УДК 81.322.2

Касымбай Айгерим, Карабаева С.Ж., Сагынбекова Гульназ
И.Арабаев атындагы КМУнун маалыматтык технологиялар кафедрасынын магистранты,
ф.и.к, Н.Исанов атындагы КМКТАУ доценти
И.Арабаев атындагы КМУнун маалыматтык технологиялар кафедрасынын магистранты
Касымбай Айгерим, Карабаева С.Ж., Сагынбекова Гульназ
Магистрант кафедры информационных технологий КГУ им. И. Арабаева,
К.ф.н., доцент КГУСТА им. Н. Исанова,
Магистрант кафедры информационных технологий КГУ им. И. Арабаева
Kassymbay Aigerim, Karabaeva S.Zh., Sagynbekova Gulnaz
Master student of the Department of Information Technologies of KSU named after I. Arabaeva,
Ph.D., Associate Professor of KSUST named after N. Isanova,
Master student of the Department of Information Technologies of KSU named after I. Arabaeva

КЫРГЫЗ ТИЛИНИН УЛУТТУК КОРПУСУНУН ЭЛЕМЕНТТЕРИН ИШТЕП ЧЫГУУ ТУУРАЛУУ

О РАЗРАБОТКЕ ЭЛЕМЕНТОВ НАЦИОНАЛЬНОГО КОРПУСА КЫРГЫЗСКОГО ЯЗЫКА

ON DEVELOPMENT OF ELEMENTS OF THE NATIONAL CORPUS OF THE KYRGYZ LANGUAGE

Аннотация: Бул макалада Интернетте ачык жеткиликтүү тексттердин негизинде жарым-жартылай автоматтык түрдө стандарттуу жазма кыргыз тилинин толук сөздүгүн түзүү аракети келтирилген. Кыргыз тилинин улуттук корпусунун тутумун өнүктүрүүнүн көйгөйлөрү жана келечеги каралды.

Аннотация: В данной статье кратко представлена попытка полуавтоматического создания полной лексики стандартного письменного кыргызского языка на основе текстов, свободно доступных в Интернете. Рассмотрены проблемы и перспективы разработки системы национального корпуса кыргызского языка.

Annotation. The present paper briefly presents an attempt to generate semiautomatically a full-form lexicon of standard written Kyrgyz based on texts freely available on the internet. Such a lexicon could be used for developing future resources, especially if it relies on widely accepted standards. Are being considered the problems and prospects of the development of the system of national corpora of the Kyrgyz language.

Негизги сөздөр: Кыргыз тили; морфология; акыркы абалдагы машина; корпуска негизделген лексика; суффиксзация.

Ключевые слова: кыргызский язык, морфология, конечный автомат, лексика основанная на корпусе, суффикация.

Keywords: Kyrgyz language, morphology, finite-state machine, corpusbased lexicon, suffixation.

Очевидно, что в мире существует около 3000 типов языков. Кыргызский язык становится все более важным вследствие экономических и политических проблем в процессе глобализации.

Кыргызский язык является родным для более семи миллионов человек. Начиная с 1989 года, он является государственным языком Кыргызской Республики [7].

Несмотря на свой официальный статус в Кыргызстане, кыргызскому языку все еще не хватает ресурсов NLP, особенно открытых.

Поскольку мы живем в кибер-веке, процесс обмена информацией набирает обороты. Таким образом, важность кыргызского языка имеет решающее значение в процессе изучения переводов для межкультурного общения и дипломатических отношений.

Начало систематическому исследованию кыргызского языка положил Касым Тыныстанов. Он поставил задачу выписать все возможные слова, которые можно построить по правилам языка. Для этого он разработал и реализовал (с помощью передвижения вертикальных бумажных полосок) алгоритм для перебора всех возможных сочетаний букв кыргызского алфавита, при помощи чего был составлен список объемом около 100000 слов [1].

Словарный запас кыргызского языка содержится в различных (в том числе электронных) словарях.

Существует множество разных методов формирования лингвистических данных, используемые в машинном переводе. Один из таких методов построения основывается на больших аннотированных данных. Эти данные являются лингвистической базой для систем машинного перевода, такими данными могут быть словари, корпуса.

Многофункциональный многоязычный Интернет сервис содержит целый набор программных модулей для компьютерной обработки тюркских языков, реализованных с использованием модели морфем: морфологический анализатор, расширенный морфологический анализатор с аналитическими формами, семантико-синтаксический анализатор, подсистему сравнительного анализа близости тюркских языков, систему машинного перевода между тюркскими языками.

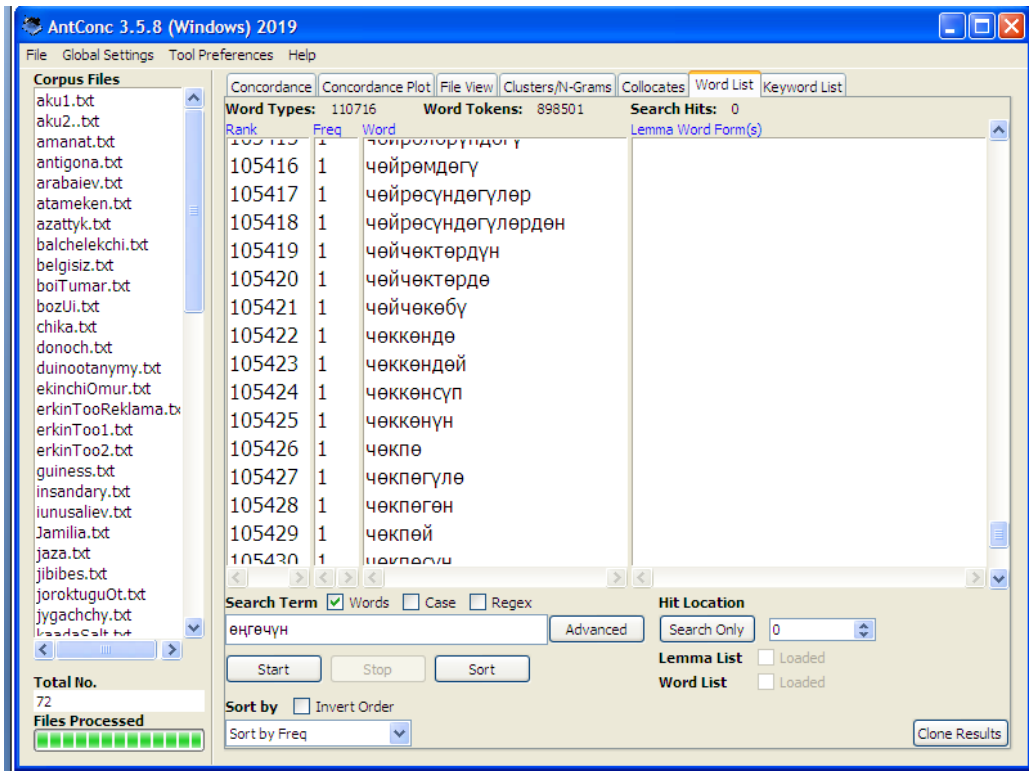
Качество машинного перевода зависит от объема языковой базы данных и от глубины описания естественных языков.

2. Корпус

Практическая значимость лингвистического корпуса пропорциональна его объему и репрезентативности языкового материала, поэтому важными являются задачи по сбору текстового материала для пополнения корпусной коллекции, очистке текстов и приведению их к единому формату.

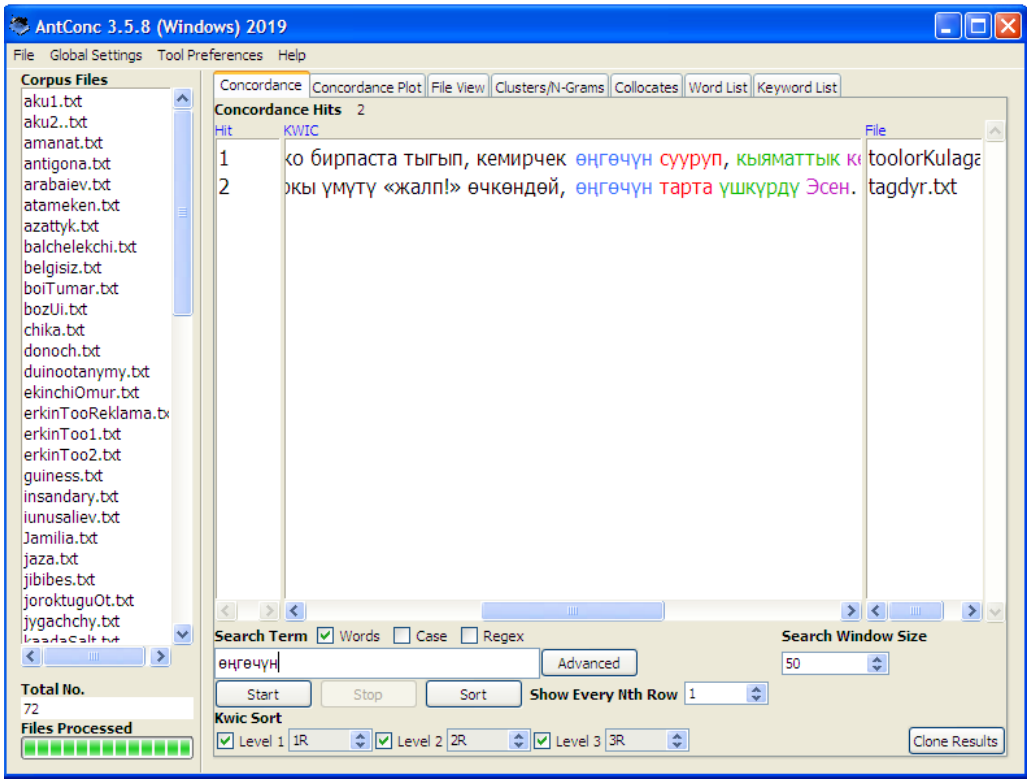
В целях содействия развитию новых свободных ресурсов для кыргызского языка в настоящей статье описывается простой полуавтоматический конвейер, который создает лексикон словоформ на основе корпуса, созданного из свободно распространяемых в Интернете текстов. Корпус, который состоит приблизительно из 1,6 миллиона слов и 170 текстов, включает в себя различные жанры, такие как литературные произведения (романы, повести, пьесы), законы и другие нормативные тексты, новости, институциональные сайты компаний, университетов, правительственных структур, статьи в кыргызской Википедии. Чтобы создать лексикон на основе текстовых материалов, а не существующих словарей, мы собрали корпус из разных интернет-источников.

Его структура (жанровая пропорция, длина текстов) неоптимальная, но наша цель – получить значительный лексический охват стандартного письменного кыргызского языка, более чем точное представление языка.



Список словоформ был автоматически извлечен из корпуса с помощью программы AntConc и отсортирован в алфавитном порядке.

Результирующий список содержит чуть менее 120 000 различных словоформ, с небольшим процентом русских слов.



Список слов был автоматически извлечен из корпуса, а элементы с не кыргызскими символами были отфильтрованы.

Мы понимаем, что размер нынешнего корпуса минимален, но разнообразие кажется приемлемым.

Полученный список из примерно 120 000 различных форм слов был проанализирован как морфемные последовательности с грамматическими значениями.

Этот морфологический анализ осуществлен простым конечным автоматом, который описывает структуру кыргызского слова. Этот конечный автомат хранится в простом текстовом файле, переходом является каждая строка. Каждый переход представляет собой единую морфемную форму поверхности, за исключением часто неправильной морфемной последовательности и принадлежности + падеж, которая представляется как один переход для простоты.

Морфологический анализ проводится в трех последовательных шагах:

- предварительная обработка входных данных;
- морфологический анализ;
- последующая обработка морфологически аннотированного вывода.

3. Предварительная обработка

В кыргызском языке, слова изменяются и новые слова образуются, в основном, путем добавления соответствующих окончаний (аффиксов) к основе слова, при этом сама основа меняется редко. Кроме того, сам аффикс при его соединении с основой слова меняется так, что его звучание в некотором смысле уподобляется звучанию последних букв (звуков) в основе слова.

В кыргызском языке буквы ю, я и ё явно бифонемны: представляют соответственно последовательности / j / + / u /, / j / + / a / и / j / + / o /.

Для упрощения морфологического анализа, ю и я переписываются как последовательность из двух букв, при этом каждая из «йотированных» букв Ё = Й * О, Ю = Й * У, Я = Й * А, Е = Й * Э. Если такие буквы находятся внутри корня, например таяк «палка», они не влияют на морфологическую структуру, но могут выходить за границу морфемы. Например, словоформа КОЮП объединяет основу КОЙ «Положить» (коюп → * койуп) с герундийским суффисом х: -УП. Буква е также может быть бифонемной, но только после гласной, и в таком контексте исправляется соответствующим образом, например ТИЕР «прикоснется» → * ТИЙЕР.

В автомате различают два типа суффиксов: словообразовательные и словоизменительные.

Таблица 1. Словообразующих и словоизменяющих аффиксов

№	Исходная форма	Свойства	Значение
1	-ЧЫ(1)	СО0, К →З	Знающий Б, умеющий человек
СЫН (экспертиза)→ СЫНЧЫ (эксперт)			
2	-ГЕР #	СО0, К →З	Человек
ЗЕР (золото (как драгоценность))→ ЗЕРГЕР (ювелир)			
3	-КӨР#	СО0, К →З	Человек
СҮТ(молоко) → СҮТКӨР			
4	-КЕЧ#	СО0, К →З	Человек
АРАБА (телега) → АРАБАКЕЧ (возчик)			
5	-МАН#	СО0, К → З	Человек
АКЫЛ (ум) → АКЫЛМАН (мудрец)			
6	-МЕН#	СО1, К →З	Человек
БИЛЕР (знающий) → БИЛЕРМЕН (знающий)			

П р и м е ч а н и е. -МАН, -МЕН совпадают по смыслу и близки по произношению к соответствующим аффиксам в английском, немецком и других европейских языках			
8	-МАР	СО0, К →З	Человек
ИШ(дело) → ИШМЕР (деятель)			
9	-КОР#	СО0, К →З	Человек
МАЛ(скот)→МАЛКОР (человек, заботливо относящийся к скоту)			
10	-ГЫЧ	СО0, Э →З	Исполнитель
АЧ(открывай) →АЧКЫЧ (ключ)			
11	-ЛЫК(1)	СО0, К→З	Обобщающий
ЖАКШЫ(хорошо) → ЖАКШЫЛЫК(добро)			
12	-ЧЫЛЫК	СО0, К →З	Обобщающий
АРЗАН (дешевый) → АРЗАНЧЫЛЫК (дешевизна)			

В глаголах в настоящем - будущем времени, 3-м лице отбрасывается последняя буква Т. Например, вопросительный : -БЫ

БОЛОТ (сталь) * БЫ → БОЛОТПУ (сталь ли)?;

БОЛОТ (будет) * БЫ → БОЛОБУ (будет ли)?

Привлечение большого объема количественных данных позволяет уточнить полученные ранее результаты о месте конструкций в лексической и грамматической системе тюркских языков и особенностях грамматикализаций глаголов.

Морфологический анализ проводится с конца слова, поскольку кыргызские морфологические процессы основаны на суффикации.

Таблица 2. Избранные UD категории для глаголов

Конечный глагол формы (3-е лицо)	UD категории
окуду	Tense=Past, Aspect=Perf
окуган	Tense=Past, Aspect=Imp
окуптур	Tense=Past, Evident=Nfh
окуур	Mood=Pot
окуса	Mood=Cond
окусун	Mood=Imp (although Mood=Opt may be better)
Причастия и герундий	UD категории
окуу	VerbForm=Inf
окуп	VerbForm=Conv, Tense=Past (Aspect=Perf might be an option)

Таким образом, этот функция анализа обеспечивает все виды анализа, формально возможные для данной входной строки (внутренне структурированной как двоичное дерево переходов), без устранения неоднозначности.

Эта работа требует много времени, учитывая размер данных. Основная лексика формируется автоматически из исправленных вручную аннотированных словоформ.

Окончательное решение о категориях, используемых для грамматической аннотации, должно привлечь многие исследователей и, вероятно, должно учитывать общие тюркские точки зрения.

В будущем планируется реализовать новый морфологический анализатор. на основе исправленной аннотированной лексики. Затем мы намерены аннотировать этот улучшенный корпус с помощью нового морфологического анализатора.

Мы сделали вывод, что в настоящее время, в связи с широким использованием различных управляемых устройств, увеличивается роль операционной функции языка, в отличие от коммуникативной и информирующей ролей.

Список использованной лллитературы:

1. Карабаева С.Ж. Единый алгоритм словоизменения и представление про-странства в кыргызском языке. – Saarbrűcken, Deutschland: Lap Lambert Aca-demic Publishing, 2016. – 62 с.

2. Карабаева С. Ж. Виртуальные геометрические объекты, создаваемые глаголами в кыргызском языке // В мире науки и искусства: вопросы филологии, искусство-ведения и культурологии: сб. статей по материалам LXVI междунар. научно-практ. конференции, № 11(66). – Новосибирск: СибАК, 2016. – С. 74-79.

3. Буазу Лоик, Мамбетказиева Д. От кыргызских текстов из интернета
до аннотированному xml-лексикону словоформ: описание несложного полуавтоматического конвейера, // Труды конференции. В 2-х томах. Т 1. – Казань: Издательство АНРТ, 2017. – С. 242-254.

4. Zadeh L. A. The concept of a linguistic variable and its application to approximate reasoning // InformationSciences, 1975, Vol. 8, pp. 199-249, 301-357; Vol. 9, pp. 43-80.

Рецензент: к.филол.н., доцент Жумалиева Г.Э.