МРНТИ 16.21.07

З.А.Сиразитдинов

Кандидат филологических наук, заведующий лаборатории лингвистики
и информационных технологий ИИЯЛ УНЦ РАН (Уфа)

УЧЕНЫЙ И НАСТАВНИК

Аннотация: Научная деятельность Аскара Кудайбергеновича Жубанова многогранна. Своими фундаментальными исследованиями (им опубликованы более 200 научных трудов, в числе которых 5 монографий, 4 учебных пособий и 5 частотных словарей) он внес значительный вклад в развитие таких направлений казахского прикладного языкознания как квантитативно-статистическое изучение и формализация содержания национального текста, автоматический анализ письменной речи, компьютерная лексикография, разработка лингвистических баз данных казахского языка.

Ключевые слова: наука, исследования, казахский язык, линвистика.

З.А.Сиразитдинов

Ресей ғылым академиясының Уфа ғылыми орталығы, ядролық физика институты, лингвистика және ақпараттық технологиялар зертханасының меңгерушісі, филология ғылымдарының кандидаты

ҒАЛЫМ ЖӘНЕ ТӘЛІМГЕР

Аннотация: Асқар Құдайбергенұлы Жұбановтың ғылыми қызметі сан қырлы. Өзінің іргелі зерттеулері арқылы (ол 200-ден астам ғылыми еңбек, оның ішінде 5 монография, 4 оқу құралы және 5 жиілік сөздік жариялады) ол қазақ қолданбалы тіл білімінің ұлттық мәтін мазмұнын квантитативті-статистикалық зерттеу және формализациялау, жазбаша тілді автоматты талдау, компьютерлік лексикография, казак тілінін лингвистикалық деректер базасын әзірлеу сияқты бағыттарын дамытуға елеулі үлес қосты.

Тірек сөздер: ғылым, зерттеу, қазақ тілі, лингвистика.

Z.A.Sirazitdinov

Candidate of Philology, Head of the Laboratory of Linguistics and Information Technologies, Institute for history, language and literature, Ufa scientific center, Russian Academy of Sciences (Ufa)

SCIENTIST AND MENTOR

Annotation. The scientific activity of Askar Kudaibergenovich Zhubanov is multifaceted. With his fundamental research (he published more than 200 scientific works, including 5 monographs, 4 textbooks and 5 frequency dictionaries), he made a significant contribution to the development of such areas of Kazakh applied linguistics as quantitative and statistical study and formalization of the content of the national text, automatic analysis of written speech, computer lexicography, development of linguistic databases of the Kazakh language.

Keywords: science, research, the Kazakh language, linguistics.

Научная деятельность Аскара Кудайбергеновича Жубанова, моего учителя и наставника, многогранна. Своими фундаментальными исследованиями (им опубликованы более 200 научных трудов, в числе которых 5 монографий, 4 учебных пособий и 5 частотных словарей) он внес значительный вклад в развитие таких направлений казахского прикладного языкознания, как квантитативно-статистическое изучение и формализация содержания национального текста, автоматический анализ письменной речи, компьютерная лексикография, разработка лингвистических баз данных казахского языка и др.

Аскар Кудайбергенович, будучи учеником Калдыбая Бектаевича, одного из основоположников квантитативно-статистической лингвистики в СССР, начал научную деятельность в 60-х годах XX века с составления частотных словарей и на их базе статистического изучения структуры казахского текста. В годы начала проникновения вычислительных машин в гуманитарные науки еще не существовало специализированных программных пакетов и приложений как для работы с текстами, так и для математического анализа результатов. Молодым ученым был оцифрован роман М. Ауэзова «Абай жолы» и на основе разработанных самим автором алгоритмов и программ про-ведено исследование материала на ЭВМ «Минск-22». Результаты исследования были опубликованы в ряде научных статей и легли в основу кандидатской диссертации молодого ученого.

Аскаром Кудайбергеновичем на материале казахских текстов было подтверждено высказанное предположение о том, что выбор нормировок (объем серии (K), количество серий (n) и объем выборки (N)) в лингвистических исследованиях существенно влияет на характер распределения изучаемой лингвистической единицы.

В этих же исследованиях автором было показано, что при малых значениях нормировок вид распределения лингвистических единиц (частей речи) зависит и от функционального стиля исследуемых текстов. Так при K>25, n>100, N>2500 в художественных текстах и при K>50, n>100, N>5000 в публицистических текстах распределение прилагательных согласуется с законами Шарлье типа А, В и Пуассона. Эти исследования позволили А.К. Жубанову и К.Б.Бектаеву предложить единый подход в составлении частотных словарей и в проведении статистико-квантитативного анализа текстов тюркских языков (с нормировкой K=200, n=1000, N=200000).

Отметим, на заре информационных технологий, когда не было ОС Windows, а существующие операционные системы в СССР поддерживали только английский и русский языки, когда еще не было самого слова компьютер, а вместо него использовалась аббревиатура ЭВМ, применение цифровой техники в лингвистических исследованиях для большинства языков Советского Союза являлось нетривиальной задачей. Пионерские работы А.К.Жубанова, посвященные проблеме кодирования для ЭВМ письменных текстов тюркских языков, составленные им алгоритмы реализации частотных, обратно-частотных, алфавитно-частотных словарей, частотных списков сочетаемости букв, буквосочетания, словосочетаний открыли возможность применения цифровой техники в тюркологии.

Предложенная автором технология обработки тюркских текстов цифровой техникой была успешно применена в работах узбекских, каракалпакских и башкирских языковедов.

Следует отметить, что Аскар Кудайбергенович многие годы оказывал советами постоянную поддержку коллегам-языковедам из тюркских республик в освоении инфор-

мационной технологии. Это отразилось, в частности, в том, что официально принятые кириллические раскладки национальных клавиатур компьютеров многих тюркских языков оказались подобными казахской.

Большинство частотных словарей по казахскому языку были реализованы при непосредственном участии А.К. Жубанова. Эти словари не только явились ценным исследовательским материалом, но и высоко подняли престиж всей казахстанской лингвистической науки. Так, частотный словарь языка поэзии Абая был в СССР вторым словарем такого типа после словаря лирики Пушкина, а словари по текстам орхонской письменности и исторического памятника «Кодекс Куманикус» явились вообще уникальными словарями такого типа в мировом масштабе. Это направление научной деятельности и сегодня остается в поле зрения ученого. Так, Аскаром Кудайбергеновичем с коллегами подготовлен большой сводный частотный словарь казахского языка, который должен увидеть свет в этом году.

Созданные ученым частотные словари, разработанная им методика статистического анализа текста, были блестяще использованы Аскаром Кудайбергеновичем в изучении авторской лексикографии и выявлении индивидуального стиля казахских писателей. Ряд работ автора посвящены изучению графемной структуры казахского текста. Ученым выявлена высокая частота употребления сонорных и смычно-взрывных звуков в казахской речи, определены возможности различных типов гласных и согласных графем дифференцировать функциональные стили. Работы в этом направлении привели ученого к решению проблемы автоматического транскрибирования казахских текстов. А.К.Жубановым был предложен алгоритм автоматического преобразования письменных текстов в фонетический текст.

Накопленный автором статистический материал по словоизменительным и формообра-зовательным категориям казахского текста позволили ученому приступить к разработке теоретических аспектов проблемы автоматического синтеза казахской словоформы. Отметим, что решение хотя бы одной задачи из связки синтезанализ, позволяет, в принципе, решить и вторую. Проблемы синтеза и анализа словоформы являются базовыми в создании автоматического морфологического анализатора языка, который открывает путь к реализации интеллектуальных систем с общением на естественном языке. К сожалению, отсутствие тесных научных связей в начале 2000-х годов не позволило скоординировать подходы тюркологов к данной проблеме. Подход А.К. Жубанова к решению проблемы на основе раздельного синтеза именных и глагольных словоформ на базе словарей основ и аффиксов с учетом комбинаторных возможностей последних был предложен значительно раньше других, что является подтверждением научной интуиции ученого. Подтверждением этому являются подходы и при синтезе башкирской и хакасской словоформ.

Фундаментальным вкладом А.К. Жубанова в казахскую прикладную лингвистику, несомненно, является его докторская диссертационная работа, посвященная проблеме многоуровневого моделирования казахского текста, основные положения которой изданы отдельной монографией. Формализация семантико-синтаксической организации текста предполагает использование некоторого искусственного языка, который был бы понятен компьютеру. Автором использован язык СЕМСИНТ, разработанный его научным руководителем А.В. Зубовым. Этот язык дает возможность описать содержание отдельного слова, предложения, абзаца и текста в целом. СЕМСИНТ содержит правила, учитывающие семантические и синтаксические отношения между текстовыми единицами, позволяющие строить их семантико-синтаксические формулы и на их основе выводить обобщенную семантико-синтаксическую формулу конкретного текста.

Отметим, что работа такого характера проделана в тюркологии впервые, в ходе которого автором:

– предложена семантическая классификация лексики знаменательных классов слов казахского языка;

– выявлены типы начальных, медиальных и конечных абзацев по их предметно-логическому содержанию в трех функциональных стилях казахского языка: художественный, публицистический, научный;

– определены типы опорных слов;

– изучены функционально-смысловые типы абзаца в зависимости от положения в тексте;

– составлены модели текстов разных функциональных стилей в виде лексико-семантических формул.

Полученные результаты являются уникальным и ценным материалом, которые в дальнейшем, безусловно, будут использованы в разработке национальной интеллектуаль-ной информационной системы, таких как: автоматическое реферирование, система автоматического сбора и классификации информации, контент анализа, автоматического порождения текста и др.

Осознавая важность филологических компьютерных баз данных как в плане источника для новых научно-теоретических изысканий, так и в плане автоматизированного справочника для широкого круга пользователей, Аскар Кудайбергенович в 2003 г. выступает в периодической печати с призывом к филологам и специалистам IT технологии объединить усилия в этой области. С этого момента А.К. Жубанов свои знания, опыт и талант организатора науки переносит в область разработки лингвистических баз данных казахского языка.

В лингвистике выделяются собственно лингвистические базы данных и полнотекстовые базы. Собственно лингвистические базы содержат информациюо лингвистических единицах различного рода: лексике, фонетике, морфологии, например, Международный компьютерный архив современного английского языка (International Computer Archive of Modern English – ICAME), базаданных русской терминологии «РОСТЕРМ» и др. Полнотекстовые базы содержат корпусы текстов, например, «Британский нацональный корпус английского языка», самый крупный в России «Национальный корпус русского языка», или «Компьютерный корпус текстов русских газет конца XX века».

В русле собственно лингвистических баз данных автором была разработана структура представления данных в лексикографической базе «Тіл – қазына». Лексикографические базы данных представляют из себя многоуровневую структуру, хранящую всю необходимую информацию, позволяющую производить различные операции поиска, доступа к различным информационным ресурсам и выводить для конечного пользователя в оптимальном для восприятии расположении компонентов данных базы. Основная трудность при разработке лексикографических баз данных заключается в том, что они создаются на базе изданных в печатной форме словарей. Структуры таких традиционных лексикографических трудов сильно различаются и не могут быть переложены напрямую в информационные базы. Автор изначально отказался от подхода создания разных форматов хранения данных для словарей с разными структурами. А.К. Жубановым были проанализированы все имеющиеся в казахском языкознании словари и выработана единая структура хранения данных в компьютерной лексикографической базе.

Сегодня «Тіл – қазына» включает: лексикографическую картотеку Института языкознания им. А. Байтурсынова КН МОН РК, данные многотомного толкового словаря и всех прочих изданных словарей казахского языка. База данных снабжена программной оболочкой, позволяющей проецировать данные на выходе в интуитивно понятные

формы толкового, синонимического, орфоэпического словарей, двуязычных общих и терминологических словарей, генеральной картотеки Института язкознания им. А. Байтурсынова КН МОН РК. Например, в проекции картотеки пользователь получает такие параметры слова как иллюстрационный материал употребления слова с указанием автора, названия произведения, года выпуска, места издания, абзаца и пр.

Данная база реализована как сетевая для научных сотрудников Института языкознания и является ценным источником для многих лингвистических исследований и лексикографических разработок. Так на его базе был подготовлен в сравнительно короткие сроки 15 томный толковый словарь казахского языка.

Отдавая много сил научно-исследовательской работе, Аскар Кудайбергенович все же находит время и для разработки новых инновационных методик в области преподавания языка. Предложенные им методы оптимизации процесса обучения казахскому языку, такие как учет наряду с частотой употребления слова и параметра важности слова K ($K= F \cdot m/N \cdot n$) и созданные им программы автоматического его подсчета, нашли широкое применение в школах и вузах республики.

Результаты многолетних исследований ученого, его идеи и открытия легли в основу разработки казахского национального корпуса, который создается в Институте языкознания им. А. Байтурсынова под руководством самого автора. Видение архитектуры национального корпуса, изложенное в его многочисленных научных трудах, разработанные им подходы к решению задач автоматического синтеза и анализа казахской словоформы, созданная база данных «Тіл – қазына» и его неустанная энергия и забота позволяют научному коллективу института, в нелегкие для науки годы, успешно его реализовывать.

Аскар Кудайбергенович самый талантливый и яркий ученик Калдыбая Бектаевича, которому учитель и основатель математической лингвистики Казахстана передал в свое время руководство созданной им группой научных сотрудников при Институте языкознания им. А. Байтурсынова КН МОН РК. За прошедшие полвека группа стала лабораторией Автоматизации в лексикографии, затем была возведена в ранг отдела прикладной лингвистик. За эти годы Аскар Кудайбергеновичне только развил школу математической лингвистики К.Б.Бектаева, но и воспиталсвою школу, школу компьютерной лингвистики.

Он является научным руководителем и консультантом и оппонентом ряда кандидатских и докторских диссертаций. Им подготовлены научные кадры не только для казахской прикладной лингвистики, но и для башкирской. Автор этой статьи проходил в 1983-1984 гг. стажировку под руководством А.К. Жубанова в группе Автоматизации в лексикографии Института языкознания им. А. Байтурсынова КН МОН РК. Аскар Кудайбергенович был в дальнейшем научным консультантом моей кандидатской диссертации.

Темы, защищенных его учениками, диссертаций охватывают разнообразные направления прикладной лингвистики. Это подтверждает энциклопедичность и глубину его научных познаний в области математики, информатики и лингвистики. С большинством своих учеников Аскар Кудайбергенович был знаком, когда они учились в ВУЗе. Он читал и продолжает читать лекции, спецкурсы, проводит практические занятия для студентов, магистрантов и PhD докторантов Казахского национального университета им. аль-Фараби и Казахского Государственного национального педагогического университета им. Абая по прикладной лингвистике и по компьютерной лингвистике.

При подготовке лингвистов-прикладников Аскар Кудайбергенович стремится не только обучать, передать свои знания, но и прививать им любовь и уважение к казахско-

му языку. Эти чувства и забота о родном языке передались ему по крови от отца Кудайбергена Куановича, первого профессора казахской лингвистики и лежат в основе его жизненной позиции.

СПИСОК ЛИТЕРАТУРЫ:

[1] Джубанов А.Х., Зубов А.В. Автоматизация некоторых лингвистических процессов// Вестник АН КазССР. Исследования молодых ученых. – Алма-Ата, 1968. №9. – С.31-36.

[2] Джубанов А.Х., Бектаев К.Б. Вероятностно-статистическое моделирование тюркских текстов//Статистика казахского текста. –Алма-Ата: Наука, 1973. – С. 299-328.

[3] Джубанов А.Х. Вероятностные законы распределения классов слов казахского текста// Труды VII Всесоюзной школы-семинара «Автоматическое распознавание слуховых образов». – Алма-Ата, 18-23 сентября, 1972 г. – Алма- Ата: Наука, 1973. – С.139-145.

[4] Лукьяненков К.Ф. Использование схем Пуассона и Гаусса в исследовании распределения лингвистических единиц текста // Вопросы лингвостатистики и автоматизации лингвистических работ. Труды ЦНИИПИ, Вып. 3. – М., 1969. – С. 5-14.

[5] Джубанов А.Х., Зубов А.В. Автоматизация некоторых лингвистических процессов // Вестник АН КазССР. – №9. 1968. – С.31-36.

[6] Джубанов А.Х., Бектаев К.Б., Зубов А.В. Автоматическое построение частотных словарей (прямого и обратного) // Вестн. АН КазССР. – №3. – 1970. – С.48-53.

[7] Жубанов А.К. Основные принципы формализации содержания казахского текста: автореф. дисс. док. филол. наук. – Алматы, 2002. – 32 с.

[8] Сиразитдинов З.А. Моделирование грамматики башкирского языка. // Словоизменительная система. – Уфа: Гилем, 2006. – 160 с.