili jsú sobě za kníže mladence jednoho* [2. half 14th c.]
and expl Poles elected aux refl-dat as ruler young-man one
'And the Poles elected a young man as their ruler.'

  - since end of the 14th century: gradually expletive
  - independent evidence: host for 2P clitics

[Trávníček 1962, no example]

subject expletives (3)

- theory of pro licensing [Lindseth/Franks 1994]
  - L allows for null referential subjects => L allows for null expletive subjects (but not <=)
  - licensing [sufficient for t without content] vs. identification [necessary for referential t]
  - relation to amount of syncretism in paradigms
  - Lindseth: *(v)on(-a,-o)* is not an unmarked subject, but functionally/emotionally charged, does not count for theory of null subjects
- German contact?

subject expletives: agenda

- frequency distributions of *(v)on(-a,-o)*, *to*, *èto* ...
- relation to 2P cliticization (Cz)
- relation to doubling with thematic subject (Cz)
- comparison to development paths of real expletives (German, English)
- what determines [± agree] - *(v)on(-a,-o)* vs. *èto*?
- Cz diachronic corpus ready to use

reflexivity (1)

[cf. Junghanns 1996]

(10) *Ivan moetsja*. [R, reflexive]
J. self washes

(11) *Anton i Nina obnimajutsja*. [R, reciprocal]
A. and N. embraced-self

(12) *Dom stroitsja (plotnikami).* [R, passive]
house build-self carpenters

(13) *Sobaka kusaetsja*. [R, detransitiv/dekausativ]
dog bites-self

(14) *Dveri otkryvajutsja*. [R, unaccusative/middle]
doors open-self

- Junghanns (1996):
  - reflexivity marker in R blocks structural acc
  - but *się* in PL may absorb external argument (=> impersonal)

reflexivity (2)

[Rivero/Sheppard 2003]

- differences within Slavic
  - Polish *się*-passive obsolete [Siewierska 1988]

(15) *Dom szybko się zbudował.*
house quickly refl built

R, Cz *se*-passive fully productive; SLN, B modal sem.

(16) *Ta kniha se mi četla dobře.*
this book refl me-dat read well

  - but *se*-anticausatives never modal in meaning:

(17) *Jankowi złamały się okulary.*
J. broke self glasses

  - differences in lexical reflexives
    - *bać się [all] - fear, ptát se [CZ]/spрašivat [R] - ask, …*

reflexivity (3)

[Rivero/Sheppard 2003]

- Intransitive impersonal in all of Slavic, transitive impersonal only in PL, UKR, SLN

(17) *Tancovalo se až do rána* [CZ, impersonal, intr. V]
danced refl up until morning
'There was dancing until the morning.'

(18) *Tę książkę się cztya z przyjemnoscią.* [PL, impersonal, tr. V]
this book-acc refl reads with pleasure
'People read this book with pleasure.'

Rivero/Sheppard: reflexive *nom* (17) vs. *implicit* (16)

- arg. structure/syntax mapping

(19) *Jankowi pracowało się dobrze*. (J.-dat worked refl badly) [PL]

sem.: topic	indef.(agens)
synt.: adjunct	subject

reflexivity (4)

[Rivero/Sheppard 2003]

- impersonals (cont'd)
  - evidence e.g.

(20) *Teraz się myśli tylko o sobie*. [PL] **Mluvilo se tam jen o sobě*. [CZ]
now self thinks only about self talked refl there only about self

  - semantic properties: human, quantificational

(21) *Jeśli się gra źle, się przegrywa*. [PL]
if refl plays badly refl loses

reflexivity: agenda

- R reflexive marker becomes a verbal enclitic in the 17th c. (Isačenko 1983)

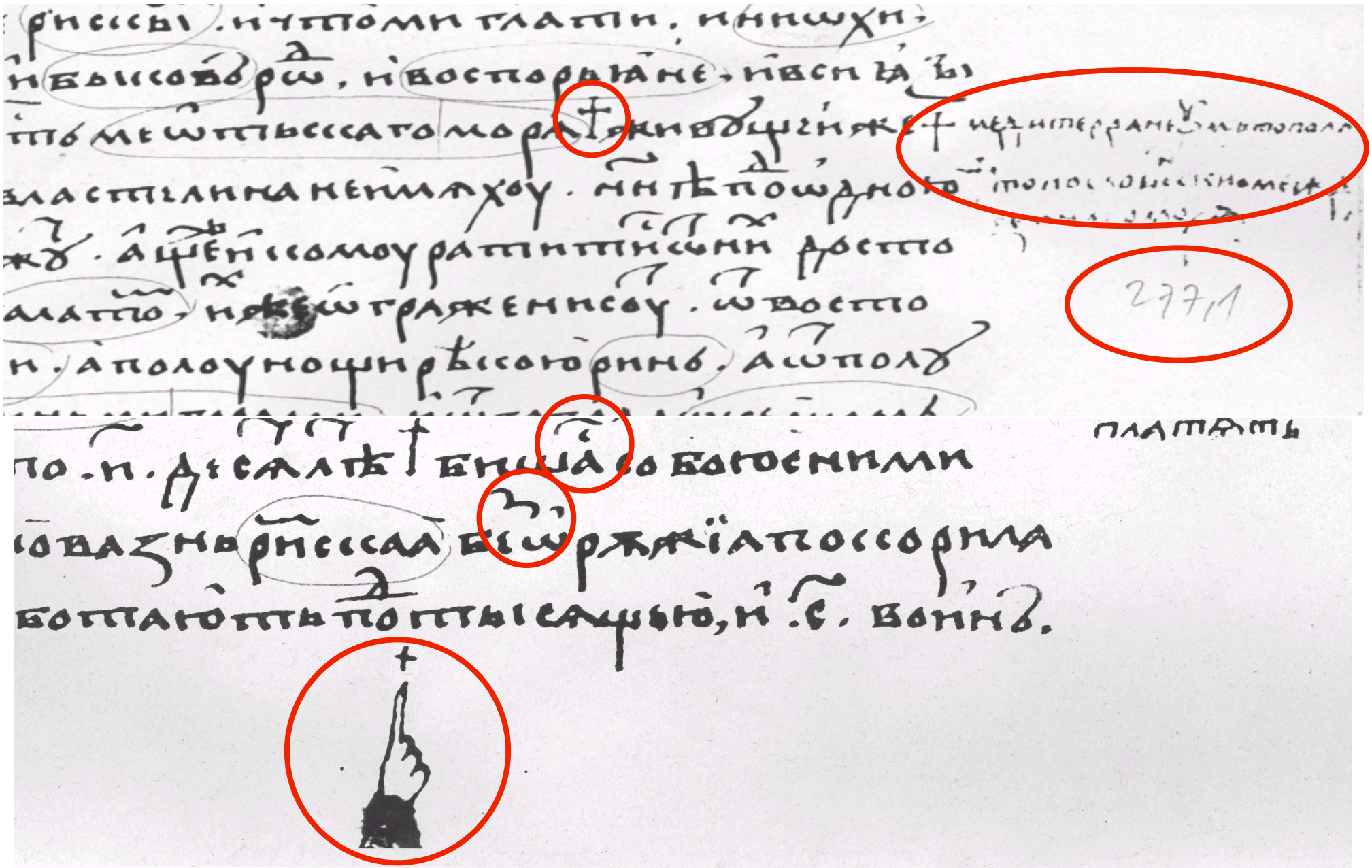
(22) *a kto sja ostalъ v gorodě* [Hypatius' chronicle 1185/1370]
and who self remained in city

- how did the variation come about?
  - transitive (PL) vs. intransitive (PL, CZ, R) impersonal
  - reflexive passive (R/CZ, ~PL, UKR?)
  - aside: lexical aspects
- cf. periphrastic passive (constant rate effect?)
- easy to search, much already done

Plan

- [x] Argument structure in diachrony
- [ ] Regensburg Diachronic Corpus
- [ ] Other diachronic corpora of Slavic languages
- [ ] First results
- [ ] Perspectives and Conclusion

What information should be encoded?



Specifics of diachronic corpora

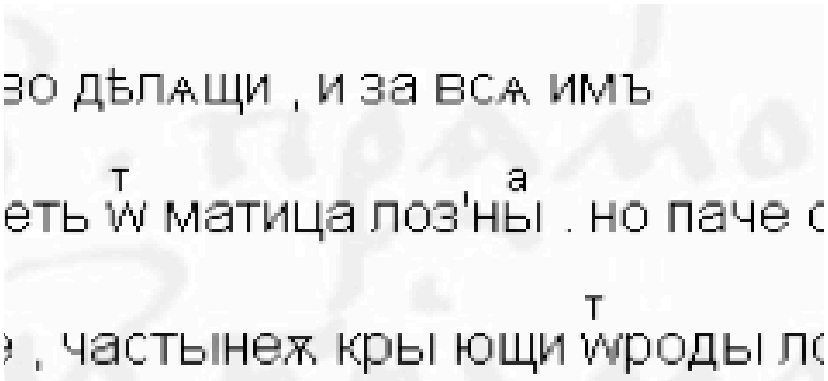
- Editorial precision
- Document preservation vs. computation
- Complicated annotation scheme
- Diachronic development vs. single annotation
  - hyperlemmas, “hyper”tags
- relatively small amounts of text
  - balancing, statistical methods?

Regensburg Diachronic Corpus

- Goal: Corpus for linguistic research on diachronic development of Slavic languages, specifically Russian
- rich structural annotation, ACT-XML & TEI-XML
- only inevitable editorial work / mainly using standard editions
- “opportunistic” corpus (now), but design schema worked out
- whole texts integrated where possible, balanced corpus for querying (prospective)

Diachronic Corpus

- concordancer
- rendering in html
- exact sources
- rich annotation
  - graphic level; graphemic level; regularizations
  - lemmatization, hyperlemmas, tagging
  - editorial remarks (mistakes/corrections/additions/ gaps/alternatives ...)
  - document structure (lines, pages, ...) vs. linguistic units (sentences, alternative word boundaries)



Issues of encoding

- 1st version CWB-based
  - fast
  - problems with Unicode
  - clumsy encoding into one table

tvar	grafémy-1	grafémy-2	morph-1	morph-2	editováno
града	*града*				
	[...]				
снвъ	*сыновъ*				*с(ы)н(о)въ*
е	*5*				
оуби^{л}	*уби*	*убил*	aor	l-part	*уби<л>*

TEI-P5

града Модеина, поемъ от своих с<u>н</u>въ е с мѣчи оуби[л]

<line n="27"> *града Модеина, поемъ от своих*
*с***<expan>***ы***</expan>***н***<expan>***о***</expan>***въ*** <expan abbr="***е***">5</expan>**
с мѣчи <choice><orig>*оуби*<add hand="later"><hi**
rend="superscript">*л***</hi></add></orig>**
<reg type="without_gloss">*оуби***</reg>**
<reg type="with_gloss">*оубил***</reg>**
</choice></line>

ACT-XML (1)

```
<originalform page="1" positioninrow="1" row="1">
  <text>свъ</text>
  <renderedform variantnumber="1">
    <text>сыновъ</text>
    <complex complex_group_refid="1"/>
  </renderedform>
</originalform>
<originalform page="1" positioninrow="2" row="1">
  <text>е</text>
  <renderedform variantnumber="1">
    <text>е</text>
    <complex complex_group_refid="1"/>
    <morphology keyword_refid= "1"/>
  </renderedform>
</originalform>

    ...
```

ACT-XML (2)

```
<complex_types>
  <complex_type name="textdiv" refid="1"/>
</complex_types>

<complex_groups>
    <complex_group complex_type_refid="1" note="chapter_1"
    refid="1"/>
</complex_groups>

<keywords>
  <keyword lemma="5" refid="1"/>
</keywords>

<pages>
  <page page="1"/>
</pages>
```

ACT -> TEI-lite

```
<pb n="1"/>
<lb n="1"/>
<seg id="1.1.1">
  <orig>свъ</orig>
  <reg ana="1" n="1">сыновъ</reg>
</seg>
<seg id="1.1.2">
  <orig>е</orig>
  <reg ana="1 24" n="1">е</reg>
</seg>
                ...

<interpGrp type="complex_types">
  <interp id="1" value="textdiv"/> </interpGrp>
<interpGrp type="complex_groups">
  <interp id="1" type="1" value="chapter_1"/> </interpGrp>
<interpGrp type="keywords">
  <interp id="24" value="5"/> </interpGrp>
```

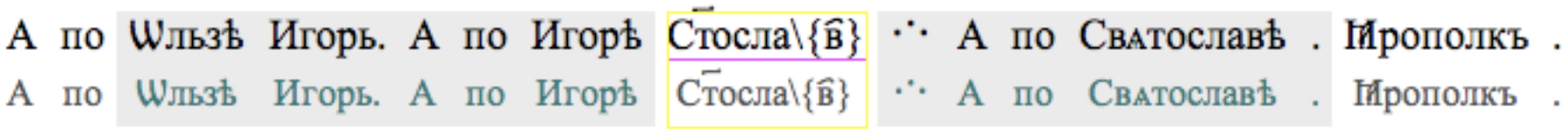
ACT

[Ribarov et al. 2004]

- designed with historical sources in mind
- flexible annotation, no concurrent hierarchies
- comfortable corpus editor (Java interface)
- XML input and output
- relational database => decent query time, easy
- PHP interface, open source
- not TEI-lite conformant
- extremely slow reading of documents into the database

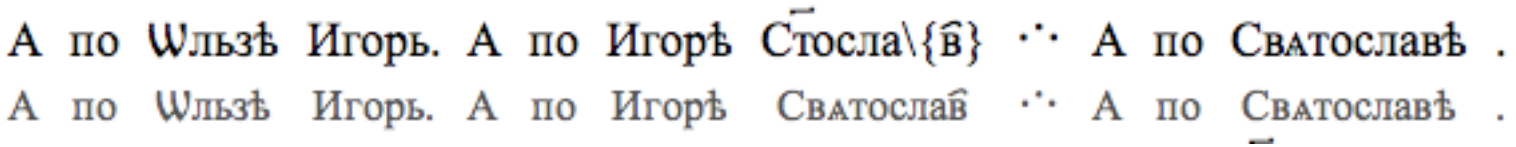
ACT: Java interface

- editor for corpus annotation (rendered forms, complexes, morphological tagging)
- rendering



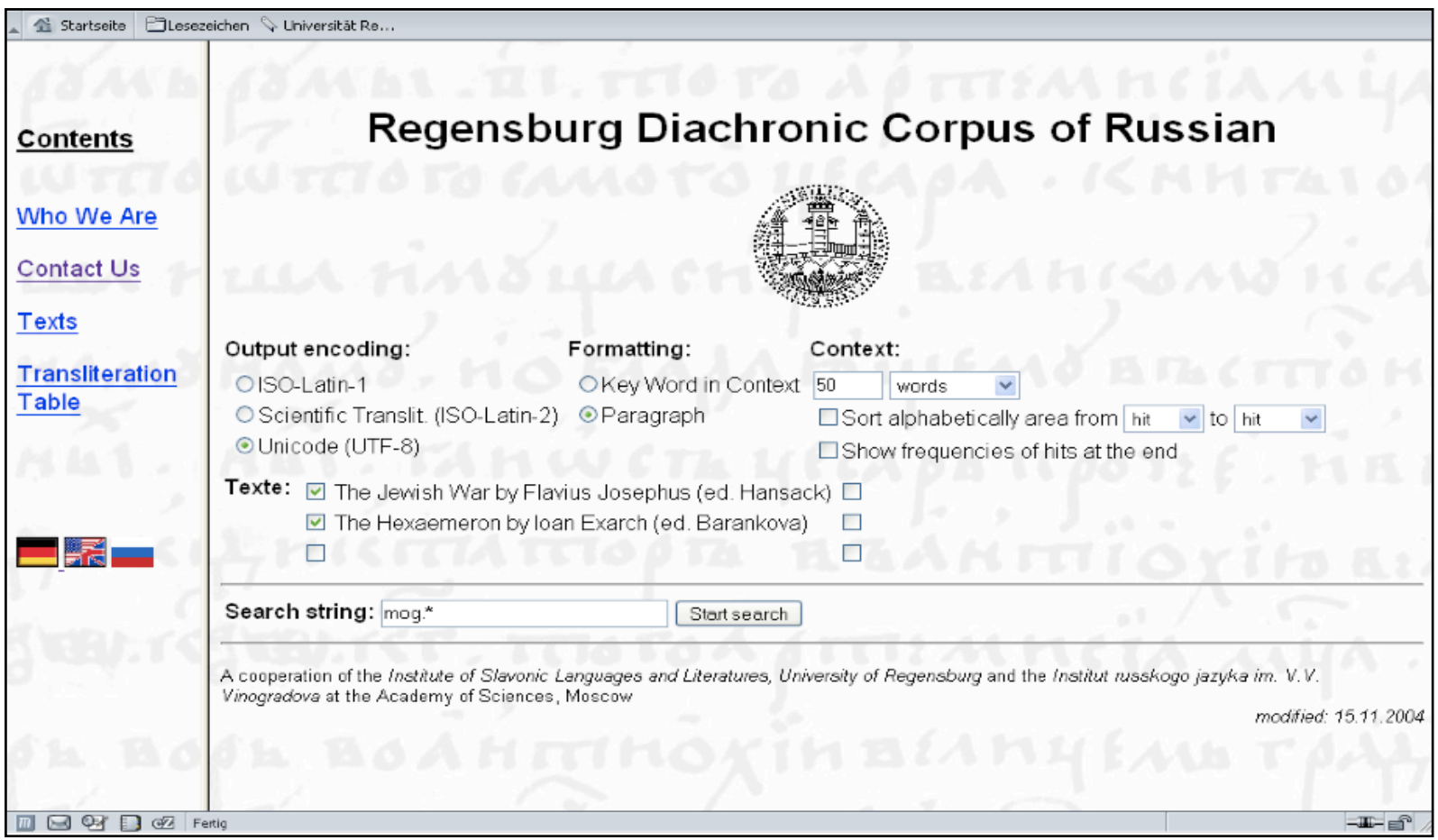
- applying regex replaces

Type	Pattern	Replace
Regular expression	([Сс])т	$1ват
Regular expression	([цЦ])\\\{ĉ}р	$1есар
Regular expression	_\\\{(.*?)\}	$1
Regular expression	(\{)?(.*?)(\})?\\\{(.*?)\}	$2$4
Exact match	#	



Query system (1st version)

- Corpus Workbench (U Stuttgart)



Query system (1st version)

Шестоднев Иоана Ексарха

Sucheingabe: mog..?t.?

Seite 88, Zeile 19, Wort 4:

6707: кде и поидѫть оубо съложенаѧ имь _ж єдинакого шьствїа не **могѫть** оустро ивъше поити . нъ на своє къждо є^с* ство

Seite 160, Zeile 2, Wort 4:

19112: оувидѣли . сии же своего не оувидѣ вше . како **могѫть** нб^с*ное , оувидѣтї . аще бо быша сами сѧ оувидѣли

Seite 173, Zeile 18, Wort 4:

21605: да сего дѣлѧ на немь же послѣд'нее wпрет'сѧ , не **могоуть** изообрѣсти . да тѣмь оуниша рещи , wбѣшена єсть на

Seite 393, Zeile 11, Wort 6:

59507: въ инацѣхь образѣхъ соуще . каци ти соуть рыбаре иже **могѫ**^т* роды ты раз'лич'ныѧ намъ начи тати . кто же можеть

Query system (2nd version)

Output

Reformatting (XSLT)

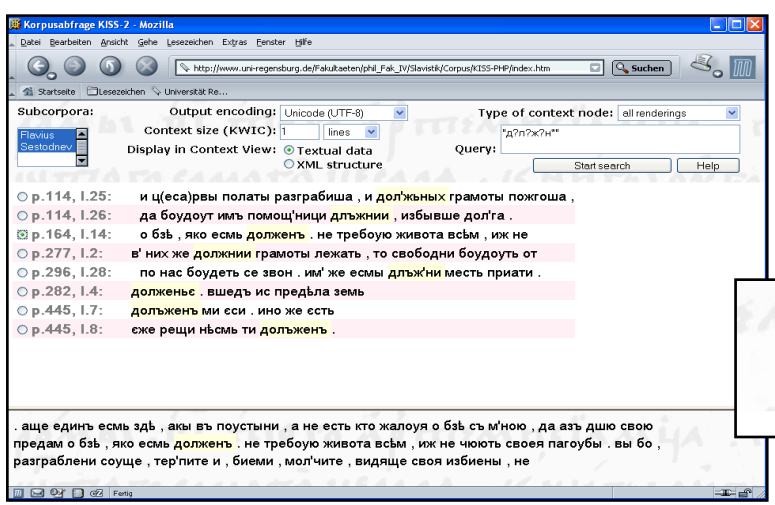
Rendering (CSS)

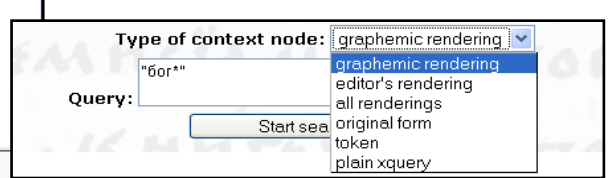
Webserver

Native XML database (eXist)

Query

Input (XML)



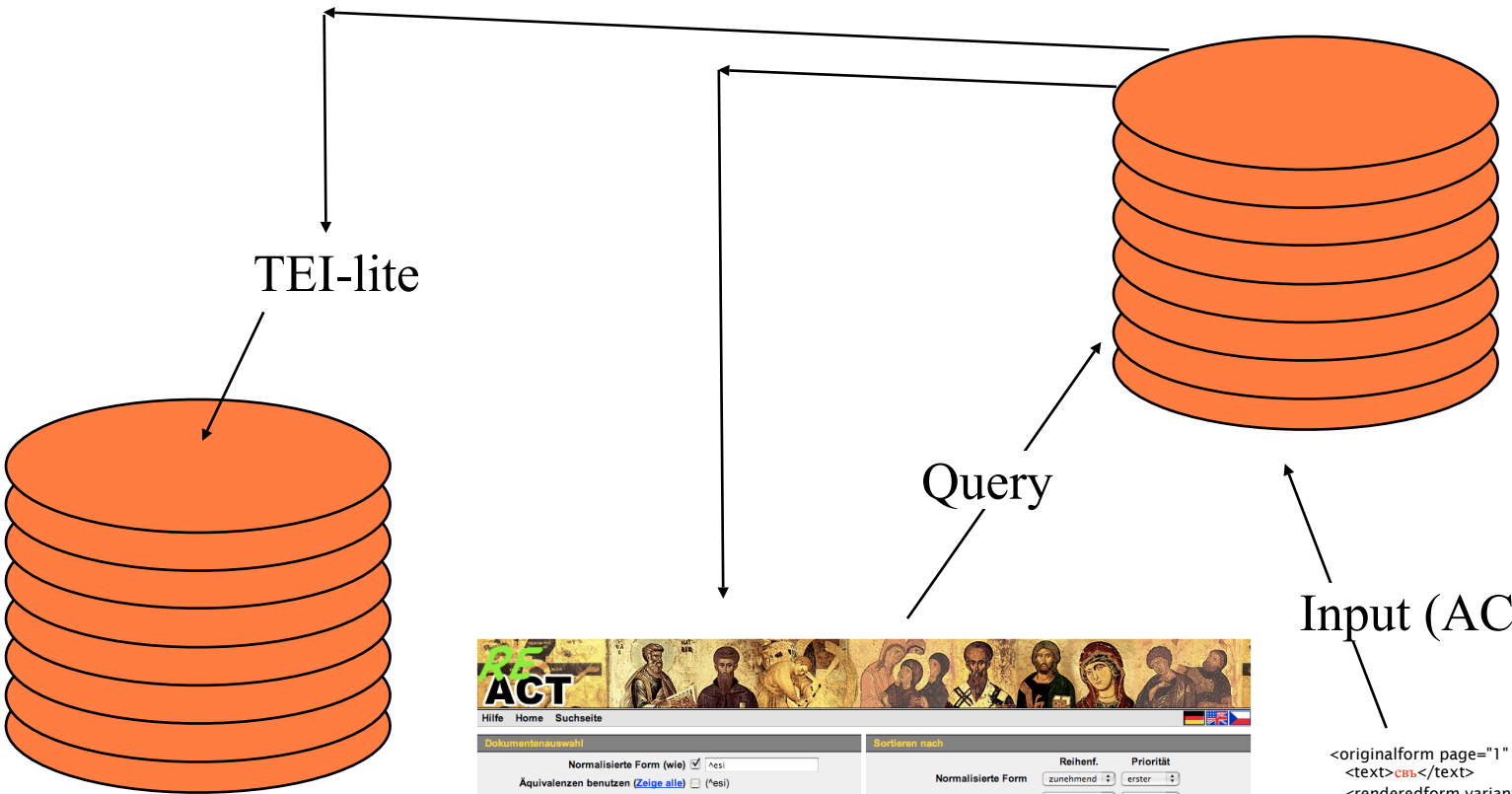



```
<seg type="token" id="1.27.1" n="226"><orig>града</orig><reg n="1"
                type="graphemic">града</reg></seg>
      [...]
<seg type="token" id="1.27.7" n="232"><orig> сцвъ </orig>
      <reg n="1" type="graphemic">сыновъ</reg></seg>
      <reg n="2" type="edited">с(ы)н(о)въ</reg></seg>
<seg type="token" id="1.27.8" n="233"><orig>е</orig><reg n="1"
          type="graphemic">5</reg></seg>
 <seg type="token" id="1.27.9" n="234"><orig>оуби<add
          hand="gloss"><hi rend="superscript">л</hi></add></orig>
      <reg n="1" type="graphemic" ana="#aor">уби</reg>
      <reg n="2" type="graphemic" ana="#lpart">убил</reg>
      <reg n="3" type="edited">уби&lt;л&gt;</reg>
</seg>
```

Present stage

- PHP interface improved
- texts: 448288 oforms (corpus positions)
  - *Iudeiskaja Vojna* (Flavius, ed. Hansack) - 15th cent.
  - Hexaemeron Ioanna Exarcha (*Šestodnev*) - 15th cent.
  - Nestor's chronicle (*Ipat'evskaja letopis'*) - 14th cent.
  - Laurentius chronicle - 14th century in prep.
- conversions in place
  - txt->ACT-XML, html->ACT-XML, ACT-XML->TEI-lite
- rendering under way
- tests with xaira, teiPublisher

Present stage (2)




```
<originalform page="1" positioninrow="1" row="1">
  <text>свъ</text>
  <renderedform variantnumber="1">
    <text>сыновъ</text>
    <complex complex_group_refid="1"/>
  </renderedform>
</originalform>
<originalform page="1" positioninrow="2" row="1">
  <text>е</text>
  <renderedform variantnumber="1">
    <text>е</text>
    <complex complex_group_refid="1"/>
    <morphology keyword_refid= "1"/>
  </renderedform>
</originalform>

  ...
```

Dokumentenauswahl

Normalisierte Form (wie) ☑ ^esi
Äquivalenzen benutzen (Zeige alle) ☐ (^esi)
Lemma ☐
Wortart ☐
Anzeige ◉ Originalformen ○ Normalisierte Formen

Sortieren nach

	Reihenf.	Priorität
Normalisierte Form	zunehmend	erster
Erstes Auftreten	nein	keine
Lemma	nein	keine

Zurücksetzen | Auswahlkriterien anwenden | Konkordanz der Treffer anzeigen

Die Suche ist beschränkt auf die Dokumente: sestodnev, flavius, Nestor

Items pro Seite: 50 Font: Litopys New Roman Größe: +0 OK

Normalisierte Form	Erstes Auftreten	Lemma	Index	Wortart	Morphologie	#	Detail	Kontext
еси	Sestodnev (1848)					295		
еси)	Nestorchronik, Ipat. (226558)					1		
е͡сѝ	Nestorchronik, Ipat. (204890)					1		

Universität Regensburg
Slavistik/Korpuslinguistik

Hilfe Home Suchseite

Die Suche ist beschränkt auf die Dokumente: sestodnev, flavius, Nestor

Seite: Seite zurück ■ 1 2 3 ■ 4 ■ 5 6 ■ Seite vor Gehe zu Seite: 4 Anzeigestil: Keyword In Context Items pro Seite: 50 Font: Litopys New Roman Größe: +0 OK

Quelle	linker Teil	Form	rechter Teil
Nestor S. 422, Z. 1	ать вси видимъ по мѣсту што ны . Бъ҃ ꙗвить . Изаславъ же рч҃е брату своему Ростиславу . ты ми	еси	брате много понуживалъ . ꙗкоже положити чс҃ть на стръи своемь и на ѿц҃и своемь се же нн҃ы Бъ҃ привелъ мѧ
Nestor S. 429, Z. 1	ищю . а Бъ҃ ти помоглъ . а ты же Киевъ собѣ . и ѣще надъ тѣмь Пересопницю и Дорогобужь	еси	оу мене ѿꙗлъ а ты мѧ тако переѡбидилъ а мнѣ еси Вышегородъ ѡдинъ далъ . ꙗ же того
Nestor S. 429, Z. 1	и ѣще надъ тѣмь Пересопницю и Дорогобужь еси оу мене ѿꙗлъ а ты мѧ тако переѡбидилъ а мнѣ	еси	Вышегородъ ѡдинъ далъ . ꙗ же того всего не правилъ . Рускыꙗ дѣлѧ землѧ . и хрс҃тьꙗнъ дѣлѧ и ѣще
Nestor S. 430, Z. 1	. а вы мене не слушаета тоти ни мнѣ еста не оуправила еже рекша но Бг҃ви и то ми	еси	молвилъ . противоу моложьшему не могу сѧ поклонити се же Изаславъ аче и двоича ступилъ слова своего се же
Nestor S. 430, Z. 1	на мнѣ положилъ и в Киевѣ мѧ посадилъ и ѿц҃мь мѧ назвалъ . а ꙗ # его сн҃омь . дажь	еси	реклъ моложьшему сѧ не поклоню . да се азъ тебе старѣи . есмь . не маломь но многомь . азъ

Plan

- [x] Argument structure in diachrony
- [x] Regensburg Diachronic Corpus
- [ ] Other diachronic corpora of Slavic languages
- [ ] First results
- [ ] Perspectives and Conclusion

Other projects (1)

- Diachronic part of ČNK (Kučera 2002, 2004):
  - maximal authenticity, ≤ 20 y around original
  - "standardized" Old Czech orthography
  - poor annotation; close to synchronic part
  - radical balancing approach: whole documents only 1250-1500, afterwards random samples
- Deutsch Diachron Digital (Dipper et al. 2004):
  - versatile corpus for all disciplines
  - grapheme-based XML annotation, stand-off
  - own object-oriented database (planned)

Other projects (2)

- Penn-Helsinki Parsed Corpus of Middle English
- TITUS Database
- Izhevsk project
  - several, partly lemmatized Old Russian texts on-line
  - (proprietory) relational database and query software
- ACT (Ribarov et al. 2004): mainly tool
- PAN/Pracownia języka staropolskiego: Korpus tekstów staropolskich (do 1500 r.)

Plan

- [x] Argument structure in diachrony
- [x] Regensburg Diachronic Corpus
- [x] Other diachronic corpora of Slavic languages
- [ ] First results
- [ ] Perspectives and Conclusion

First results

- auxiliary in R past tense and overt pronouns
- rise in use of 2nd person (ty + aux/copula)
  - Flavius:

	+pron	-pron
auxiliary	2	52
copula	4	8

  - Nestor's chronicle:

	+pron	-pron
auxiliary	35	139
copula	21	18

Examples

- ty + copula

  (23) *otce, ty esi uže staръ*
  father you are already old

. и на Василка Дюргевича . и ѡкоже стаста на Лтѣ . и Изаславъ Вячеславу . ѡц҃е # ты	рѣ҃ч҃ еси	оуже старъ а тобѣ не достоить трудитиса . а со мною пусти .

- ty + auxiliary

  (24) *a ty sja esi ešče s ljudmi Kievě ne outverdilъ*
  and you self aux already with people in-Kiev not confirmed

же боранахуть . ему поити Чернигову . рекучи ему се Бъ҃ поѡлъ. строѧ твоего Вячеслава . а ты са	еси	еще с людми Киевѣ не оутвердилъ . людми оутвердиса . да

Plan

- [x] Argument structure in diachrony
- [x] Regensburg Diachronic Corpus
- [x] Other diachronic corpora of Slavic languages
- [x] First results
- [ ] Perspectives and Conclusion

Summary/Conclusion

- first argument realisation in diachrony
- to a large extent word-form based
- regularization/rendition
- problems with XML (large documents, query)
- need for statistical data
- linguistic analysis vs. editorial work
- more data, sampling, more languages