ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА – БАН

***Секция "Телекомуникации"***

---

**Милена Петрова Добрева**

# Моделиране на вариативност в старобългарски текстове

ДИСЕРТАЦИЯ

за присъждане на научна и образователна степен "Доктор"

Научен консултант:
ст.н.с. II ст. Пл. Матеев

София, 1999

# СЪДЪРЖАНИЕ

# СПИСЪК НА ФИГУРИТЕ

# СПИСЪК НА ТАБЛИЦИТЕ

# У В О Д

## Актуалност на проблема

Славянското средновековно ръкописно наследство наброява стотици хиляди ръкописи и ръкописни фрагменти, пръснати в хранилища в различни страни. Изследванията на старобългарските текстове в синхронен и диахронен аспект все още се извършват с традиционните методи без широко приложение на съвременните информационни технологии. Същевременно, използването на методи за компютърно представяне и анализ на средновековните ни текстове би позволило след постепенното натрупване на електронни корпуси от старобългарски (и в по-широк обхват – средновековни славянски) текстове да се облекчат някои рутинни задачи, свързани с извличането на първични изследователски данни за паметниците, и да се премине към съпоставителни изследвания на по-широка изворова основа.

До момента обаче не съществуват стандарти или поне широко приети системи за компютърно представяне на старобългарските текстове. Адекватното компютърно представяне на средновековните текстове е възможно след разработката на специализирани информатични модели, които да отразяват във възможно най-пълна степен особеностите на моделирания обект. Старобългарските текстове, към които досега се е подхождало с методите за представяне и анализ на текстове на съвременни езици, имат няколко такива особености: те са *нелинейни,* с изобилие от *ненормирани съкращения* и с *висока степен на вариативност* на всички езикови нива.

Разработването на информатичен модел, който отчита тези особености, ще позволи създаването и използването на електронни

ресурси в областта на старобългаристиката да се извършват в среда *ad modum* вместо чрез съществуващите досега многобройни подходи *ad hoc.* При подобен подход става възможно да се обменят текстове и изследователски данни за старобъпгарските текстове между различни колективи.

## Цели на дисертационната работа

Основната цел на настоящата работа е да спомогне за подобряване на използването на съвременните информационни технологии в работата на старобългаристите.

За осъществяването на тази цел са извършени *изследване, моделиране* и *практическа реализация* на използването на съвременните информационни технологии за представянето и анализа на старобългарски текстове в компютърен вид.

За постигането на тези цели трябваше да се решат следните задачи:

- да се изследват особеностите на старобългарската писмена традиция и да се предложи модел, който да дава възможност за адекватно компютърно представяне на текстовете;
- да се разработи модел за представяне на текстовите особености в старобългарски текстове;
- да се предложи методика за анализиране на вариативност в старобългарски текстове с помощта на компютърни средства;
- да се проведат експерименти с текстове за оценка на предложената методика;
- да се разработи мултимедиен модел на специализирано работно място на старобългариста, който да отразява

особеностите на предметната област.

## Кратко съдържание на работата и анотация на получените резултати

В Глава 1. на дисертационния труд е направен обзор на приложенията на информационните технологии в старобългаристиката и на българския опит в тази област. Показано е, че най-голям интерес сред специалистите-старобългаристи до момента предизвиква областта на компютърно представяне и анализ на текстове. Същевременно не е решен задоволително проблемът за създаване на подходящо компютърно кодиране на средновековните славянски текстове. Това води до ограничено използване на възможностите на новите информационни технологии в областта.

Глава 2. е посветена на проблемите, свързани с компютърното кодиране на старобългарски текстове. В нея се разглеждат в съпоставителен план най-разпространените компютърни кодирания на старобългарски текстове, вкл. транслитерация на латиница и използване на декларации за описание на писмени системи със средствата на SGML (Standard Generalized Mark-up Language). Посочени са особености, свързани с графичното разположение на някои по-особени елементи в старобългарските текстове, които не са заложени в съществуващите стандарти за представяне на текстове. Направено е предложение за разширение на TEI.transcr (раздела на предписанията на Text Encoding Initiative за транскрипция на първични източници в процеса на създаване на електронни версии на подобни текстове). Предложените елементи включват `<position>`, с чиято помощ могат да се маркират подписани, надписани, вписани и ротирани букви; `<ligature>`, с чиято помощ се маркира плетено писмо, `<rubric>`, с чиято помощ се отбелязва червенослов, и `<initial>`, с чиято помощ

се отбелязват инициали.

В Глава 3. е разгледан въпросът за компютърното представяне и анализ на вариативността в средновековните славянски текстове. Възможностите за компютърно-подпомогнато изследване на вариативността са от особена важност при работата със средновековни текстове. Затова в работата е отделено най-обстойно внимание именно на този проблем. Представени са статистически експерименти за изследване на правописна вариативност. При експериментите са използвани данни за честотата на употреба на всички букви от старобългарската азбука, както и на честотата на употреба на качествените признаци, които характеризират различни правописни редакции на старобългарския език. За пръв път в настоящата работа са представени количествени данни за употребата на буквите и качествените признаци за правописа в българска, руска и сръбска правописни редакции на основата на проведени експерименти с текстове от Псалтира. В изследването се представят данни за количествените характеристики на буквените употреби за текстове от българска, руска и сръбска редакции. Представено е и приложение на клъстерния анализ за изследване и илюстриране на близостта между текстове от различни редакции. Въз основа на проведените експерименти се предлага методика за по-нататъшни изследвания на правописната вариативност в средновековни славянски текстове.

В Глава 4. е представен модел на работно място на старобългариста, ориентиран към работа в средата на Интернет. Разгледани са данните, които участват при изграждането на компютърно представяне на средновековни славянски ръкописи. Представени са типичните дейности при въвеждане на данни за ръкописи, както и при анализа на въведените данни. Предимство на предложения модел е, че освен на представянето на средновековни

текстове и структурирани изследователски данни за тях, включва и все още малко популярните в областта на старобългаристиката количествени методи за изследване на данни, свързани с ръкописното наследство. Разгледани са възможни приложения на работното място в изследователската работа и при обучението на студенти.

Представеното изследване илюстрира как могат да се представят в ненормализиран вид старобългарски текстове и как те могат да се използват за получаване на количествени данни за старобългарски текстове. Това изследване може да служи като основа за постепенно натрупване на старобългарски текстове в електронна форма и да допринесе за извършването на съпоставителни изследвания върху тях. Публикуването на резултатите от тези дейности със средствата на Интернет-технологиите също е засегнато в работата.

## Практическа приложимост

Изследването на системите за кодиране и предложеното разширение на TEI с елементи, отразяващи разположението на символите в старобългарските паметници, са принос към по-доброто представяне на старобългарски текстове в компютърен вид. При подобно представяне се намалява традиционно провежданата при компютърно въвеждане на текстове нормализация и се дава възможност при бъдещо пълнотекстово търсене да се използват и такива подробности като графичното разположение на определени текстови елементи, както и участието им в украсата на паметника.

Методиката за анализ на вариативност в старобългарски текстове, приложена върху по-широк кръг от текстове, позволява да се получат нови данни за количествените характеристики на средновековните славянски писмени паметници. Тези данни биха

помогнали да се осветлят допълнително различни становища за развитието на българския език и другите славянски езици.

Предложеният модел на специализирано работно място на старобългариста може да се използва успешно при конкретни изследователски задачи и при обучението на студенти в областта на старобългаристиката. Подобен модел в областта на средновековната славистика се създава за пръв път.

# Глава 1. Обзор на приложенията на информационните технологии в областта на старобългаристиката

## 1.1. Състояние на изследванията в областта

През последните две десетилетия у нас и в чужбина расте интересът към приложенията на информатиката в работата на специалистите-старобългаристи. Първоначалните опити в тази област са свързани с приложенията на бази от данни за събиране на кодикологично описание на ръкописи [Geurts et al. 87]. Почти едновременно с това започва използването на текстови редактори за въвеждане на старобългарски текстове, като първият редактор, използван за представяне на текстовете в нетранскрибиран на латиница вид, е ChiWriter. От тези начални опити и до днес не е решен въпросът със създаването на общоприета система за кодиране на старобългарските текстове.

В последните пет години интересът към приложенията на информационните технологии се засили значително. На практика във всички български изследователски центрове и университетски катедри, провеждащи старобългаристични изследвания, се използват компютърни текстообработващи системи. В Института за литература при БАН от 1995 г. се използва специализирана система за описание на ръкописи, създадена в рамките на съвместен българо-американски проект [Miltenova 98b ]; в катедрата по Кирило-методиевистика при СУ "Св. Кл. Охридски" са правени експерименти с използването на системата COLLATE за текст-критично представяне и изследване на текстове [Guergova 96]; в Кирило-Методиевския научен център при БАН се използва системата DBT за лингвистичен анализ на старобългарски текстове [Пики и др., 97].

В нашата страна са проведени и два специализирани международни форума, свързани с приложенията на информационните технологии в старобългаристиката – конференцията *Компютърна обработка на средновековни текстове* (м. юли 1995 г., Благоевград, вж. [BBDM 96]) и семинарът *Текстова вариативност в средновековните текстове* (м. септември 1997 г., София, вж. [Dobreva 98b]).

Табл. 1.1. Брой публикации, свързани с приложения на ИТ в старобългаристиката (по тематични направления)

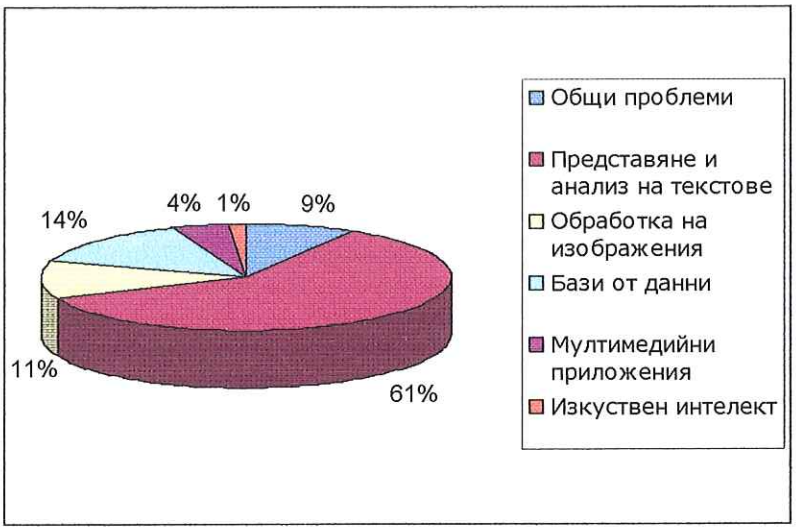
| **НАПРАВЛЕНИЕ** | **БРОЙ публикации** |
|---|---|
| Общи проблеми | 6 |
| Кодиране | 3 |
| Представяне и анализ на текстове | 19 |
| Приложение на SGML и TEI | 10 |
| Текст-критичен апарат, колация на ръкописи | 5 |
| Работа върху конкретни паметници | 5 |
| Шрифтове | 2 |
| Обработка на изображения, OCR | 4 |
| Бази от данни | 10 |
| Мултимедийни приложения | 3 |
| Методи от областта на изкуствения интелект | 1 |
| Консервация | 2 |
| **ОБЩО** | **70** |

Изследването е направено на базата на всички издирени публикации у нас и от наши специалисти в чужбина, направени в периода 1993-98 г. То показва ясно предпочитанията на специалистите към различните възможности за използване на ИТ (вж. Табл. 1.1.). Най-значим е интересът към компютърното

представяне и анализ на текстове, на което са посветени около 61% от всички изследвани публикации (вж. Фиг. 1.1.).

Направления като *приложения на изкуствения интелект, дизайн на специализирани шрифтове, мултимедийни приложения* и *проблеми на кодирането на текстове* се появяват в изолирани публикации. Докато за част от тези области това не е учудващо поради факта, че става дума за нови технологии, следва да се отбележи малкият брой публикации в областта *кодиране на текстове*, тъй като този проблем все още не е намерил своето общоприето решение, а е съществен за работата във всички останали области.

Фиг 1.1. Приложения на ИТ в старобългаристиката




Областта, която заема второ място по брой публикации след компютърното представяне и анализ на текстове, е тази на *приложения на системи за управление на бази от данни*. Въпреки че тази област е първата, в която са разработени конкретни приложения за въвеждане и анализ на данни за ръкописи, след навлизането на SGML-съвместимо представяне на старобългарски текстове, се наблюдава известен отлив от интереса към

приложенията на системи за управление на бази от данни. Публикациите, свързани с използване на SGML и TEI, заемат равен дял с тези, свързани с приложения на бази от данни. Това е лесно обяснимо с необходимостта да се въведат максимално изчерпателни данни, отразяващи научното описание на кодирания паметник. При това не се накърнява естественото желание на специалистите от предметната област да не разделят текста от научното му описание в процеса на компютърното му въвеждане.

Общите проблеми в областта на приложенията на информационните технологии в старобългаристиката са очертани в няколко публикации. В публикацията [Birnbaum 96a] се отделя внимание на такива естествени изисквания към създаването на компютърни ресурси в областта на старобългаристиката като многоцелевото използване, структурираността на данните, тяхната преносимост и съхранението им, независимо от компютърната платформа. В публикацията [Paskaleva, Dobreva 96] е направен кратък обзор на приложенията на компютърната лингвистика в работата с текстове на съвременен български език. Посочени са основни нерешени проблеми, свързани с компютърната обработка на старобългарски текстове, като кодирането на текстове и представянето на знания, необходими за обработката на текстовете на различни езикови нива. Важността на тази проблематика е подчертана и в [Янакиев 98], където е направена съпоставка на традиционните и компютърно-подпомогнатите подходи към текст-критичното издаване на средновековни паметници. В статията [Dobreva 98a] е направен обзор на българските начинания в областта на приложенията на информатиката за обработка на данни, свързани с културно-историческото наследство, вкл. средновековните ръкописи. Подчертава се необходимостта от изграждането на национална стратегия за дигитализация на античните и

средновековни паметници.

Публикациите, посветени на общите проблеми при прилагане на информационните технологии в старобългаристиката, очертават една широка област, в която все още не са решени основни проблеми.

Най-съществения подобен проблем е този за намиране на подходящо и общоприето кодиране на старобългарските текстове в компютърен вид. В [Kempgen 96] са разгледани проблемите при създаване на специализирани шрифтове за представяне на старобългарски текстове, като основно се има предвид работата в среда на Apple Macinstosh. В [Grünberg 96 ] са предложени правила за транскрибиране на старобългарски текстове на латиница, в които буквите с особени позиции в текста се кодират чрез използването на специални символи. И двете посочени по-горе статии подхождат към решаването на проблема за намиране на подходящо кодиране по едностранчив начин, без да предлагат основа, която би послужила за изчерпателното му решаване. Най-значим принос в това направление има работата на проф. Д. Бирнбаум от Университета в Питсбърг, САЩ.

В публикацията [Birnbaum 96b] се разглежда въпросът за най-малките единици, които подлежат на кодиране в старобългарските текстове. WSD (декларациите за писмена система) на SGML са посочени като задоволително решение при кодирането на старобългарски текстове. Конкретно предложение за подходящото кодиране на старобългарските текстове на ниво графеми е направено в статията [Birnbaum 96c].

Предложената там съвкупност от елементи е представена по-подробно в част 2.1. на настоящия труд. Като продължение на работата на Д. Бирнбаум за кодиране на текстове от старопечатни книги могат да се разглеждат предложенията на Р. Клеминсън в

[Cleminson 98]. Публикациите в областта на компютърното кодиране на старобългарски текстове показват, че отделни изследователи търсят подходящо решение на проблема. Когато то, обаче, се обвързва с конкретна хардуерна платформа или с транскрибиране на текстовете на латиница, не може да се очаква да бъде прието единодушно от старобългаристичната колегия и да послужи за безпроблемна размяна на текстове и първични изследователски данни по електронен път. Най-перспективно е обвързването на кодирането с международно приет стандарт за кодиране на текстове, независим от компютърната платформа, какъвто е SGML. В тази област е несъмнен приносът на Д. Бирнбаум, който е разработил в своите изследвания списък на елементите, които трябва да бъдат представени в едно изчерпателна система за кодиране. Все още не е разработено представянето на особените местоположения на графеми в старобългарските текстове. Този въпрос е разгледан в част 2.2. на дисертационния труд.

Най-много публикации в областта на приложенията на ИТ в старобългаристиката са посветени на представянето и анализа на текстове. Пръв опит да се систематизират особеностите в тази област е направен в публикацията [Добрева, Бояджиев 93]. В [Dobreva 94b, 95a] са разгледани текстовите редактори, използвани за въвеждане на старобългарски текстове, като са посочени онези особености на средновековната ни писмена традиция, които не могат да бъдат отразени при въвеждането на текстовете в тези редактори. Някои приложения на информатиката в текстологията са разгледани в [Янакиев 96]. Втази пубпикация е очертан преходът от използването на индекси на словоформите и конкорданси към директното търсене на изследваните елементи със средствата на съвременните текстови редактори. Важността на приложението на компютърните технологии при търсене на средновековни изонорми на различните езикови нива

е подчертана в [Miltenova 98a]. В [Buzzetti, Rehbein 98] са представени общите проблеми при компютърно подпомогнатата подготовка на текст-критични издания. В Санкт-Петербургския клон на Руското библейско общество се работи върху текстkритично издание на Евангелието от Йоан [Azarova 96]. Този амбициозен проект цели съпоставителното изследване на над 1200 паметника, като при изследването се използва клъстерен анализ. Не ни е известно да са публикувани резултатите от работата по този проект, нито са ни достъпни данни за конкретните компютърни програми, които се използват при изследването.

В рамките на Машинния фонд на руския език се планира създаването на текстов корпус от средновековни текстове [Молдован 96].

В [Пики и др., 97] е представена разработената в Института по компютърна лингвистика в гр. Пиза, Италия, система за обработка на текстове DBT. Първият опит за приложение на DBT при изучаването на средновековни славянски текстове е направен от М. Камуля върху материал от Псалтира. Основната цел на това изследване е изследването на лексикалната вариативност в избраните паметници (вж. [Camuglia 96a и 96b] и [Camuglia, Picchi 98]). В Кирило-Методиевския научен център при БАН продължават експериментите за използване на DBT при представяне и анализ на средновековни текстове.

Разработката на компютърен речник на средновековните славянски текстове е разгледана в [Ribarova, Ribarov 96]. В тази публикация, както и в [Ribarov, Ribarova 98], е представено програмното средство STINO, разработено специално за индексиране на средновековни славянски текстове от К. Рибаров.

В [Лотошко 98] се разглежда сричковият състав на руски текстове от XVII в. Със специално разработена за целта програма,

RPS-305, освен данните за употребата на различни срички се извличат и данни за поредността на появата им в думите. Представени са емпирични данни за наличие на корелация между повторяемостта на сричките и лексикалния състав на изследвания текст.

Изследване на правописната вариативност в процеса на промяна на писмената система на чешкия език през периода XIII-XX в. е представено в [Kučera 98].

Някои проблеми при компютърното представяне и обработка на глаголически текстове са представени в [Fetkova 98]. В тази публикация отново се набляга на важността на предварителното планиране и постигането на консенсус за отразяването на явления, свързани с вариативността, при компютърното представяне на текстове. В този конкретен случай се е стигнало до решение да се отразява лексикалната вариативност. При подобен подход се губи възможността за изследване на правописната вариативност на базата на създадения корпус от текстове.

Създаването на корпус от евангелски текстове от XVI-XVII в. на литовски език е представено в [Lucinskiene 98]. Интересно е да се отбележи, че и в този случай за изследването на текстовете са разработени специализирани програмни средства. С тяхна помощ са съставени индекси на словоформите на преводи на Евангелието на литовски език от различни периоди, съдържащи общо над 180 000 словоформи и конкорданс.

На изследвания на конкретни типове текстове и паметници са посветени редица публикации (менологиите са обект на представяне и изследване в [Vakareliyska 96 ], химнографските сборници - в [Guergova 96], евхологиите - в [Shniter 96]; Светостефанската хризовулия - в [Salter 96]).

Прегледът на публикациите, които представят приложения

на компютърни средства за анализ на средновековни славянски текстове, показва, че различните изследователски колективи работят разединено. В отделни организации се създават специализирани програмни средства, които решават конкретни задачи, но в подобна обстановка не може да се очаква разпределеното създаване на значими по обем ресурси, тъй като използваните подходи и предпочетените хардуерни и софтуерни платформи се различават съществено. Друга особеност на част от разработките в областта е предварителното взимане на решение за нормализация на текстовете. Това ограничава бъдещата им употреба от други изследователи при работата на по-ниско езиково ниво. Обикновено се предпочита отразяването на лексикалната вариативност, като графичната и правописната не се взимат предвид при въвеждането на текстовете в компютърен вид.

При това положение, особено перспективна за решаването на изброените проблеми изглежда работата на онези изследователи и колективи, които са се ориентирали към приложение на SGML и TEI за въвеждане на средновековни славянски текстове. Най-значими усилия в това направление са положени в нашата страна. Разработваният от 1994 г. насам проект "Компютърен репертоар на старата българска литература" в Института за литература при БАН е представен в [Miltenova 98b]. В [Miltenova 96] е отделено обстойно внимание на използването на данни от научното описание на ръкописи, представени в TEI формат, за анализ на макроструктурата и типологията на средновековните ръкописи. В [Bojadzhiev 96] са разгледани трудностите при представянето на палеографски данни и при отразяването на правописните особености в подобни описания. В [Dobreva 96a] са описани проблемите при разработка на подобен интердисциплинарен проект, като в [Dobreva 96b] в детайли са разгледани въпросите, свързани с обучението на специалисти

филолози за работа в SGML среда. Усвояването на подобно сложно описание на ръкописи, което включва над 120 различни елемента, е сложна задача. По време на работата над проекта са проведени психологически експерименти за степента на усвояването на описанието и за удобството на работа с него, които също са засегнати в последната публикация. В [Stoykova 98], [Dimitrova 98] и [Radoslavova 98] се представени някои проблеми, свързани с представянето на съдържанието на средновековните ни ръкописи при въвеждането на данни за тях в съответствие с предписанията на TEI.

В [Gagova 98] е даден изчерпателен анализ на възможностите за въвеждане на данни за приписки с препоръки, които могат да послужат при следващи допълнения на TEI. Интересен е опитът за използване на TEI, представен в [MacRobert 96], където също се отделя основно внимание на представянето на структурата на ръкописа с апарата на TEI. Отново обаче става дума за представяне на макроструктурата на текста, като се оставя без внимание представянето на графичното разположение на елементите при написването на текста, което не е задоволително решено в TEI. На този въпрос е посветена част 2.3. на дисертационния труд, в която е направено предложение за разширение на TEI за отразяване на графичното разположение на елементите в средновековните ни текстове.

Едно друго обособено направление на работа, свързано с изследване на текстове, е свързано с компютърно-подпомогнатото изследване на множество варианти на един и същи текст (колация на паметници) и създаването на текст-критичен апарат. Още от началото на прилагането на информационните технологии в старобългаристиката, това направление се развива активно от школата на проф. У. Федер, Университет в Амстердам, Холандия. На

общите принципи в областта е посветена публикацията [Veder 96 ], като в статията [Veder 98] може да се намери конкретно изследване на текстовете на преписите на "За буквите" от Черноризец Храбър. Приложението на подхода е представено в случая на текстовете от Евангелието в [Bakker 96a и 96b], и за текста на Апостола в [van derTak 98]. На улесняването на търсенето на варианти на конкретни текстови части е посветена публикацията [Ivanova et al., 96].

На автоматичното разпознаване на средновековни текстове са посветени публикациите [Stoyanov 96], [Klasinc, Kurent 96] и [Majstoroska 96]. Постигнатата точност на разпознаване при различните изследователи все още не е обнадеждаваща за използването на оптическо разпознаване на символи при въвеждането на текстове. В [Mitrevski 96] се разглеждат въпроси, свързани с разработката на старобългарски шрифтове.

Направлението, което се е развивало в началния етап на приложение на информационните технологии в старобългаристиката, е свързано с приложение на базите от данни за каталогизиране на ръкописи. Подобни разработки са представени в [Jankoska et al. 96], [Лампе 96], [Лабынцев 96], [Щавинская 96]. Правени са опити да се приложи специализираната база от данни за историци Κλειω при представянето на структурирани данни от архивни документи (вж. [Тяжелникова 96], [Tikhonov 96], [Юшин 96]). По-нестандартни приложения на Κλειω са тези, свързани с изследване на типологията на гадателните книги [Angusheva 96] и на гръцките и латинските заемки в един конкретен ръкопис - Ню Йоркския мисал [Dimitrova 96].

Използването на бази от данни в старобългаристичните изследвания е свързано основно със съхранението и обработката на каталожни данни за ръкописи. Изключение правят приложенията на системата Κλειω, които са ориентирани към обработка на данни от ръкописи и архивни материали.

Мултимедийните приложения са все още рядкост в областта на старобългаристиката, въпреки че съчетаването на различни представяния на ръкописите (във вид на изображения, текст и структурирани данни) води до най-пълноценно представяне на изследвания материал. В [Bratanova et al. 96] е представен проектът "Света София", чиято цел е събиране на изворов и илюстративен материал, свързан с култове към жени-светици. Първите два опита да бъдат събрани изображения на ръкописи от сбирките на Националната ни библиотека на компакт-дискове са представени в [Hristova 96]. В [Dobreva, Ivanov 98] са обобщени данни за електронни публикации у нас в областта на средновековните славистични изследвания.

Опит за приложение на методи от областта на изкуствения интелект за датиране на средновековни ръкописи е представен в [Koychev, Dobreva 94J. На базата на набор от признаци като тип на писмото, цветове на използваните мастила, правописни характеристики и стил на орнаментация чрез приложение на алгоритми от областта на машинното обучение се генерират правила за датиране, които дават над 97% точност при определяне на датировката на конкретни паметници.

В областта на консервацията на ръкописни паметници дигитализацията е посочена като възможно решение в [НПОБФ 1997]. За разлика от други страни, където се разработват програми за дигитализация в национален мащаб, у нас подобно начинание все още не е предприето.

## 1.2. Нерешени проблеми

Въз основа на прегледа на литературата в областта на приложенията на информационните технопогии в старобългаристиката, могат да бъдат изброени следните незадоволително решени

проблеми:

1. Не е създаден общоприет стандарт за компютърно кодиране на двете старобългарски азбуки - кирилицата и глаголицата;
2. Не е разработена система за представяне на особените елементи в средновековните текстове като надписани, долуподписани и вписани букви и плетено писмо.
3. Не са разработени модели за представяне на вариативността на различните езикови нива в старобългарски текстове.
4. Съществуват значителни различия във възгледите на специалистите-старобългаристи за структурирането на данните от предметната област, както и за връзките между тях.

Нерешаването на тези основни проблеми води до съществено намаляване на възможностите за облекчаване на работата на специалистите-старобългаристи чрез приложение на новите информационнитехнологии.

*Първият проблем* води до затруднения в обмяната на текстове и първични изследователски данни между специалистите от областта, като последицата в дългосрочна перспектива е забавянето на създаване на значителни ресурси в електронен вид (например текстови корпуси и речници).

*Вторият проблем* води до масова практика текстовете да се представят в нормализиран вид[1] вместо във вид, максимално близък до оригинала.

Това намалява възможностите за изчерпателен анализ на лингвистични признаци, които се обявяват за несъществени a priori

---

[1] Нормализацията обикновено включва въвеждане на основния ред на буквите, които се явяват извън него, както и пълно изписване на съкратените думи.

още при въвеждането на текстовете. До момента представянето на вариативността на правописно ниво бива игнорирано, въпреки че отразяването на различията в написването на текстовете може да послужи като източник за интересни наблюдения върху развитието на средновековната писмена традиция.

*Третият проблем* води до ограничаване на възможностите за пълнотекстово търсене. Особено силно това проличава на правописното текстово ниво.

*Четвъртият проблем* се появява като следствие от силната специализация на старобългаристите и различните им изследователски интереси. Но без неговото задоволително решаване отделните опити за въвеждане и обработка на структурирани данни ще си остават изолирани и ефектът от положените усилия ще е нисък.

## 1.3. Обосновка на необходимостта от създаване на специализирано работно място на старобългариста

Като се има предвид огромният брой средновековни славянски ръкописни паметници (само в български хранилища се съхраняват над 8000 ръкописа, голяма част от които принадлежат към старобългарското ръкописно наследство) и разпръснатостта на средновековните славянски ръкописи в хранилища и сбирки в десетки държави, е наложително създаването на международно общоприети стандарти и модели в областта . По този начин ще бъде възможно разпределеното създаване на електронни ресурси в областта на средновековното славянско ръкописно наследство в различни страни по унифициран начин.

Особеностите на предметната област налагат естественото изискване за интеграция на представяне на текст във вид максимално близък до оригинала, изображение и структурирани

данни (научно описание на изследваните паметници). Подобно интегрирано представяне може да се постигне чрез прилагането на мултимедийни технологии.

Създаваната среда трябва да предоставя възможности за извличане на първични изследователски данни, което обичайно отнема много време. В областта на работата със средновековни паметници подобни дейности са ексцерпирането, подредбата по азбучен ред и пр. Ценно допълнение към извличането на първични изследователски данни предоставяват възможностите за анализ на събраните данни. В областта на хуманитарните науки се работи с огромен обем данни и необходимостта от сравняване на голямо количество изходни данни в много случаи забавя изследователската работа.

Важна стъпка към получаването на нови данни за предметната област е възможността за извличане и анализ на количествени характеристики на изследваните обекти . В тази област до сега почти не е работено, а при наличието на вариативност прилагането на количествени методи за изследване биха могли да доведат до получаване на неизвестни досега характеристики на развитието на средновековните славянски текстове в синхронен и диахронен аспект.

# Глава 2. Компютърно представяне на старобългарски текстове

## 2.1. Проблемът за състава на компютърното кодиране на старобългарските азбуки

Използването на различни начини на компютърно кодиране на старобългарските текстове създава трудности при обмена на текстове по електронен път между специалистите, работещи в областта.

Причините за използване на различни системи за кодиране са следните:

- различната вътрешна организация на текстовите редактори (сред най-активно използваните за въвеждане на старобългарски текстове са освен WINDOWS-приложенията и редакторите ChiWriter и $T^3$);
- различните изисквания на специалистите към степента на детайлност при въвеждане на старобългарските текстове; ненормираността на азбуката през Средновековието в сравнение със съвременното понятие за норма .

Сред най-разпространените системи за кодиране, основани на представяне на старата кирилица чрез транслитерация на латиница, е предложената от проф. Уилям Федер от Амстердамския университет (вж. [Geurts et al. 87] и първата колонка от таблицата в Приложение 1.) . Тази система за транслитерация вкпючва 112 символа, употребявани в старобългарските ръкописи. Тя е с променлива дължина на транслитериращите последователности. В редица случаи различни транслитериращи низове се съдържат един в друг.

Старобългарската кирилица все още няма задоволително представяне в стандартите на ISO (International Organization for Standardization - Международната организация по стандартизация). Обичайна практика в тези стандарти е създаването на общ стандарт за азбуките на множество езици - вж. например [ISO 9:1995 (E)]. Този принцип е приложен и в стандарта Unicode [Unicode 91], въпреки очевидните неудобства при кодиране на текст, съдържащ едновременно съвременен български и старобългарски език.

Съвременният подход за избягване неудобството от еднакво кодиране на съвременни и средновековни буквени знаци е да се прилага кодиране, което съдържа показател за азбуката. В съответствие с този принцип проф. Дейвид Бирнбаум от Университета в Питсбърг, САЩ, в статията [Birnbaum 96c], предлага система за кодиране, включваща 205 елемента.

Следва да се подчертае голямата разлика в обхвата на системите за кодиране на проф. Уилям Федер и проф. Дейвид Бирнбаум, която се дължи на различното им отношение към въпроса за детайлизираността на въвеждането на средновековни текстове.

Кодирането на проф. У. Федер е подходящо в случаите, когато се изследва текстът като цяло без да се обръща внимание на лингвистичните детайли на правописно ниво, докато кодирането на проф. Д. Бирнбаум е подходящо за изследвания на по-детайлизирано ниво.

Като още един пример за кодиране е представена таблицата, ползвана при един от старославянските шрифтове, разработени за WINDOWS среда (нетърговски шрифт Evangelie, създаден във Франция). Тази система за кодиране е най-бедна в сравнение с първите две. В нея се наблюдават и недоработени елементи като използването на 2 кода за представянето на един и същи символ, или използването на главна и малка латинска буква,

които представят различни кирилски букви (обикновено се очаква в системите за кодиране да има някаква вътрешна логика, която в случая не е спазена последователно).

Следното обобщение показва кои са прилаганите досега подходи за представянето на старата кирилица в компютърен вид:

- *Транслитерация на латиница*

  Слабост на този подход е, че предполага използването на специални средства за нормална визуализация на старобългарските текстове. Освен това неестествената среда за работа е предпоставка за допускане на грешки при въвеждане на текстовете.

- *Разработка на специализирани шрифтове с включване в стандартните кодови таблици*

  Този подход е силно ориентиран към конкретна платформа и затруднява обмена на текстове. Освен това поради ограничения в кодовото пространство представените графеми не винаги удовлетворяват изискванията на старобългаристите за пълнота на представянето.

- *Използване на специални кодиращи последователности от типа на декларациите за писмена система в SGML*

  Този подход дава най-добри възможности за детайлно представяне на особеностите на изписването и разположението на текстовите елементи в източника. Но и при него, подобно на използването на транслитерацията, се изисква разработката на специални програмни средства, които да правят възможно нормалното визуализиране на текста.

Отделихме място на системите за кодиране, тъй като от доброто познаване на вътрешното представяне на текстовете зависи успешната им по-нататъшна обработка.

Системите за кодиране са организирани по различен начин, но обикновено включват елементи с нефиксирана дължина на кодовете.

Това означава, че в случаите на наличие на взаимно съдържащи се последователности, за трансформирането на текст от представяне в една система в друга трябва да се подбере алгоритъм, при който най-дългите кодирания се търсят първи в низа за прекодиране. Тази особеност прави сложно прилагането на някои алгоритми за обработка на низове, тъй като има кодове, които се съдържат взаимно (вж. напр. кодирането на Уилям Федер).

## 2.2. Нелинейността в старобългарските текстове и компютърното й представяне и обработка

Разгледаните в част 2.1. проблеми с кодирането на старобългарски текстове не изчерпват всички трудности, които трябва да се преодолеят за създаването на добро компютърно представяне на старобългарски паметници.

Използването в тях на надписани, долуподписани и вписани букви изисква прилагането на такива начини за кодиране, които да не затрудняват следващата обработка на текстовете. При търсенето на определена лексема в текст на машинен носител една изместена спрямо основния ред буква ще доведе до ненамиране на подобни появи на лексемата в текста. Например търсенето на ѿтъ и ѿ трябва да се извърши на два пъти, като се зададат двата варианта на изписване.

Всички използвани до сега текстови редактори (приложения на Windows, $T^3$ и ChiWriter) решават описания проблем по начин, който не позволява лесно търсене в последствие.

При ChiWriter вътрешното представяне на текстовете включва един основен ред и три допълнителни реда, които служат за

записване на надписани и подписани букви. В $T^3$ в записа на файла се вмъкват служебни маркери за особеното положение на буквите.

Част от разработените шрифтове, използвани в Windows-среда, включват като отделни позиции в кодовата таблица надписани варианти на буквите от азбуката. Това също не решава проблема за пълнотекстово търсене, защото при търсенето на една и съща буква трябва да се зададат и двата кода, отразяващи възможностите за нейното изписване (като буква от основния ред и като надписана буква).

При никой от тези подходи освен това не е възможно точното отразяване на надписаните букви – между две букви от основния ред или точно върху буква от основния ред. В никой от посочените редактори не може да се създават вписани букви.

Подобно състояние на нещата изисква създаването на такова кодиране, което позволява освен кода на съответния символ да се съхранява и код за относителното местоположение на символа спрямо околните букви. Една от целите на настоящата работа е предлагането на решение на този проблем.

Фиг. 2.1. Илюстрация за разнообразни
графични елементи в старобългарски текст,
Иван-Александрово Четвероевангелие (XVI в.)

Представянето на местоположението на буквите спрямо основния писмен ред в процеса на компютърно кодиране на текста е съществено за изучаване на особеностите в развитието на средновековната славянска писмена традиция. Тенденцията за постепенно увеличаване на броя на буквите, изписани извън

основното редово пространство, не е изследвана досега с количествени методи. Същевременно, старобългаристите използват като помощен признак при датиране на средновековните паметници общите наблюдения за появата на надписани, допуподписани и вписани букви.

## 2.3. Предложение за разширение на TEI

Въпросите, свързани с представянето на оригинални (в т.ч. и ръкописни) текстове са запожени в дяла на TEI *Transcription of Primary Sources,* като формалното дефиниране на елементите се съдържа в множеството *teitran2.dtd.* В множеството от дефинирани елементи в *teitran2.dtd* липсват елементи, които да позволят да се опише точното местоположение на графемите в текста . Подобни елементи са необходими при въвеждането в електронен вид на средновековни текстове, тъй като при тях често се срещат съкращения. Тази част от стандарта регламентира:

- представянето на изменени, коригирани и грешни текстови пасажи;
- означаване на позицията на смяна на писарите;
- отбелязването на повредени участъци в оригинала.

`<abbr>` е маркерът , който се използва в TEI P3 за отбелязване на съкращения в кодирания текст. При него атрибутът `type` може да приема стойност `superscriptions`, която се използва в случаите, когато съкращението включва в себе си надписани букви. При това обаче кодирането не включва експлицитни указания кои букви се явяват извън основния ред.

Друг елемент в TEI P3, при който се отчита по-особената позиция на определени символи по отношение на основния писмен ред, е елементът `<AddSpan>`, който служи за маркиране на добавки към основния текст, направени от автор, писар ипи коректор.

Атрибутът `place` на този елемент може да приема стойности като: `supralinear`, `infralinear` или `inline`, които показват местоположението на добавката спрямо основния писмен ред. Този елемент също обаче не може да се използва за представяне на особеното местоположение на букви в старобългарски текстове, тъй като е създаден за маркиране на *по-късни добавки* към текстовете, а не за маркиране на символи, които са изписани извън основния ред в процеса на създаването на паметника.

TEI не предоставя също така и апарат за отбелязване на подробности около използването на инициали, плетено писмо и червенопис. Елементът `<hi>` може да се използва за маркиране на всякакви текстове, изписани по различен начин от основния текст, но за да може да се посочват конкретните особености, трябва да се разработи множество от допълнителни стойности за атрибутите му.

За представяне на особените позиции на букви спрямо основния писмен ред, предлагаме следната дефиниция на нов елемент в TEI:

```
<!ELEMENT %n.position;    - - (%phrase.seq)                          >
<!ATTLIST %n.position;        %a.global;
          type                (superscript|subscript|inscript)
          placement           (exact|between)
          rotation            (0|45|90|135|225|270|315)              >
```

Стойността на елемента представя кодираното представяне на самата графема, а стойностите на атрибутите представят позицията спрямо предходния символ (надписана, подписана, или вписана, зададено чрез атрибута `type`), вида на разположението (точно над, под или във, както и с отместване, зададено чрез атрибута `placement`) и ъгъла на завъртане на буквата, зададен чрез атрибута `rotation`. Така се избягва необходимостта от кодиране на букви, които са ротирани под определен ъгъл, като отделни елементи от компютърното кодиране (подобни случаи има в

кодирането, предложено от Дейвид Бирнбаум, вж. Приложение 1.), и се постига по-голяма универсалност.

За представяне на плетено писмо предлагаме елемента `ligature` със следната дефиниция:

```
<!ELEMENT %n.ligature;    - - (%phrase.seq)                    >
<!ATTLIST %n.ligature;       %a.global;
          source             #IMPLIED                          >
```

В този случай стойността на атрибута представлява нормализиран запис на плетеното писмо. Като стойност на атрибута `source` се внася името на файла, който съдържа изображението на съответната част от текста .

За представяне на червенослов предлагаме следната дефиниция:

```
<!ELEMENT %n.rubric;     - - (%phrase.seq)                     >
<!ATTLIST %n.rubric;         %a.global;                        >
```

За представяне на инициали предлагаме следната дефиниция:

```
<!ELEMENT %n.initial;    - - (%phrase.seq)                     >
<!ATTLIST %n.initial;        %a.global;
          type               (Balkan | neo-Byzantine | Anthropomorphic |
                              Islamic | Geometric | Interlaced)
                             #IMPLIED
          source             #IMPLIED
          span               #IMPLIED                          >
```

Стойността на елемента представлява съответната буква, кодирана като елемент от декларацията за писмена система (WSD). Атрибутът `type` служи за уточняване на стила на украса на инициала. Атрибутът `source` служи за внасяне на наименованието на файла, съдържащ изображение на инициала. Атрибутът `span` служи за внасяне на обхвата на инициала в брой редове.

С така описаните допълнения teitran2.dtd добива следния вид:

```
<!-- teitran2.dtd:  written by OddDTD 1994-09-09              -->

<!-- 18:  Transcription of Primary Sources                    -->
<!-- Text Encoding Initiative: Guidelines for Electronic      -->
```

```
<!-- Text Encoding and Interchange. Document TEI P3, 1994.     -->
<!-- Copyright (c) 1994 ACH, ACL, ALLC. Permission to copy     -->
<!-- in any form is granted, provided this notice is           -->
<!-- included in all copies.                                   -->
<!-- These materials may not be altered; modifications to      -->
<!-- these DTDs should be performed as specified in the        -->
<!-- Guidelines in chapter "Modifying the TEI DTD."            -->
<!-- These materials subject to revision. Current versions     -->
<!-- are available from the Text Encoding Initiative.          -->
<!-- 18.1.4:  Added and Deleted Spans                          -->
<!ENTITY % addSpan 'INCLUDE' >
<![ %addSpan; [
<!ELEMENT %n.addSpan;    - O  EMPTY                                >
<!ATTLIST %n.addSpan;         %a.global;
          type                CDATA                #IMPLIED
          place               CDATA                #IMPLIED
          resp                IDREF                %INHERITED
          cert                CDATA                #IMPLIED
          hand                IDREF                %INHERITED
          to                  IDREF                #REQUIRED
          TEIform             CDATA                'addSpan'    >
]]>

<!ENTITY % delSpan 'INCLUDE' >
<![ %delSpan; [
<!ELEMENT %n.delSpan;    - O  EMPTY                                >
<!ATTLIST %n.delSpan;         %a.global;
          type                CDATA                #IMPLIED
          resp                IDREF                %INHERITED
          cert                CDATA                #IMPLIED
          hand                IDREF                %INHERITED
          to                  IDREF                #REQUIRED
          status              CDATA                'unremarkable'
          TEIform             CDATA                'delSpan'    >
]]>

<!-- (end of 18.1.4)                                           -->
<!-- 18.1.6:  Cancelled Deletions                              -->
<!ENTITY % restore 'INCLUDE' >
<![ %restore; [
<!ELEMENT %n.restore;    - O  (%phrase.seq;)                       >
<!ATTLIST %n.restore;         %a.global;
          wit                 CDATA                #IMPLIED
          cause               CDATA                #IMPLIED
          varSeq              NUMBER               #IMPLIED
          type                CDATA                #IMPLIED
          desc                CDATA                #IMPLIED
          resp                IDREF                %INHERITED
          cert                CDATA                #IMPLIED
          hand                IDREF                %INHERITED
          TEIform             CDATA                'restore'    >
]]>

<!-- (end of 18.1.6)                                           -->
<!-- 18.1.7:  Supplied Text                                    -->
<!ENTITY % supplied 'INCLUDE' >
<![ %supplied; [
<!ELEMENT %n.supplied;   - O  (%paraContent;)                      >
<!ATTLIST %n.supplied;        %a.global;
          reason              CDATA                #IMPLIED
          resp                CDATA                %INHERITED
          hand                IDREF                %INHERITED
```

```
          agent                CDATA                #IMPLIED
          source               CDATA                #IMPLIED
          TEIform              CDATA                'supplied'     >
]]>

<!-- (end of 18.1.7)                                              -->
<!-- 18.2.1:  Hand Shifts                                         -->
<!ENTITY % hand 'INCLUDE' >
<![ %hand; [
<!ELEMENT %n.hand;       - O  EMPTY                               >
<!ATTLIST %n.hand;
          id                   ID                   #IMPLIED
          n                    CDATA                #IMPLIED
          rend                 CDATA                #IMPLIED
          hand                 CDATA                #REQUIRED
          scribe               CDATA                #IMPLIED
          style                CDATA                #IMPLIED
          lang                 CDATA                #IMPLIED
          ink                  CDATA                #IMPLIED
          character            CDATA                #IMPLIED
          first                CDATA                #IMPLIED
          resp                 CDATA                %INHERITED
          TEIform              CDATA                'hand'         >
]]>

<!ENTITY % handShift 'INCLUDE' >
<![ %handShift; [
<!ELEMENT %n.handShift; - O  EMPTY                               >
<!ATTLIST %n.handShift;      %a.global;
          new                IDREF                #IMPLIED
          old                IDREF                #IMPLIED
          style              CDATA                #IMPLIED
          ink                CDATA                #IMPLIED
          character          CDATA                #IMPLIED
          resp               IDREF                %INHERITED
          TEIform            CDATA                'handShift'    >
]]>

<!ENTITY % handList 'INCLUDE' >
<![ %handList; [
<!ELEMENT %n.handList;   - O  ((%n.hand)*)                       >
<!ATTLIST %n.handList;       %a.global;
          TEIform            CDATA                'handList'     >
]]>

<!-- (end of 18.2.1)                                             -->
<!-- 18.2.3:  Damage and Illegiblity                              -->
<!ENTITY % damage 'INCLUDE' >
<![ %damage; [
<!ELEMENT %n.damage;     - O  (%paraContent;)                    >
<!ATTLIST %n.damage;         %a.global;
          type               CDATA                #IMPLIED
          extent             CDATA                #IMPLIED
          resp               IDREF                %INHERITED
          hand               IDREF                %INHERITED
          agent              CDATA                #IMPLIED
          degree             CDATA                #IMPLIED
          TEIform            CDATA                'damage'       >
]]>

<!-- (end of 18.2.3)                                             -->
<!-- 18.2.5:  Spaces in the source                                -->
```

```
<!ENTITY % space 'INCLUDE' >
<![ %space; [
<!ELEMENT %n.space;      - O  EMPTY                                    >
<!ATTLIST %n.space;           %a.global;
          dim                 (horizontal | vertical)
                                                  #IMPLIED
          extent              CDATA               #IMPLIED
          resp                CDATA               #IMPLIED
          TEIform             CDATA               'space'           >
]]>

<!-- (end of 18.2.5)                                                 -->
<!-- 18.3:  Headers and footers                                       -->
<!ENTITY % fw 'INCLUDE' >
<![ %fw; [
<!ELEMENT %n.fw;         - O  (%phrase.seq;)                          >
<!ATTLIST %n.fw;              %a.global;
          type                CDATA               #IMPLIED
          place               CDATA               #IMPLIED
          TEIform             CDATA               'fw'              >
]]>

<!-- (end of 18.3)                                                   -->
<!-- (end of 18)                                                     -->


<!-- Additions by Milena Dobreva                                     -->
<!-Draft, 1999

<!ELEMENT %n.position;   - - (%phrase.seq)                            >
<!ATTLIST %n.position;        %a.global;
          type                (superscript|subscript|inscript)
          placement           (exact|between)
          rotation            (0|45|90|135|225|270|315)               >

<!ELEMENT %n.ligature;   - - (%phrase.seq)                            >
<!ATTLIST %n.ligature;        %a.global;
          source              #IMPLIED                                >

<!ELEMENT %n.rubric;     - - (%phrase.seq)                            >
<!ATTLIST %n.rubric;          %a.global;                              >

<!ELEMENT %n.initial;    - - (%phrase.seq)                            >
<!ATTLIST %n.initial;         %a.global;
          type                (Balkan | neo-Byzantine | Anthropomorphic |
                               Islamic | Geometric | Interlaced)
          source              #IMPLIED
          span                #IMPLIED                                >
```

# Глава 3. Вариативност в старобългарските текстове и нейното компютърно моделиране

## 3.I. Нива на езикова вариативност. Състояние на изследванията в областта

### *3.I.I. Значение на езиковата вариативност*

Изследванията в областта на езиковата вариативност са обект на нарастващ интерес в последните години. Традиционно тази област обединява в себе си методи от областите на лингвистиката, социологията и психологията, като напоследък в нея все по-активно навлизат и методи от областта на информатиката.

Основните интереси при изследването на езиковата вариативност са насочени към изучаването на различни *типове вариативност* и *причините*, които водят до тяхната проява. На този етап усилията на изследователите са насочени към анализ на явленията в областта и тяхната класификация.

Компютърните модели и приложения обаче не са все още много популярни за моделиране на езиковата вариативност, въпреки че някои от типовете вариативност могат успешно да се опишат по формализиран начин. Освен за подпомагане на работата на изследователите по традиционната им парадигма, такива модели понякога водят до появата на нови методи на изследване.

Като типичен пример за промяна в парадигмата можем да посочим областта на компютърните корпуси от текстове. Появата на първите корпуси от текстове в компютърен вид е продиктувана от желанието да се съставят конкорданси или индекси на произведенията на определен автор. Тъй като съставянето на конкорданс "на ръка" е изключително трудоемко и отнема много

време, идеята е била с помощта на компютъра да се облекчи рутинният процес на многократно преписване и сортиране. С течение на времето корпусите стават мощно средство за проверка на лингво-статистически хипотези и сега е немислимо да се формулират хипотези, без те да се апробират върху корпус. Това е промяна на парадигмата, защото преди появата на корпусите за апробиране на хипотеза е било приемливо да се използва скромна по обем текстова извадка или само да се посочат няколко подходящи примера.

Работата с компютърни корпуси от текстове е вероятно първата област, която навежда изследователите в областта на компютърната обработка на естествен език в чужбина да обърнат внимание и на проблемите, породени от вариативността на естествените езици и, респективно, текстовете. Наличието на правописна вариативност например пречи на получаването на точни резултати при търсене на определена дума или фраза. Това означава, че изследователите трябва да избират между две възможности - да нормализират текстовете в корпуса или да разработят допълнителни средства, които ще подпомогнат обработката на случаи на вариативност.

Например при създаването на *The Century of Prose Corpus* (корпус, в който са събрани текстове на автори от Великобритания от периода 1680-1780 г.) [Milic 95], основното предназначение на корпуса е формулирано като "ресурс за изучаващите езика на епохата" (стр. 327, цит. публ.). Въпреки ясното разбиране на автора на корпуса, че това е една от малкото диахронни колекции от текстове (при диахронните колекции от текстове правописната вариативност например се проявява в по-голяма степен, отколкото при синхронните), в корпуса е направена нормализация на правописа. Тази нормализация при дадения примерен текст на стр. 330-331 от цит. съч . е проведена в около 3% от думите, като не е

посочено дали това съотношение важи за целия корпус. Представена е ясна мотивация за провеждането на нормализацията и тя е, че "основната цел на корпуса е да улесни изследването на *стила и езика*, а не на *работата на печатарите и издателите*" (курсив мой - М.Д.). Учудващо е това причисляване на правописните различия единствено към областта на компетенция на печатарите и издателите, като имаме предвид, че авторите обикновено са контролирали на някакъв етап от процеса на издаване как изглежда текстът им.

Може да се предположи, че разликите в правописа в корпуса се дължат или на неустановени твърдо норми в английския език от XVII-XVIII в., или на промени на същите в диахронен аспект . Тази хипотеза би могла да бъде обект на изследване за специалистите по английски език, за което обаче *The Century of Prose Corpus* не може да бъде използван поради описаните по-горе принцип за нормализация, следван при създаването му.

След като разгледахме един типичен случай на създаване на компютърен корпус, за който е взето решение за нормализиране на определени типове вариативност, нека сега се спрем на основните направления на изследвания в областта на езиковата вариативност. Както вече споменахме, тази област се радва на нарастващ интерес през последните години. Косвено потвърждение за този интерес е провеждането на ежегодни международни форуми (вж. напр. [Dennlng et al. 87], [Ferrara et al. 88], [Edmondson et al. 9О]) и издаването на списанието *Language Variation and Change* от 1989 г . насам [LVC] .

Основните тематични направления, в които се провеждат изследвания (изброените направления са обобщени от томовете на цитираните по-горе конференции, списанието *Language Variation and Change* и сборниците [Fasold 83], [Fasold, Schiffrin 89]), са следните:

- фонетека и фонология,
- морфология,
- синтаксис,
- местни диалекти,
- английски език на бялото население в САЩ,
- английски език на чернокожото население в САЩ,
- езикови заемки,
- езикови интерференции,
- влияние на пола върху езика,
- промени в изговора,
- езикови контакти,
- усвояване на майчин и чужд език,
- исторически аспекти,
- лексикални изследвания.

Можем да отбележим, че при групирането на публикациите по области се наблюдава изследване на няколко *езикови нива* (като напр. фонология, лексикология и т.н.) и на *фактори*, причиняващи вариативност (напр. *произход* при изследването на езика на чернокожото или бялото население, влияние на *пола* върху езика и т.п.).

Както може да се забележи, сред тези направления има такива, които засягат определени слоеве от населението в САЩ и представляват интерес главно за изучаващите английски език. Има и по-общи направления, в които могат да се провеждат изследвания и за други езици освен английския. На практика обаче за момента изследванията са силно езиково зависими и не се търсят по-общи методики, които да могат да се приложат при изследвания на други

езици. Сред малкото публикации, свързани с методиката в областта на моделирането на езиковата вариативност, можем да посочим [Sankoff 78], където са представени някои вероятностни модели за описание на езикови промени [Rousseau, Sankoff 78] .

Част от проблемите, свързани с изучаване на вариативността, могат да послужат за изследване на типични грешки на определено езиково ниво. Като пример можем да посочим изследването върху когнитивни процеси при усвояването на английския правопис [Frith 80], където слабо е засегната връзката между вариативността и някои фактори от областта на когнитивната наука, които водят до погрешно изписване на определени думи. Този труд е ориентиран към английския език, който несъмнено спада към езиците с трудно усвояем дори за носителите на езика правопис. Особено внимание е отделено на изследване на зависимостта между доброто усвояване на прочетен текст и правописа.

Важността на тематиката, свързана с езиковата вариативност проличава от факта, че в съвременното обучение на студенти в областта на лингвистиката се включва разглеждането и на езиковата вариативност (вж. напр. [Москов, Бояджиев 77], където особено подробно е разгледана фонетичната вариативност, и [Jannedy et al. 94]), където е дадена обща трактовка на проблематиката в областта на езиковата вариативност и са разгледани някои нейни специфични прояви на различни езикови нива,

### *3.1.2. Класификация на изследванията в областта на езиковата вариативност*

Въз основа на изложения преглед на проблематиката в областта на езиковата вариативност в ч. 3.1.1., предлагаме следната класификация на изследванията в областта на езиковата вариативност според обекта, който се изучава:

1. Изследвания, свързани с *лингвистичното ниво, на което*

се проявява вариативност (напр., фонетично, фонологично, морфологично, лексикално, синтактично, семантично);

2. Изследвания на *фактора/ите, причиняващи вариативност* (напр. произход; регионални различия; различия, причинени от половата принадлежност, и т.н.). Като по-специфичен случай можем да посочим изучаването на влиянието на фактора *време* върху езика, което намира сериозна проява в трудовете по диахронна лингвистика.

Основните фактори, които до момента са обект на изследователски интерес, са:

- време;
- пол;
- социална група;
- етническа група;
- възраст;
- жанр/стил

3. *Многомерни изследвания,* включващи изследване на комбинация от фактори и/или лингвистични нива (напр. *диахронно* изследване на *лексиката* в прозата на *жени-писателки* – пример на изследване на действието на два фактора на лексикално ниво).

Нека да приемем, че вариативността на всяко езиково ниво може да се опише като съвкупност от наблюдавани признаци $P=\{p_1, p_2, \ldots, p_k\}$. За всеки изучаван фактор F може да се проследи как неговите съставляващи компоненти $\{f_1, f_2, \ldots, f_j\}$ влияят върху всеки признак $p_i$, който изследваме.

В най-простия случай може да се посочи булева стойност, за да се опише влияе ли даден фактор върху конкретен признак. Като резултат от такъв тип изследване се получава булева матрица.

Подобни матрици могат да се съпоставят. Това означава, че

изследването на съвкупност от текстове може да се улесни, като се използват съответните на текстовете матрици, а такъв тип представяне е компактен и нагледен.

В по-сложни изследвания можем да търсим числови стойности, които да ни ориентират за интензивността на проявяване на определен признак. Когато такива стойности могат да бъдат получени (на правописно ниво например е възможно измерване на честотата на употреба на буквите от азбуката; на лексикално ниво е възможно определянето на честотата на употреба на определени думи и пр.), можем да търсим и някакви метрични зависимости за поведението на наблюдаваните признаци. Подобна атрибуция се отнася към областта на лингвостатистиката.

При тази постановка, ако дадена област на езикова вариативност се опише чрез набор от елементарни признаци, които да се проследяват сравнително лесно, може да се направи значима стъпка към формализирането на изследвания, включващи проследяване на езикова вариативност. Основен изследователски проблем е изясняването на признаците, които трябва да отговарят на следните няколко условия:

- *да са значими,* за да имат съществена роля при характеризирането на изследвания тип вариативност (напр. при изследванията на правописната вариативност на старобългарски текстове голям интерес представлява поведението на еровите гласни и носовките, докато данните за честотата на употреба на съгласните букви не са толкова характерни, що се отнася до правописните особености на изследвания текст).
- *да са представителни* - трябва да се търсят такива съвкупности от признаци, които да позволяват добиването на възможно най-ясна картина на вариативността. В този

смисъл изследването на еровите гласни и носовките не е достатъчно при работата със старобългарски текстове, защото съществени за правенето на изводи за правописа на текста са и поведението на йотуваните гласни, на епентетичното ѧ и др.

- *да са измерими* - трябва да се търсят признаци, които да могат да бъдат охарактеризирани по начин, който да позволява съпоставянето на числови данни за различни изследвани обекти.

Тъй като подобен тип изследвания все още не са популярни във филологическите среди, една от основните задачи по време на работата върху настоящото изследване е търсенето на подобни признаци на различните езикови нива в старобългарските текстове.

Тази постановка на въпроса за моделиране на езиковата вариативност ни позволява да представим следната фигура, илюстрираща обсега на изследванията в областта на езиковата вариативност в момента:

Фиг. 3.1. Фактори и нива на езикова вариативност

| Езиково ниво / ФАКТОР | Фонетика | Правопис | Лексика | Синтаксис | Семантика |
|---|---|---|---|---|---|
| Пол | | | | | |
| Социална група | | | | | |
| Етническа група | | | | | |
| Времеви период | | | | | |
| Жанр/стил | | | | | |
| Без определен фактор | | | | | |

Защрихованите части показват наличието на изследвания в съответното направление. Например, изследването на фонетичната вариативност без изучаването на факторите, които й влияят, може да се представи по следния начин:

Фиг. 3.2. Фонетична вариативност

| Езиково ниво / ФАКТОР | Фонетика | Правопис | Лексика | Синтаксис | Семантика |
|---|---|---|---|---|---|
| Пол | | | | | |
| Социална група | | | | | |
| Етническа група | | | | | |
| Времеви период | | | | | |
| Жанр/стил | | | | | |
| Без определен фактор | | | | | |

Както вече споменахме, изследователската работа в областта на езиковата вариативност има главно описателен характер. Това означава, че тя все още се намира на етапа на събиране на данни, като в бъдеще може да се очаква появата на формализирани модели.

Не са описани набори от признаци, нито количествени критерии за тяхното изследване. Развитието на подобни модели, пригодени за работата със старобългарски текстове в компютърна среда, е една от основните цели на настоящата работа.

### *3.1.3. Компютърно моделиране на езикова вариативност*

Създаването на компютърни средства в областта на обработката на естествени езици включва много малко знания от областта на езиковата вариативност. Обикновено подобни модели се ограничават с вариативността в правописа (напр. при разликите в

правописа между британски и американски английски).

Поради унифицирането на стиловете и широкото навлизане на типови документи (специално за английски език всеки съвременен текстов редактор има включени разнообразни шаблонни документи), при създаването на търговски софтуерни продукти няма голям интерес към моделиране на явления от различни нива на вариативност на съвременните езици. Показателно за това състояние е отсъствието на каквото и да било споменаване на проблема за вариативността в обзорната статия за компютърната обработка на естествен език [Church, Rau 95] .

Тази особеност обяснява донякъде специфичното положение, в което остават онези изследователи, които работят със средновековни текстове. При всичките затруднения, през които те преминават при доставянето, инсталацията и често следващата лична преработка на шрифтовете за азбуката, с която те работят , те достигат в един момент до извода, че въпреки че са в състояние да наберат даден текст във вида, в който биха искали да го ползват, това не решава проблемите по използването на текста. Често тази констатация идва с осъзнаването на невъзможността компютърът да се използва за автоматично търсене на всички словоформи на дадена дума без създаването на специална програма.

Ако това е в състояние да бъде преодоляно чрез търсене на основната форма на словоформата, още по-неприятен е случаят, когато дадена дума не може да бъде намерена в текста поради разлики в правописа.

Нека разгледаме един пример на старобългарски език. Лексемата **дѫбъ** може да се появи в поне три варианта поради особеностите на използване на еровите гласни: **дѫбъ, дѫбь** и **дѫб᾽**. Като се вземат предвид особеностите на използване на голямата носовка, се очертава възможната употреба на формите **дѫбъ** и **доубъ**.

По този начин, като отчетем вариативността при еровете и при голямата носовка, ще имаме общо шест варианта на написване на една четирибуквена дума на стробългарски език. Ако добавим и различни графични варианти на ѡ҃ (напр. ѵ҃ и ꙋ҃) при положение че те могат да се използват при набирането на текста, можем да стигнем до поне 12 правилни начина на написване на една четирибуквена дума. Тук разгледахме само един пример, за да дадем представа за основния проблем при наличие на правописна вариативност: всички компютърни операции, свързани с търсене на дума, част от дума или фраза в текст ще водят до изопачени резултати, ако вариативността не бъде моделирана по подходящ начин.

Това положение изисква разработването и прилагането на модели, които да се съобразяват с езиковата вариативност. Възникнала първоначално като съществен подраздел на социолингвистиката и психолингвистиката, днес езиковата вариативност постепенно се преориентира към използването на инструментариума на лингвостатистическите изследвания.

Тук следва да подчертаем съществената разлика в изследванията на вариативността в съвременните езици и в средновековните писмени паметници. Докато силното влияние на социолингвистиката води до повишен интерес към такива фактори, причиняващи вариативност в съвременните езици, като социалната група, расата или пола (вж. Фиг. 3.2), то вариативността в средновековните текстове се причинява от по-различни фактори . Те са представени на Фиг. 3.3.

На Фиг. 3.3 вляво са оцветени областите, за които изследването на вариативността на средновековните писмени паметници е съществено. В светлосиньо са оцветени традиционните области на изследване на вариативност. Добавили сме два фактора, причиняващи вариативност. Нарекли сме ги "история на текста",

като тук се има предвид процесът на създаване и последователно преписване на определен писмен паметник, и "'графика и украса", като в този случай имаме предвид онези компоненти, които се отнасят до начина на изписване на буквите, орнаментите, инициалите и т.п. Тяхното изследване, съотнесено към дадено езиково ниво, е представено чрез оцветяване в червено.

Фиг. 3.3 . Фактори и нива на езикова вариативност в средновековните писмени паметници

| Езиково ниво / ФАКТОР | Фонетика | Правопис | Лексика | Синтаксис | Семантика |
|---|---|---|---|---|---|
| Пол | | | | | |
| Социална група | | | | | |
| Етническа група | | | | | |
| Времеви период | | | | | |
| Жанр/стил | | | | | |
| Без определен фактор | | | | | |
| История на текста | | | | | |
| Графика и украса | | | | | |

Някои основни различия между изследванията на средновековни и съвременни текстове се състоят в следното:

- Фонетичното езиково ниво не може да се постави в групата на активно изследваните при средновековните писмени паметници. Такива социолингвистични фактори като *пол* и *етническа група* представляват по-голям интерес за съвременните езици. За тях съществуват сравнително малко данни при средновековните текстове, затова не сме ги отбелязали като фактори, подлежащи на изследване.
- Има и още една съществена особеност на вариативността в

средновековните текстове. Тя е обект на по-силен интерес от страна на изследователите в сравнение със съвременните езици поради самата специфика на изучавания обект, като ударението пада не толкова върху *факторите*, които я причинявят, а върху *проявите* на вариативността в текстовете. За съжаление в тази област няма почти никакви компютърни средства, които да могат да облекчат изследователската работа. Можем да посочим поне две причини за това:

- Вариативността в съвременните езици не е навлязла в традиционните компютърни средства и поради това няма опит в създаването на такива средства.
- Няма търговски интерес в създаването на подобни компютърни средства, защото те са насочени към ограничена потребителска група, която при това е с недостатъчни финансови ресурси. Това означава, че създаването на търговски компютърни системи е нерентабилно, т .е. ако се очаква подобно моделиране, то ще дойде от страна на изследователите, които имат нужда от подобни средства, а не от фирмите, които произвеждат програмни продукти.

Досега са правени два опита за моделиране на явления от средновековните английски текстове с помощта на компютър. В [Barnbrook 92] е разгледан проблемът за правописната вариативност в "Кентърберийски разкази" на Чосър. Представен е модел на различията в правописни варианти . В [Robertson, Willet 94] е разгледан същият проблем от гледна точка на използването на бази от данни за думите в стари английски текстове. Докато Барнбрук се занимава с моделиране на различията в коректни правописни форми, във втората цитирана публикация се изследват

информатични подходи от областта на коригирането на правописа, които могат да послужат при моделирането на различни правописни форми в старите текстове.

Бегло е разгледан проблемът за вариативността в старофренски и старохоландски текстове в [Huber 89]. Там не са предложени средства за анализ на вариативността, но при въвеждането на текстовете всяка дума получава трицифрен код, който носи граматическа и семантична информация. Този подход въпреки предимствата при следващата обработка на текстовете е много трудоемък при въвеждането на текстове. Нямаме данни за неговото развитие в следващите години.

Някои проблеми на вариативността в ръкописите са разгледани в изследванията, публикувани в сборника [van Reenen 88]. Сред търсенето на информатични модели може да се поставят изследванията, свързани с автоматично създаване на стема[1] на ръкопис.

След този кратък преглед на състоянието на изследванията в областта на езиковата вариативност, можем да отбележим следната тенденция: докато интересът към изследване на езиковата вариативност нараства, компютърните средства все още се прилагат много слабо в разглежданата област. Една от причините за това състояние в момента е слабото развиване на формализирани модели в областта.

### *3.1.4.Компютърни корпуси и вариативност*

Бихме искали да се спрем отделно на някои проблеми, свързани с моделирането на вариативността в компютърни корпуси от текстове. В тази част ще направим кратък преглед на най-използваните корпуси от текстове в областта на античните и

---

[1] Представане на историята на разпространение на даден ръкописен текст във вид на дърво.

средновековни текстове (вж. Табл. 3.1. и 3.2.).

Таблица 3.1. Корпуси от текстове

| Данни / Наименование | Съкратено наименование | Източник |
|---|---|---|
| 1. Thesaurus Linguae Graecae | TLG | [Helgerson 88] |
| 2 . Isocrates | I | [Helgerson 88] |
| 3. Библия на CD-ROM | B | [Helgerson 88] |
| 4. Corpus dei Manoscritti Copti Letterarl | CMCL | [Orlandi 90] |
| 5. Perseus | P | [Helgerson 88] |
| 6. Изследователски компютър Ibycus | ISC | [Helgerson 88] |
| 7. Работно място на филолога[2] | PW | [Caligaris 92] |
| 8. Текст-критика с компютър[2] | TC | [Bozzi et al. 86] |
| 9. The Century of Prose Corpus | CPC | [Milic 95] |
| 10. Старофренски и старохоландски корпус[2] | OFD | [Huber 86] |

Сред най-значимите подобни корпуси са TLG - Thesaurus Linguae Graecae [Helgerson 88] и CMCL - Corpus dei Manoscritti Copti Letterari [Orlandi 90]. Обикновено в подобни корпуси се съхраняват само текстове, но не и изображения на ръкописи. Тъй като началото на работата по посочените два корпуса е поставено през 60-те години, текстовете са набирани ръчно чрез клавиатура (от средата на 80-те години насам се предпочита сканиране на текстовете, вж. напр. [Bozzi et al. 86]).

Важна стъпка при проектирането на корпусите е било

взимането на решение за кодиране на текстовете. В TLG например с оглед на по-лесното търсене на думи и фрази в бъдеще всички текстове са въведени, като е използван т.нар. бета-код, който представлява транслитерация на гръцката азбука на латиница. Всички букви се въвеждат като главни латински букви според таблица за транскрибиране на старогръцките текстове на латиница. В CMCL също е използвана латинска транскрипция за предаване на коптската азбука.

Таблица 3.2. Данни за компютърни корпуси
на антични и средновековни текстове

| Корпус | Размер | Създаден | Институция |
|---|---|---|---|
| TLG | 62 млн думи<br>212 мв | 1972-прод. | Университет в Ървин, САЩ |
| I | TLG +125 мв<br>индекси | 1985-прод. | Университет в Браун, САЩ |
| CMCL | | 1968-прод. | Университет "La Sapienza" – Рим |
| P | 100 MB | 3-4 г. | Харвардски университет |
| TKA, PW | | 1975-прод. | Институт за компютърна лингвистика, Пиза |
| CPC | 500 хил.<br>Думи | Няма данни | Кливлънд, САЩ |

От изброените корпуси само CMCL и TC са създавани с оглед на отразяване на вариативността. При набора на текстовете от TLG са използвани издания с нормализирани текстове. Текст-критичният апарат не е въведен.

При CMCL са въвеждани текстовете по оригинални ръкописи

---

[2] Това наименование е въведено от мен, М.Д., тъй като в цитираната публикация няма специално предложено име.

във вид максимално близък до това, което може да се разчете на ръкописа. По време на работата по проекта ТС са изследвани произведенията на римския поет Клавдиан, като е взета предвид вариативността в преписите. В момента в Института по компютърна лингвистика в Пиза, Италия, продължава разработването на среда за подпомагане на работата с текст-критичен апарат [Camuglia 96b].

Представените данни за отразяването на вариативността в корпуси от текстове свидетелстват за особеното положение с моделирането на вариативността в компютърен вид. Пред опасността да не може да се използват възможностите за пълнотекстово търсене поради наличието на вариативност, дизайнерите на компютърни корпуси предпочитат създаването на осакатени в определен смисъл корпуси, където е проведена нормализация. В някои случаи това решение е продиктувано от особеностите на класическата писмена традиция, където вариативността има скромни измерения в сравнение със славянските текстове. Особеностите на компютърното обработване на славянски текстове е разгледано в следващата част на настоящия труд.

### *3.1.3. Езикови нива и фактори при моделирането на вариативност в старобългарски текстове*

Старобългарските текстове, както всеки тип текстове на естествени езици, могат да се изучават на различни езикови нива. Например датирането на даден ръкопис се извършва въз основа на набор от признаци, които се проследяват от изследователите.

При изучаването на старобългарски текстове интерес представлява и влиянието на някои особени фактори, сред които диахронното развитие; влиянието на различни езици върху старобългарския език; индивидуалните промени, внасяни от различни писари и др.

Особеностите на даден ръкопис се описват при неговото

издание. Така чрез наблюденията над голям брой издания могат да се обобщят признаци на определени езикови нива, които да покажат влиянието на определени фактори.

На ниво *правопис* може да се изследва употребата на определено множество от букви в даден ръкопис (т.е. може да се опише азбуката на всеки един ръкопис). Освен това може да се проследи употребата на определени букви или групи от букви, които са характерни за дадено време или място на създаване на ръкопис.

Най-често решаваните задачи с помощта на такива наблюдения са класификационни – въз основа на употребата на буквите в определен ръкопис да се формулира хипотеза за времето и мястото на създаването му. Тези въпроси са предмет на разглеждане в част 3.2.

По подобна методика могат да се извършват и изследвания на лексикално и синтактично ниво. Подобни изследвания биха били възможни при натрупването на повече текстове в електронен вид, тъй като прилагането на количествени методи е обвързано с наблюдения върху извадки, които трябва да отговарят на определени изисквания за обем.

## 3.2. Представяне и изследване на правописна вариативност

### *3.2.1. Постановка на задачата*

Досега не са извършвани широкообхватни изследвания на средновековните славянски текстове със статистически методи. (Единствените данни, които можахме да намерим, се отнасят за употребата на буквата ***А*** в българския език в диахронен аспект. Изследването е на проф. М. Янакиев и включва данни за някои средновековни паметници. Сериозна поредица от лингвостатистически изследвания у нас е правила Л. Бонева, но те не се отнасят до

старобългарския езиков материал.). Същевременно средновековните славянски текстове се отличават с редица особености, които до момента се отбелязват от изследователите, без да се посочват точни количествени данни. Ето някои от тези особености:

- Буквеният състав на ръкописи, писани на стара кирилица[3], не е еднакъв. Нямаме сведения да са провеждани мащабни изследвания, свързани със съпоставянето на буквения състав на средновековните славянски паметници. Същевременно до момента не е създаден общоприет стандарт за компютърно представяне на старата кирилица, като една от причините е липсата на единно мнение на изследователите за всички буквени и небуквени символи, които трябва да се включат в компютърното кодиране на старата кирилица и глаголицата.
- В различни старобългарски граматики са представени правописните особености, характерни за редакции или ръкописни школи. Най-често посочваните особености се отнасят за употребата на еровите гласни, носовките, йотуваните гласни и някои буквени съчетания, но няма пълно единство на мненията и общоприета класификация в областта,

Представеното тук изследване цели да се съберат за пръв път и изследват със специализирани статистически методи данните за употребата на буквите от старата кирилица в паметници от различно време, характеризирани като българска, руска и сръбска редакции. Само по себе си, събирането на подобни данни е новост в изследванията на средновековните славянски паметници. Подобно изследване, проведено върху широк кръг от ръкописи, би

спомогнало например да се уточнят графемите, които трябва да присъстват в компютърното кодиране на старата кирилица. Освен това то би помогнало при разработването на алгоритми за автоматично разпознаване на средновековни текстове, при които се използват вероятностите за поява на определени букви и буквени съчетания.

По-съществено обаче в нашата работа е търсенето на нови данни за употребата на онези букви и буквени съчетания, които се посочват като характерни признаци на посочените три редакции.

Целта е да бъде намерено минималното множество от характеризиращи признаци въз основа на обработката на количествени данни в съпоставка с посочваните в езиковедските изследвания качествени признаци. В настоящото изследване сме използвали като такива качествени признаци особеностите, посочени в раздел "Разпространение на старобългарския книжовен език като официален език у други славянски народи, Редакции (изводи) на старобългарския книжовен език" [Граматика 93, стр.36-38].

### *3.2.2. Постановка на експеримента*

#### *3.2.2.1. Извадков материал*

За провеждането на експеримента бяха подбрани по различен начин следните ръкописи, съдържащи 15 еднакви псалма от 7 различни псалтира:

- Синайски (българска редакция; 10 век);
- Болонски (българска редакция, 13 век);
- Норовски (руска редакция, 13 век);
- Сръбски (сръбска редакция; 13 век);

---

[3] В това изследване няма да разглеждаме ръкописи, писани на глаголица, с изключение на един транскрибиран на кирилица текст.

- Киевски (руска редакция; 14 век);
- Генадиевски (руска редакция, 15 век);
- Църковнославянски (съвременна версия).

Работихме с извадка с обем 15 приблизително еднакви по големина ексцерпти (текстови фрагменти) от всяка от изследваните редакции, като големината на ексцерптите вариираше от 1 до 14 случайно подбрани псалма. Като предпочитани и оптимални за изследване големини се оформиха ексцерптите с големина един, два и пет псалма. Тези големини бяха подбрани на базата на предишни изследвания с конструиране на многобройни кутийки с мустачки за един и същи текст, но с различни по големина ексцерпти.

Най-малките ни ексцерпти, тези с големина 1 псалм, включваха едни и същи псалми от 7-те ръкописа (номера 1, 39, 40; 41, 44, 45, 64, 73, 74, 75, 89, 91, 98, 102, 134). Изключение прави Сръбският псалтир, в който не е запазен псалм N 1. Псалмите са подбрани специално така, че да не са строго последователни.

Тази големина на минималните изследвани ексцерпти се доближава до предпочитаната при някои филологически изследвания големина с размер 1 килолитера. Същевременно, при нашата постановка се изследват единни текстови цялости, за разлика от килолитерите, при които извадката се взима по формален признак (1000 букви). Работата с единни текстови цялости моделира достатъчно добре предметната област, в която и ръкописните фрагменти от 1 лист обикновено съдържат фрагмент с подобна дължина. При тази постановка изследванията се извършват на базата на относителните, а не абсолютните честоти на употреба на различните изучавани елементи. Съвременните компютърни статистически средства облекчават в изключителна степен подобни експерименти. В нашата работа използвахме статистическия пакет Statistica for Windows.

Изследваните текстове - псалтирите, принадлежат към богослужебните книги. При тях текстовата вариативност се смята за по-ниска в сравнение с небогослужебните произведения поради специалните функции на текста.

*3.2.2.2. Извличане на данни от текстовете*

Първоначално проведохме изследването, като използвахме резултатите от статистическата обработка на текстови ексцерпти с програмата TACT (Textual Analysis Computing Tool), която е разработена в Университета вТоронто и се разпространява безплатно като инструмент за изследователска работа.

Необходимостта да се обработят 104 ексцерпта и получените данни да се импортират в статистически пакет доведе до някои неудобства при използването на TACT. Затова беше разработен програмен инструмент за преброяване на употребите на различните букви и предварително зададени от изследователя буквени съчетания  Предимствата му са: лесна работа с нестандартни азбуки; възможност за задаване на буквени съчетания (нещо изключително важно за филологическите изследвания) и обработка на текстовете в пакетен режим [Dobreva, Dobrev 98].

Като изходен резултат се получават броят употреби на всяка буква или буквено съчетание в началото надума, в края на дума и общо. Освен тези данни се получава и списък на всички символи, които не са представени в зададената от изследователя азбука, което позволява да се контролира зададената азбука. Катосе използват данните за общия брой букви в текстовия ексцерпт , се изчисляват *относителните честоти*, които се използват в собствено статистическото изследване.

*3.2.2.3. Проведени изследвания*

**И1. Изследване на честотите на употреба на буквите в**

**текстовете.**

Целта е да се определи минимална текстова извадка, при която се получават устойчиви резултати за честотата на употреба на буквите.

Резултатите са съществени при решаване на следната практическа задача: колко голям фрагмент от ръкопис трябва да се изследва, за да са достоверни резултатите.

**И2. Изследване на правописа на текст, базирано на честотата на употреба на характерните за дадена правописна школа особености.**

Целите при провеждането на тази серия от експерименти са:

- Да се установи при какъв обем на текстовия фрагмент данните за буквеното разпределение представят резултат , съпоставим с разпределението за целия текст.
- Да се изследва възможно ли е разделянето на буквите от азбуката на 2 групи: такива, които дават съществени отклонения в честотата на употреба в различни текстове, и такива, които имат устойчиво поведение.
- Да се изследва специално поведението на еровите гласни, носовките, йотуваните гласни и техните възможни заместители (в тази част сме изследвали поведението на група от признаци, която отразява описаните в [Граматика 93] признаци, характеризиращи редакциите, а именно:
    - носовите гласни ѫ и ѧ и техните възможни промени в оу и ѣ в руска и в оу и ѥ в сръбска редакция;
    - употреба на една ерова гласна в сръбска редакция;
    - замяна на групите -рд- -лд- -рѣ- лѣ с -оро- -ѥлѥ- -оро- -ѥлѥ- в руска редакция;

- преминаване на средисловно -въ- между съгласни в -ъв- в сръбска редакция,
- замяна на шт и жд с ч и ж в руска редакция,
- замяна на начално ѥ с о в руска редакция.

За всички букви и низове, които се появяват в тези групи, бяха пресметнати също честотите на употреба в текстовете, използвани при експериментите.

*3.2.2.4. Данни за разпределението на буквените честоти по редакции*

В Табл. 3.3. са представени за пръв път данни за разпределение на буквените честоти и качествените признаци според редакциите на паметниците, получени при експериментите ни.

Представените резултати показват ясно, че най-силно варира употребата на гласните букви. Фактът, че стандартното отклонение за употребата им, получено при изследване на текстови извадки от всички редакции, обикновено е два пъти и повече по-високо от стандартното отклонение при изследването на текстове от една и съща редакция, показва, че при различните редакции се наблюдава нееднородност при употребата на гласните букви.

Ще разгледаме особеностите на данните за честотата на употреба на различните букви по редакции. При разглеждането сме се спрели на букви с честота на употреба, чието стандартно отклонение е по-високо от 0,3.

| Буква/низ | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | РУСКА РЕДАКЦИЯ | | | | БЪЛГАРСКА РЕДАКЦИЯ | | | | СРЪБСКА РЕДАКЦИЯ | | | | ВСИЧКИ РЕДАКЦИИ | | | |
| а | 6.74 | 6.11 | 7.61 | 0.22 | 6.04 | 5.21 | 7.35 | 0.38 | 6.49 | 5.67 | 7.07 | 0.24 | 6.46 | 4.41 | 8.65 | 0.55 |
| б | 1.75 | 1.29 | 2.08 | 0.11 | 1.69 | 1.38 | 2.17 | 0.11 | 1.78 | 1.36 | 2.20 | 0.12 | 1.73 | 0.84 | 2.79 | 0.20 |
| в | 5.70 | 4.94 | 6.55 | 0.22 | 5.43 | 4.59 | 6.32 | 0.22 | 5.66 | 5.11 | 6.28 | 0.22 | 5.59 | 3.11 | 7.21 | 0.41 |
| г | 2.56 | 1.83 | 3.62 | 0.25 | 2.45 | 1.59 | 3.30 | 0.23 | 2.54 | 1.92 | 3.36 | 0.25 | 2.53 | 1.27 | 4.27 | 0.40 |
| д | 3.35 | 2.81 | 3.88 | 0.18 | 3.27 | 2.66 | 3.81 | 0.17 | 3.56 | 0.00 | 4.08 | 0.35 | 3.35 | 0.00 | 4.92 | 0.34 |
| е | 6.37 | 5.62 | 7.22 | 0.21 | 5.47 | 4.83 | 5.99 | 0.23 | 6.12 | 5.07 | 6.75 | 0.25 | 6.03 | 3.95 | 9.36 | 0.57 |
| ж | 1.13 | 0.57 | 1.59 | 0.13 | 0.96 | 0.63 | 1.35 | 0.11 | 0.99 | 0.56 | 1.41 | 0.12 | 1.05 | 0.28 | 1.84 | 0.21 |
| ѕ | 0.04 | 0.00 | 0.14 | 0.03 | 0.06 | 0.00 | 0.19 | 0.05 | 0.01 | 0.00 | 0.02 | 0.01 | 0.04 | 0.00 | 0.44 | 0.05 |
| з | 1.77 | 1.20 | 2.20 | 0.12 | 1.70 | 1.43 | 2.11 | 0.11 | 1.80 | 1.40 | 2.18 | 0.12 | 1.75 | 0.64 | 2.78 | 0.22 |
| и | 8.35 | 7.41 | 9.21 | 0.30 | 5.44 | 1.65 | 9.30 | 3.15 | 8.96 | 8.38 | 9.63 | 0.23 | 7.46 | 0.16 | 10.13 | 2.38 |
| і | 0.45 | 0.09 | 1.02 | 0.16 | 2.51 | 0.04 | 6.12 | 2.40 | 0.20 | 0.06 | 0.34 | 0.04 | 1.11 | 0.00 | 7.17 | 1.74 |
| ї | 0.34 | 0.00 | 0.76 | 0.19 | 0.50 | 0.00 | 1.85 | 0.51 | 0.01 | 0.00 | 0.02 | 0.01 | 0.35 | 0.00 | 4.35 | 0.40 |
| н | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.07 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.55 | 0.02 |

| Буква/низ | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | РУСКА РЕДАКЦИЯ | | | | БЪЛГАРСКА РЕДАКЦИЯ | | | | СРЪБСКА РЕДАКЦИЯ | | | | ВСИЧКИ РЕДАКЦИИ | | | |
| ѫ | 0.00 | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.01 |
| к | 2.12 | 1.62 | 2.50 | 0.13 | 2.07 | 1.68 | 2.48 | 0.12 | 2.13 | 1.82 | 2.39 | 0.12 | 2.11 | 1.03 | 3.22 | 0.21 |
| л | 3.12 | 2.38 | 3.75 | 0.20 | 2.97 | 2.35 | 3.63 | 0.18 | 3.18 | 2.78 | 3.64 | 0.16 | 3.07 | 1.09 | 4.35 | 0.33 |
| м | 3.24 | 2.14 | 4.11 | 0.27 | 3.19 | 2.36 | 4.10 | 0.27 | 3.37 | 2.57 | 3.98 | 0.24 | 3.25 | 0.91 | 6.34 | 0.49 |
| н | 4.64 | 4.00 | 5.20 | 0.20 | 4.52 | 3.86 | 5.09 | 0.18 | 4.57 | 3.97 | 5.13 | 0.18 | 4.60 | 3.22 | 7.59 | 0.35 |
| о | 8.93 | 7.42 | 10.21 | 0.56 | 8.87 | 7.74 | 10.22 | 0.45 | 8.82 | 7.82 | 9.88 | 0.30 | 8.87 | 6.29 | 12.35 | 0.67 |
| ѡ | 0.79 | 0.34 | 1.24 | 0.20 | 0.43 | 0.02 | 0.97 | 0.36 | 0.76 | 0.59 | 0.89 | 0.06 | 0.66 | 0.00 | 1.80 | 0.33 |
| п | 2.76 | 2.41 | 3.16 | 0.12 | 2.77 | 2.48 | 3.07 | 0.10 | 2.85 | 2.66 | 3.10 | 0.08 | 2.78 | 1.89 | 4.10 | 0.20 |
| р | 3.44 | 2.67 | 4.47 | 0.26 | 3.41 | 2.72 | 4.18 | 0.25 | 3.48 | 2.39 | 4.29 | 0.28 | 3.43 | 1.85 | 5.40 | 0.42 |
| с | 6.31 | 5.42 | 7.02 | 0.26 | 6.17 | 5.41 | 6.93 | 0.22 | 6.40 | 5.91 | 7.45 | 0.22 | 6.29 | 3.67 | 8.18 | 0.42 |
| т | 5.40 | 4.16 | 6.38 | 0.30 | 5.42 | 4.43 | 6.39 | 0.28 | 5.17 | 4.19 | 6.07 | 0.31 | 5.37 | 3.07 | 7.38 | 0.52 |
| оу | 2.12 | 1.10 | 3.14 | 0.55 | 1.25 | 0.79 | 1.78 | 0.14 | 2.54 | 2.02 | 3.54 | 0.23 | 1.91 | 0.46 | 4.73 | 0.70 |
| ѳ | 0.01 | 0.00 | 0.09 | 0.02 | 0.00 | 0.00 | 0.02 | 0.01 | 0.02 | 0.00 | 0.05 | 0.01 | 0.01 | 0.00 | 0.26 | 0.02 |

| Буква/низ | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | РУСКА РЕДАКЦИЯ | | | | БЪЛГАРСКА РЕДАКЦИЯ | | | | СРЪБСКА РЕДАКЦИЯ | | | | ВСИЧКИ РЕДАКЦИИ | | | |
| ѵ | 0.00 | 0.00 | 0.02 | 0.00 | 0.03 | 0.00 | 0.11 | 0.02 | 0.01 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.19 | 0.02 |
| Ф | 0.02 | 0.00 | 0.09 | 0.02 | 0.02 | 0.00 | 0.06 | 0.01 | 0.04 | 0.00 | 0.09 | 0.02 | 0.02 | 0.00 | 0.28 | 0.03 |
| ф | 0.06 | 0.00 | 0.13 | 0.03 | 0.03 | 0.00 | 0.08 | 0.02 | 0.02 | 0.00 | 0.06 | 0.01 | 0.04 | 0.00 | 0.28 | 0.04 |
| Х | 1.09 | 0.71 | 1.42 | 0.10 | 1.07 | 0.76 | 1.39 | 0.10 | 1.06 | 0.73 | 1.36 | 0.11 | 1.08 | 0.30 | 1.95 | 0.18 |
| Ш | 0.59 | 0.23 | 0.92 | 0.13 | 0.38 | 0.02 | 0.72 | 0.09 | 0.62 | 0.44 | 0.85 | 0.07 | 0.52 | 0.00 | 1.37 | 0.19 |
| Ц | 0.63 | 0.37 | 0.99 | 0.11 | 0.62 | 0.39 | 0.89 | 0.09 | 0.61 | 0.41 | 0.79 | 0.08 | 0.62 | 0.18 | 1.73 | 0.16 |
| ч | 0.51 | 0.31 | 0.79 | 0.07 | 0.50 | 0.38 | 0.66 | 0.05 | 0.51 | 0.37 | 0.68 | 0.06 | 0.51 | 0.21 | 1.27 | 0.12 |
| ҵ | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.01 |
| ш | 1.20 | 0.69 | 1.65 | 0.14 | 1.31 | 0.00 | 1.76 | 0.15 | 1.16 | 0.90 | 1.58 | 0.11 | 1.24 | 0.00 | 2.72 | 0.24 |
| ъ | 3.31 | 1.82 | 5.52 | 0.88 | 7.57 | 6.74 | 8.38 | 0.27 | 0.18 | 0.07 | 0.33 | 0.04 | 4.28 | 0.00 | 11.09 | 2.71 |
| ъі | 0.00 | 0.00 | 0.03 | 0.01 | 1.11 | 0.76 | 1.31 | 0.08 | 0.01 | 0.00 | 0.02 | 0.01 | 0.38 | 0.00 | 1.67 | 0.53 |
| ъи | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ьи | 0.00 | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 | 0.15 | 0.01 |

| Буква/низ | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | РУСКА РЕДАКЦИЯ | | | | БЪЛГАРСКА РЕДАКЦИЯ | | | | СРЪБСКА РЕДАКЦИЯ | | | | ВСИЧКИ РЕДАКЦИИ | | | |
| Ъї | 1.32 | 1.04 | 1.64 | 0.09 | 0.14 | 0.00 | 0.44 | 0.08 | 0.89 | 0.65 | 1.11 | 0.08 | 0.86 | 0.00 | 2.02 | 0.55 |
| Ы | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.07 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.01 |
| Ь | 2.58 | 1.40 | 3.71 | 0.69 | 1.54 | 0.20 | 2.73 | 0.75 | 6.39 | 6.10 | 7.00 | 0.13 | 2.82 | 0.00 | 7.53 | 1.76 |
| Ѣ | 2.57 | 2.06 | 3.40 | 0.30 | 3.57 | 2.70 | 4.80 | 0.69 | 2.74 | 2.48 | 3.02 | 0.10 | 2.94 | 1.61 | 5.41 | 0.68 |
| Ю | 0.70 | 0.07 | 1.38 | 0.39 | 0.28 | 0.07 | 0.48 | 0.08 | 0.91 | 0.69 | 1.09 | 0.08 | 0.58 | 0.00 | 1.99 | 0.38 |
| Ій | 0.88 | 0.23 | 1.95 | 0.56 | 0.49 | 0.00 | 1.21 | 0.50 | 1.20 | 0.90 | 1.46 | 0.09 | 0.80 | 0.00 | 2.33 | 0.56 |
| Є | 0.05 | 0.00 | 0.11 | 0.02 | 0.26 | 0.00 | 1.14 | 0.30 | 2.19 | 1.76 | 2.96 | 0.19 | 0.45 | 0.00 | 3.45 | 0.79 |
| Ѧ | 2.20 | 1.62 | 3.18 | 0.40 | 0.88 | 0.00 | 1.98 | 0.86 | 0.02 | 0.00 | 0.05 | 0.01 | 1.41 | 0.00 | 4.00 | 1.03 |
| Ѫ | 0.78 | 0.00 | 2.70 | 1.02 | 1.77 | 0.80 | 2.74 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 4.18 | 1.02 |
| Ѭ | 0.01 | 0.00 | 0.06 | 0.01 | 1.24 | 0.00 | 2.86 | 1.24 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.00 | 3.86 | 0.93 |
| Ѯ | 0.01 | 0.00 | 0.05 | 0.01 | 0.50 | 0.00 | 0.94 | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.00 | 2.26 | 0.29 |
| Ѱ | 0.05 | 0.00 | 0.13 | 0.03 | 0.03 | 0.00 | 0.14 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.42 | 0.04 |
| ь | 2.58 | 1.40 | 3.71 | 0.69 | 1.54 | 0.20 | 2.73 | 0.75 | 6.39 | 6.10 | 7.00 | 0.13 | 2.82 | 0.00 | 7.53 | 1.76 |

| Буква/низ | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | РУСКА РЕДАКЦИЯ | | | | БЪЛГАРСКА РЕДАКЦИЯ | | | | СРЪБСКА РЕДАКЦИЯ | | | | ВСИЧКИ РЕДАКЦИИ | | | |
| Ъ | 3.31 | 1.82 | 5.52 | 0.88 | 7.57 | 6.74 | 8.38 | 0.27 | 0.18 | 0.07 | 0.33 | 0.04 | 4.28 | 0.00 | 11.09 | 2.71 |
| Ѫ | 0.78 | 0.00 | 2.70 | 1.02 | 1.77 | 0.80 | 2.74 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 4.18 | 1.02 |
| Ѹ | 2.12 | 1.10 | 3.14 | 0.55 | 1.25 | 0.79 | 1.78 | 0.14 | 2.54 | 2.02 | 3.54 | 0.23 | 1.91 | 0.46 | 4.73 | 0.70 |
| Ѳ | 0.01 | 0.00 | 0.09 | 0.02 | 0.00 | 0.00 | 0.02 | 0.01 | 0.02 | 0.00 | 0.05 | 0.01 | 0.01 | 0.00 | 0.26 | 0.02 |
| Ѵ | 0.00 | 0.00 | 0.02 | 0.00 | 0.03 | 0.00 | 0.11 | 0.02 | 0.01 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.19 | 0.02 |
| Ѧ | 2.20 | 1.62 | 3.18 | 0.40 | 0.88 | 0.00 | 1.98 | 0.86 | 0.02 | 0.00 | 0.05 | 0.01 | 1.41 | 0.00 | 4.00 | 1.03 |
| Ꙗ | 0.88 | 0.23 | 1.95 | 0.56 | 0.49 | 0.00 | 1.21 | 0.50 | 1.20 | 0.90 | 1.46 | 0.09 | 0.80 | 0.00 | 2.33 | 0.56 |
| Ѣ | 2.57 | 2.06 | 3.40 | 0.30 | 3.57 | 2.70 | 4.80 | 0.69 | 2.74 | 2.48 | 3.02 | 0.10 | 2.94 | 1.61 | 5.41 | 0.68 |
| Ѥ | 6.37 | 5.62 | 7.22 | 0.21 | 5.47 | 4.83 | 5.99 | 0.23 | 6.12 | 5.07 | 6.75 | 0.25 | 6.03 | 3.95 | 9.36 | 0.57 |
| ВЬ | 0.40 | 0.00 | 1.06 | 0.28 | 0.31 | 0.00 | 0.87 | 0.29 | 1.48 | 1.26 | 1.78 | 0.09 | 0.54 | 0.00 | 2.60 | 0.51 |
| ЬВ | 0.01 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.05 | 0.01 | 0.05 | 0.00 | 0.11 | 0.02 | 0.02 | 0.00 | 0.35 | 0.03 |
| ШТ | 0.07 | 0.00 | 0.50 | 0.11 | 0.22 | 0.02 | 0.51 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.89 | 0.15 |
| Щ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Ү | 0.51 | 0.31 | 0.79 | 0.07 | 0.50 | 0.38 | 0.66 | 0.05 | 0.51 | 0.37 | 0.68 | 0.06 | 0.51 | 0.21 | 1.27 | 0.12 |
| ЖД | 0.08 | 0.00 | 0.17 | 0.04 | 0.09 | 0.02 | 0.16 | 0.02 | 0.10 | 0.04 | 0.15 | 0.02 | 0.08 | 0.00 | 0.35 | 0.05 |
| Ж | 1.13 | 0.57 | 1.59 | 0.13 | 0.96 | 0.63 | 1.35 | 0.11 | 0.99 | 0.56 | 1.41 | 0.12 | 1.05 | 0.28 | 1.84 | 0.21 |

| Буква/низ | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | РУСКА РЕДАКЦИЯ | | | | БЪЛГАРСКА РЕДАКЦИЯ | | | | СРЪБСКА РЕДАКЦИЯ | | | | ВСИЧКИ РЕДАКЦИИ | | | |
| ѓя- | 1.01 | 0.70 | 1.38 | 0.10 | 0.96 | 0.65 | 1.28 | 0.09 | 1.03 | 0.75 | 1.32 | 0.09 | 0.99 | 0.31 | 1.73 | 0.16 |
| ѓо- | 0.38 | 0.15 | 0.69 | 0.08 | 0.42 | 0.13 | 0.69 | 0.08 | 0.48 | 0.21 | 0.69 | 0.08 | 0.41 | 0.00 | 1.07 | 0.14 |
| ля- | 0.70 | 0.42 | 0.95 | 0.08 | 0.63 | 0.38 | 0.87 | 0.08 | 0.71 | 0.43 | 0.92 | 0.09 | 0.67 | 0.00 | 1.40 | 0.16 |
| ло- | 0.50 | 0.23 | 0.80 | 0.08 | 0.52 | 0.28 | 0.76 | 0.07 | 0.51 | 0.23 | 0.74 | 0.08 | 0.51 | 0.00 | 1.25 | 0.14 |
| -ѓя- | 1.01 | 0.70 | 1.38 | 0.10 | 0.96 | 0.65 | 1.28 | 0.09 | 1.03 | 0.75 | 1.32 | 0.09 | 0.99 | 0.31 | 1.73 | 0.16 |
| -оро- | 0.01 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.03 | 0.01 | 0.02 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.12 | 0.01 |
| -ля- | 0.70 | 0.42 | 0.95 | 0.08 | 0.63 | 0.38 | 0.87 | 0.08 | 0.71 | 0.43 | 0.92 | 0.09 | 0.67 | 0.00 | 1.40 | 0.16 |
| -оло- | 0.02 | 0.00 | 0.06 | 0.01 | 0.02 | 0.00 | 0.06 | 0.01 | 0.03 | 0.00 | 0.07 | 0.01 | 0.02 | 0.00 | 0.16 | 0.02 |
| -рѣ- | 0.42 | 0.21 | 0.85 | 0.13 | 0.63 | 0.47 | 0.82 | 0.06 | 0.57 | 0.34 | 0.76 | 0.06 | 0.51 | 0.00 | 1.41 | 0.17 |
| -єре- | 0.01 | 0.00 | 0.05 | 0.01 | 0.01 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.16 | 0.01 |
| -лѣ- | 0.18 | 0.05 | 0.38 | 0.06 | 0.26 | 0.12 | 0.47 | 0.06 | 0.25 | 0.11 | 0.36 | 0.05 | 0.22 | 0.00 | 1.00 | 0.10 |
| -єлє- | 0.03 | 0.00 | 0.08 | 0.01 | 0.04 | 0.00 | 0.07 | 0.01 | 0.02 | 0.00 | 0.06 | 0.01 | 0.03 | 0.00 | 0.16 | 0.02 |
| ıє- | 0.05 | 0.00 | 0.11 | 0.02 | 0.26 | 0.00 | 1.14 | 0.30 | 2.19 | 1.76 | 2.96 | 0.19 | 0.45 | 0.00 | 3.45 | 0.79 |
| о- | 8.93 | 7.42 | 10.21 | 0.56 | 8.87 | 7.74 | 10.22 | 0.45 | 8.82 | 7.82 | 9.88 | 0.30 | 8.87 | 6.29 | 12.35 | 0.67 |

Табл. 3.3. Основни статистически данни за честотите на старобългарските букви и качествени признаци в българска, руска и сръбска редакции

Табл. 3.4. Букви с най-високи стандартни отклонения при честотата на употреба - сръбска редакция

| Буква | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|
| д | 3.56 | 0.00 | 4.08 | 0.35 |
| т | 5.17 | 4.19 | 6.07 | 0.31 |
| о | 8.82 | 7.82 | 9.88 | 0.30 |

В Табл. 3.4. са представени данни за буквите с най-високо стандартно отклонение при текстовете от сръбска редакция. Изследваните текстове от сръбската редакция принадлежат на един и същ паметник (Сръбския псалтир от 13 в.). Честотата на употреба на различните букви дава най-устойчиви резултати именно в този случай - има едва три случая на букви, чиято честота на употреба показва стандартно отклонение между 0,3 и 0,5. Интересно е да се отбележи, че за разлика от българската и руската редакции, където данните за употребата на гласните показват най-съществените различия, в случая на сръбските текстове има разлики в употребата на съгласните д, т и гласната о.

Данните за най-високите стандартни отклонения при честотата на употреба на буквите в изследваните текстове, принадлежащи към руската редакция, са представени в Табл. 3.5.

Интересно е да се отбележи, че в руските текстове, които произхождат от три различни ръкописа, се наблюдават съществени различия при употребата на 11 букви, от които само една е съгласна - т. Най-висока стойност (1,02) има стандартното отклонение при данните за честотата на употреба на голямата носовка.

Табл. 3.5. Букви с най-високи стандартни отклонения при честотата на употреба - руска редакция

| Буква | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|
| ѫ | 0.78 | 0.00 | 2.70 | 1.02 |
| ъ | 3.31 | 1.82 | 5.52 | 0.88 |
| ь | 2.58 | 1.40 | 3.71 | 0.69 |
| о | 8.93 | 7.42 | 10.21 | 0.56 |
| ѣ | 0.88 | 0.23 | 1.95 | 0.56 |
| оу | 2.12 | 1.10 | 3.14 | 0.55 |
| ѧ | 2.20 | 1.62 | 3.18 | 0.40 |
| ю | 0.70 | 0.07 | 1.38 | 0.39 |
| и | 8.35 | 7.41 | 9.21 | 0.30 |
| т | 5.40 | 4.16 | 6.38 | 0.30 |
| ѣ | 2.57 | 2.06 | 3.40 | 0.30 |

Интересни са получените данни за двата български ръкописа. В сравнение с данните за честотата на употреба на буквите в руските ръкописи, употребата на буквите н, і и ѩ показва много висока нееднородност (вж. Табл. 3.6.).

В двата български ръкописа най-силно варира употребата на тринадесет гласни букви. Между създаването на Синайския и Болонския псалтир са минали около три века, в които постепенно се е променяло използването на графичната система (тук нямаме предвид използването на различни азбуки, а постепенното изменение на буквения състав на средновековното славянско писмо). Синайският псалтир е писан на глаголица и при изследването беше използвана транскрипция на кирилица. Получените от нас данни потвърждават, че постепенно е намаляло разнообразието в използването на трите графеми, представящи

различни варианти на "и" (і, ѝ, ї), както и на характерната за глаголическите ръкописи малка йотувана носовка (ѩ).

Табл. 3.6. Букви с най-високи стандартни отклонения при честотата на употреба - българска редакция

| Буква | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|
| и | 5.44 | 1.65 | 9.30 | 3.15 |
| і | 2.51 | 0.04 | 6.12 | 2.40 |
| ѩ | 1.24 | 0.00 | 2.86 | 1.24 |
| ѧ | 0.88 | 0.00 | 1.98 | 0.86 |
| ь | 1.54 | 0.20 | 2.73 | 0.75 |
| ѣ | 3.57 | 2.70 | 4.80 | 0.69 |
| ж | 1.77 | 0.80 | 2.74 | 0.53 |
| ї | 0.50 | 0.00 | 1.85 | 0.51 |
| ѫ | 0.49 | 0.00 | 1.21 | 0.50 |
| о | 8.87 | 7.74 | 10.22 | 0.45 |
| а | 6.04 | 5.21 | 7.35 | 0.38 |
| ѡ | 0.43 | 0.02 | 0.97 | 0.36 |
| ѥ | 0.26 | 0.00 | 1.14 | 0.30 |

Данните за стандартните отклонения при честотата на употреба на различните букви при текстовете от всички изследвани редакции (вж. Табл. 3.7.) показват силна нееднородност при 28 букви и съчетанието ѣь. В деветнадесет случая отново силно варира употребата на гласни букви. На практика това са всички гласни от средновековната кирилица.

Табл. 3.7. Обобщени данни за букви с най-високи стандартни отклонения при честотата на употреба във всички редакции

| Буква | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|
| ъ | 4.28 | 0.00 | 11.09 | 2.71 |
| и | 7.46 | 0.16 | 10.13 | 2.38 |
| ь | 2.82 | 0.00 | 7.53 | 1.76 |
| і | 1.11 | 0.00 | 7.17 | 1.74 |
| ѧ | 1.41 | 0.00 | 4.00 | 1.03 |
| ж | 1.00 | 0.00 | 4.18 | 1.02 |
| ѣ | 0.42 | 0.00 | 3.86 | 0.93 |
| ѥ | 0.45 | 0.00 | 3.45 | 0.79 |
| оу | 1.91 | 0.46 | 4.73 | 0.70 |
| ѫ | 2.94 | 1.61 | 5.41 | 0.68 |
| о | 8.87 | 6.29 | 12.35 | 0.67 |
| е | 6.03 | 3.95 | 9.36 | 0.57 |
| ѯ | 0.80 | 0.00 | 2.33 | 0.56 |
| а | 6.46 | 4.41 | 8.65 | 0.55 |
| ъї | 0.86 | 0.00 | 2.02 | 0.55 |

| Буква | Средна стойност | Минимална стойност | Максимална стойност | Стандартно отклонение |
|---|---|---|---|---|
| ъі | 0.38 | 0.00 | 1.67 | 0.53 |
| т | 5.37 | 3.07 | 7.38 | 0.52 |
| вь | 0.54 | 0.00 | 2.60 | 0.51 |
| м | 3.25 | 0.91 | 6.34 | 0.49 |
| р | 3.43 | 1.85 | 5.40 | 0.42 |
| с | 6.29 | 3.67 | 8.18 | 0.42 |
| в | 5.59 | 3.11 | 7.21 | 0.41 |
| г | 2.53 | 1.27 | 4.27 | 0.40 |
| ї | 0.35 | 0.00 | 4.35 | 0.40 |
| ю | 0.58 | 0.00 | 1.99 | 0.38 |
| н | 4.60 | 3.22 | 7.59 | 0.35 |
| д | 3.35 | 0.00 | 4.92 | 0.34 |
| л | 3.07 | 1.09 | 4.35 | 0.33 |
| ѡ | 0.66 | 0.00 | 1.80 | 0.33 |

Отбелязаните различия при наблюденията от текстове от различни редакции и в съвкупност позволяват да се направят следните изводи:

- При изследването на текстове от един ръкопис стандартните отклонения за честотата на употреба на буквите достигат до ≈0,35.
- По-високи числови стойности на стандартните отклонения се появяват при изследване на ексцерпти от различни редакции.
- При текстове с голяма разлика във времето на написване,

количествените показатели за използване на графемите за **И** и носовките показват високи отклонения.

- При изследване на текстове от различни редакции данните за употреба на **всички** гласни букви показват съществени отклонения.
- Качествените признаци, използвани от специалистите филолози за определяне на редакцията на определен текст , са с ниска употреба и при количествени изследвания не са чувствителни, с изключение на носовките и еровите гласни.

При прилагането на методи за количествено изследване могат да бъдат игнорирани данните за употреба на:

- Букви, които представят типични за гръцкия език звукове, които не са били необходими за представяне на звуковата система на славянските езици и постепенно са отпаднали от азбуката ни: ѳ, ѫ, ѱ с изключение на ѡ. Те са ценен качествен признак, но в повечето ръкописи не се употребяват или употребата им се ограничава само в имената от гръцки произход.
- Графичните варианти на оу (ꙋ, ꙗ) и ъı (ьі, ън, ъі), чиято употреба също намалява постепенно във времето.
- Букви с по-рядка употреба като ф, ѵ и ѱ.

В Табл. 3.8. е направена съпоставка на състава на буквите, които показват отклонения при изследване на текстове само от една редакция и при изследване на смесени текстови извадки. Появата на определена буква в колонката за съответната редакция показва, че данните за употребата й са показали стандартно отклонение над 0,3 за съответните текстове.

Табл. 3.8. Списък на буквите с най-високи стандартни отклонения в различните редакции

| Всички редакции | Руска редакция | Българска редакция | Сръбска редакция |
|---|---|---|---|
| ѳ | ѳ | ѳ | ѳ |
| ѣ | ѣ | ѣ | |
| ѧ | ѧ | ѧ | |
| ѣ | ѣ | ѣ | |
| ѫ | ѫ | ѫ | |
| т | т | | т |
| і | | і | |
| ж | | ж | |
| ѣѧ | | ѣѧ | |
| ює | | ює | |
| а | | а | |
| ї | | ї | |
| ѡ | | ѡ | |
| оу | оу | | |
| ъ | ъ | | |

| Всички редакции | Руска редакция | Българска редакция | Сръбска редакция |
|---|---|---|---|
| и | и | | |
| ю | ю | | |
| д | | | д |
| є | | | |
| ъї | | | |
| ъі | | | |
| вь | | | |
| м | | | |
| р | | | |
| с | | | |
| в | | | |
| г | | | |
| н | | | |
| л | | | |

Единствено употребата на буквата ѳ варира както в текстовете от всички отделни редакции, взети и поотделно, и заедно. Интересно е да се отбележи, че тази буква не спада към качествените признаци за определяне на редакцията на средновековни славянски паметници.

Следващите букви с високо вариране в честотата на употреба са ѣ, ѧ, ѣ и ѫ.

*3.2.2.5. Приложения на клъстерния анализ за изследване на различията в средновековните редакции на старобългарския език*

Наблюденията над основните статистически характеристики, представени в част 3.2.2.4. показват, че са налице съществени различия в количествените данни за употреба на буквите в руската, сръбската и българската редакция. При това положение е обосновано прилагането на методи за групиране на изследваните обекти, в случая – текстове от различна правописна редакция, за да се провери доколко количествените данни могат да послужат за коректно класифициране на текстовете според произхода им.

Клъстерният анализ на данните за буквените честоти и на честотите на употреба на предварително описаните групи от качествени признаци води до интересни резултати.

В Табл. 3.9. са посочени номерата на случаи от обработвания със STATISTICA файл (тези номера са съществени при разглеждането на представените по-долу фигури).

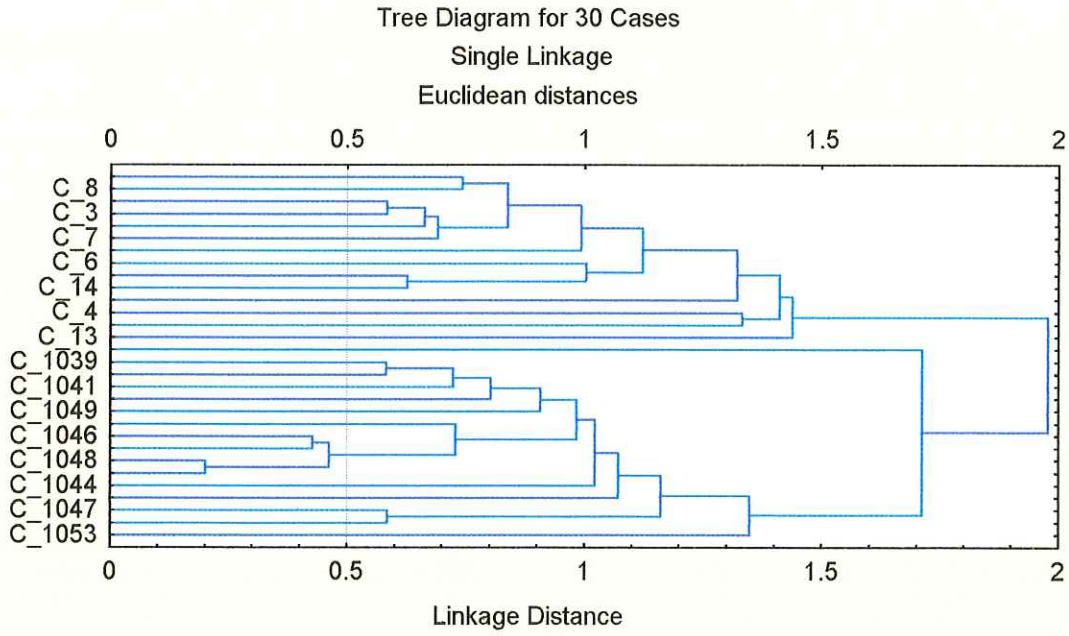
Табл. 3.9. Показалец на номера на случаи в използвания мегафайл в Statistica for Windows

| **Псалтир** | Номера на случаи в STATISTICA, представящи данни за ексцерпт от ръкописа |
|---|---|
| Синайски (българска редакция, 10 век) | 1O39-1249 |
| Болонски (българска редакция, 13 век) | 1-211 |
| Норовски (руска или атонска редакция, 13 в.) | 634-844 |
| Сръбски (сръбска редакция, 13 век) | 845-1038 |
| Киевски (руска редакция, 14 век) | 423-633 |
| Генадиевски (руска редакция, 15 век) | 212-422 |
| Църковнославянски (съвременна версия) | 1250-146O |

В поредица от експерименти изследвахме дали при използване на данни за буквените честоти на ексцерпти с различна датировка или от различна редакция (българска, руска и сръбска) се получава разпределение на текстовете по групи според произхода им.

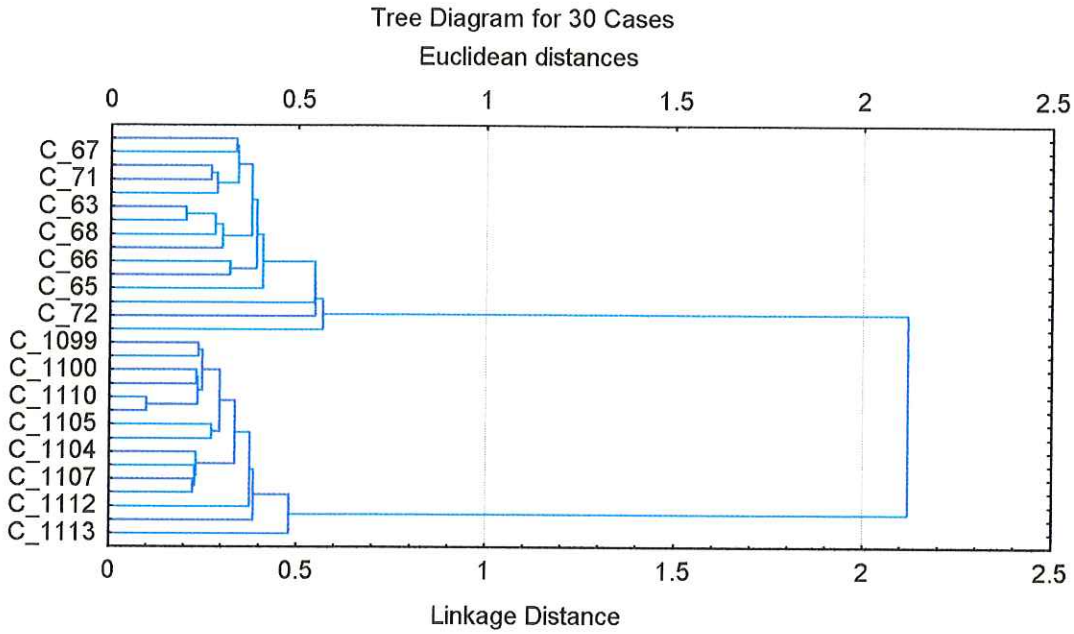
Първоначално изследвахме текстовете от Синайския (10 в.) и Болонския псалтир (13 в.). Двата ръкописа са български, но Синайският псалтир е писан на глаголица и изследването се осъществява върху кирилската транскрипция на текста.

Фиг. 3.4. Синайски и Болонски псалтири, извадка по 1 псалм, изследване на група от качествени признаци (носовки)

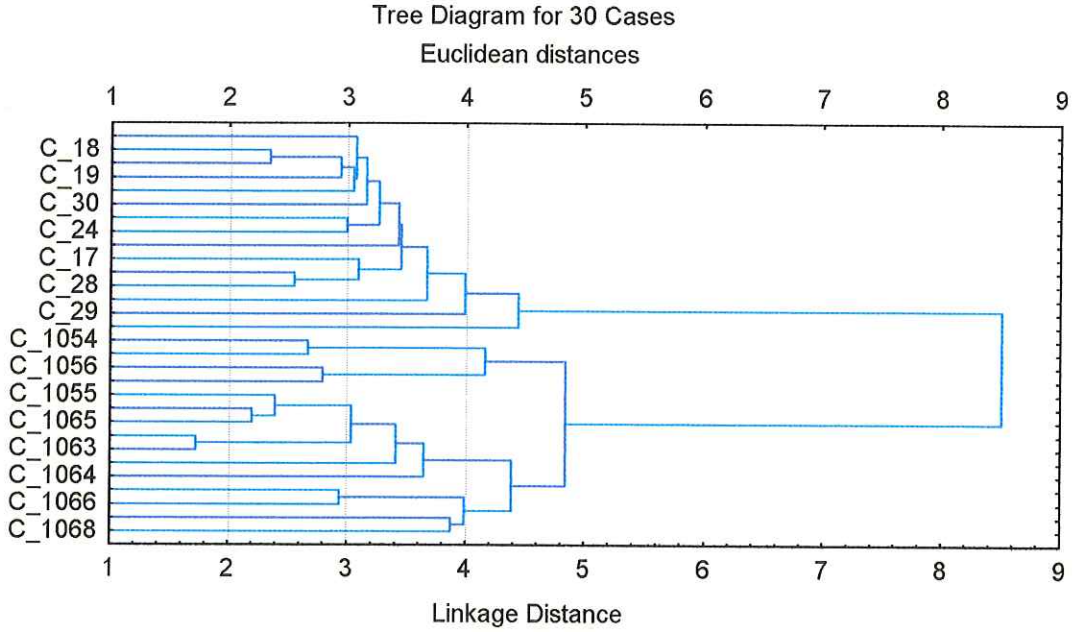



Различията между двата текста, единият от които е старобългарски, а вторият - среднобългарски, са достатъчно значими, за да доведат до съвсем ясно разграничаване на двете групи от ексцерпти при големини на изследваните обекти 1 псалм (около 1 килолитера) и изследване на поведението само на носовите гласни (вж. Фиг. 3.4.).

Фиг. 3.5. Синайски и Болонски псалтири,
извадка с големина 5 псалма, изследване на
група от качествени признаци (носовки)




Фиг. 3.6. Изследване на извадка с големина 2
псалма от Синайски и Болонски псалтир:
Всички букви

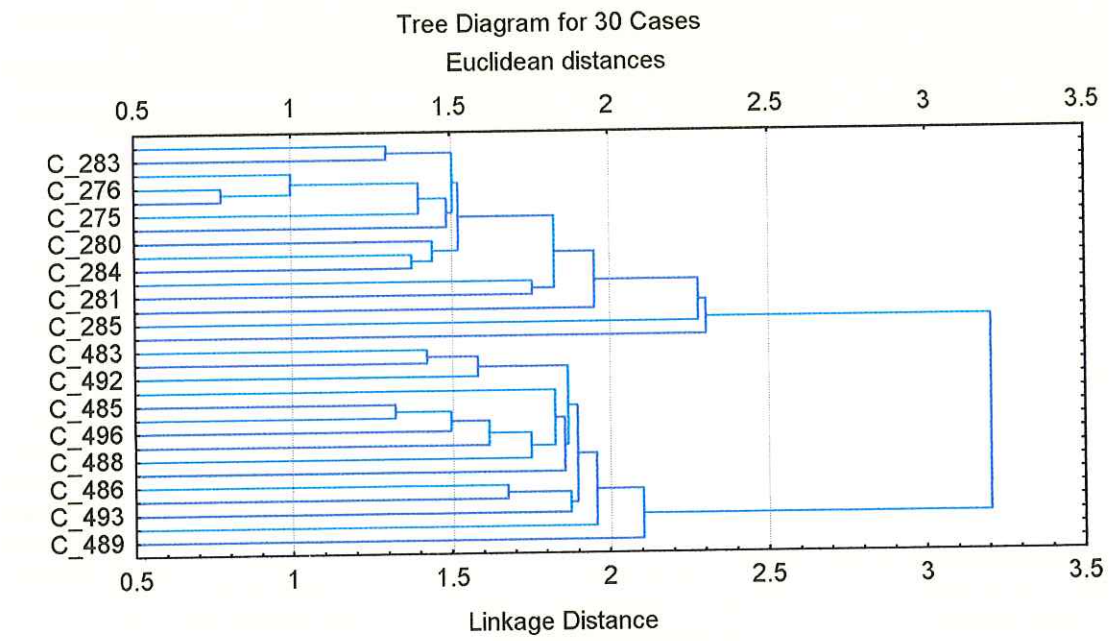



При големина 5 псалма (вж. Фиг. 3.5.) разстоянията между двете групи се увеличават като същевременно вътрешногруповите

разстояния намаляват, което потвърждава значимостта на правописното различие между двата ръкописа.

За съпоставка изследвахме и клъстеризирането при включване на данните за честотите на всички букви от азбуката (вж. Фиг. 3.6.). При това изследване също се наблюдава ясно разграничаване на текстовете от двата ръкописа, като междугруповото разстояние се увеличава, но и вътрешногруповото разстояние нараства в сравнение с изследванията, проведени само върху честотите на употреба на носовите гласни.

Фиг. 3.7. Изследване на извадка с големина 5 псалма от Киевски и Генадиевски псалтири: Всички букви

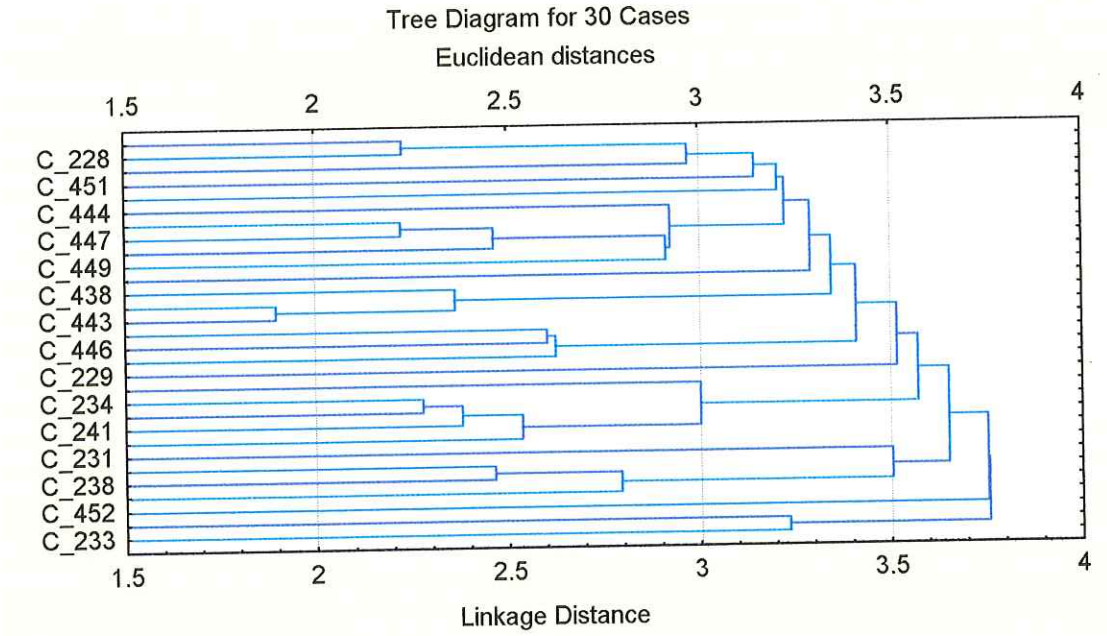



Продължихме експериментите с изследване на два руски ръкописа, чиято разлика в датировките е по-малка в сравнение с датировките на двата български паметника. Киевският и Генадиевският псалтир са написани съответно през XIV и XV в. Разликата от един век, през който не са предприемани съществени правописни реформи, означава, че процесът на промени в употребата на графемите в средновековните паметници не е бил

интензивен.

Предположението при изследване на подобни паметници е, че ще са необходими по-големи текстови ексцерпти, за да се получи правилно разпределение по клъстери на ексцерптите от различните паметници.

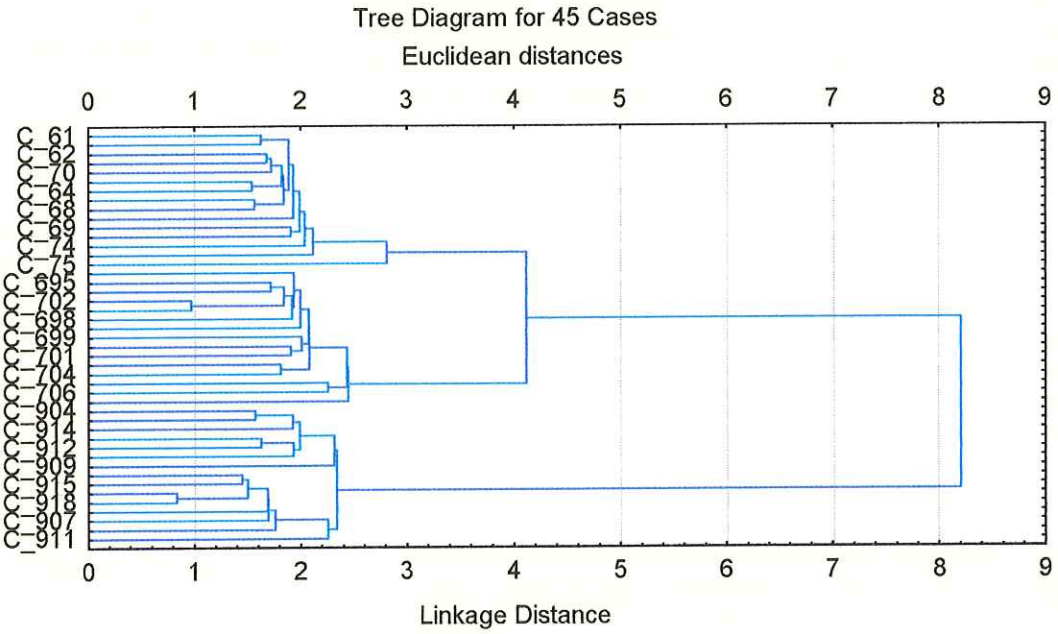
Фиг. 3.8. Изследване на извадка с големина 2 псалма от Киевски и Генадиевски псалтири: Всички букви



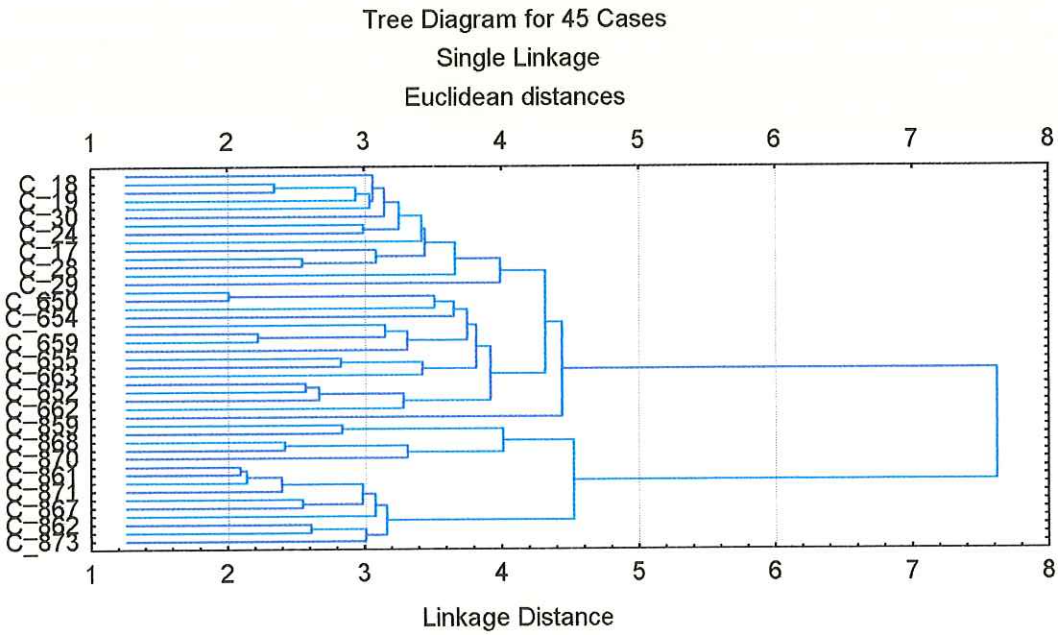

Резултатите от експериментите потвърждават това предположение. При изследване с достатъчно големи извадки от по 5 псалма (около 5 килолитери) и включване на всички букви, текстовете се групират съответно на произхода им (вж. Фиг. 3.7.). Но при по-малък обем на извадката (2 псалма или около 2 килолитери, вж. Фиг. 3.8.), текстовете от двата паметника се смесват.

Описаните по-горе експерименти целяха да се изследва доколко разликите между употребите на буквите позволяват да се групират правилно текстове от една и съща редакция в диахронен аспект.

Фиг. 3.9. Извадка с големина 5 псалма за текстовете от 13 век Норовски, Сръбски, Болонски: Всички букви



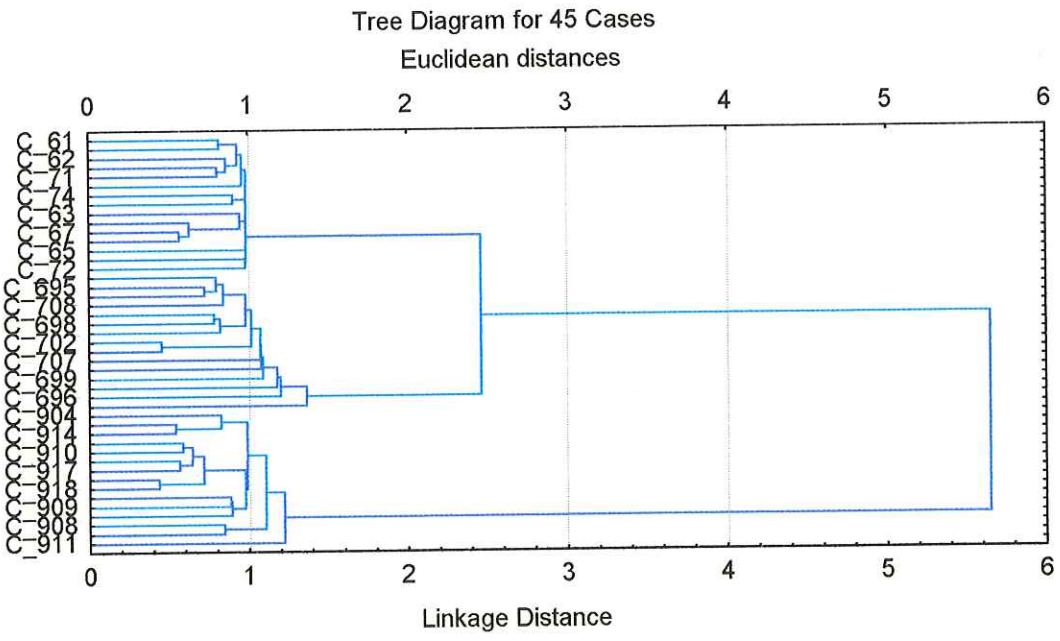

Фиг. 3.10: Изследване на текстови ексцерпти с големина 2 псалма за текстовете от 13 век - Норовски, Сръбски, Болонски: Всички букви



В следващата серия от експерименти изследвахме текстове, написани през един и същи век (в случая – XIII в.), но от три различни редакции – българска (Болонски псалтир), руска (Норовски псалтир) и сръбска (Сръбски псалтир).

Резултатите от част от тези опити са представени на Фиг. 3.9, където в изследването са участвали честотите на употреба на всички букви при големина на текстовите ексцерпти 5 псалма (около пет килолитери) и Фиг. 3.10 (отново всички букви, но с големина на ексцерпта 2 псалма).

Фиг. 3.11. Изследване на текстови ексцерпти с големина 5 псалма за текстовете от 13 век – Норовски, Сръбски, и Болонски псалтири. Всички качествени признаци




Най-добри показатели за вътрешногрупово и междугрупово разстояние се наблюдават при изследването, извършено въз основа на най-големите текстови извадки. При намаляването на големината на ексцерптите от 5 на 2 килолитери текстовете продължават да се разпределят правилно в клъстери според произхода им, но вътрешногруповото разстояние се увеличава, а междугруповото

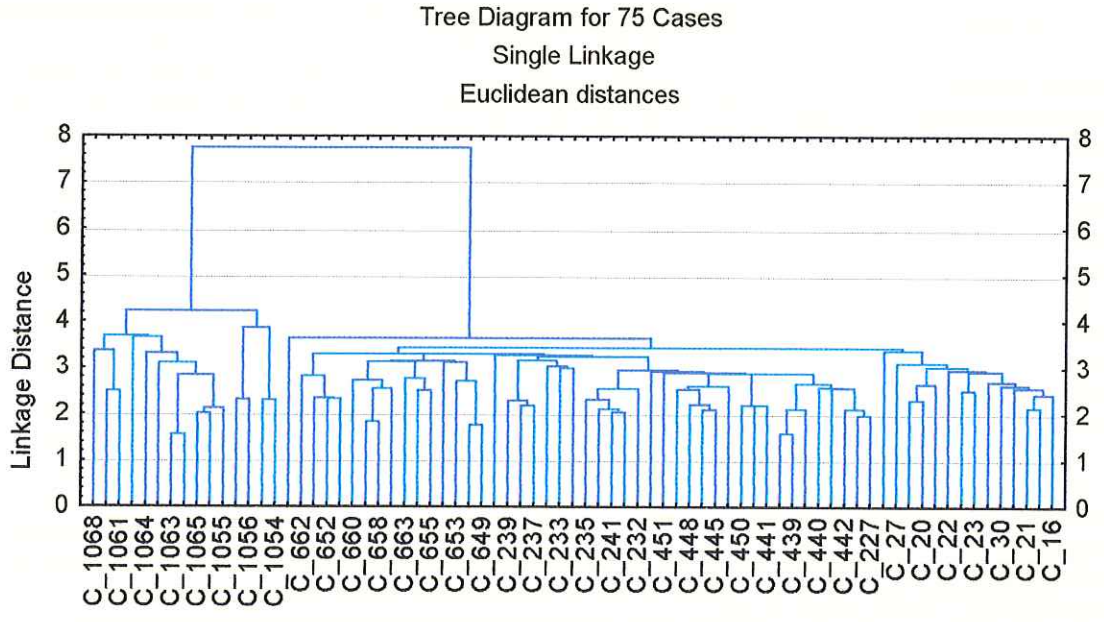
намалява.

При следващото намаляване на големината на текстовата извадка на 1 килолитера изследваните обекти престават да се групират по клъстери съответно на произхода си.

За да илюстрираме ролята на подбора на признаците, които участват в изследването, на Фиг. 3.11. представяме резултатите от клъстеризиране на същите паметници при големина на извадката 5 псалма, проведено само върху данните за честота на употреба на качествените признаци. В сравнение с изследването, направено върху честотите на употреба на всички букви, се получава по-малко междугрупово разстояние.

Пример за изследване на повече текстове е представен на Фиг. 3.12.

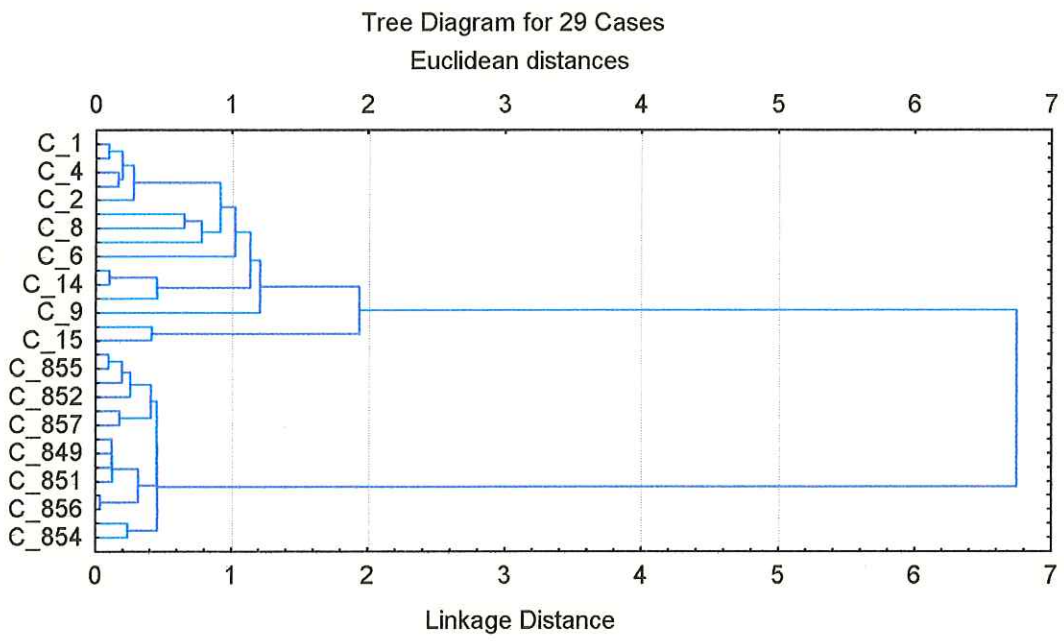
Фиг. 3.12. Изследване с големина на извадката 2 псалма,
всички букви, клъстери отляво надясно:
Синайски (10 в.), Норовски (13 в.),
Генадиевски (15 в.), Киевски (14 в.), Болонски (13 в.)



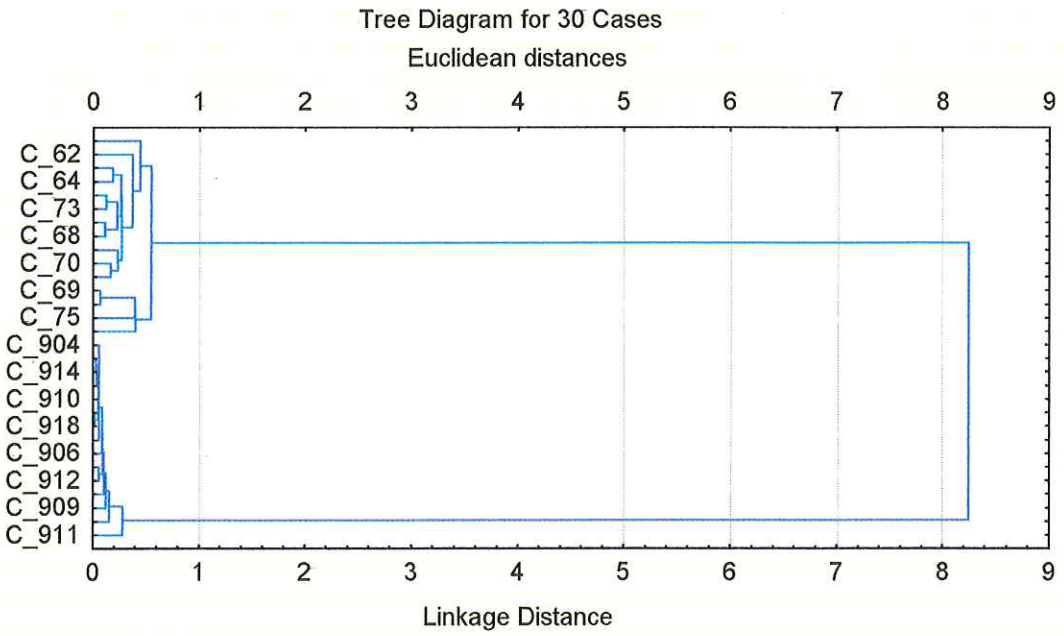

Като последна илюстрация на важността на подбора на буквените честоти, върху които да се извършва клъстерен анализ,

представяме два случая, чиито графики са представени на Фиг. 3.13 и Фиг. 3.14.

Фиг. 3.13. Изследване на текстове с големина 1 псалм, Група ерови гласни. Сръбски и Болонски Псалтири




Фиг. 3.14. Изследване на текстове с големина 5 псалма, Група ерови гласни. Сръбски и Болонски Псалтири



При тези изследвания бяха включени само текстове от Болонския и Сръбския псалтири. И двата текста са от XIII в., съответно българска и сръбска редакция. При изследванията участват само данните за употреба на еровите гласни .

Поради това, че сръбската редакция е едноерова, дори и при големина на текстовите извадки 1 псалм текстовете се разпределят правилно в две групи. При увеличаване на големината на извадката (Фиг. 3.14) отново вътрешногруповото разстояние се намалява, а междугруповото се увеличава.

Анализът на резултатите от серията описани експерименти ни дава основания да формулираме следните изисквания за големина на текстовите ексцерпти и подбор на данни, участващи при прилагането на статистически методи за класификация:

- За изследване на текстове от различни векове и/или редакции минималните текстови ексцерпти е препоръчително да са с обем поне 2 килолитери.
- При изследване на текстове от една и съща редакция и време минималните текстови ексцерпти е препоръчително да са с обем поне 5 килолитери.
- Изследванията върху данните за всички буквени честоти водят до получаването на по-добри междугрупови и вътрешногрупови разстояния, отколкото изследванията на честотите на качествените признаци.
- При предварително известен произход на текстовете могат да се прилагат специално подбрани множества от признаци. Например при текстове от сръбска редакция може да се използват данните само за честотите на употреба на еровите гласни. Уточняването на подобни множества от най-характерни признаци, които да се прилагат при количествени изследвания от различен тип е предмет на

бъдеща експериментална работа.

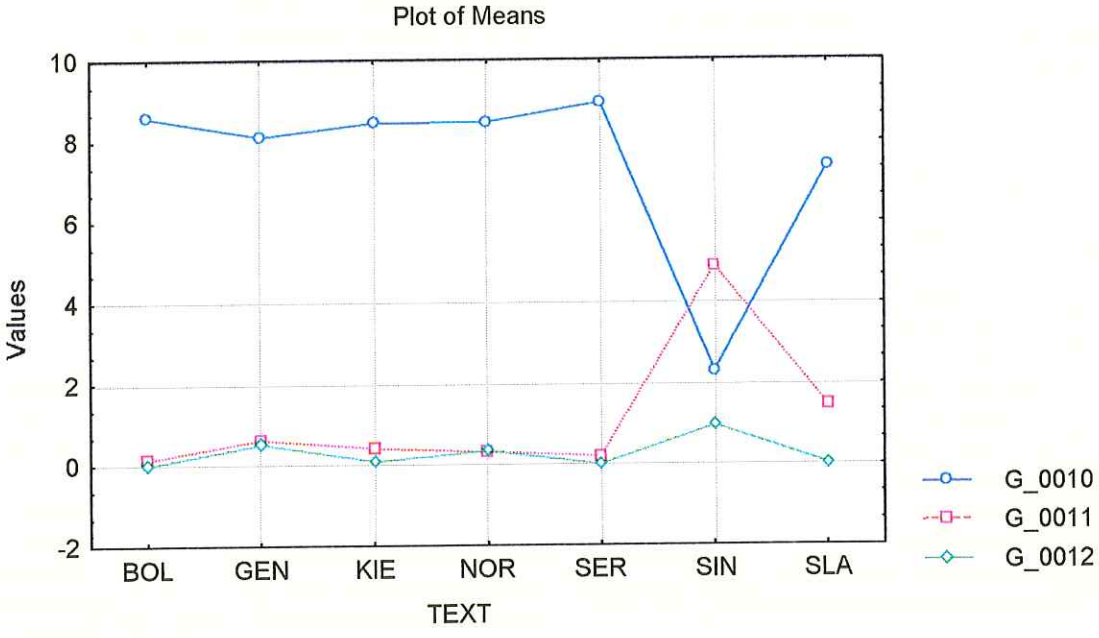
Графиките, получени при прилагането на клъстерен анализ, могат успешно да се прилагат при обучаването на студенти, защото помагат нагледно да се представи кои данни при подобни експерименти водят до ясно разграничаване на текстовете на групи.

Подобни илюстрации могат да онагледят и значението на големините на извадките, с които се работи, при различни видове съпоставки (текстове от една и съща редакция или от различи редакции; текстове от един и същи период или от различни периоди).

*3.2.2.5. Анализ на буквени честоти: съпоставки по групи*

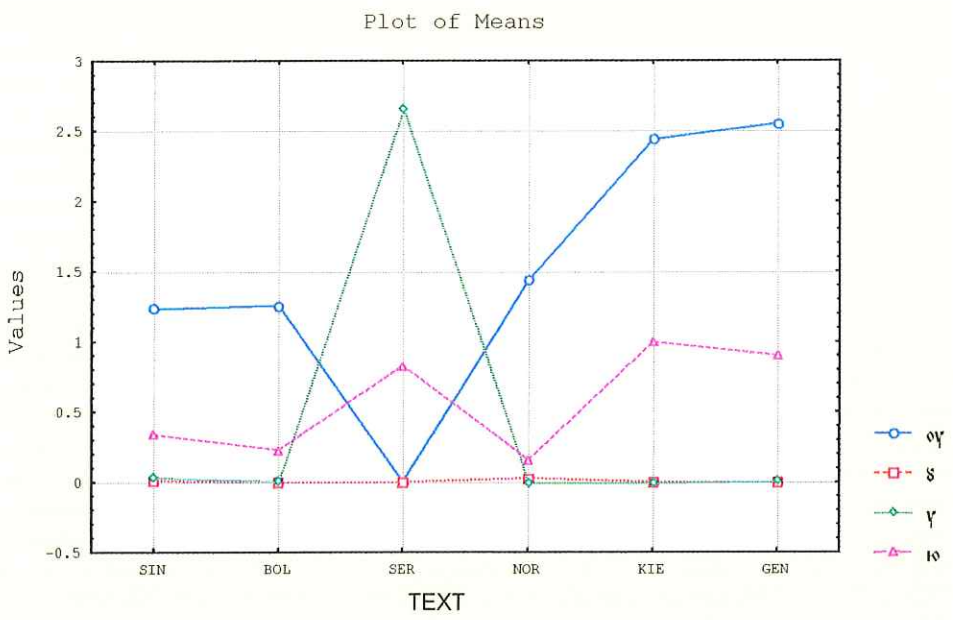
В тази част от изследването са представени диаграми за употребата на групи графеми, които представляват варианти на изписване на дадена буква. Освен това са изследвани някои групи от буквени съчетания, посочени в [Граматика 93] като характеризиращи определена редакция.

Фиг. 3.15. Употреба на три графеми за "и"
(син – и, червен – і, зелен - ї)



Интересно е да се отбележи как употребата на трите графеми, представящи и в глаголическия текст на Синайския псалтир е довело до употреба на 3 графеми в кирилската транскрипция. При другите ръкописи е явно предпочитанието към една графема (вж. Фиг. 3.15.). Това може де се използва като илюстрация на тенденцията да се преминава към употреба на една графема в случаите, когато е имало няколко символа със сходна звукова стойност. Изследването на текстове от повече паметници би могло да покаже по-ясно как е протекъл във времето този процес.

Фиг. 3.16. Употреба на ѹ и ꙋ




Очакваната тенденция при употреба на различните видове изписвания на ѹ е да се наблюдава нарастване на употребата в руските и сръбскитетекстовеза сметка на замяната на голямата носовка с "у" (вж. Фиг . 3.16). Наистина в Болонския и Синайския псалтири, които са български, се наблюдава по-малка употреба в сравнение състаналите ръкописи. Отклонение се наблюдава в съвременния текст на Псалтира (ръкописа, означен със SLA), където се използва съвременно У.

При еровите гласни ясно се наблюдава употребата на един

вид ерова гласна в сръбския ръкопис (означен със SER). Освен това е интересно да се наблюдава как тенденцията към намаляване на употребата на ерови гласни се отразява в руските ръкописи (вж. Фиг. 3.17).

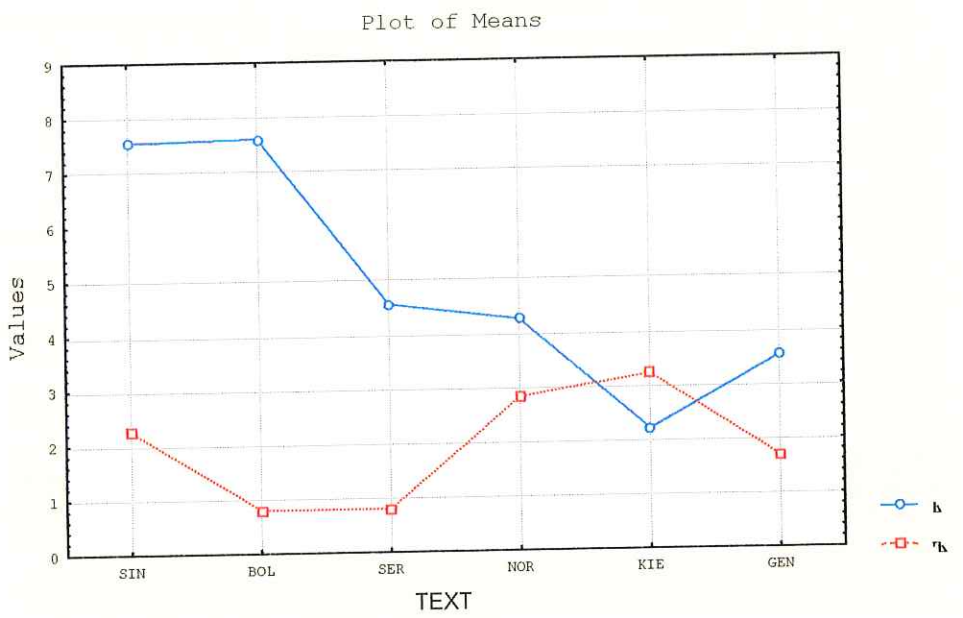
Характерно за употребата на носовките е, че не се наблюдава отчетлива тенденция според описаните граматични наблюдения (вж. Фиг. 3.18).

Решихме да изследваме употребата на нейотуваните гласни и йотуваните им аналози, като търсехме дали изследването няма да покаже интересни отлики за някои от ръкописите. При изследването на йотуваните и нейотуваните гласни интересни резултати се получават за носовките.

На фиг. 3.18 е представена употребата на малката носовка. От графиката ясно личи, че йотуваният вариант на носовката се употребява само в Синайския псалтир, а във всички останали паметници той не се среща. В Сръбския псалтир не се употребява и нейотуваната малка носовка, което е естествено за сръбската правописна редакция. Вижда се, че Норовският псалтир се доближава по употреба на голямата носовка до Болонския псалтир и силно се отличава от другите руски паметници.

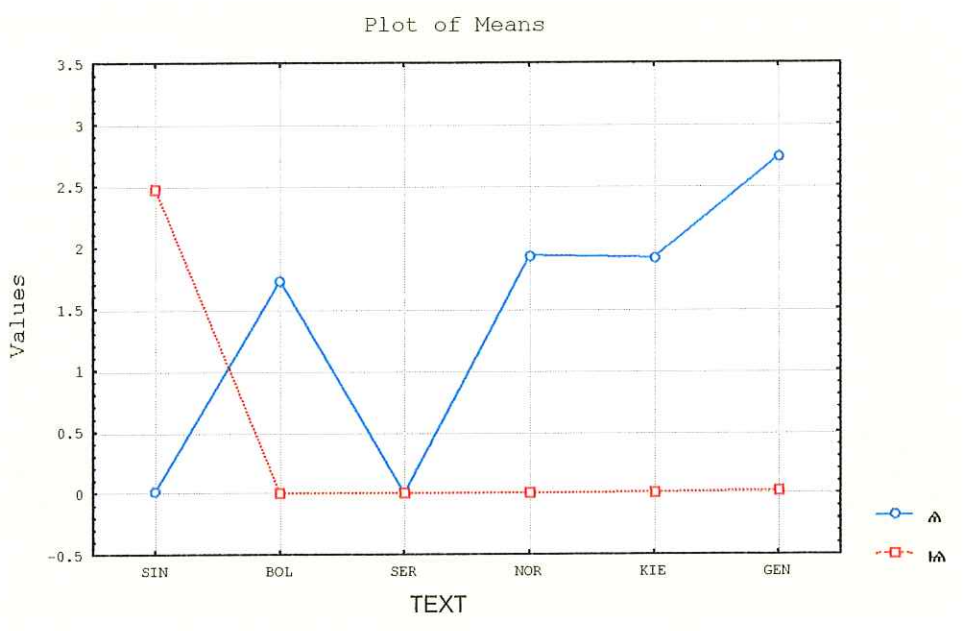
Успоредното разглеждане на Фиг. 3.18. и 3.16. дава ясна представа за това, как замяната на голямата носовка с ѹ води до по-интензивна употреба на ѹ в руските и сръбски ръкописи. Тук Норовският псалтир отново заема нетипично за редакцията си място със значително по-ниската употреба на у в сравнение с другите руски ръкописи.

Фиг. 3.17. Употреба на двете ерови гласни

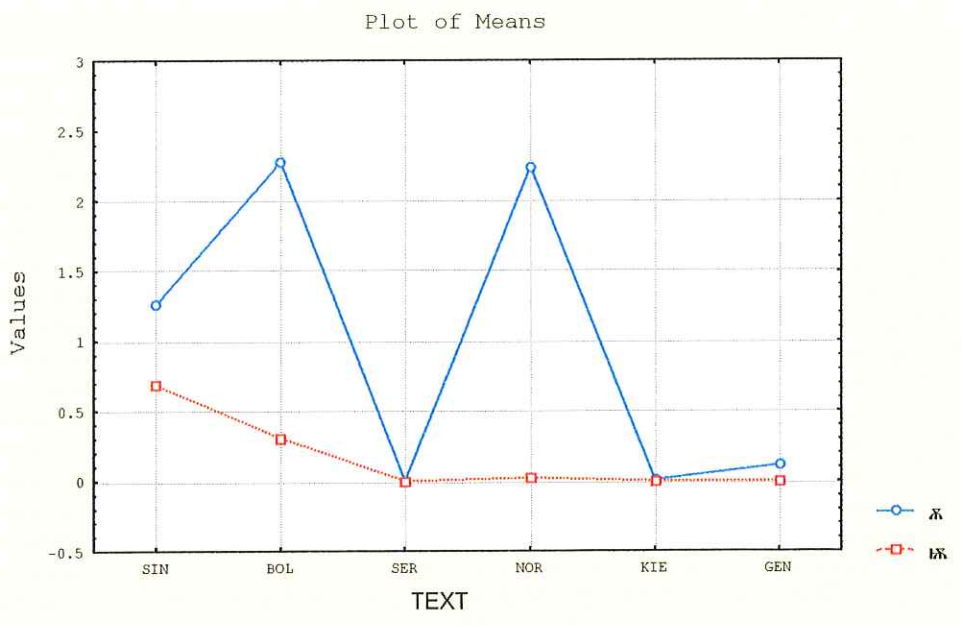



Фиг. 3.17. представя добра илюстрация за едноеровата редакция - вж. SER. Подобни илюстрация дават ясна представа за интензивността на употреба на двата ера в съпоставка както в рамките на един ръкопис, така и в рамките на цялата изследвана група от ръкописи. Ясно личи тенденцията за намаляване на употребата на еровите гласни в Киевския и Генадиевския псалтири, защото при събиране на стойностите за употребите на двата вида ера (Киевски - около 5,4%, Генадиевски - около 6,4 %) се получава обща стойност, по-ниска от стойностите за по-употребявания вид ер в Синайския (7,6%), Болонския (7,6%) и Сръбския (6,6%) псалтири. Норовският псалтир със 7,2% кумулативна стойност за употребата на еровите гласни отново заема особено място в групата от изследвани ръкописи.

Фиг. 3.18. Употреба на носовки




3.18.а – малки носовки




3.18.б – големи носовки

Анализите на буквените честоти ни дават интересни резултати, които тепърва ще се сравняват с данни от други текстове,

за да се получи по-ясна картина. Традиционно в досегашните изследвания е специалистите да се обръщат към качествените характеристики като към по-леснодостъпно средство за анализ. Необходимо е да се проведат по-прецизни статистически изследвания над по-разнородни множества от средновековни славянски текстове, за да се даде достатъчно обоснован отговор на въпроса кои са представителните количествени характеристики за славянските средновековни текстове.

*3.2.2.6. Заключения*

Представените първоначални данни показват, че изследването на употребата на различни букви показва неизвестни досега данни и зависимости. Ще посочим накратко някои изводи и наблюдения:

- Извадката от 1 псалм, чиято големина гравитира към 1 килолитера, обикновено не е достатъчно представителна за изследване, основано на буквените честоти. Извадки с големина една килолитера са предпочитани в статистическите изследвания, правени върху текстове на съвременен български език. За средновековните текстове те не са достатъчни.
- Резултатите от клъстерния анализ показват, че приликата между текстовете от отделните ръкописи е достатъчно висока, за да се извърши разпределение по клъстери, което отговаря на произхода на текстовете, при подходящ избор на големината на текстовите извадки според типа на текстовете (минимум 2 килолитери при текстове от различно време и/или редакция и минимум 5 килолитери при текстове от една и съща редакция и времеви период).
- Основните количествени характеристики на буквените

честоти в старобългарските кстове от различни редакции, посочени в Табл. 3.3., са първите публикувани данни, които са ни известни до момента. Те могат да служат като отправна точка за бъдещи изследвания върху по-широк кръг от текстове.

### *3.2.3. Общи препоръки за експерименти за изследване на вариативност*

Въз основа на проведените и представените в работата експерименти, препоръчваме следното за поставяне на експерименти за изследване на вариативност в средновековни славянски текстове.

1. *Подбор на текстове съобразно с целите на изследването.*
   Подборът на текстове обикновено зависи от интересите на изследователите. Ако целта на изследването не е изучаване на конкретен текст, при подбора трябва да се имат предвид регионите и датиранията на паметниците. Изборът дали ще се изследват преписи на един и същи текст, или различни текстове, също зависи от целите на изследователя.
   Тъй като в нашето изследване искахме да проверим доколко количествените характеристики на един и същи текст от различни региони и времеви периоди варира, подбрахме едни и същи текстови ексцерпти от различни преписи на еднакъв текст.
   Общо съображение е също, че религиозните текстове са подлежали на по-малко промени отколкото текстовете от светската литература.
2. *Подбор на големината на ексцерптите.*
   Големината на ексцерптите е съществена и е обвързана със спецификата на изследваните текстове. Колкото по-близки са текстовете по произход, толкова по-големи

ексцерпти трябва да се включат в изследването. При текстове от една и съща редакция при нашите експерименти препоръчителната големина е 5 килолитери. При текстове от различни редакции и/или времеви периоди препоръчителната големина е 2 килолитери.

3. *Подбор на признаците за изследването*

   Подборът на това кои данни ще доведат до добри резултати при изследването е от съществено значение. В нашите изследвания използвахме съвкупности от данни за употребата на всички букви; на качествените признаци, както и множества от отделни подбрани букви. Използването на честотите на употреба на буквите ѣ, ъ, ѧ, ѫ показа добри резултати, но при всяко конкретно изследване трябва да се търси най-подходящата съвкупност.

4. *Подбор на илюстративен материал за представяне на резултатите*

   Едно от сериозните предимства на използването на съвременни статистически пакети е възможността получените резултати лесно да се представят в графичен вид. Подобни графики могат успешно да се използват за онагледяване, особено при обучението на студенти.

# Глава 4. Специализирано работно място на старобългариста

## 4.1. Специфика на работното място на старобългариста

Основното предназначение на специализираното работно място на старобългариста е да предоставя възможности за организиране на средновековни ръкописи в електронен вид и изследователски данни. Тази среда трябва да позволява осъществяването на взаимен обмен на подобни данни между различни специалисти.

Тъй като средновековните ръкописи представляват обекти, сложни за формализирано описание, подобна среда трябва да бъде достатъчно гъвкава, за да може да отразява различните гледни точки на отделните специалисти.

Досегашните приложения на информационните технологии в медиевистиката обикновено са ориентирани към решаването на една определена изследователска задача: например, подобряване на четливостта на изображението на ръкопис с цел уточняване на неясни пасажи. Обикновено усилията на изследователите се насочват към работата с изображения, текстове или структурирани данни от научни описания, като все още са изолирани случаите на създаване на интегрирани среди, включващи възможности за комбиниране на тези различни представяния.

Сред съществуващите интегрирани среди интересен пример представлява BAMBI [Calabretto, Rumpler 98], която дава възможност за проследяване на текста на даден паметник и за справки с изображението на съответния лист от ръкописа. Тази среда е ориентирана към достъп чрез Интернет и при разработването й голямо внимание е било отделено на правата на достъп до

съхраняваните данни за различни групи от потребители.

Работата на старобългаристите има съществени отличителни черти от работата на медиевистите, работещи с други видове текстове. Най-съществената отличителна черта на средновековните славянски текстове е високата степен на вариативност на всички езикови нива, която трябва да бъде взета предвид при проектирането на подобна среда. От една среда, тази вариативност трябва да се отразява възможно най-точно при представянето на текстовете. От друга страна, компютърната среда трябва да предоставя инструменти за изследването на проявленията на вариативността, тъй като една от основните ползи от приложенията на информационната технология в случая са възможностите за разширяване на обхвата на съпоставителните изследвания и за получаването на количествени данни за различни неизследвани характеристики на средновековното ни ръкописно наследство.

Предлаганият модел на специализирано работно място на старобългариста е съобразен с тези основни отличителни черти на предметната област.

## 4.2. Модели на данните

Компютърното моделиране на средновековни ръкописи се извършва като изследваните обекти се представят и изследват с инструментариума на четири различни направления на компютърната информатика:

- Методите и средствата от областта на *компютърната графика и обработката на изображения* се използват за представяне и обработка на изображения на средновековни ръкописи.
- Методите от областта на *компютърната лингвистика и лингвостатистика* се използват за представянето и

обработката на текстовете на ръкописите. В някои случаи представянето на текстовете включва и научното описание на кодирания текст.

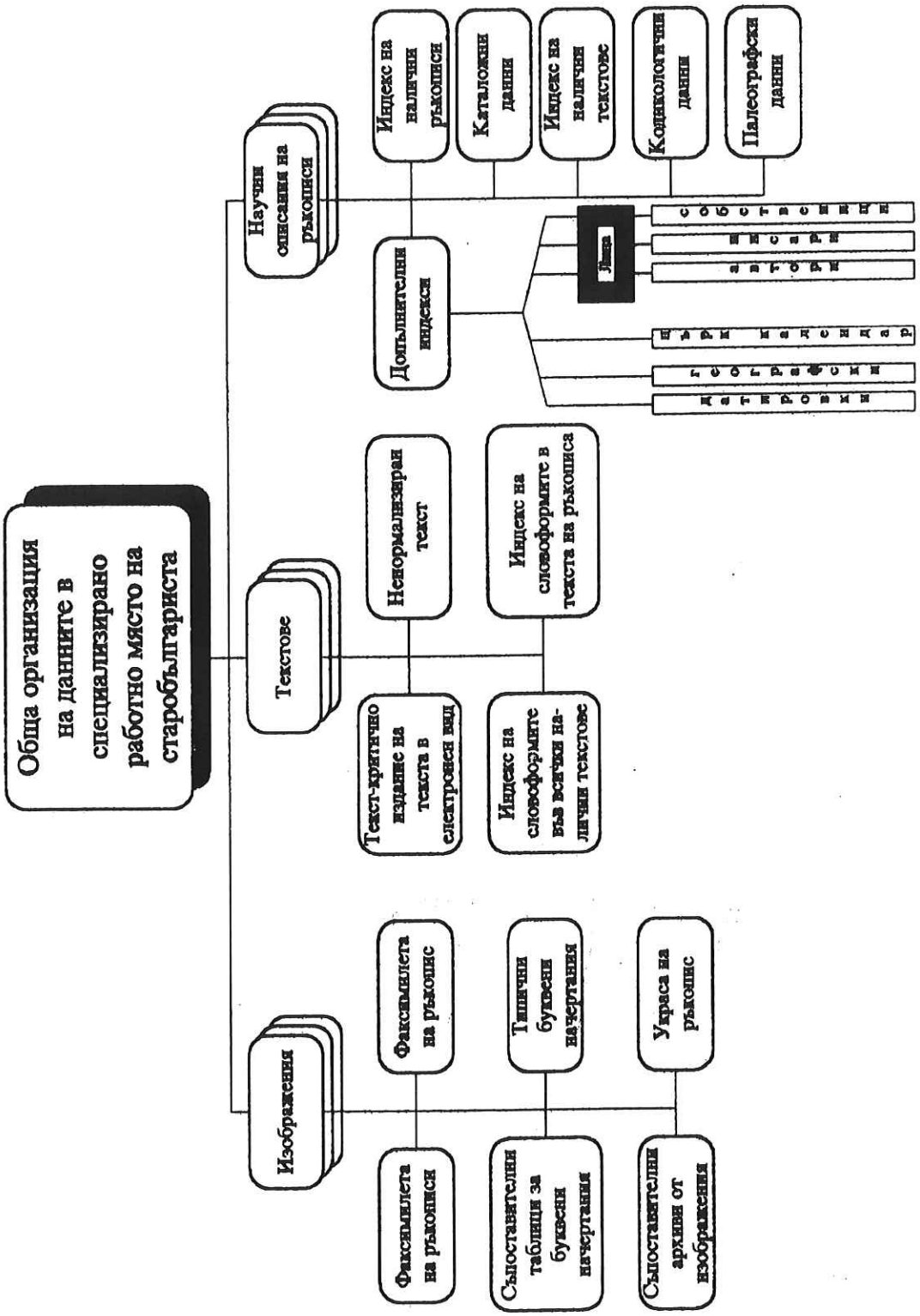
- Методите от областта на *системите за управление на бази от данни* се използват за представяне на научните описания на средновековните ръкописи.
- Методите от областта на *мултимедийните технологии* могат да се използват за интегрирано представяне на обекти от трите горни вида.

В настоящия труд е представен модел от последния вид, при който се предполага съхраняването и обработката на изображения, текстове и научни описания в интегрирана среда, каквато съвременните технологии за работа в Интернет дават възможност да се изгради. При практическата реализация е използван HTML 4.0.

На Фиг. 4.1. е представена общата организация на данните в специализираното работно място. Моделът е построен на базата на предположението, че за включените ръкописи ще се съхраняват изображения, текстове и научни описания. Изображенията представляват файлове в графичен формат (в практическата реализация е предпочетен JPG поради съображения за икономичност на представянето, което води до възможност за по-бърз достъп).

Освен факсимилета на ръкописите, като отделни изображения се съхраняват характерни образци на букви от паметниците, както и образци от орнаментацията на ръкописите. Защрихованите елементи показват, че освен изображенията, свързани с конкретен паметник, в работното място ще могат да се използват и изображенията от всички други въведени паметници, което позволява да се правят сравнения на орнаментацията, буквените начертания или на общия вид на ръкописите.

Фиг. 4.1. Обща организация на данните в специализирано работно място на старобългариста



Текстовете и научните описания са представени като отделни структури на Фиг. 4.1., въпреки че в конкретната реализация на прототипа на работното място на старобългариста се предполага, че те представляват за всеки паметник един електронен документ в TEI формат. В случая е използвана тази особеност на SGML, че при кодирането на текст в електронен вид освен самия текст електронният документ съдържа и данни за създаването на този текст. Това позволява при кодирането на старобългарски текстове да се въведат и данни, свързани с научното описание на изследвания паметник. По този начин всеки текст се кодира заедно с научното си описание в един електронен документ в TEI формат. Самият текст се вмества в елемента <TEXT>, а данните, свързани с научното описание – в елемента <TEIHeader>.

Освен текста, специализирано място от този тип трябва да включва и индекс на словоформите от всички налични паметници. В областта на старобългаристиката разработването на компютърна морфология, която би улеснила създаването на речник, е все още нерешен проблем. Затова в настоящия труд само сме маркирали необходимостта от включването на подобни данни и разработката на съответния софтуерен инструментариум.

В Приложение 2. е представена общата структура на дефиницията за тип документ, който позволява да се извършва въвеждане на научните описания и на текста за ръкописи[1]. Използването на стандарт от типа на SGML за кодиране на текста и данните от научното описание позволява в системата да се натрупват текстове, които са описани с различна степен на детайлизация и с разлики в използваните дефиниции за тип документ (DTD). Въпреки че подобен разнобой в използваните дефиниции не е препоръчителен, все

---

[1] Дефиницията е създадена за целите на работата по българо-американски съвместен проект "Computer Processing of Medieval Slavlc Manuscripts" [Dobreva 96b], [Miltenova 98].

пак използваната обща основа е предпоставка за възможен взаимен обмен между различни институции.

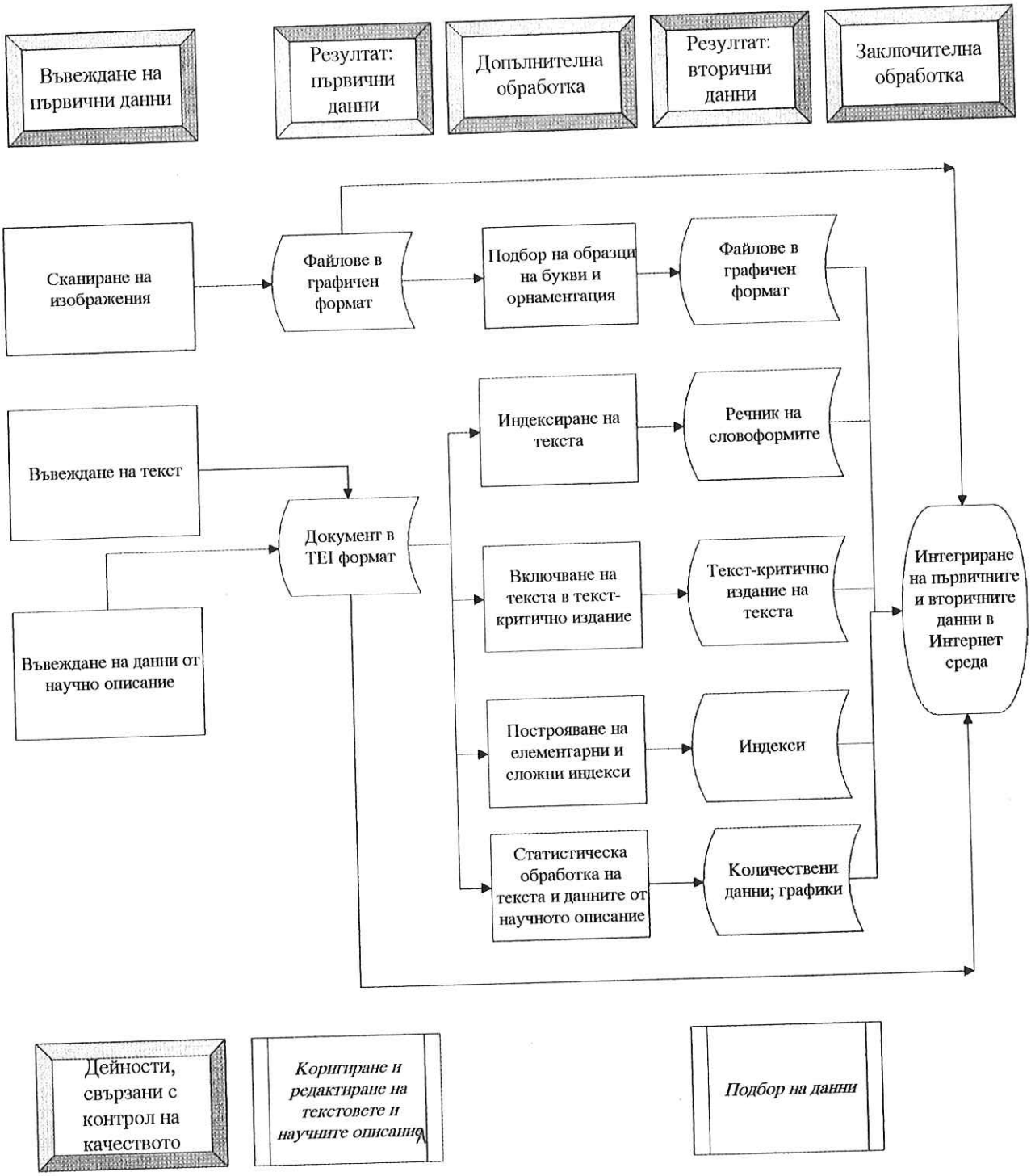
На Фиг. 4.1. като отделна подгрупа от данни са представени и индекси, извлечени на базата на научните описания на електронните документи. Едно от основните предназначения на работното място е търсенето на данни за конкретен паметник или търсенето на паметници, които да са сходни с него по определени критерии. За осъществяване на бързо търсене по различни критерии се налага извличането на индекси на онези данни от научните описания, които улесняват търсенето на конкретни паметници. В конкретната реализация индексите са извличани от описанията на ръкописи в TEI формат чрез специално разработени за целта програми на езика ICON. Тези индекси се получават като резултат от допълнителна обработка на научните описания (вж. като илюстрация Фиг. 4.4).

## 4.3. Дейности при въвеждане на данни в работното място на старобългариста

На Фиг. 4.2. е представена последователността от дейности, свързани с въвеждането на данни от различен вид за работното място на старобългариста.

На първия етап от работата се въвеждат първични данни и изображения на всички листове от ръкописа; въвеждат се текстът във вид, максимално близък до оригинала и данните от научното описание. Резултатът от работата на този етап са група от файлове в графичен формат, съдържащи изображенията на листовете от ръкописа, и един документ в TEI формат. Ако тези данни не се подложат на допълнителна обработка, резултатът от работата ще бъде просто един електронен архив от изображения и електронни документи, който обаче не може да послужи за нищо друго, освен за съхранение и разпространение на въведените ръкописи.

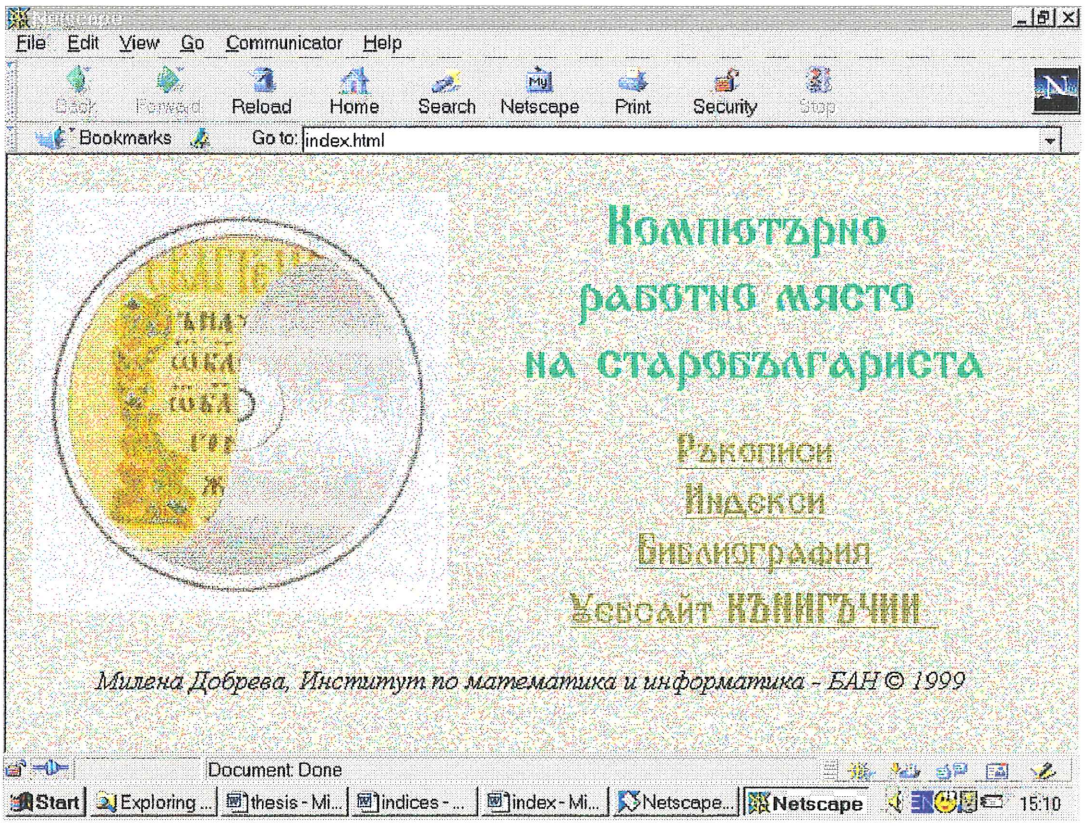
Фиг. 4.2. Дейности, свързани с въвеждане и обработка на данни за работното място на старобългариста



За да може да се подпомогне работата на старобългаристите, въведените първични данни се подлагат на допълнителна обработка. В случая на въведените изображения на ръкописи, такава обработка може да бъде насочена към обработка на изображенията чрез специализирани програмни продукти, които подобряват четливостта на текста . Полезно е също така и да се натрупат нови изображения, съдържащи характерни буквени начертания и орнаменти от всеки въведен ръкопис.

Текстът на ръкописа също се подлага на допълнителна обработка. Целта е да се получи списък на всички словоформи. Когато един ръкопис представлява препис на текст, за който вече са въведени други преписи, може да се пристъпи към колация на текста с цел подготвяне на текст-критично издание.

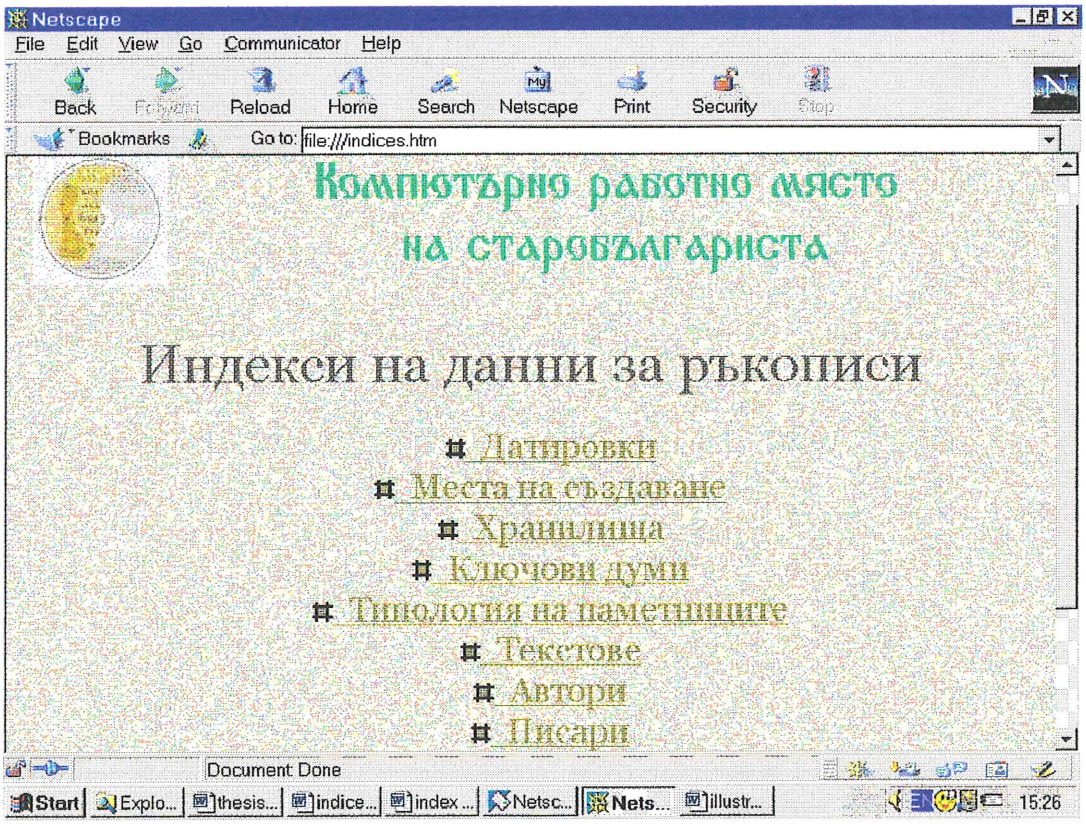
Фиг. 4.3. Начален екран на специализирано работно място на старобългариста



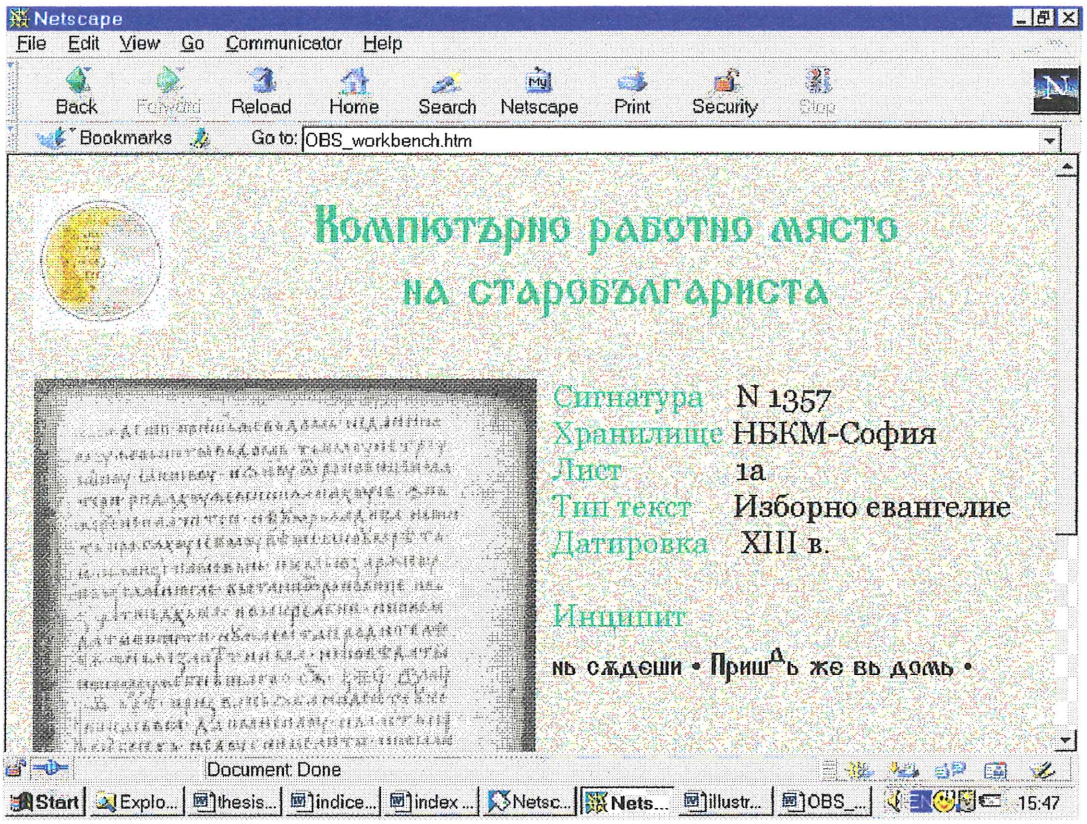
При решаването на определени изследователски задачи, на този етап също така се натрупват статистически характеристики (подобно на примера с изследването на количествени характеристики на правописа, разгледан в Глава 3. на настоящия труд). За натрупването на тези характеристики в нашите експерименти беше използвана специално разработена програма за пребройване на елементи и продуктът Statistica for Windows.

Досегашният опит показва, че изключително трудоемки при въвеждане на данни за старобългарски ръкописи са дейностите, свързани с контрола на качеството на електронните ресурси.

Фиг. 4.4. Част от индекси в специализирано работно място на старобългариста




Фиг. 4.5. Общо представяне на ръкопис в специализирано работно място на старобългариста




На фиг. 4.3, 4.4 и 4.5. са показани примери за използването на съвременните Интернет технологии за достъп до наличните данни в специализираното работно място. Фиг. 4.4. илюстрира част от възможностите за търсене на данни от научни описания на ръкописи, а Фиг. 4.5. – основното представяне на ръкопис.

## Апробация на резултатите

Повечето от резултатите в дисертацията са получени самостоятелно и са публикувани в [Dobreva 94a], [Dobreva 94b], [Dobreva 95a], [Dobreva 96b], [Dobreva 98a], [Dobreva 99]. Останалата част са получени съвместно със Ст. Керпеджиев [Dobreva, Kerpedjiev 92], Е. Паскалева [Paskaleva, Dobreva 96], С. Иванов [Dobreva, Ivanov 98] и Д. Добрев [Dobreva, Dobrev 98a]. При съвместните изслевания участието на авторите е равноправно.

Основните резултати от настоящия труд са докладвани пред:

- *Научни семинари и отчетни сесии* на секция "Телекомуникации", ИМИ-БАН, и Катедрата по Кирило-Методиевистика, ФСлФ при СУ "Св. Кл. Охридски".
- *Конференции у нас:*
  - Първа национална конференция Информатика'94, София.
  - Международна работна среща Text Variety in the Witnesses of Medieval Texts, Sofia, 21-23 September 1997.
- *Конференции и лекции по покана в чужбина:*
  - Digital Resources for the Humanities, Oxford, July 1996.
  - SGML Belux conference, Brussels, October 1996.
  - Electronic Publishing'98: ICCC/IFIP Conf. Budapest, April 1998.
  - Sixth DELOS Workshop Preservation of Digital Information, Tomar, Portugal, June 1998.
  - ALLC-ACH'98, July 5-10 1998, Debrecen, Hungary.
  - ALLC-ACH'99, June 8-12 1999, Charlottesville, USA.
  - Част от материала в докторския труд е представена по време на лекционните курсове "Компютърно представяне на данни за средновековни славянски ръкописи" (юли 1996 г.) и "Дигитализация на средновековни ръкописи" (юли 1997 г.) в рамките на Летните школи на Департамента по средновековни изследвания при Централно-Европейския Университет, Будапеща, както и в лекция по покана в Института по компютърна лингвистика в Пиза, Италия, през м. октомври 1996 г.

Части на труда са докладвани по време на 5-месечна специализация в Департамента по приложение на информатиката в хуманитарните науки в Университета в Грьонинген, Холандия, 1994 г., и по време на 3-месечна специализация в Департамента по славистика при Университета в Питсбърг, САЩ, 1995 г.

Проекти, ръководени от авторката, в областта на приложенията на ИТ в хуманитарните науки, са финансирани от Национален фонд "Научни изследвания" (МУ-ИС-6/94, МУ-ИС-2/95, МУ-О-2/96)и от Централно-Европейския Университет - Прага (RSS 125/91 и RSS 481/97). Авторката е координирала провеждането на три международни форума (с подкрепата на UNESCO през 1997 г. и Институт "Отворено общество" – Будапеща през 1998 и 1999 г.) по тематика, свързана с приложения на ИТ в медиевистиката.

Признание за изследователската работа на авторката е и академичната награда за млади учени, присъдена през 1998 г. *за*

*оригинални постижения при компютърни представяния на средновековните славянски текстове и за заслуги за утвърждаване на българщината*.

## Заключение

Представените в работата изследвания са насочени към по-плодотворното използване на възможностите на съвременните ИТ в работата на старобългаристите.

Един от проблемите в областта на използването на приложенията на компютрите за представянето и анализа на средновековни славянски текстове е създаването на подходящо компютърно представяне. То трябва да дава възможност точно да се кодират такива специфични особености като нелинейното писане, използването на съкращения и елементите на орнаментация на паметниците, които са съществени при текстологичните изследвания. Към решаването на този проблем са насочени представените в Глава 2 на настоящия труд резултати.

Следваща стъпка при приложението на съвременните ИТ е те да бъдат използвани там, където решаването на определена задача изисква прекомерно много човешки усилия. Типична такава област е количественото изследване на текстове. Средновековните славянски текстове не са били обект на подобни изследвания, въпреки че получените резултати несъмнено биха допринесли за по-доброто разбиране на разпространението на писмената култура през Средновековието. В Глава 3 на дисертационния труд са представени статистически експерименти, в резултат на които са получени съпоставителни данни за употребата на буквите и качествените признаци, характеризиращи различните редакции. Резултатите включват данни за текстове от българска, руска и сръбска редакция.

Използването на количествени методи е илюстрирано и чрез серия от експерименти, използващи клъстерен анализ. Тези

експерименти покажат, че количествените данни за текстовете от различни редакции позволяват да се извърши класифициране по клъстери съответно на произхода на текстовете при препоръчителен обем на текстовите извадки от минимум 2 килолитери за текстове от различна редакция и 5 килолитери за текстове от една и съща редакция.

В работата е разгледан и въпросът за създаване на компютърно работно място на старобългариста, в което се ползват възможностите на съвременните технологии за работа в Интернет. Предимство на ориентацията към интернет е възможността за широко разпространение на натрупваните ресурси в областта.

## Научни приноси

Старобългарските текстове като обект за компютърно представяне имат специфични особености, които трябва да се вземат предвид при разработването на информатични модели. Приложението на ИТ в старобългаристиката досега се е извършвало без да са разработени специализирани модели за представяне на старобългарските текстове в компютърен вид. Основните приноси на дисертационния труд са в следните области:

1. *Компютърно кодиране на старобългарски текстове.*

   Направен е съпоставителен анализ на три системи за кодиране на стара кирилица в компютърен вид. Систематизирани са специфични особености на старобългарските текстове, които трябва да бъдат отразявани при компютърното представяне на текстовете. Предложено е множество от елементи, които позволяват внасянето на данни за местоположението на графемите и орнаментацията в старобългарски текст според

предписанията на TEI. Разработени са дефиниции за предложените елементи със средствата на SGML.

2. *Компютърно-подпомогнат количествен анализ на старобългарски текстове.*

   Направени са изследвания с количествени методи на употребата на проявите на качествени признаци, характеризиращи основните редакции на старобългарския език (българска, руска и сръбска). Проведени са експерименти за изследване на различията между редакциите на старобългарския език чрез клъстерен анализ. Въз основа на направените изводи от резултатите от експериментите са предложени общи методически препоръки за планиране и извършване на компютърно-подпомогнати количествени изследвания на старобългарски текстове.

3. *Разработка на модел на мултимедийно специализирано работно място на старобългариста.*

   Разработен е модел на мултимедийно специализирано работно място на старобългариста, ориентиран към използване в Интернет с цел препоръки за разработчиците. За практическа демонстрация на работното място е реализиран експериментален прототип, като са използвани възможностите на HTML 4.0.

* * *

Предлаганият дисертационен труд е в област, в която сме свидетели на бурно начално развитие. Докато в други страни тематиката, свързана с приложения на ИТ за представяне на културно-историческото наследство е национален приоритет, у нас няма национална стратегия за работа в областта. Подобни начинания изискват влагането на сериозни ресурси - и като техника, и като високо квалифициран труд на специалисти.

Тази работа е извършена с желанието да се допринесе за бъдещото развитие на достойното представяне на нашето културно-историческо наследство чрез възможностите, предлагани от новите информационни технологии.

## Литература


[Велчева 80] Б. Велчева, Праславянски и старобългарски фонологически изменения, С., Изд. на БАН, 1980.

[Граматика 93] Авторски колектив под ред. на Ив. Дуриданов, *Граматика на старобългарския език*, Издателство на БАН, 1993 г.

[Добрева, Бояджиев 93] М. Добрева, А. Бояджиев, Проблеми, свързани с компютърното обработване на текстове, написани на православните славянски езици. В. Сборник от международен симпозиум Методология на математическото моделиране, Велико Търново, 1993, стр. 180-185.

[Иванова 80] Кл. Иванова, Правописно езикови проблеми при описването на южнославянските кирилски ръкописи, В: Славянска палеография и дипломатика, С., 1980, 102-109 стр.

[Лабынцев 96] Ю. Лабынцев, Международный исследовательский проект "Slavia Orthodoxa et Slavja Romana : Взаимодействие славянских миров: Духовная культура Подляшья", В: D. Birnbaum, A. Bojadzhlev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, стр. 168-l76.

[Лампе 96] Р. Лампе, Создание единного сводного каталога славянского письменного наследия, В: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference .'Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, стр. 159-l67.

[Лотошко 98] Ю. Лотошко, Компьютерный анализ, графическая и фонетическая вариативность русских текстов XVII века в сравнении с современныч русским языком, В: M. Dobreva (ed.) Text Variety jn the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S. 1998, pp. 70-76.

[Минчева 87] А. Минчева, Българският език през XIII в., В: Българската литература и книжнина през XIII в., С., 1987 , 39-45 стр.

[Молдован 96] А. Молдован, О создании фонда письменных источников русского языка XI-XII вв. в Институте русского языка РАН. В: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, стр. 94-95.

[Москов, Бояджиев 77] М. Москов, Ж. Бояджиев, Увод в езикознанието, С., 1977.

[НПОБФ 1997] Национална програма за опазване на библиотечните фондове (проект), София, СБИР, 1997, 71 стр.

[Пики и др., 97] Е. Пики, Д. Камуля, М. Йовчева, М. Камуля, Програмата DBT - перспективи при компютърната обработка на средновековни текстове, В: Palaeobulgarica/ Старобългаристика XXI (1997), 1, стр. 3-21.

[Тяжелникова 96] В. Тяжелникова, Русская версия источнико-ориентированного программного обеспечения КЛИО: новые возможности обработки источников, В: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, стр. 257-266.

[Щавинская 96] Л. Щавинская, Основные задачи и принципы построения банка данных Международной славяноведческой программы "История книжной культуры Подляшья", В: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, стр. 177-184.

[Юшин 96] И. Юшин, Источнико-ориентированный подход и проблемы социальной истории, В: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts" Blagoevgrad, July 1995, S., 1996, стр. 277-290.

[Янакиев 96] М. Янакиев, Компютърните технологии и текстологията, В: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International

Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, стр. 315-322.

[Янакиев 98] М. Янакиев, Приветственное слово к участником рабочего совещания по теме Текстовая вариативность в списках Средневековья , В : M . Dobreva ( ed . ) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Vvorkshop, Sofia, 21-23 September 1997, S., 1998, pp. 7-10.

[Angusheva 96 ] A. Angusheva, Computer Investigation of Medievaf Prognostic Books by Means of KLEIO, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedlngs of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 291-297.

[Azarova 96] I. Azarova, Processing Slavonic BibIe Texts in the Russian BibIe Society (Sankt-Petersburgh branch), In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 89-93.

[Bakker 96a] M. Bakker , Computer Collation of Manuscript Transcription, In : D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 96.

[Bakker 96b] M. Bakker, Towards a Critical Edition of the Old Slavic New Testament (A Tranparent and Heuristic Approach), PhD Thesis, University of Amsterdam, 1996.

[Barnbrook 92] G. Barnbrook, ComputerAnalysis ofspelling Variants in Chaucer's Canterbury Tales, In: J. Svartvik et al. (eds.), New Directions in English Language Corpora, 1992, p. 277-287.

[BBDM 96] O. Birnbaum, A. Bojadzhiev, M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference Computer Processing of Medieval Slavic Manuscripts, 24-28 July 1995, Blagoevgrad, S., 1996, 336 pp.

[Biber 88] D. Biber, Variation across speech and writing, Cambridge University Press, 1988, 299 pp.

[Birnbaum 96a] D. Birnbaum, How Slavic Philologists Should Use Computers, In: D. Birnbaum, A. Bojadzhiev, M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts". Blagoevgrad July 1995, S., 1996, pp. 19-28.

[Birnbaum 96b] D. Birnbaum, Informational and Presentatlonal Units in Early Cyrillic Writings. In: D. Birnbaum, A,. Bojadzhiev M. Dobreva, A.Miltenova (eds.) Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S. 1996, pp. 41-49

[Birnbaum 96c] Standardizing Characters, Glyphts and SGML Entities for Encoding Early Cyrillic Writing, In: Computer Standards & Interfaces 18 (199-6)., pp. 201-252.

[Bojadzhiev 96] A,. Bojadzhiev, Paleographic and Orthographic Features in the Description of Slavic Manuscripts, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Mjltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S. 1996: pp. 140-146.

[Bozzi et al. 86] Andrea Bozzi, A. Nikolova, G. Capeiii, G. Giuliani, Il trattamento delle varianti nello spoglio elettronico di un testo, Una prova sui Carmina di Claudiano. In: Materiali e discussioni per analisi dei testi classicii N 16, Pisa, 1986, pp. 155-179.

[Bratanova et al. 96] V. Bratanova, E, Kotseva, D. Petrov, The St. Sophia Project. A Stage in its Realization and the Propects for its Completion, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.). Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 238-256.

[Buzzetti, Rehbein 98] D. Buzzetti, M. Rehbein, Textual Fluidity and Digital Editions , In: M. Dobreva (ed.) Text Variety In the Witnesses of .Medieval Texts. Proceedings of the Int. Wokshop, Sofiaf 21-23 September 1997, S., 1998, pp. 14-39.

[Calabretto, Rumpler 98] S. Calabretto, B. Rumpier, Distributed Multimedia Workstation for Medieval Manuscripts, In. F. Rowiand and J.Smith (eds.), Electronic Publishing'98:

Towards the Information-Rich Society, Proceedings of an ICCC/IFIP Conference, Budapest, Hungary , 20-22 April 1998, pp. 166-178,

[Caligaris et al. 92] Caligaris C., et al., An Integrated Environment for Lexjcaf Analysis, In.: Proc. of COLING-92, Nantes, pp. 935-939.

[Camugiia 96a] M. Camuglia, The Psalter, Its Tradition and the Computer: a New Method of Textual Anarysis. In: PalaeoЬulgarica/Старобългаристика, xx (1996), vol.1, p. 3-13.

[Camuglia 96b] M. Camuglia, The Psalter: from the Oral to the Informatic tradition. In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Mjltenova (eds.), Proceedings of the First International Conference 'Computer Processing of Medieval Slavic Manuscripts',. Blagoevgrad, July 1995, S. 1996: pp. 185-196

[Camuglia, Picchi 98] M. Camuglia, E. Picchi, Towards an Ancjent Slavonic Electronic Dictionary, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997, S., 1998, pp. 116-122.

[Church, Rau 95] Church, K. L. Rau, Commercial Appllcations of Natural Language Processing, In: New Horizons in Commercial and Industrial Artificial Intelligence, a special Issue of Communications of the ACM, vol. 38 (11), 1995, pp. 71-79.

[Cleminson 98] R.M. Cleminson, The Early Printed Book as a Textual Variant, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 50-6O.

[Damerau 64] F. Damerau, Techniques for Computer Oetection and Correction of Spelling Errors, In: Communications of the ACM, vol. 7, pp.171-176.

[Dees 87] Atlas des formes Linguistjques des textes Litterairs de Ancien francais, Tübingen, 1987.

[Denning et al. 87] K. Denning, S. Inkelas, F. macNair-Knox, J. Rickford, Variation in Language: 15th Annual Conference on New Ways of Analyzing variation in Language, Stanford, CA, 1987.

[Dimitrova 96] M. Dimitrova, Loanwords in the New York Missal and the KLEIO Computer Program, In: D. Birnbaum, A. Bojadzhiev, M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts" , Blagoevgrad, July 1995, S., 1996, pp. 298-314.

[Dimitrova 98] D. Dimitrova, On the Role of Computer Description of Manuscripts with the TSM Project in the Research of Compiled Texts, In: M . Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997, S., 1998, pp. 157-160.

[Dobreva 94a ] M . Dobreva , Applications of Image Processing to the Study of Medieval Manuscripts, B: Сборник от първа национална конференция Информатика'94, С., 1994, стр. 90-97.

[Dobreva 94b] M. Dobreva, Word Processing of Texts Written in Medieval Slavic Languages: Tools and ProbIems, In: International Journal Information Theories and Applications Vol. 2, No. 9 (1994), pp. 3240,

[Dobreva 95a] M. Dobreva, Applications of Computer Tools in Studying Medieval Slavonic Manuscripts, Sofia, 1995, 66 pp.

[Dobreva 96a] M. Dobreva, Problems in Design and Use of TEI-based Repertoire of Slavic manuscripts, In: Int. Conference Digital Resources for the Humanities, Oxford, 1-3 July 1996, pp. 10-11.

[Dobreva 96b] M. Dobreva, Use of SGML by Philologists, In: Proceedings of SGML-Belux conference, Brussels, October 3O-31, pp. 39-53. http://sgmlbelux.be/96/dobreva.htm

[Dobreva 98a] M. Dobreva, The First Steps in Creating Cultural Heritage Digital Resources in Bulgaria, In : Sixth DELOS Workshop Preservation of Digital Information, Tomar, Portugal, 17-19 June 1998, ERCIM, pp. 61-65.

[Dobreva 98b] M. Dobreva (ed.) Text Varjety in the Witnesses ofMedieval Texts, Proceedings of the workshop held in Sofia, 21-23 September 1997 , S., 1998, 168 pp.

[Dobreva, Dobrev 98a] M . Dobreva, D. Dobrev , Application of Quantitative Study of the Orthographic Variety in Medieval Slavic Texts to Research and Teaching, In: M. Dobreva (ed.) Text Varlety In the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 85-98.

[Dobreva, Dobrev 98b] M. Dobreva, D. Dobrev , Orthographic Variety in Medieval Slavic Texts: How to Study and Model It? In : ALLC-ACH'98 Conference abstracts, July 5-10 1998, Debrecen, Hungary, pp.36-38.

[Dobreva, Ivanov 98] M. Dobreva, S. Ivanov, Issues in Electronic Publishing on the Medieval Slavic and Byzantine World, In: F . Rowland and J . Smith (eds.), Electronic Publishing'98: Towards the Information-Rich Society, Proceedings of an ICCC/IFIP Conference, Budapest, Hungary , 20-22 April 1998, pp. 55-64.

[Dobreva, Kerpedjiev 92] M. Dobreva, S. Kerpedjiev, Automatic Conversion of Encylcopedia Entries into a Hypertext, In: SERDICA Math. Publ., 1992, pp. 367-86.

[Edmnodson et al. 90] J. Edmondson, C. Feagini P. Muhlhauser (eds), Development and Diversity: Language Varlation across Time and Space, Summer Institute of Linguistics, 1990.

[Fasold 83] R. Fasold (ed.) Variation in the Form and Use of Language, Washington, 1983.

[Fasold, Schiffrin 89] R, Fasold, D. Schiffrin (eds.), Language Change and Variation, In: Series: Current Issues in Llngulstic Theory, Vol. 52, Amsterdam Studies in the Theory and History of Linguistic Science, John Benjamins, 1989.

[Ferrara et al. 88] K. Ferrara, B. Brown, K. Walters, J. Baugh (eds.), Linguistic Change and Contact, Proceedings of the 16th Annual Conference on New Ways of Analyzing Variation, Austin, Texas, 1988,

[Fetkova 98] P. Fetkova, Problems in Editing Medieval Slavonic Texts, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 123-125.

[Frith 80] U. Frith (ed.), Cognitive Processes in Spelling, Academic Press, 1980.

[Gagova 98] N. Gagova, TSM: The Element NOTE: Improvement of the Encoding Possibilities: A Philological Proposal, In: M. Dobreva (ed,) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int, Workshop, Sofia, 21-23 September 1997, S. 1998, pp. 161-164.

[Geurts et al. 87] A.J. Geurts, A. Gruijs, J. van Krieken, W.R. Veder, Codicography and Computer, In: Полата книгописъната. Vol. 17-18 (87), pp. 4-29.

[Grünberg 96] K. Grünberg, Transcription Rules for Old Church Slavonic Writing, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 50-67.

[Guergova 96] E. Guergova, Computational Analysis of Hymnographic Compendia, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 207-214.

[Havermans 98] J. Havermans, TNO's Research on Paper Conservation and Future Possibilities, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedlngs of the Int. Workshop, Sofia, 21-23 September 1997, S.. 1998, pp. 133-136.

[Helgerson 88] Linda W . Helgerson, CO-ROM and Scholarly Research in the Humanities. In : Computers and the Humanities, Vol. 22 (1988), pp. 111-116.

[Hristova 96] B. Hristova, The First Bulgarian CO-ROM: Bulgarian Letters in the Context of Balkan Literature, Tetraevangella and Qurans, In : D.Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference .'Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 332-335.

[Huber 89] Onno Huber, The construction of Lexically Analyzed Text Corpora on the Computer, Vrije Universiteit, Amsterdam, 1989, 116 pp.

[ISO 8879, 86] International Organization for Standardization, ISO 8879: Information processing - Text and office systems - Standard Generalized Markup Language (SGML), Geneva, ISO, 1986.

[ISO 9: 1995 (E)] Information and Documentation - Transliteration of Cyrillic characters int Latin characters - Slavic and non-Slavic Ianguages, ISO, 1995-02-15.

[Ivanova et al., 96] T. Ivanova, N. Shojleva, I. Bolcheva, Advanced Searching, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 97-102.

[Jankoska et al. 96] S. Jankoska, D. Mihajlov, L. Josifovski, Oatabase of Slavonic Manuscripts in Macedonia, In: O. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the Flrst International Conference "Computer Processing of Medieval Slavic Manuscripts'., Blagoevgrad, July 1995f S., 1996, pp. 147-152.

[Jannedy et al. 94] S. Jannedy, R. Poletto, T. Weldon (eds.), Language Files, 6th edition, Columbus, 1994, see p. 36l-398.

[Kempgen 96] S. Kempgen, Complex Script-Systems on Today's Personal Computer, In : D. Birnbaum, A. Bojadzhiev . M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 68-78.

[Klasinc, Kurent 96] P. Klasinc, V. Kurent, First Results ofTesting the Computer Recognition ofManuscripts, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 323-331.

[Koychev, Dobreva 94] I. Koychev, M. Dobreva, Appllcation ofLearning Ьy Example Method to Dating Medieval Bulgarian Manuscripts, In : International Journal Information Theories and Applications, vol. 2 (1994), No. 5, pp. 38-44.

[Kruskal, Sankoff 83] J. Kruskal, D. Sankoff (eds.) Time warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison, Addison-Wesley I 1983.

[Kučera 98] K. Kučera, Some Aspects of Orthographic Variety in a Changing Writing System, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997, S., 1998, pp. 77-84.

[Leib 93] H.-H. Leib, Linguistic Variables: Towards a Unified Theory ofLinguistlc Variation, In : Current Issues of Linguistic Theory N 108, John Benjamins, 1993, 258 pp.

[Lucinskiene 98] M. Lucinskiene, Computerlsation of Old Lithuanian Scripts of XVI-XVII Centuries, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedlngs of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 126-132.

[LVC] Language Variation and Change (journal), eds. D, Sankoff, W. Labov, Д. Kroch, Cambridge University Press. 1989 (vol. 1)-1995 (vol. 7).

[MacRobert 96] M. MacRobert, A TEI View ofMS Pec 68, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 79-88.

[Majstoroska 96] M. Majstoroska, A Church Slavonic Alphabet or Reprinting Old Manuscripts Using a Microcomputer, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 103-113.

[Milic 95] L.T. Milic, The Century of Prose Corpus: A Half-Million Word Historical Oata Base, In: Computers and the Humanities, Vol. 29, N 5, 1995, pp. 327-337.

[Miltenova 96] A. Miltenova, Computer Assisted Analysis of the Macrostructure and Typology of Medieval Slavic Miscellanies, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 129-129.

[Miltenova 98a] A. Miltenova, Variety of Wltnesses of Slavic Written Tradition (welcome address), In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 11-12.

[Miltenova 98b] A. Miltenova, Computer Repertory of Medieval Literature and Letters, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997, S., 1998, pp. 138-149.

[Mitrevski 96] L. Mitrevski, Fund, Fonts and Character Sets of Church Slavic Graphemes, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 114-118.

[Orlandi 85] Shenute contra Origenistas, Testo con Introduzione e traduzione a cura di Tito Orlandi, Roma, CIM - 1985, pp. 143.

[Orlandi 90] Tito Orlandi, The Corpus dei Manoscritti Copti Literali- In: Computers and the Humanities, VoI. 24 (1990)1 pp. 397-405.

[Paskaleva, Dobreva 96] E. Paskaleva, M. Dobreva, New Tools for Old Language: Computer Processing of Bulgarian Texts, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts.., Blagoevgrad, July 1995, S., 1996, pp. 30-39.

[Radoslavova 98] D. Radoslavova, SGML Tagging of Complex Texts in Old Slavic Liturgical Manuscripts, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 165-168.

[van Reenen 88] P. van Reenen, K. van Reenen-Steln (eds.), Distribution spatiales et temporelles, constellations des manuscrits. John Benjamins B.V., 1988.

[Ribarov, Ribarova 98] K. Ribarov , Z. Ribarova, A Time for Corpus of Old Church Slavonic, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 61-68.

[Ribarova, Ribarov 96] Z. Ribarova, K. Ribarov, Computer Processlng of Old Church Slavonic Manuscripts: Results and Prospects, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 153-158.

[Robertson, Willett 93] A. Robertson, P. Willett, A Comparlson ofSpelling Correction Methods {or the Identification of Word Forms in Historical Text Databases, In: Literary and linguistic Computing, vol. 8, no. 3, 19931 pp. 143-152.

[Rousseau, Sankoff 78] P. Rousseau, D. Sankoff, Advances in Variable Rule Methodology, In: D. Sankoff (ed.), linguistic Varuiation: Models and Methods, Academic Press, 1978,

[Salter 96] F. Salter, Some suggestions Arising from a Personal Computer Analysis of the svetostefansko Hrlsovuljer In : D. Birnbaum, A. BoJadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts.', Blagoevgrad, July 1995, S., 1996, pp. 228-237.

[Sankoff 78] D. Sankoff (ed.), Linguistic Variation: Models and Methods. Academic Press, 1978.

[Shniter 96] M. Shniter, A System {or Encoding Euchological (mainly noncalendar) Texts within the Parameters of the IST Computer Program , In : D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 215-227.

[Stoyanov 96] I. Stoyanov, Optical Character Recognition of Historical Documents, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp, 119-127.

[Stoykova 98] A. Stoykova, Variety and standardization of Old Slavonic Texts (Problems of Creating a Data Base), In: M. Dobreva (ed.) Text Variety in the Witnesses of

Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997, S., 1998, pp. 150-156.

[van der Tak 98] J. van der Tak, The Handling of Variation in Old Bulgarian Apostolos Texts, In: M. Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997, S., 1998, pp. 40-49.

[Tikhonov 96] V. Tikhonov, The Problems of Categorlsation in Content Analysis, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 267-276.

[Unicode 91] The Unicode Standard, World Wide Character Encoding. Version 1.0, Vol, 1. Addison-Wesley, 1991.

[Vakareliyska 96] C. Vakareliyska, Medieval Slavic Menologies Оп Line, In: D. Birnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 197-206.

[Veder 96] W. Veder, Transcription and Edition, In: O. Blrnbaum, A. Bojadzhiev. M. Dobreva, A. Miltenova (eds.), Proceedings of the First International Conference "Computer Processing of Medieval Slavic Manuscripts", Blagoevgrad, July 1995, S., 1996, pp. 96.

[Veder 98] W. Veder, Вариация в кругу семьи О писменехъ, In: M, Dobreva (ed.) Text Variety in the Witnesses of Medieval Texts, Proceedings of the Int. Workshop, Sofia, 21-23 September 1997 , S., 1998, pp. 99-115.

# Приложение 1.

## Примери на компютърно представяне на стара кирилица

| У. Федер | Д.Бирнбаум | Evangelie | Символ |
|---|---|---|---|
| **1. БУКВИ**<br>**(112 символа)** | **(205 символа)** | **(58 символа)** | |
| A | &Aos; | A | |
| a | &aos; | A | |
| | &A1os; | | |
| | &a1os; | | |
| B | &Bos; | B | |
| b | &bos; | B | |
| | &B1os; | | |
| | &b1os; | | |
| V | &Vos; | V | |
| v | &vos; | v | |
| | &V1os; | | |
| | &v1os; | | |
| | &V2os; | | |
| | &v2os; | | |
| G | &Gos; | G | |
| g | &gos; | g | |
| | &HARDGos; | | |
| | &hardgos; | | |
| D | &Dos; | D | |
| d | &dos; | d | |
| | &SOFTDos; | | |
| | &softdos; | | |
| E | &Eos; | E | |
| | &E1os; | | |
| | &e1os; | | |
| E2 | &E2os; | | |
| e2 | &e2os; | | |
| | &E3os; | | |
| | &e3os; | | |
| | &E4os; | | |
| | &e4os; | | |
| | &E5os; | | |
| | &e5os; | | |
| | &E6os; | | |
| | &e6os; | | |
| ZH | &ZHos; | > | |
| zh | &zhos; | < | |

| У. Федер | Д.Бирнбаум | Evangelie | Символ |
|---|---|---|---|
| | &DZHRos; | |  |
| | &dzhros; | |  |
| Z2 | &DZos; | |  |
| z2 | &dzos; | |  |
| | &DZ1os; | |  |
| | &dz1os; | |  |
| | &DZ2os; | |  |
| | &dz2os; | |  |
| Z | &Zos; | Z |  |
| z | &zos; | z |  |
| | &IOCTos; | |  |
| | &ioctos; | |  |
| I | &IOCT1os; | I |  |
| i | &ioct1os; | i |  |
| 1* | &IDECos; | J |  |
| | &IDEC1os; | |  |
| | &idec1os; | |  |
| | &IDEC2os; | |  |
| | &idec2os; | |  |
| 1h* | &IDEC3os; | |  |
| 1h | &idec3os; | |  |
| J | &Jos; | |  |
| j | &jos; | |  |
| | &J1os; | |  |
| | &j1os; | |  |
| K | &Kos; | K |  |
| k | &kos; | k |  |
| L | &Los; | L |  |
| l | &los; | l |  |
| | &SOFTLos; | |  |
| | &softlos; | |  |
| M | &Mos; | M |  |
| m | &mos; | m |  |
| | &SOFTMos; | |  |
| | &softmos; | |  |
| N | &Nos; | N |  |
| n | &nos; | n |  |
| | &N1os; | |  |
| | &n1os; | |  |
| | &SOFTNos; | |  |
| | &softnos; | |  |
| | &SOFTN1os; | |  |
| | &softn1os; | |  |
| O | &Oos; | O |  |
| o | &oos; | o |  |
| | &BROADOos; | |  |

| У. Федер | Д.Бирнбаум | Evangelie | Символ |
|---|---|---|---|
| | &broadoos; | | |
| | &OOCos; | | |
| | &oocos; | | |
| | &OBOCos; | | |
| | &obocos; | | |
| | &OBOC1os; | | |
| | &oboc1os; | | |
| | &omocos; | | |
| P | &Pos; | P | Π |
| p | &pos; | p | π |
| R | &Ros; | R | Ρ |
| r | &ros; | r | ρ |
| | &rros; | | |
| S | &Sos; | S | C |
| s | &sos; | s | c |
| T | &Tos; | T | Τ |
| t | &tos; | t | τ |
| | &T1os; | | |
| | &t1os; | | |
| | &t2os; | | |
| | &DJos; | | |
| | &djos; | | |
| U | &Uos; | | |
| u | &uos; | | |
| Ou | &Ouos; | | |
| OU | &OUos; | U | |
| O4 | &Ou1os; | | |
| O4* | OU1os; | | |
| o4 | ou1os; | | |
| | &U1os; | | |
| | &u1os; | | |
| | &USos; | | |
| | &usos; | | |
| F | &Fos; | F | Φ |
| f | &fos; | f | φ |
| TH | &THos; | } | |
| th | &thos; | { | |
| X | &Xos; | X | Χ |
| x | &xos; | x | χ |
| OH | &OMos; | ] | |
| oh | &omos; | [ | ω |
| | &OMTITos; | | |
| | &omtitos; | | |
| OH+T | &OMTOS; | ) | |
| oh+t | &omtos; | ( | |
| | | ^ | |
| WT | &SHTos; | W | |

| У. Федер | Д.Бирнбаум | Evangelie | Символ |
|---|---|---|---|
| wt | &shtos; | w | Ѱ |
| C | &TSos; | C | Ц |
| | &tsos; | | ц |
| | &DZos; | | Ѕ |
| | &dzos; | | ѕ |
| CH | &CHos; | Q | Ꙗ |
| ch | &chos; | q | ꙗ |
| | &KOPos; | | Ҁ |
| | &kopos; | | ҁ |
| SH | &SHos; | | Ш |
| sh | &shos; | | ш |
| 7* | &BJERos; | | Ъ |
| 7 | &bjeros; | | ъ |
| 6*1* | &JERYos; | | Ы |
| 61 | &jeryos; | | ы |
| | &JERY1os; | | Ы́ |
| | &jery1os; | | ы́ |
| 7*1* | &JERYBos; | | Ъы |
| 71 | &jerybos; | | ъы |
| 7*1h* | &JERYB1os; | | Ъ́ы́ |
| 71h | &jeryb1os; | | ъ́ы́ |
| 7*I | &JERYB2os; | | ЪИ |
| 7*i | &jeryb2os; | | ъи |
| 6* | &FJERos; | Y | Ѣ |
| 6 | &fjeros; | | ѣ |
| | &NJERos; | | Ь |
| | &njeros; | | ь |
| | &paerokos; | | Ҍ |
| | &erikos; | | ҍ |
| EH | &JATos; | | Ѣ |
| eh | &jatos; | | ѣ |
| 1H*EH | &JATJos; | | Ѣ |
| 1heh | &jatjos; | | ѣ |
| ju | &juos; | | Ю |
| | &JUROs; | | Ѻ |
| | &juros; | | ѻ |
| JA | &JAos; | | Я |
| ja | &jaos; | | я |
| JE | &JEos; | | Е |
| je | &jeos; | | е |
| 8* | &JUS1os; | | Ѧ |
| 8 | &jus1os; | | ѧ |
| | &JUSJ2os; | | Ѩ |
| | &jusj2os; | | ѩ |
| 8H | &JUS3os; | | Ѫ |
| 8h | &jus3os; | | ѫ |
| | &JUS4os; | | Ѭ |

| У. Федер | Д.Бирнбаум | Evangelie | Символ |
|---|---|---|---|
| | &jus4os; | |  |
| | &JUS5os; | |  |
| | &jus5os; | |  |
| Q | &JUSBos; | |  |
| q | &jusbos; | |  |
| | &JUSBLos; | |  |
| | &jusblos; | |  |
| J8* | &JUSJLos; | |  |
| j8 | &jusjlos; | |  |
| JQ | &JUSJBos; | |  |
| jq | &jusjbos; | |  |
| J8H | &JUSJ3os; | |  |
| j8h | &jusj3os; | |  |
| K3 | &KSIos; | |  |
| k3 | &ksios; | |  |
| P2 | &PSIos; | |  |
| p2 | &psios; | |  |
| 4 | &izhos; | y |  |
| | &IZH1os; | |  |
| | &izh1os; | |  |
| | &IZHKos; | |  |
| | &izhkos; | |  |
| | &YNos; | |  |
| | &ynos; | |  |
| | | |  |
| **2. ДИАКРИТИЧНИ ЗНАЦИ** за числови означения | | | |
| | &mult4os; | |  |
| | &mult5os; | |  |
| | &mult6os; | |  |
| | &mult7os; | |  |
| | &mult8os; | |  |
| | &mult9os; | |  |
| | | | |
| **3. СЪКРАЩЕНИЯ** | | | |
| | | |  |
| // | &titos; | |  |
| | &tit1os; | |  |
| | &tit2os; | |  |
| | &pokos; | |  |
| | | | |
| **4. УДАРЕНИЯ** | | | |
| | &palos; | |  |
| | &roughos; | |  |
| | &smoothos; | |  |
| | &acuteos; | |  |
| | &graveos; | |  |
| | &dacuteos; | |  |

| У. Федер | Д.Бирнбаум | Evangelie | Символ |
|---|---|---|---|
| | &longaos; | | ᾱ |
| | &circos; | | ᾶ |
| | &ibreveos; | | ᾰ |
| | &breveos; | | ᾰ |
| | &macronos; | | ᾱ |
| | &odotos; | | ᾽ |
| | &diaeros; | | ῎ |
| | | | |
| **5. ПУНКТУАЦИОННИ ЗНАЦИ** | | | |
| ' | &apostos; | | ᾽ |
| + | &plusos; | | + |
| , | &commaos; | | , |
| . | &periodos; | | . |
| | &colonos; | | · |
| ; | &semios; | | ; |
| | &refmkos; | | ※ |
| | &middotos; | | · |

# Приложение 2.

## Елементи в DTD за представяне на средновековни славянски ръкописи

<TEI.2>
<TEIHEADER>
<FILEDESC>
<TITLESTMT>
<TITLE>
<AUTHOR>
<EDITOR>
<FUNDER>
<PRINCIPAL
<SPONSOR>
<PUBLICATIONSTMT>
<PUBLISHER>
<PUBPLACE>
<DATE>
<SOURCEDESC>
<CATALOGUESTMT>
<MANUSCRIPTNAME>
<MANUSCRIPTLOCATION>
<REPOSITCOUNTRY>
<REPOSITCITY>
<REPOSITORY>
<REPOSITSIGNATURE>
<CATALOGNR>
<RELATEDPERSON>
</MANUSCRIPTLOCATION>
</CATALOGUESTMT>

```
            </SOURCEDESC>
</FILEDESC>
<ENCODINGDESC></ENCODINGDESC>
<PROFILEDESC>
            <LANGUSAGE>
            <CODICOLOGY>
                  <NUMFOLIO>
                  <QUIRESTRUCTURE>
                        <QUIRE>
                              <NUM>
                              <COMPOSITIONQUIRE>
                  <PAGINATION>
                  <PRICKING>
                  <BINDING>
                  <MATERIALDESC
                  TYPE=paper|vellum|papyrus
                  EXTENT=general|partial
                                    missing>
                        <LAYOUT>
                              <NUMFOLIO>
                              <SIZEMATERIAL
                              TYPE=vertical|horizonal
                              RANGE=material|written
                                          area
                              UNIT=cm|inch>
                              <RULELINE>
                              <NUMBCOLUMN>
                              <NUMBLINES>
                        </LAYOUT>
                        <INK>
```

```
                <WATERMARK>
                <GREGORYRULE>
                <ORNAMENT
                TYPE=borders|cadels|
                calendars|capitals|
                initials|illustrations|
                linefillers|vjaz>
                <MISCOBSERVAT
                TYPE=miscdamage|
                restoration|palimpsest>
                <ALPHABET>
        </MATERIALDESC>
</CODICOLOGY>
<SCRIBE>
        <NAME>
        <PAGERANGE>
                <STARTINGPAGE>
                <ENDINGPAGE>
        <ORTHOGRCHARACT>
        <PALAEOCHARACT>
</SCRIBE>
<MANUSCRIPTCONTENTDESC
TYPE=compilation|original|
translation
STYLE=narrative|non-narrative>
<NUMBERTEXTS>
<MANUSCRIPTCREATION>
        <MANUSCRIPTDATE>
        <MANUSCRIPTPLACE>
<SOURCE TYPE=Greek|other>
```

```
                <TRANSLATION>
                    <NUM>
                    <DATE>
                <PROTOGRAPH>
                <ANTIGRAPH>
                <LITREDACTION>
            </MANUSCRIPTCONTENDESC>
            <ARTICLECONTENTDESC
                ID
                TYPE=compilation|original|
                translation
                STYLE=narrative|non-narrative>
                <NUMBERTEXTS>
                <ARTICLENAME>
                <ARTICLEAUTHOR>
                <SOURCE TYPE=Greek|other>
                <TRANSLATION>
                    <NUM>
                    <DATE>
                <ANTIGRAPH>
                <APOGRAPH>
                <CHURCHCALENDAR>
            </ARTICLECONTENDESC>
        </PROFILEDESC>
        <REVISIONDESC></REVISIONDESC>
        </TEIHEADER>
         <TEXT>
            <BODY>
                <DIV>
                    <HEAD></HEAD>
```

```
                    <INCIPIT>
                        <NORMINCIPIT>
                        <NONNORMINCIPIT>
                    </INCIPIT>
                    <P></P>
                    <EXPLICIT>
                        <NORMEXPLICIT>
                        <NONNORMEXPLICIT>
                    </EXPLICIT>
                </DIV>
            </BODY>
        </GROUP>
    </TEXT>
    </TEI.2>
```