Лемматизаторы для Национального корпуса русского языка

Проблема

НКРЯ включает тексты разных периодов и разновидностей русского языка:

– древнерусский (11-14 век = 880 тыс + 23 тыс БГ)
– старорусский (15-17 век = 9.3 млн)
– церковнославянский (17-18 век = 5.3 млн)
– современный русский (19-21 век)
– диалектный etc.

Разные методы лемматизации:

– современный – автоматическая (mystem) + ручная правка
– старорусский – полуавтоматическая (parser) + ручная правка
– церковнославянский – словарь словоформ (dicgram)
– древнерусский – ручная лемматизация

Проблема: орфография vs. орѳографія (старорусский, ЦС, современный)

Определения

Лемматизатор

Программа, выполняющая лексико-грамматический анализ текста:
1) токенизация (слово, знак препинания, цифры, тег разметки)
2) лемматизация для слов, имеющихся в словаре (лемма+граммемы)
3) построение гипотез для нераспознанных (несловарных) слов
4) снятие омонимии для нескольких вариантов (если возможно)

Грамматическая модель

Формальное описание словоизменения.

Грамматический словарь

Список лексем с приписанной информацией о словоизменении:
1) основа с указанием чередований
2) постоянные признаки лексемы (часть речи, род, вид, переходность etc.)
3) код словоизменительного типа (парадигмы).

Словоизменительный тип (парадигма)

Набор флексий (включая схему чередования), общий для некоторого множества лексем.

Грамматический словарь

Пример	Основа	Граммемы	Парадигма
день	де*н+ь	N,m,inan	N12*
кольцо	кол(е|ь)ц+о	N,n,inan	N24*
сказка	сказо*к+а	N,f,inan	N33*
белеть	беле+ть	V,ipf,intr	V1
велеть	вел+еть	V,pf-ipf,intr	V5
читать	чита+ть	V,ipf,tr	V1
лепетать	лепе(т|ч)+ать	V,ipf,intr	V6
трепетать	трепе(т|щ)+ать	V,ipf,tr	V6tj
врать	вр+ать	V,ipf,intr	V6o
брать	б(|е)р+ать	V,ipf,tr	V6o*
держать	держ+ать	V,ipf,tr	V5
сжать	с(жа|ожм)+ть	V,pf,tr	V14
сжать	с(жа|ожн)+ть	V,pf,tr	V14
ходить	хо(д|ж)+ить	V,ipf,intr	V21j
носить	но(с|ш)+ить	V,ipf,tr	V21j
бить	б(и|ь|е)+ть	V,ipf,tr	V13
сбить	с(би|обь|бе)+ть	V,pf,tr	V13
жечь	ж(г|ж|ег|е)+чь	V,ipf,tr	V8*g
сжечь	с(ожг|ожж|жег|же)+чь	V,pf,tr	V8*g

Таблицы парадигм

Существительные

Пара-дигма	N11	N11*	N12	N12*	N13	N13*	N14	N14b
Пример	стол, новосел	посо*л, осе*л	кон+ь, олен+ь	ого*н+ь, де*н+ь	поток, сапог	дымо*к, кул(е|ь)к	плач, муж	врач, нож
sg,nom	#	2#	ь	2ь	#	2#	#	#
sg,gen	а	а	я	я	а	а	а	а
sg,dat	у	у	ю	ю	у	у	у	у
sg,ins	ом	ом	ем	ем	ом	ом	ем	ом
sg,loc	е	е	е	е	е	е	е	е
pl,nom	ы	ы	и	и	и	и	и	и
pl,gen	ов	ов	ей	ей	ов	ов	ей	ей
pl,dat	ам	ам	ям	ям	ам	ам	ам	ам
pl,ins	ами	ами	ями	ями	ами	ами	ами	ами
pl,loc	ах	ах	ях	ях	ах	ах	ах	ах

Прилагательные

Парадигма	**A1 – A1b**	**A1***	**A2**	**A3 – A3b**	**PA6**	**PN1in**	**PN1**
Пример	нов+ый жив+ой	уме*н+ый, бу(й|е)ный	син+ий	строг+ий сух+ой	мо+й	дядин	Попов, Фомин
sg,m,nom/acc	ый – ой	ый	ий	ий – ой	й	#	#
sg,n,nom/acc	ое	ое	ее	ое	е	о	о
sg,m/n,gen	ого	ого	его	ого	его	ого/а	а
sg,m/n,dat	ому	ому	ему	ому	ему	ому/у	у
sg,m/n,loc	ом	ом	ем	ом	ем	ом	е
sg,m/n,ins	ым	ым	им	им	им	ым	ым
sg,f,nom	ая	ая	яя	ая	я	а	а
sg,f,acc	ую	ую	юю	ую	ю	у	у
sg,f,gen/dat/loc	ой	ой	ей	ой	ей	ой	ой
sg,f,ins	ой/ою	ой/ою	ей/ею	ой/ою	ей/ею	ой/ою	ой/ою
pl,nom	ые	ые	ие	ие	и	ы	ы
pl,gen/loc	ых	ых	их	их	их	ых	ых
pl,dat	ым	ым	им	им	им	ым	ым
pl,ins	ыми	ыми	ими	ими	ими	ыми	ыми
predic,sg,m	#	2#	ь	#	–	–	–
predic,sg,f	а	а	я	а	–	–	–
comp	ее	ее	ее	–	–	–	–

Глаголы

Парадигма	V1	V6t	V4t	V11	V8k, V8g	V8g*
Пример	дела+ть, гре+ть, ду+ть	пи(с|ш)+ать пря(т|ч)+ать пла(к|ч)+ать	но(с|ш)+ить тра(т|ч)+ить пу(ст|щ)ить	б(и|ь|е)+ть с(би|обь|бе) +ть	пе(к|ч|)+чь стри(г|ж|)чь	ж(г|ж|ег|е)чь с(ожг|ожж|ж ег|же)+чь
ind,pres,sg,1	ю	2у	2у	2ю	у	у
ind,pres,sg,2	ешь	2ешь	ишь	2ешь	2ешь	2ешь
ind,pres,sg,3	ет	2ет	ит	2ет	2ет	2ет
ind,pres,pl,1	ем	2ем	им	2ем	2ем	2ем
ind,pres,pl,2	ете	2ете	ите	2ете	2ете	2ете
ind,pres,pl,3	ют	2ут	ят	2ют	ут	ут
ptcp,pres,act	ющ+$A4	2ущ+$A4	ящ+$A4	2ющ+$A4	ущ+$A4	ущ+$A4
ptcp,pres,pass	ем+$A1	2ем+$A1	им+$A1	2ем+$A1	ом+$A1	ом+$A1
cvb,pres	я	2а	я	2я	я	я
imp,sg,2	й	2ь/2и	ь/и	3й	и	и
imp,pl,2	йте	2ьте/2ите	ьте/ите	3йте	ите	ите
ind,past,sg,m	л	ал	ил	л	#	3#
ind,past,sg,f	ла	ала	ила	ла	ла	ла
cvb,past	в/вши	ав/авши	ив/ивши	в/вши	ши	3ши
ptcp,past,act	вш+$A4	авш+$A4	ивш+$A4	вш+$A4	ш+$A4	3ш+$A4
ptcp,past,pass	нн+$A1*2	анн+$A1*2	2енн+$A1*2	т+$A1	2енн+$A1*2	2енн+$A1*2
inf	ть	ать	ить	ть	3чь	4чь

Алгоритм работы

Разбить слово на основу+флексию и проверить в словаре основ и флексий

нашей = на+шей, наш+ей, наше+й, нашей+#

1) на=PREP + шей=ptcp,past,act,sg,f,gen/dat/loc/ins (пек+шей) => нет
2) наш=A-PRO=PA4 + ей=sg,f,gen/dat/loc/ins
   => совместимы => да
3) наш(и|ь|е)+ть=V,pf,tr=V11 + й=imp,sg,2=3
   => совместимы, правильный вариант основы (3) => да
4) нашей+# => основы нет

соловей = сол+овей, солов+ей, солове+й, соловей+#

1) сол+ь + овей=pl,gen | сол+ить + овей=pl,gen
   => парадигмы несовместимы => нет
2) солов+ый=A=A1 + ей=comp
   => совместимы => да
3) солове+ть=V,ipf,intr=V1 + й=imp,sg,2
   => совместимы => да
4) солов(ь|е)+й=N,m,anim=N12* + й=sg,nom=2
   => совместимы, правильный вариант основы (2) => да

Лемматизатор для современного русского языка (mystem)

Mystem (Яндекс) включает грамматический словарь (на базе Зализняка) с указанием чередований и список парадигм.
Возможности: токенизация, лемматизация, построение гипотез.
Частично понимает старую орфографию по нашей просьбе для ФЭБ (http://feb-web.ru, словарь Грибоедова, конкордансы).

Проблемы:
1. Словарь и парадигмы встроены в программу и их невозможно изменить (*-аго, -яго, -ыя, -ия* vs. *-ыи, -ии*)
2. Лишние (маловероятные) разборы для многих частотных слов:
для = деепр. от *длить*, *при* = род. ед. от *пря*, *ли* = китайская мера длины,
их = междометие, *он* = имя буквы, *а, в, и, к, с, у* = имя буквы.
(Отсекаются с помощью фильтров.)
3. Не понимает некоторые нестандартные формы (исправлено):
деепр. сов. вида от основы презенса – *прийдя, увидя, взгромоздясь*

4. Продуктивные словообразовательные модели.
1) Субстантивированные прилагательные и местоимения (*все, это, каждый, больной, старое, белые, черные*...), которые (случайно) попали в словарь Зализняка, а другие потенциальные субстантиваты нет.
2) Отадъективные наречия (*весело, громко, легко*...) образуются почти от любого прилагательного, но в словаре есть самые частотные, а остальные трактуются как кр. ф. ср. рода (*забавно, задумчиво, интригующе*...).

5. Гипотезы выдаются в непредсказуемом порядке, правильный вариант часто стоит в конце.
Для разных форм слова выдаются разные наборы лемм:
Мономах => пр.мн. от *моном* или *монома* (~~ *домах, дамах, проблемах*)
Мономаха => *мономах* или *мономаха* (~~ *монах, росомаха*)
Байкал, *Байкала* => прош. от *байкать*
Байкалу => от сущ. *байкал* или *байкала*
салями => тв.мн. от *саль* или *саля* (~~ *королями, солями*)
интернет => наст. вр. от *интернуть*
Программа разбивает слово на основу+флексию и выбирает самые частотные комбинации: *сал+ями*, *моном+ах*.

6. Грамматическая омонимия во флексиях, особенно именных:
супруг–супруга, гарнитур–гарнитура, физик–физика,
Евгений–Евгения, германий–Германия, франций–Франция,
Пушкин–Пушкина–Пушкино, Бородин–Бородина–Бородино
→ Невозможность предсказать лемму и форму, лишние варианты.

7. Невозможно предсказать семантические признаки лексемы (одушевленность, имя собственное, вид, переходность):
полковник=одуш vs *половник*=неодуш
счетчик=одуш/неодуш, *телятник*=одуш/неодуш
гусеница=одуш/неодуш, *старица*=одуш/неодуш

8. Невозможно предсказать словоизменительный тип:
повелевать=V1 vs. *горевать*=V2
хлебать=V1 vs *колебать*=V6p
чихать=V1 vs *махать*=V6t
нарезáть=V1 vs *нарéзать*=V6t
болеть=V1 vs *болеть*=V5n

Альтернативные лемматизаторы

Dialing (Сокирко, aot.ru), sphinxsearch, pymorphy.
Словарь дополнен тысячами новых слов (имена людей, организаций etc.)
Словарь и парадигмы встроены, изменить невозможно.
Грамматическая модель примитивна, нет чередований, основа минимальна (ж+ечь, ж+ать, пи+сать), отсюда тысячи парадигм.

Лемматизатор для дореформенной русской орфографии (parser)

Проблема.
Лингвистические процессоры, ориентированные на современный русский язык, непригодны для анализа старых текстов (18–19 века), особенно в старой орфографии.

1. Спелчекер – считает ошибкой
2. Программа распознавания (FineReader) – не распознает
3. Лемматизатор (mystem, dialing) – не разбирает или неверные гипотезы

Примеры.
дома, дому, новый, милость, бездушный, ходить = совпадают (+)
домъ, домѣ, новаго, милостію, безсильныя, ходити = отличаются (–)

Примеры

Александръ Сергѣевичъ Пушкинъ
Евгеній Онѣгинъ (1837)

«Мой дядя самыхъ честныхъ правилъ,
Когда не вшутку занемогъ,
Онъ уважать себя заставилъ,
И лучше выдумать не могъ;
Его примѣръ другимъ наука:
Но, Боже мой, какая скука
Съ больнымъ сидѣть и день, и ночь,
Не отходя ни шагу прочь!
Какое низкое коварство
Полуживаго забавлять,
Ему подушки поправлять,
Печально подносить лѣкарство,
Вздыхать и думать про себя:
Когда же чортъ возметъ тебя!»

М. В. Ломоносов

Ода на победу над Турками и Татарами и на взятие Хотина 1739 года

Восторг внезапный ум пленил,
Ведет на верьх горы высокой,
Где ветр в лесах шуметь забыл;
В долине тишина глубокой.
Внимая нечто, ключь молчит,
Которой завсегда журчит
И с шумом в низ с холмов стремится.
Лавровы вьются там венцы,
Там слух спешит во все концы;
Далече дым в полях курится...
Корабль как ярых волн среди,
Которыя хотят покрыти,
Бежит, срывая с них верьхи,
Претит с пути себя склонити...
К Российской силе так стремятся,
Кругом объехав, тьмы Татар;
Скрывает небо конской пар!
Чтож в том? Стремглав без душ валятся.

Орфография+морфология?

Дореформенная	Современная
ѣ, і, ѳ, ѵ	е, и, ф, и
-ъ	–
без-, воз-, из-, низ-, ра(о)з-, ч(е)рез- + глухие	-с
ъ: съиграть, съузить, съэкономить, съзади, близълежащий	ъи=>и, –
-аго, -яго, -ыя, -ія	-ого, -его, -ые, -ие
ея, онѣ, однѣ, однѣхъ	ее, они, одни, одних
-ію (*милостію*)	
-іе (*счастіе*),-ье (*вниманье, занятье*)	
е/о после шипящих и ц (*лице, значекъ, волченокъ, чортъ)*	
слитно/раздельно/дефисно (*то-есть, повидимому, кто нибудь)*	

Словоизменение

1. Адъективные флексии (*-аго/-яго, -ыя/-ія*).
2. Усеченные формы прилагательных (*красна/о/ы/у*), которые совпадают с краткими формами, но имеют и другие падежи (*красну*).
3. Особые формы местоимений (*ея, онѣ, однѣ, однѣхъ*).
4. Творительный падеж 3-го склонения на *-ію* (*милостію, помощію*).
5. Сравнительная степень на *-яй, -яе* (*сильняй*, *сильняе, скоряе*).

6. Вариант частицы *-ся* после гласных (*валюся, валилася*), который вполне употребляется в современном языке
7. Деепричастия совершенного вида от основы презенса (*прийдя, увидя, взгромоздясь*), которые вполне употребляются в современном языке.
8. Глагольные флексии *-ти* и *-ши* (*ходиши, ходити*).

Орфографическая вариативность

1) Мягкость шипящих (*мечь–меч*, *идешь–идеш, ночьной–ночной*)

2) *Ё/о* после шипящих и *ц* (*чёрный–чорный, дьячёк–дьячок, шёл–шол, душою–душею, лицо–лице, зайцов–зайцев*)

3) Глухие/звонкие согласные в позиции нейтрализации (*возсиять–воссиять, восток–возток, дерзкий–дерзский–дерсский–дерский*)

4) Фонетические варианты *и/й/ь* (*дыхание–дыханье, счастие–счастье, вариант–варьянт, италиянский–итальянский*)

5) Прописные/строчные (*Английский, Англичане, Славенский*)

6) Слитно/раздельно/дефисно (*кто нибудь, повидимому, по русски, до тла, вшутку, то-есть, кто-бы–ктоб, небойся, долголь)*

Mystem

Mystem частично понимает старую орфографию путем приведения ее к современной и знает некоторые старые флексии (*-аго, -ыя, онѣ, онѣхъ*).
Эта функция была добавлена по нашей просьбе для проекта ФЭБ (http://feb-web.ru), где многие тексты даны в старой орфографии.
Она неплохо работает для 19 века, но для 17–18 века нужна серьезная переделка словаря и грамматических таблиц, а mystem закрыт.

При приведении старой орфографии к новой часть информации теряется:
1) Смешиваются квази-омонимы (ѣ–е):
слѣзъ – слезъ, сѣлъ – селъ, свѣдѣніе – сведеніе, Вѣна – вена, ѣсть – есть, объѣмъ – объемъ, тѣстѣ – тестѣ, тѣмъ – темъ, всѣ – все, чѣмъ – чемъ, морѣ – море
2) Ошибочные написания = правильные:
ѣлка=елка, өѵзика=фізіка=физика, історія=исторія

Parser

Собственный лемматизатор с возможностью настройки и адаптации к различным вариантам языка и орфографии.
Основные принципы:

1. Нормализация

Текст преобразуется во внутреннее (нормализованное) представление, которое унифицирует орфографические различия:
1) заменяет старые буквы на современные эквиваленты (*ѣ=е і=и ѳ=ф ѵ=и*)
2) отсекает конечный -ъ;
3) заменяет начальные *без-/в(о)з-/из-/низ-/раз-/роз-/ч(е)рез-* => -с перед глухими (NB);
4) заменяет недопустимые сочетания букв на правильные:
– убирает лишние еры внутри слова (*съиграть=>сыграть, съузить, съэкономить, съагитировать, съзади, близълежащий*);
– нормализует гласные после шипящих (*чюжой=>чужой, жывой=>живой, ночьной=>ночной*). etc.

2. Грамматическая модель

Вся конкретно-языковая информация (леммы, основы, парадигмы, флексии, граммемы) не фиксирована в коде программы, а вынесена в отдельные таблицы, которые легко редактировать. (См. выше)

Основа – словарь Зализняка, парадигмы переработы в формальный вид. Добавлены старинные или неучтенные формы (флексии), которые имеют специальные пометы и их можно включать/отключать.

old1=19 век
1) Адъективные флексии (*+аго/яго, +ыя/ия*).
2) Творительный падеж 3-го склонения на *-ию* (*милост+ию, помощ+ию*).
3) Особые формы местоимений (*ея, он+ѣ, одн+ѣ, одн+ѣхъ*).
4) Мест. ед. среднего рода на -и (*о копь+и, варень+и, здань+и*).

old2=18 век
1) Усеченные формы прилагательных (*красн+а/о/ы/у*).
2) Сравнительная степень на *-яе/ае* (*сильн+яе, чуж+ае*), *-яй/ай* (*сильн+яй, чуж+ай*), *-ея* (*сильн+ея*).

3) Глагольные флексии *-ти* и *-ши* (*ходи+ши, ходи+ти*).
4) Множ. число среднего рода на *-ы/-и* (*озер+ы, войск+и, лиц+ы*).
5) Множ. число пригательных на *-ыи/-ии* (*нов+ыи, син+ии, живущ+ии*).

old3=церковнославянский (частично)
1) Множ. число существительных (*град+и, град+ом, град+ѣх, град+ми, кон+ѣх, кон+ьми*)
2) Множ. число прилагательных (*нов+ии*)
3) Флексия *-а* вместо *-я* после *-и* (*Ефреми+а, здани+а*)
4) Формы аориста (*ходи+х, ходи, ходи+хом/сте/ша, пек+ох, печ+е, пек+охом/осте/оша*)
и имперфекта (*любл+ях/яше/яхом/ясте/яху, печ+ах/аше/ахом/асте/аху*)
5) Инфинитив на *-щи* (*пе+щи*)

Не сделано:
1) ЦС палатализация, отсутствующая в русском (*отроц+ѣ, отроц+ы, отроч+е, хожд+аху*).
2) Словарь современный, ЦС слова не отделены от русских, смешанные формы (*бере+щи, берег+ох, город+и* вм. *бре+щи, брег+ох, град+и*).

3. Грамматические фильтры

1. Глаголы сов. вида не имеют форм презенса (**сделающий,*сделаемый*), соотв. формы означают будущее (*делаю*=ind,pres vs. *сделаю*=ind,fut).

2. Непереходные глаголы не имеют форм пассива (*варимый, варенный* vs **веримый, *веренный*), кроме конструкций (*хожено, сижено*), и не имеют медиальных форм (-*ся*), кроме (*не сидится, не гуляется ему*).

3. Многократные глаголы (*сиживать*) не имеют форм презенса (**сиживаю*).

4. Безличные глаголы (*тошнить*) имеют только формы pres,sg,3 (*тошнит*) и past,sg,n (*тошнило*).

5. Относительные прилагательные (*вчерашний*) не имеют кратких форм и сравнительной степени.

6. Аккузатив обычно совпадает с генитивом для одуш. и с номинативом для неодуш., кроме некоторых парадигм (*сестру, лошадь, новую, его, их*).

Реализация

Модель лемматизатора можно потестировать:
http://dic.feb-web.ru/russian/parser/parser.htm
Эта модель написана на Javascript и работает прямо в браузере.

Реальный лемматизатор написан на Python и используется для анализа текстов 17-18 века.

Лемматизатор для церковнославянского корпуса

Церковнославянский воспринимается как архаичный книжный русский.

Церковнославянизмы глубоко пропитали русский язык:
вообще, восток, восторг, восхищение, власть, враг, вред, глава, награда, здравствуй, изверг, исключить, привлечь, существовать...

Некоторые грамматические формы имеют ЦС происхождение:
1) активные причастия (*делающий, сделавший* vs *делая, сделав*)
2) превосходная степень (*сильнейший, высочайший* vs *сильнее, выше*)

Словарь Академии Российской (1789–1794) дает ЦС и русский вместе:
брада=борода глава=голова гласъ=голосъ градъ=городъ здравъ=здоровъ кладязь=колодезь крава=корова краткий=короткий младый=молодой млеко=молоко езеро=озеро есень=осень
мощь=мочь нощь=ночь свѣща=свѣча жещи=жечь влещи=влечь двигнути=двинуть излияти=изливать изыти=изойти отъяти=отнять
Словарь ЦС и русского языка (1847) дает отдельно с пометой *церк*.

Кодировка

НІР

_о='ч~е на'шъ, и='же _е=си` на нб~сјь'хъ, да ст~и'тся и='мя твое`: да пріи'детъ цр\ствіе твое`: да бу'детъ во'ля твоя`, ја='кw на нб~си`, и= на земли`: хлјь'бъ на'шъ насу'щный да'ждь на'мъ дне'сь: и= w=ста'ви на'мъ до'лги на'шя, ја='кw и= мы` w=ставля'емъ должникw'мъ на'шымъ: и= не введи` на'съ въ напа'сть, но и=зба'ви на'съ w\т лука'вагw: ја='кw твое` _е='сть цр\ствіе и= си'ла и= сла'ва во вјь'ки. А=ми'нь.

Unicode (упрощенный)

о́ч҃е на́шъ, и́же є̀сѝ на нб҃сѣ́хъ, да ст҃и́тсѧ и́мѧ твоѐ:
да прі́идетъ црс̃тві҃е твоѐ: да бу́детъ во́лѧ твоѧ̀, я́кѡ
на нб҃сѝ, и на землѝ: хлѣ́бъ на́шъ насу́щный да́ждь
на́мъ дне́сь: и ѡста́ви на́мъ до́лги на́шѧ, я́кѡ и мы̀
ѡставля́емъ должникѡ́мъ на́шымъ: и не введѝ
на́съ въ напа́сть, но изба́ви на́съ ѿ̀ лука́ваг҃ѡ: я́кѡ
твоѐ є́сть црс̃тві҃е и си́ла и сла́ва во вѣ́ки. Ами́нь.

Отче нашъ, иже еси на небесѣхъ, да святится имя твое:
да пріидетъ царствіе твое: да будетъ воля твоя, якш
на небеси, и на земли: хлѣбъ нашъ насущный даждь
намъ днесь: и остави намъ долги наша, якш и мы
оставляемъ должникомъ нашымъ: и не введи насъ
въ напасть, но избави насъ отъ лукаваго: якш твое
есть царствіе и сила и слава во вѣки. Амінь.

ѻ́ч҃е на́шъ, и́же є́сѝ на нб҃сѣ́хъ, да с҃ти́тсѧ и́мѧ твоѐ: да прїи́детъ цр҃твїе твоѐ: да бу҃детъ во́лѧ твоѧ̀, ꙗ́кѡ на нб҃сѝ, и́ на землѝ: хлѣ́бъ на́шъ насу҃щный да́ждь на́мъ днє́сь: и́ ѡ́ста́ви на́мъ до́лги на́шѧ, ꙗ́кѡ и́ мы̀ ѡ́ставлѧ́емъ должникѡ́мъ на́шымъ: и́ не введѝ на́съ въ напа́сть, но и́зба́ви на́съ ѿ лу҃ка́ваго: ꙗ́кѡ твоѐ є́сть цр҃твїе и́ си́ла и́ сла́ва во вѣ́ки. Ѧ́ми́нь.

Орфография

1. Дополнительные буквы

И=І=Ѵ, О=Ѻ=Ꙋ, Ѿ=От, Е=Є, Оу=Ꙋ, Ꙗ=Ѧ, Ф=Ѳ, S=З, Ѯ=Кс, Ѱ=Пс, Ъ=’

Греческие слова

Сохраняют этимологическое написание *и–і–ѵ, о–ѡ, ф–ѳ, ѯ, ѱ*:
аѵгустъ–авраамъ, акѵ̋ндінъ, анѳраѯъ, вавѵ̋лѡнъ, ідѡлъ, ікѡна, ісаа́къ–иса́въ, скиніа–скѵменъ, сімѡнъ–сѵ̋меѡнъ, ѳѵ̋міамъ.

И–І

Позиционно: І+гласная/й vs. И в прочих случаях:
пріиду–пришедъ, пріяти–призвати, тихій–тихимъ.
NB. Различение корней: *миръ* (pax) vs *міръ* (mundus), *вино* vs *віна*;
миротворецъ–міродержецъ–мѵроносица, винный–вінный.

Ꙋ–О, Ѿ—От

ѡ, ѿ пишется в префиксах (особенно глагольных), иначе *о, от*:
ѡблачати – облакъ, общій; ѡсмотрити – осмь; ѿродіе – отрокъ; ѿцвѣсти – отцы; ѿчаятися – отчій; ѿрада – ѡтра́ва; ѿрасль – ѡтребіе; ѿтрясти.
Колебания ѡ/о: *ѡбразъ=образъ* (но *ѡбразовати*).

Ѡ–О

Различение форм:
1) ед. vs. множ. число (*милости* vs. *милѡсти*, *домомъ* vs. *домѡмъ*);
2) вин. vs. род. падеж адъектив. (*новаго* vs. *новагѡ*, *моего* vs. *моегѡ*);
3) прилагательное vs. наречие (*сильно* vs. *сильнѡ*).

Ѻ–О

ѻ (широкое) в начале слова и корня (*ѻтецъ, праѻтецъ*), иначе *о* (узкое).

Є–Е

є (широкое) в начале слова, иначе *е* (узкое).
NB. Различение форм ед. vs. множ. числа: *конемъ* vs. *конємъ, іерее* vs. *іереє, єлени* vs. *єлєни*.

Оу–Ꙋ

оу в начале слова, иначе Ꙋ (*оучити–наꙊчити, оумный–безꙊмный*).

Ꙗ–Ѧ

ꙗ в начале слова, иначе ѧ (*ꙗвити–ѡбъѧвити, ꙗритисѧ–разъѧритисѧ*)
NB. Различение корней *ꙗзыкъ* (народ) vs. *ѧзыкъ* (орган).
Можно было бы использовать з vs. s (*языкъ* vs. *ѧзыкъ*)

2. Ударение

Оксия (') заменяется на варию (`) в конце фонетического слова, не перед клитикой: *жена̀ – жена́ же, ты̀ – ты́ бо, повелѝ – повели́ ми*.
Камора (^) служит для различения форм ед. vs. множ./дв. числа: *благи́мъ* vs. *благи̂мъ*, *дѣ҃а́нiа* vs. *дѣ҃а̂нiа*, *раба́* vs. *раба̂*, *рабы́* vs. *рабы̂*, но здесь с ней конкурирует различие е–є и о–ѡ (см. выше).
Придыхание ставится автоматически над первой гласной слова.

3. Титла и буквотитла.

Служат для сокращения сакральных слов:
аг҃гл҃ъ, ап҃слъ, бг҃ъ, блгд҃ть, бл҃гослови́ти, блж҃енный, бцд҃а, влд҃ка, воскр҃нiе, гдс҃ь, дв҃а, дх҃ъ, дш҃а, єv҃гліе, iерс҃лимъ, iи҃ль, iи҃съ, крс҃тити, крс҃тъ, кр҃щеніе, млс҃рдiе, млс҃ть, мл҃тва, мрд҃ый, мт҃и, мцс҃ъ, нб҃о, нбс҃ный, нн҃ѣ, оц҃ъ, оч҃е, оч҃ества, првд҃ный, прд҃теча, про҃рокъ, прс҃толъ, ржс҃тво, сн҃ъ, сп҃сеніе, сп҃съ, ст҃ити, ст҃ый, сщ҃е́нникъ, тро҃ца, хрс҃тосъ, хс҃ъ, цр҃квь, црс҃тво, цр҃ь, чл҃къ, чл҃ческій, учн҃къ, учт҃ль.
Буквотитло обычно не ставится над начальной буквой, поэтому при чтении нужно его нужно сдвинуть влево:
бцд҃а→бд҃ца, гдс҃ь→гс҃дь, мрд҃ый→мд҃рый, мцс҃ъ→мс҃цъ.

Орфография в корпусе

Не нужно воспроизводить точный типографский вид, особенно элементы, затрудняющие чтение и поиск (придыхания, буквотитла, оу/ѹ=>у, ꙗ/ѧ=>я).
Но нужно сохранить существенные элементы, служащие для различения лексем или грамматических форм (*о–ѡ, и–і–ѵ, ф–ѳ, з–ѕ*).

Словоформы в корпусе даются в точном виде (*е–є, о–ѻ*, титла, ударения):
єлень, ѻтецъ, мт҃и, млс̃ть.
Леммы даются в нормализованном виде (*е, о*, титла раскрыты):
елень, отецъ, мат҃и, милость + *августъ, анѳраѯъ, вавѵлѡнъ*

Для поиска можно использовать три орфографические системы:
1) Точная – для поиска точных словоформ (*е–є, о–ѻ*, титла, ударения)
2) Упрощенная – для поиска лемм, сохраняет лексические различия (*о–ѡ, и–і–ѵ, ф–ѳ, з–ѕ*), игнорирует нелексические (*е=є, о=ѻ*, титла).
3) Модернизированная – для поиска лемм и словоформ в современной орфографии (*е=ѣ, и=і=ѵ, ф=ѳ, з=ѕ, о=ѡ*), не нужно знать точное написание:
лѣто=лето, ікѡна=икона, вавѵлшнъ=вавилон, анѳраѯъ=анфракс,
но: *езеро ≠ озеро*, *дѣлати ≠ делать, агглъ ≠ ангел*.

Лемматизатор для ЦС

Задуман как расширение лемматизатора для старой орфографии:
1) Правила перевода реального написания во внутреннее (унифицированное) представление, чтобы упростить анализ.
2) Таблицы парадигм на базе описаний в грамматиках ЦС языка.
3) Мини-словарь с парадигмами для обучения модели.
4) Предсказатель для анализа новых словоформ на основе существующих.
Напустить предсказатель на список словоформ, проверить результат, исправить и дополнить словарь и парадигмы, опять напустить etc.

Проблемы.
1) Предсказатель выдает много лишних разборов, правильные часто оказываются в конце списка. Аорист и имперфект типа *дрѣлахъ, дрѣла* вторгаются в склонение существительных и порождают лишние варианты.
2) Титла и буквотитла искажают форму слова и не позволяют привести его к лемме (*аггл҃ъ=>ангелъ, апс̃лъ=>апостолъ, гл҃ати=>глаголати*).
3) Грамматики ЦС языка описывают идеальную картину, которая не соответствует реальному состоянию в текстах. Масса орфографических и словоизменительных вариантов, слово пишется по-разному в соседних строках, есть явные русизмы.

Вариативность.

1. Множ. число = камора (^) или замена о→w, е→є.
высоты́ – высоты̂ (9), высwты́ (17) или высwты̂ (1)
ми́лости – ми́лwсти (93), ми̂лости (2)

2. Множ. число = и (кони), е (іереє), іе (царіе)
дѣ́латель – дѣ́латели (36), дѣ́телє (13), дѣ́телеіе (18)

3. Церковнославянские vs. русские формы:
закẃны (15), зако́нми (1), зако́нами (9)
зако́нwмъ (59), зако́намъ (4)
зако́нѣхъ (15), зако́нахъ (1)
зако́нwвъ (71), закẃнъ (10)

Результат.
Состав текстов ограничен, новые тексты почти не появляются.
Просмотреть список словоформ (150 тыс) и вручную выбрать правильный вариант, иногда два (*рабъ–раба*).
Главное поставить правильную лемму, а остальные признаки потом.
Получен список словоформ с разборами, настроенный на существующий корпус текстов. При расширении корпуса неизбежно появятся новые словоформы и список придется расширять.

Словарь словоформ можно посмотреть здесь:
http://dic.feb-web.ru/slavonic/dicgram/index.htm.

Лемматизатор работает примитивно – выделяет слово из текста, преобразует во внутреннее (нормализованное) представление, находит слово в списке и подставляет разборы во входной текст.

Результаты работы лемматизатора можно посмотреть здесь:
http://ruscorpora.ru/search-orthlib.html.

Грамматические таблицы (парадигмы)

Парадиг-ма	N1t	N1t*	N1j	N1k	N1x	N1k*
Пример	раб+ъ	осе*л+ъ, со*н+ъ	кон+ь, цар+ь	отро(к|ц|ч)+ъ	ду(х|с|ш)+ъ	свит(к|ок|ц|ч)+ъ
sg,nom	Ъ	2ъ	ь	ъ	ъ	2ъ
sg,acc	=nom/gen	=nom/gen	=nom/gen	=nom/gen	=nom/gen	=nom/gen
sg,gen	а	а	я	а	а	а
sg,dat	у	у	ю	у	у	у
sg,loc	ѣ	ѣ	и	2ѣ	2ѣ	3ѣ
sg,ins	омъ	омъ	емъ	омъ	омъ	омъ
sg,voc	е	е	ю	3е	3е	4е
pl,nom/voc	и	и	и/ie	2ы	2и	3ы
pl,acc	ы/=gen	ы/=gen	и/=gen	и/=gen	и/=gen	и/=gen
pl,gen	wвъ/ъ^	wвъ/ъ^	ей	wвъ	wвъ	wвъ
pl,dat	wмъ	wмъ	ємъ	wмъ	wмъ	wмъ
pl,loc	ѣхъ	ѣхъ	ехъ	2ѣхъ	2ѣхъ	3ѣхъ
pl,ins	ы	ы	и/ьми	и	и	и
du,nom/acc	а^	а^	я^	а^	а^	а^
du,gen/loc	у^	у^	ю^	у^	у^	у^
du,dat/ins	ома	ома	ема	ома	ома	ома

Парадигма	N1s	N1sj	N1c*	N1a	N1i	N1e
Пример	муж+ъ	врач+ь	от(ц|ец|ч)+ъ	кра+й	агапі+й	іере+й
sg,nom	ъ	ь	2ъ	й	й	й
sg,acc	=nom/gen	=nom/gen	=nom/gen	=nom/gen	=nom/gen	=nom/gen
sg,gen	а	а	а	я	а	а
sg,dat	у	у	у	ю	ю	ю
sg,loc	и	и	ѣ!	и	и	и
sg,ins	емъ	емъ	емъ	емъ	емъ	емъ
sg,voc	у	у	3е	ю	ю	ю/е
pl,nom/voc	и/ie!	и/ie!	ы	и^	и^	є
pl,acc	ы/=gen	и!/=gen	ы!/=gen	и^/=gen	и^/=gen	и^/=gen
pl,gen	ей	ей	євъ/ъ^	євъ	євъ	євъ
pl,dat	ємъ	ємъ	ємъ	ємъ	ємъ	ємъ/ѡмъ
pl,loc	ахъ	ахъ	ѣхъ!	ехъ	ехъ	ехъ
pl,ins	ы	и^!	ы^!	и^	и^	и^
du,nom/acc	а^	а^	а^	я^	я^	а^
du,gen/loc	у^	у^	у^	ю^	ю^	ю^
du,dat/ins	ема	ема	ема	ема	ема	ема/ома

Грамматический словарь

лексема	часть	словофор-ма	граммемы
благі́й	A	бла́гъ	brev,sg,m,nom/acc
благі́й	A	бла́же	brev,sg,m,voc
благі́й	A	бла́зи	brev,pl,m,nom
благі́й	A	бла́зѣ	brev,sg,f,dat/loc | brev,sg,m/n,loc
благі́й	A	бла́зѣй	plen,sg,f,dat/loc
благі́й	A	бла́зѣмъ	plen,sg,m/n,loc
благі́й	A	бла̑ги	brev,pl,ins | brev,pl,m,acc | brev,pl,f,nom/acc
благі́й	A	блага̀	brev,sg,f,nom | brev,sg,m/n,gen
благі́й	A	блага́го	plen,sg,m,acc
благі́й	A	блага́гw	plen,sg,m/n,gen
благі́й	A	блага́я	plen,sg,f,nom
благі́й	A	блага̑я	plen,pl,n,nom/acc | plen,du,m,nom/acc
благі́й	A	бл҃гі́я	~,plen,sg,f,gen
благі́й	A	бл҃гі̑я	~,plen,pl,m,acc|plen,pl,f,nom/acc
благі́й	A	бл҃жа́йшая	~,comp,plen,sg,f,nom
благі́й	A	бл҃жа́йшіи	~,comp,plen,pl,m,nom|comp,plen,du,n/f,nom/acc
благі́й	A	бл҃же	~,brev,sg,m,voc
благі́й	A	бл҃зі́и	~,plen,pl,m,nom
благі́й	A	бл҃зѣ	~,brev,sg,f,dat/loc | brev,sg,m/n,loc

Ссылки

Русский

Парсер
http://dic.feb-web.ru/russian/parser/parser.htm

Сводный исторический словарь:
http://dic.feb-web.ru/rusdict/

Церковнославянский

http://dic.feb-web.ru/slavonic/

Корпус:
http://dic.feb-web.ru/slavonic/corpus/
http://dic.feb-web.ru/slavonic/corpus/0/bible1581/

Грамматический словарь:
http://dic.feb-web.ru/slavonic/dicgram/