УДК 81

КОРПУС ТАТАРСКОЙ ХУДОЖЕСТВЕННОЙ ЛИТЕРАТУРЫ

М. Р. Сайхунов[1]

Т. И. Ибрагимов[2]

К. Р. Галиуллин[3]

Аннотация. Работа посвящена описанию процесса создания корпуса татарской художественной литературы в Институте языка, литературы и искусства Академии наук Республики Татарстан. Рассматриваются такие вопросы, как сбор и обработка текстовых материалов, морфологическая аннотация, выбор поисковой системы и др.

Ключевые слова: татарский язык, корпусная лингвистика, художественная литература, морфологическая разметка.

В ходе корпусных лингвистических исследований возникает необходимость использования так называемых подкорпусов. Они позволяют ограничить материал определенной тематикой или сферой использования. Существуют различные виды подкорпусов: художественной литературы, фольклора, публицистики, языка деловых бумаг, диалектов и др.

Работа над корпусом художественных текстов в Институте языка, литературы и искусства Академии наук Республики Татарстан началась в конце 2016 года и включала несколько этапов:

I. **Сбор материала.** Руководство Института приложило большие усилия в целях содействия процессу сбора материала для создания данного

[1] Институт языка, литературы и искусства им. Г. Ибрагимова АН РТ, канд. филол. наук, Казань, e-mail: 6688000@gmail.com.

[2] Институт языка, литературы и искусства им. Г. Ибрагимова АН РТ, канд. филол. наук, доцент, Казань, e-mail: tavzikh.ibragimov@kpfu.ru.

[3] Институт языка, литературы и искусства им. Г. Ибрагимова АН РТ, доктор филол. наук, проф., Казань, e-mail: galiullin.kamil@mail.ru.

корпуса. При посредничестве Кабинета министров Республики Татарстан были достигнуты договоренности с различными издательствами о предоставлении электронных версий имеющейся печатной продукции. Таким образом, в течение 2017 года разработчикам удалось собрать большое количество текстов художественной литературы на татарском языке общим объемом около шести миллионов словоупотреблений.

II. **Обработка материала.** Этап обработки исходных текстов можно разделить на несколько частей:

1) *Сегментация отдельных художественных произведений из книг и сборников.* Помимо указания границ отдельного произведения, здесь подразумевается и присвоение ему определенной мета-информации:

a) Автор
b) Название
c) Дата написания
d) Жанр: роман, повесть, рассказ, поэма, стихотворение и др.
e) Тип: поэзия или проза.
f) Служебная информация, необходимая для поисковой системы.

Данный этап является одним из наиболее трудоемких. Поэтому были использованы наработки Письменного корпуса татарского языка [5] в сфере создания различных программных средств автоматизации ручной работы, которые позволили существенно ускорить процесс.

2) *Морфологическая разметка.* Эта процедура включает в себя определение части речи и всех грамматических категорий, характеризующих каждое слово в корпусе. В связи с тем, что ручная морфологическая разметка требует большого количества человеческих ресурсов и времени, было принято решение на начальном этапе воспользоваться системой автоматической разметки.

Для выполнения морфологической разметки корпуса использовали

наиболее актуальную версию международной системы машинного перевода Apertium [1], которая поддерживает огромное число языков, в том числе и татарский. Перечислим некоторые из используемых на сегодняшний день в системе Apertium тегов:

a) Части речи

<n> Существительное

<np> Имя собственное

<adj> Прилагательное

<num> Числительное

<prn> Местоимение

<det> Детерминатив

<v> Глагол

<vaux> Вспомогательный глагол

<adv> Наречие

<post> Послелог

<postadv> Посленаречие

<cnjcoo> Сочинительный союз

<cnjsub> Подчинительный союз

<cnjadv> Наречие-союз

<ij> Междометие

<abbr> Аббревиатура

<cop> Копула

<ideo> Звукоподражательное слово

b) Типы существительных

<top> Топоним

<ant> Антропоним

<cog> Фамилия

<pat> Отчество

<org> Название организации

c) Число

<sg> Единственное число

<pl> Множественное число

<sp> Единственное/Множественное: *Сез миңа игътибар итмәгез.*

d) Категория принадлежности

<px1sg> 1 лицо, единственное число

<px2sg> 2 лицо, единственное число

<px3sp> 3 лицо, единственное/множественное число

<px1pl> 1 лицо, множественное число

<px2pl> 2 лицо, множественное число

<px3pl> 3 лицо, множественное число

<px> н{I}к{I}: *-ныкы/-неке*

e) Падеж

<nom> Именительный

<gen> Родительный

<dat> Дательный

<acc> Винительный

<abl> Творительный

<loc> Местно-временной

f) Степени сравнения прилагательных

<comp> Сравнительная степень

g) Виды местоимений

<pers> Личные местоимения

h) Типы числительных

<ord> Порядковые

<coll> Собирательные: *берәү, икәү*

<dist> Разделительные: *берәр, икешәр*

i) Наклонение

<imp> Повелительное

<opt> Желательное

<evid> Оттенок неочевидности: *Бу шүрәленең дә теле бар* ***икән****!*

j) Залог

<caus> Понудительный

<pass> Страдательный

<coop> Взаимно-совместный

k) Времена глаголов

<pres> Настоящее -{E}

<past> Прошедшее неопределенное -{G}{A}н

<ifi> Прошедшее категорическое -{D}{I}

<fut> Будущее неопределенное -{I}р

<fut2> Будущее категорическое -{A}ч{A}к

<fut_plan> -м{A}кч{I}: *Моны 2018 елга кадәр тормышка* ***ашырмакчылар****.*

l) Причастие / Деепричастие

<prc_perf> -{I}п: ***Йоклап*** *яткан мәче авызына тычкан үзе* ***килеп*** *керми.*

<prc_impf> -{E}: *ул* ***уйный*** *алмады; мин* ***яза*** *башладым.*

<prc_vol> -{E}с{I}: ***эчәсем*** *килә.*

<prc_cond> -с{A}: *моны* ***алсаң*** *була* (в значении "*можешь взять*").

<prc_fplan> -м{A}кч{I} перед вспомогательным глаголом: - *Нәрсә* ***әйтмәкче идең****? - диде хатыны Маһирә. Бакчага* ***бармакчы*** *идем.*

m) Глагольные наречия

<gna_perf> -{I}п: *...ул вакытта инде кояш* ***баеп****, йолдызлар күренә башлаган иде...* (Ф.Хөсни)

<gna_cond> -с{A}: *...кайда икәнен **белсә**, миңа моның турында сөйләр иде…*

<gna_until> -{G}{A}нч{I}: *Авылның басу капкасына **җиткәнче** эңгер-меңгердә карлы юлдан озак кайта ул.*

<gna_after> -{G}{A}ч: *Берәү, патша йортын күреп **кайткач**, үз өенә ут төрткән, ди.*

n) Глагольные причастия

<gpr_past> -{G}{A}н: ***килгән** кеше; **укылмаган** китап.*

<gpr_impf> -{A} торган

<gpr_pot> -{U}ч{I}: ***сөйләүче** кеше; үз урынын **белмәүче**.*

<gpr_ppot> -{I}рл{I}к/-{A}рл{I}к: *Казанга **алып барырлык** товар бу.*

<gpr_fut> -{I}р/-{A}р: ***барыр** җир; **сөйләр** сүз.*

<gpr_fut2> -{A}ч{A}к: ***әйтеләчәк** фикер; **эшләнәчәк** эш.*

<gpr_fut3> -{E}с{I}: ***Үләсе** күбәләк ут күзенә керер.*

o) Имя действия

<ger> -{U}

<ger_past> -{G}{A}н

<ger_perf> -{G}{A}нл{I}к: *Барысының да мәсьәләне үз башында **йөрткәнлеге**, теге яктан да, бу яктан да үлчәп **караганлыгы** сизелеп тора.* (Ф. Хөсни)

<ger_ppot> -{I}рл{I}к/-{A}рл{I}к: *Арыдым, мунчага **барырлыгым** калмады.*

<ger_abs> -{U}ч{I}л{I}к: *Менә бу юллар Мәҗит Гафуриның бу дәвердә әле мәсьәләне бөтен революцион кискенлеге белән куя алмаганын, аңарда билгеле бер вак буржуа **чикләнүчелеге** һәм каршылыклары урын алганын күрсәтәләр.* (М. Җәлил)

<ger_fut> -{I}р/-{A}р: *Ярым кем **булырын** белмим, мин әле ялгыз йөрим.* (Ш.Галиев)

<ger_fut2> -{A}ч{A}к

<ger_fut3> -{E}с{I}: ***Күрәселәре** алда әле.*

<ger1> -м{A}к

<inf> Инфинитив -{A/I}рг{A}

p) Переходность

<tv> Переходный

<iv> Непереходный

q) Лицо

<p1> 1 лицо

<p2> 2 лицо

<p3> 3 лицо

<frm> Формальность

r) Модальные частицы

<qst> -м{I}

<emph> -ч{I}, -с{A}н{A}

<mod_ass> Модальная частица (*ич, ләбаса, бит*): ***Шул** күчтәнәчләрнең тәме... Чаннарны урлаганнар **бит**!*

<mod_ind> Выражение неоднозначности -д{I}р: *Хәер, анысы язмыш **эшедер** инде. Дусларын **озатадыр**.*

s) Пунктуация

<sent> Маркер предложения

<guio> Дефис

<cm> Запятая

<apos> Апостроф

<rquot> Кавычка закрывающая

<lquot> Кавычка открывающая

<rpar> Скобка закрывающая

<lpar> Скобка открывающая

III. **Выбор поисковой системы.** На сегодняшний день в сети Интернет свободно доступно несколько корпус-менеджеров и корпусных поисковых систем: The IMS Open Corpus Workbench (CWB/CQP) [3], NoSketch Engine (Manatee/Bonito) [2], Fastmorph [4] и др. Авторы настоящей статьи являются непосредственными разработчиками корпусной поисковой системы Fastmorph. Эта система хорошо зарекомендовала себя в Письменном корпусе татарского языка общим объемом 116 млн. слов как в плане скорости поиска, так и в плане функциональности и гибкости. В силу названных причин было решено использовать именно эту программу. К тому же система Fastmorph потребовала минимальных изменений для адаптации под требования корпуса художественных текстов.

Со всеми возможностями корпусной поисковой системы Fastmorph можно ознакомиться в статье «Система сложного морфологического поиска в Письменном корпусе татарского языка» [6].

IV. **Сервер и адрес.** Корпус художественных текстов размещен на сервере Академии Наук Республики Татарстан и доступен по адресу: http://litcorpus.antat.ru.

Планы.

1) Увеличение объема корпуса путем включения в него максимально возможного количества доступных художественных произведений на татарском языке, в том числе переводов с других языков и некоторых фольклорных жанров.

2) Устранение с привлечением сотрудников ИЯЛИ АН РТ случаев неоднозначности (омонимии), возникших в результате автоматической морфологической разметки системой Apertium.

3) Добавление опций поиска по определенным авторам, типам (проза, поэзия) и периодам.

Литература

1. Apertium - Открытая платформа машинного перевода [Электронный ресурс]. Режим доступа: http://wiki.apertium.org/wiki/Publications, свободный.

2. Корпус-менеджер NoSketch Engine (Manatee/Bonito). Режим доступа: https://nlp.fi.muni.cz/trac/noske, свободный.

3. Корпус-менеджер The IMS Open Corpus Workbench (CWB/CQP). Режим доступа: http://cwb.sourceforge.net, свободный.

4. Корпусная поисковая система Fastmorph. Режим доступа: https://github.com/mansayk/fastmorph, свободный.

5. Сайхунов М. Р., Ибрагимов Т. И., Хусаинов Р. Р. Письменный корпус татарского языка [Электронный ресурс]. Казань, 2012. Режим доступа: http://corpus.tatar, свободный.

6. Сайхунов М. Р., Хусаинов Р. Р., Ибрагимов Т. И. Система сложного морфологического поиска в Письменном корпусе татарского языка. // Традиционная культура тюркских народов в изменяющемся мире: материалы I Международной научной конференции (12-15 апреля 2017 г.). – Казань: Изд-во «Ак Буре», 2017. – С. 382-385.

TATAR BELLETRISTIC LITERATURE CORPUS

M. R. Saykhunov[4]

T. I. Ibragimov[5]

K. R. Galiullin[6]

Summary. The work is devoted to the description of creating the Tatar Belletristric Literature Corpus in the G. Ibragimov Institute of Language, Literature and Art of the Tatarstan Academy of Sciences. We consider such issues as collecting and processing text materials, morphological annotation, choosing the search system, etc.

Keywords: Tatar language, corpus linguistics, belletristics, morphological markup.

[4] G. Ibragimov Institute of Language, Literature and Art of the Tatarstan Academy of Sciences, Candidate of Philological Sciences, Kazan, e-mail: 6688000@gmail.com.

[5] G. Ibragimov Institute of Language, Literature and Art of the Tatarstan Academy of Sciences, Candidate of Philological Sciences, Docent, Kazan, e-mail: tavzikh.ibragimov@kpfu.ru.

[6] G. Ibragimov Institute of Language, Literature and Art of the Tatarstan Academy of Sciences, Doctor of Philological Sciences, Prof., Kazan, e-mail: galiullin.kamil@mail.ru.