Computer Processing of Written Cultural Heritage Sources

Jiří Bubník, Jiří Čelák, Vojtěch Janota, Alexandr Kára, Václav Novák, Kiril Ribarov, Tomáš Vondra

INFORUM 2004

We will focus on:

Lexical and corpus processing of written cultural sources

We place the written cultural sources in an electronic contextual (e-context) field with the following connecting elements:

- source image along with language based contextual structure of the word mass present in the sources

- connections (inner and outer links) among various types of written cultural sources within wider cultural environment.

We aim at:

Development of technologies and tools

for

large-scale activities

aimed towards

multi-aspectual presentation

of

written cultural heritage

in a

highly distributed manner

Our approach is:

Experimental

↓

development of various tools
and their testing and modification
based on user-specialist input.

↓

Collecting experience
from past models and experiments

Developed as individual work and recently also as student project at the Center for Computational Linguistics, Charles University.

Corpus Processing

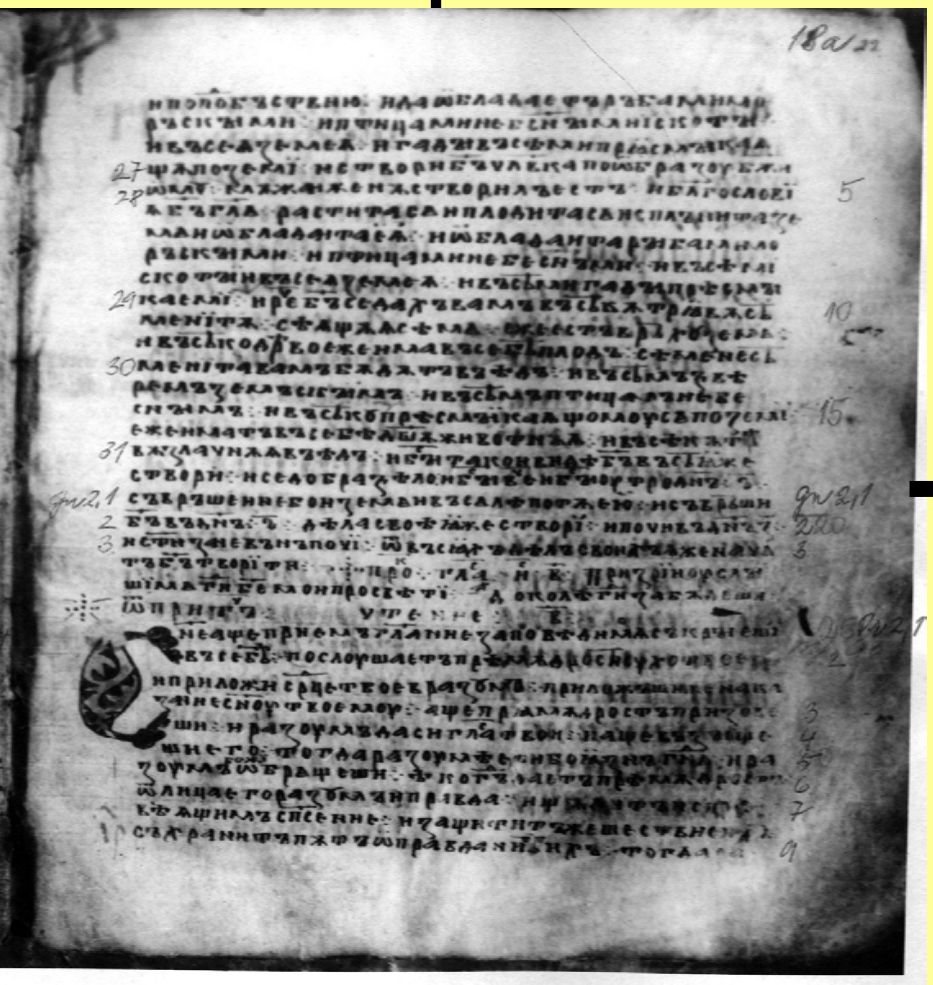
- Developed for contemporary languages
- Existence of various corpus managers

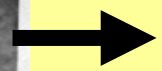
but

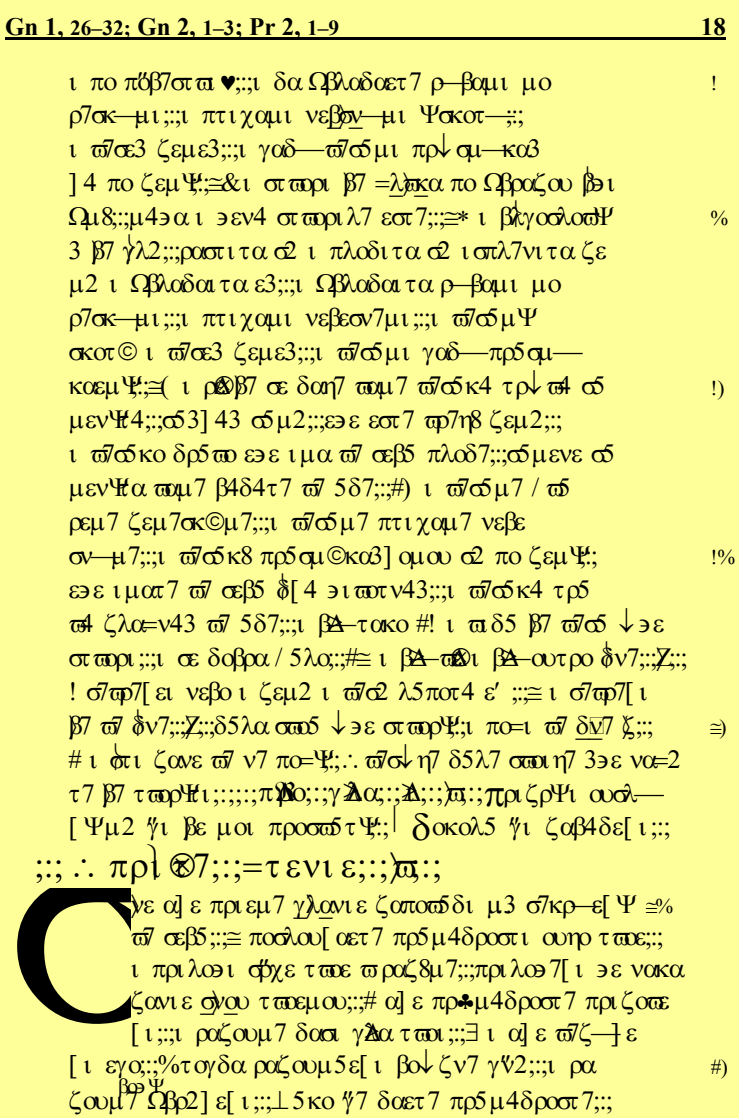
None of them covers specialties present in processing of old sources (dead languages).

We need special approaches,
but whenever possible,
we prefer the use existent modules.

The input







ι πο πόβ7στπι ♥;:;ι δα Ωβλοδοετ7 ρ—βομι μο !
ρ7σκ—μι;:;ι πτιχομι νεβ̲ν̲—μι Ψσκοτ—;:;
ι πσ7σε3 ζεμε3;:;ι γοδ—πσ7σδμι πρ↓σμ—κσ3
]4 πο ζεμΨ;:≅&ι στπορι β7 =λ̲σκ̲α πο Ωβροζου β̲ι
Ωμι8;:;μ4ε α ι ε εν4 στπορι λ7 εστ7;:;≅* ι β̲γοσλοπΨ %
3 β7 γ̲λ2;:;ροστιτα σ2 ι πλοδιτα σ2 ισπλ7νιτα ζε
μ2 ι Ωβλοδαιτα ε3;:;ι Ωβλοδαιτα ρ—βομι μο
ρ7σκ—μι;:;ι πτιχομι νεβεσν7μι;:;ι πσ7σδμΨ
σκοτ© ι πσ7σε3 ζεμε3;:;ι πσ7σδμι γοδ—πρδσμι—
κοεμΨ;:≅(ι ρ⊗β7 σε δοη7 πσμ7 πσ7σδκ4 τρ↓πσ4 σδ !)
μενΨ4;:;σδ3]43 σδμ2;:;εσε εστ7 πρ7η8 ζεμ2;:;
ι πσ7σδκο δρδπσ εσε ιμα πσ7 σεβ5 πλοδ7;:;σδμενε σδ
μενΨα πσμ7 β4δ4τ7 πσ7 5δ7;:;#) ι πσ7σδμ7 / σδ
ρεμ7 ζεμ7σκ©μ7;:;ι πσ7σδμ7 πτιχομ7 νεβε
σν—μ7;:;ι πσ7σδκ8 πρδσμ©κσ3]ομου σ2 πο ζεμΨ;; !%
εσε ιμοτ7 πσ7 σεβ5 $[4 ειπστν43;:;ι πσ7σδκ4 τρδ
πσ4 ζλοε=ν43 πσ7 5δ7;:;ι β̲Α—τσκο #! ι πιδ5 β7 πσ7σδ ↓σε
στπορι;:;ι σε δοβρα / 5λσ;:;#≅ ι β̲Α—πσ⊗ι β̲Α—ουτρο $ν7;:;Ζ;:;
! σ7πρ7[ει νεβο ι ζεμ2 ι πσ7σ2 λ5ποτ4 ε′ ;:;≅ ι σ7πρ7[ι
β7 πσ7 $ν7;:;Ζ;:;δ5λα στπσδ ↓σε στπορΨ;:ι πο=ι πσ7 δ̲ν̲7 ζ̲;:; ≅)
ι φτι ζονε πσ7 ν7 πο=Ψ;:∴ πσ7σ↓η7 δ5λ7 στπσιη7 3σε νο=2
τ7 β7 τπορΨι;:;:;:;:;πβ⊗ο;:;γ̲Α̲α;:;Α̲;:;}π;:;πριζρΨι ουσλ—
[Ψμ2 ψι βε μοι προστπδτΨ;:⌋ δοκολ5 ψι ζοβ4δε[ι;:;
;:;∴ πρ⌋⊗7;:;=τενιε;:;}π;:;
Cνε σ]ε πριεμ7 γλ̲ονιε ζοπστπδδι μ3 σ7κρ—ε[Ψ ≅%
πσ7 σεβ5;:;≅ ποσλου[οετ7 πρδμ4δροστι ουπρ τπτε;:;
ι πριλοει φχε τπτε πρσζ8μ7;:;πριλοε7[ι σε νοκα
ζονιε σ̲νου τπτεμου;:;# σ]ε πρ♣μ4δροστ7 πριζοπτε
[ι;:;ι ροζουμ7 δοστι γ̲Α̲α τπτοι;:;∃ ι σ]ε πσ7ζ—]ε
[ι εγσ;:;%τσγδα ροζουμ5ε[ι βο↓ζν7 γψ2;:;ι ρα #) INFORUM 2004
βσ̲ Ψ
ζουμ7 Ωβρ2]ε[ι;:;⊥5κο ψ7 δοετ7 πρδμ4δροστ7;:;
∴ λιχα εγο ροζ8μ7 ι προπδσ;:;&ι]4διτ7 ισπρα

Original forms:

,!*,

ΞGn, ≅⊥ ι πο πόβ7στπι ♥;:;ι δα Ωβλοδοετ7 ρ—βομι μο?
ρ7σκ—μι;:;ι πτιχομι νεβ̲ν̲—μι Ψσκοτ—;:;
ι πσ7σε3 ζεμε3;:;ι γοδ—πσ7σδμι πρ↓σμ—κσ3?
]4 πο ζεμΨ;:≅&ι στπορι β7 =λ̲σκ̲α πο Ωβροζ̲ου β̲ι
Ωμι8;:;μ4ε α ι ε εν4 στπορι λ7 εστ7;:;≅* ι β̲γοσλοπΨ
3 β7 γ̲λ2;:;ροστιτα σ2 ι πλοδιτα σ2 ισπλ7νιτα ζε?
μ2 ι Ωβλοδαιτα ε3;:;ι Ωβλοδαιτα ρ—βομι μο?
ρ7σκ—μι;:;ι πτιχομι νεβεσν7μι;:;ι πσ7σδμΨ
σκοτ© ι πσ7σε3 ζεμε3;:;ι πσ7σδμι γοδ—πρδσμι—?
κοεμΨ;:≅(ι ρ⊗β7 σε δοη7 πσμ7 πσ7σδκ4 τρ↓πσ4 σδ

.....

,!*,

ΞGn, ≅⊥ι πο πΦβ7στω♥;:;ι δα Ωβλοδστ7 ρ—βομι μο?
ρ7σκ—μι;:;ι πτιχομι νεβ̲σ̲ν—μι Ψσκοτ—;:;
ι ϖ7σ3 ζεμε3;:;ι γοδ—ϖ7σδμι πρ↓ομ—κο3?
]4 πο ζεμΨ;:≅&ι στϖρι β7 =λ̲π̲κ̲α πο Ωβρ̲α̲ζ̲ου β̲ι
Ωμ8;:;μ4ε α ι ε εν4 στϖρι λ7 εστ7;:;≅* ι β̲γ̲ο̲λ̲οϖΨ
3 β7 γλ2;:;ρ̲ο̲σ̲τ̲ι̲τ̲α̲σ2 ι πλοδιτασ2 ισπλ7νιτα ζε?
μ2 ι Ωβλοδαιτα ε3;:;ι Ωβλοδαιτα ρ—βομι μο?
ρ7σκ—μι;:;ι πτιχομι νεβεσν7μι;:;ι ϖ7σδμΨ
σκοτ© ι ϖ7σ3 ζεμε3;:;ι ϖ7σδμι γοδ—πρδομ—?
κοεμΨ;:≅(ι ρΦβ7 σε δοη7 ϖομ7 ϖ7σδκ4 τρ↓ϖ4 σδ

18r1,3:

πο<δο>β7στωιφο

18r4,5: β<ογ>7

18r5,7:

βλ<α>γοσλοωι?3

.....

18r10,2: ρε<=ε>

Original forms: → **Rendered** forms:

Multiple meanings, Multiple Understanding, Multiple.....

ι||εγοɜε||ϖιδι[ι||πλοδα||σε||σ7τϖορι||ϖ7||μν5
(*and the fruit you see created in me*),

2 variants!

the string σ7τϖορι||ϖ7||μν5

could also be divided as σ7τϖοριϖ7||μν5

(where σ7τϖοριϖ7 is the past participle - active mood of *create*)

both are grammatically correct

...multiple...multiple...multiple...

o_f_1 o_f_2 . . . o_f_{i-1} o_f_i o_f_{i+1} . . . o_f_n

Complex used also for translation

$r\_f_i^{1,1}$ C $r\_f_i^{1,M}$

$r\_f_i^{I,J}$

$r\_f_i^{N,1}$ $r\_f_i^{N,M}$

Part of:
1. Context
2. Multiple Complexes
3. Correlations to other sources

$_1m\_t_i^{I,J}$ ... $_Km\_t_i^{I,J}$

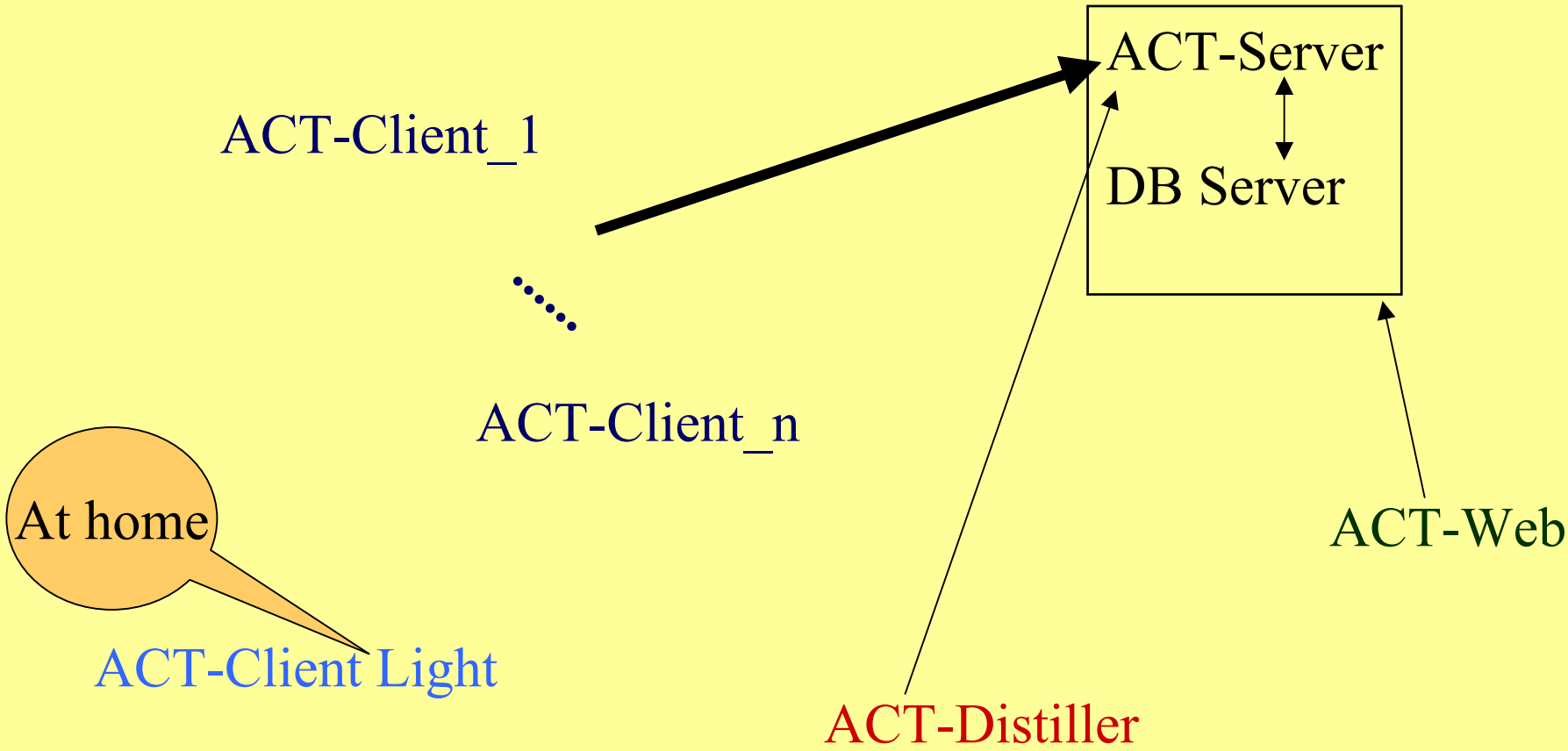
POS

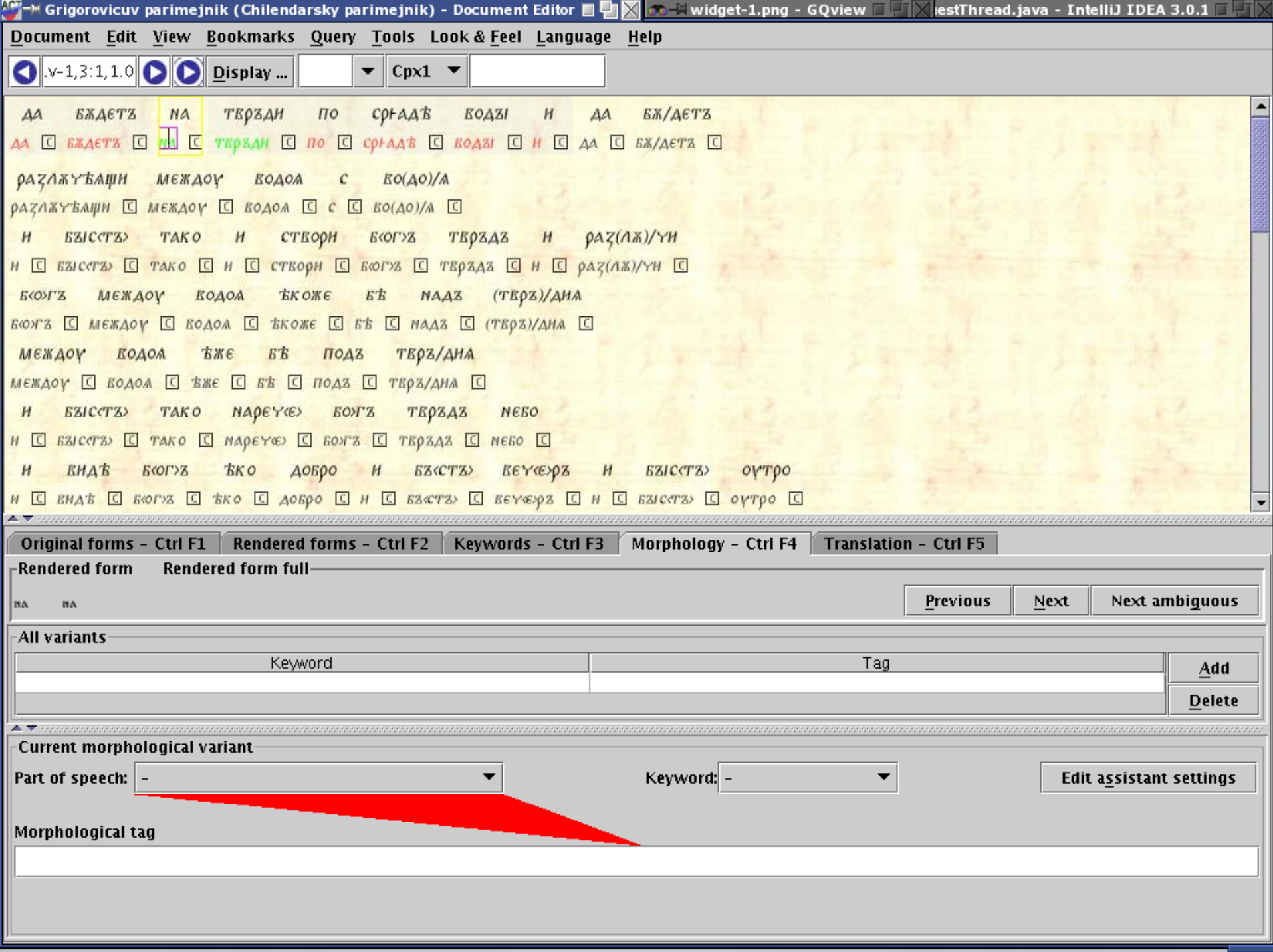
$_1h\_s(IR)_i^{I,J}$ ... $_Lh\_s(IR)_i^{I,J}$

$_1h\_s(R_1)_i^{I,J}$

$_1h\_s(R_R)_i^{I,J}$

ACT Structure




Grigorovicuv parimejnik (Chilendarsky parimejnik) - Document Editor
widget-1.png - GQview
estThread.java - IntelliJ IDEA 3.0.1
Document Edit View Bookmarks Query Tools Look & Feel Language Help
.v-1,3:1,1.0
Display ...
Cpx1
Original forms - Ctrl F1
Rendered forms - Ctrl F2
Keywords - Ctrl F3
Morphology - Ctrl F4
Translation - Ctrl F5
Rendered form
Rendered form full
Previous
Next
Next ambiguous
All variants
Keyword
Tag
Add
Delete
Current morphological variant
Part of speech:
Keyword:
Edit assistant settings
Morphological tag

Inputs and Outputs

- **Input:** RTF, XML, TXT
- **Output** (for any subset of documents) HTML, XML, RTF, TXT:
  - various search outputs
  - complete index (index verborum)
  - concordance index
  - word frequences
  - bigrams
  - retrograde indexes
  - etc...

Documents in ACT

Catalogue: language, sorting, morphology

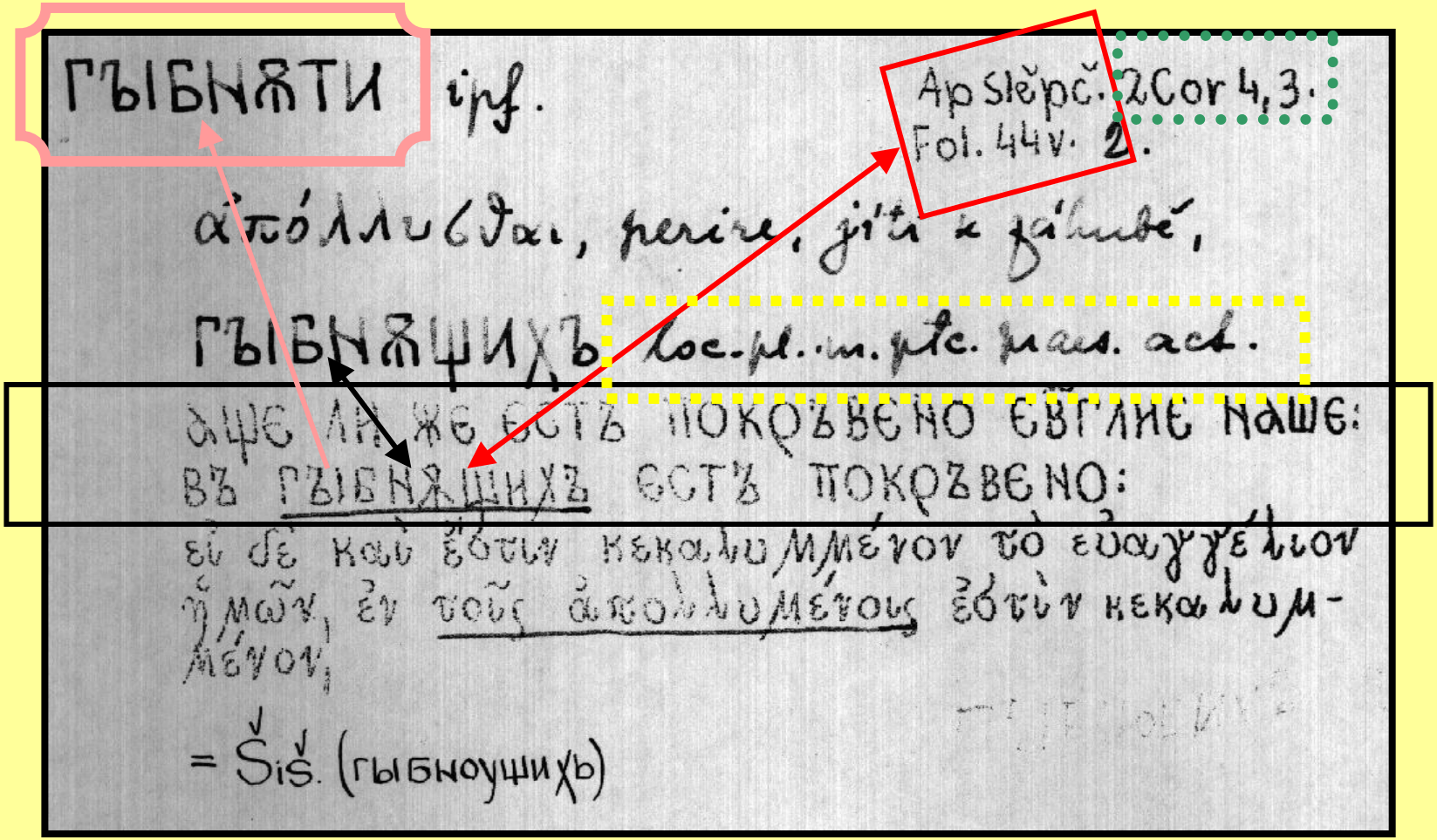
Text: ***Document*** $_1$ ***..... Document*** $_n$

Pictures: ***Document*** $_1$ ***..... Document*** $_n$

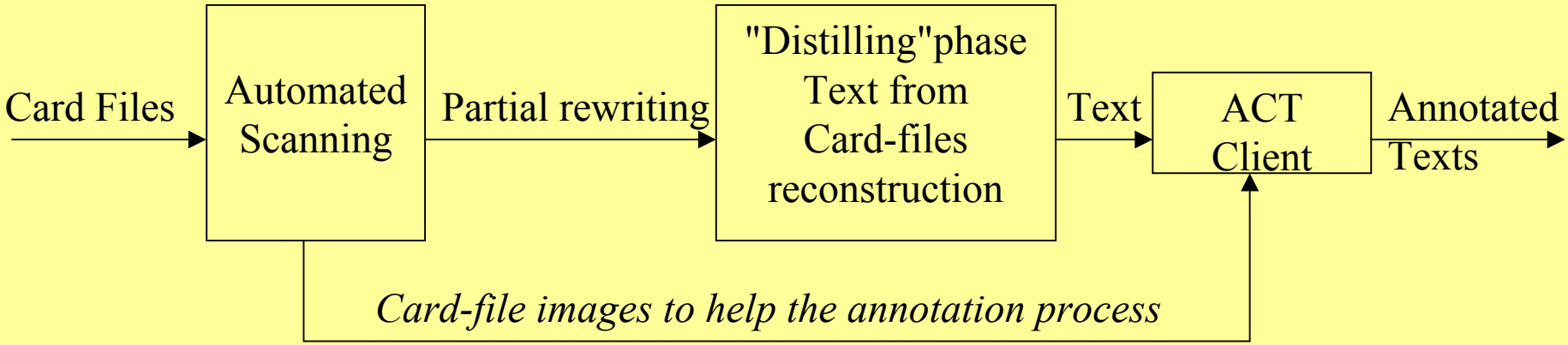
Pictures: ***Card File*** $_1$ ***..... Card File*** $_{15\,000\,000}$

Location+ Oform --> guess context

A single card




ACT


Card Files
Automated Scanning
Partial rewriting
"Distilling"phase Text from Card-files reconstruction
Text
ACT Client
Annotated Texts
Card-file images to help the annotation process

Search criteria

Rendered form (like) ☐

Use equivalences (show all) ☐

Lemma ☑ v%

Part of speech ☐

Index ☐

Morphology (_ is a wildcard) ☑ S _ m _ _ _ _ _ _ _ _ _ _ _

morphology help

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Order by

	order	priority
Rendered form	ascending	first
First occurence	no	none
Lemma	no	none
Index	no	none
Part of speech	no	none
Morphology	no	none

Reset Search

Search is restricted to the following documents: Tr, Orb, VK, Mkd, Lesn, Grig, GreekTrans, Hlud, BT, ST, KR, Bon

page: 1 ▪ 2 3 4 5 6 7 8 9 10 11 ▪ Next page

go to page: 1 items per page: 50 font: Kliment size: +0 OK

Rendered form	First occurence	Lemma	Index	Part of speech	Morphology	#	Detail	Context
ѢАВИЛА	ST (1577)	ѢАВИЛА	o	S	S-m----------	1		
ѢАВИЛО	ST (1652)	ѢАВИЛА	o	S	S-m----------	1		
ѢАВИЛЫ	ST (1566)	ѢАВИЛА	o	S	S-m----------	3		
ѢАВИЛОНЬ	ST (27992)	ѢАВИЛОНЪ	o	S	S-m----------	3		
(ѢАВИЛОНА	ST (30806)	ѢАВИЛОНЪ	o	S	S-m----------	1		
ѢАВИЛОНОР	ST (30791)	ѢАВИЛОНЪ	o	S	S-m----------	1		
ѢАВИ/ЛОНѢ	ST (30781)	ѢАВИЛОНЪ	o	S	S-m----------	1		
ѢАВИ/ЛА	Orbelsky triod (8486)	ѢАВѴЛА		S	S-m----------	1		
ѢАВИЛОНѢ	BT (9138)	ѢАВѴЛОНЪ		S	S-m----------	4		

Home Catalog ACT Web Guide

Search is restricted to the following documents: Tr, Orb, VK, Mkd, Lesn, Grig, GreekTrans, Hlud, BT, ST, KR, Bon

page: 1 ▪ 2 ▪ Next page

go to page: 1 items per page: 50 font: Kliment size: +0

left part	form	right part
(–д– възвеличимь вси чл<овѣк<о>любие ти х<рист>е б<ож>е на шь слава твоихь	рабь	и вѣнець вѣрнымь възвеличавы ѫтробѫ рождьшѫѫ та ра ѡт(врьсть быс(<ть>
вза(ѣжѫще чистоѫ м<о>л(<и>твы слаще впиемь х<рист>у при/ми м<о>л(бы сп<а>с(е твоихь	рабь	и м<о>л(ениꙗ азь тебѣ са м<о>л(а м<а>ти х<ристо>ваꙗ не прѣстаи
(п)оите и прѣвьзносит+е иже подвигомь добрымь подвигьше са с<ва>ти м<ѫ>ч(<е>н<и>ци	рабь	недостоинь ко х<рист>у ходатаи не забѫдѣте да достоин течение постное
ꙗко/же г<оспо>д(<ь>нь аврамїе ѡт(вѣщав же клиросни/ци реко шѫ емоу тьи ꙗес(<т>	рабь	б<о>жїи и чрьньць сьврьшень еп<и>с(<ко>пь реч(<е> кь нимь его
(зла ньи̅ богы вьса раско/пальꙗ̅ес(<т> и немогошѫ его ѡзлобити вьи̅сти/нѫ	рабь	ꙗес(<т> б<ог>а живаго и всѣ реченѿ̓наа ѡт (него и/стинна
д<ь>нѣр г<осподь>нѣ потьщимса̅ оубо н<и>нѣ и подвигнѣмса̅ да сане̅ѡбра/щемь ꙗко	рабь	ѡнь егоже пришьд (г<осподь> его ѡбращеть питаѫща̅са и нерадаща тѣ/мжео
ꙗко ни iѡсифоу оуподо(<о>биса̅ ни бл<а>женнѣи соуса/нѣ ащ	рабь	продань̅бы̅с(<т> егуптѣнинѫ ради нѫ непокорйса̅ льще/ни

Current State

- 8 manuscripts, 700 000 forms
  - with keywords (lemma, POS, hom), correlation to other sources, translations
- Annotation tool - ACT
  - distributed annotation
  - direct web exporting
- Created conditions for application of NLP tools
- Possibility to process enormous card files