# Об утверждении Правил формирования и ведения Национального корпуса казахского языка

Приказ Министра науки и высшего образования Республики Казахстан от 30 апреля 2025 года № 227

В соответствии с пунктом 5 статьи 24-4 Закона Республики Казахстан "О культуре" и подпунктом 161-1) пункта 15 Положения о Министерстве науки и высшего образования Республики Казахстан, утвержденного постановлением Правительства Республики Казахстан от 19 августа 2022 года № 580 "О некоторых вопросах Министерства науки и высшего образования Республики Казахстан", ПРИКАЗЫВАЮ:

1. Утвердить прилагаемые Правила формирования и ведения Национального корпуса казахского языка.

2. Комитету языковой политики Министерства науки и высшего образования Республики Казахстан в установленном законодательством порядке Республики Казахстан обеспечить:

1) государственную регистрацию настоящего приказа в Министерстве юстиции Республики Казахстан;

2) размещение настоящего приказа на интернет-ресурсе Министерства науки и высшего образования Республики Казахстан.

3. Контроль за исполнением настоящего приказа возложить на курирующего вице-министра науки и высшего образования Республики Казахстан.

4. Настоящий приказ вводится в действие после дня его первого официального опубликования.

*Министр* *С. Нурбек*

Утверждены приказом
Министра науки и
высшего образования
Республики Казахстан
от 30 апреля 2025 года № 227

## Правила формирования и ведения Национального корпуса казахского языка

### Глава 1. Общие положения

1. Настоящие Правила формирования и ведения Национального корпуса казахского языка (далее – Правила) разработаны в соответствии с пунктом 5 статьи 24-4 Закона Республики Казахстан "О культуре" и подпунктом 161-1) пункта 15 Положения о Министерстве науки и высшего образования Республики Казахстан, утвержденного постановлением Правительства Республики Казахстан от 19 августа 2022 года № 580 "

О некоторых вопросах Министерства науки и высшего образования Республики Казахстан", и устанавливают общие требования и процедуры для создания, пополнения , обработки, хранения и использования Национального корпуса казахского языка (далее – Национальный корпус), который служит основой для научных, образовательных и практических исследований в области казахского языка и лингвистики.

2. Администратор Национального корпуса – Министерство науки и высшего образования Республики Казахстан.

3. Работу по координации Национального корпуса осуществляет Республиканское государственное предприятие на праве хозяйственного ведения "Институт языкознания имени Ахмета Байтурсынулы" Комитета науки Министерства науки и высшего образования Республики Казахстан (далее – Институт).

4. В настоящих Правилах используются следующие понятия:

1) аннотация – описание данных в корпусе, включающее информацию о источнике заданного элемента поиска, такую как автор текста, его тема, жанр, контекст, структура и содержание, а также фонетические, морфологические, просодические, лексические, семантические, синтаксические и другие лингвистические характеристики;

2) разметка – метод кодирования и систематизации лингвистической и структурной информации в текстах корпуса, обеспечивающий их анализ и обработку;

3) научные стандарты качества – совокупность критериев, которым соответствуют тексты и их аннотации в корпусе для обеспечения точности, достоверности и пригодности для лингвистических исследований;

4) унифицированные форматы данных – стандартные структуры и правила, которые делают данные совместимыми и удобными для обработки и объединения, устанавливая требования к кодировке, структуре и разметке, что упрощает обмен и анализ данных, снижая риск ошибок и потребность в доработках;

5) жанр – категория текста, определяющая его тип, назначение и устоявшуюся форму обладающую определененными стилистическими чертами, что помогает классифицировать материалы для анализа и исследования;

6) частотность – числовой показатель, отражающий, как часто языковые элементы встречаются в определенном тексте или корпусе;

7) интерфейс – программно-аппаратная система корпуса, позволяющая пользователю искать, извлекать, просматривать результаты, фильтровать и анализировать данные;

8) система кодирования и классификации – упорядоченная структура, позволяющая присваивать текстам уникальные коды и классифицировать их по ключевым характеристикам, обеспечивая удобство поиска, хранения и анализа данных в корпусе;

9) корпус – часть Национального корпуса, состоящая из отдельных подкорпусов и сформированная по определ Ұнным критериям для проведения целенаправленного лингвистического или статистического анализа;

10) база данных корпуса – электронные версии письменных и устных текстов на естественном языке, которые предварительно обработаны и размечены для включения в корпус;

11) национальный корпус казахского языка – информационно-справочная система, содержащая тексты во всех стилях и жанрах казахского языка, оснащенная системой поиска и другими средствами работы с текстом;

12) тексты – языковые единицы, которые составляют основу корпуса;

13) метаразметка – краткое источниковедческое описание текста, включенного в корпус и его содержания, которое предоставляет ключевую информацию о его тематике, авторе, жанре, цели, стиле и других характеристиках;

14) морфологическая разметка – процесс автоматического или ручного снабжения текстов в корпусе морфологическими характеристиками и определения их грамматических признаков;

15) синтаксическая разметка – процесс снабжения и описания синтаксической структуры предложений в тексте, включая определение связей между словами и их грамматических ролей, для изучения грамматической структуры языка и его правил;

16) семантическая разметка – процесс аннотирования смысла, значений, контекстуальных оттенков слов, выражений и предложений в тексте, а также их взаимосвязей, с целью создания инструмента для естественно-языковой обработки, корпусной лингвистики, машинного обучения и цифровых технологий;

17) словоупотребление – совокупность характеристик использования слова в текстах, входящих в корпус;

18) стиль – способ выражения в языке, который определяется такими характеристиками, как тон, словарный запас, грамматическая структура и другие особенности, отражающие цель и особенности текста.

## Глава 2. Формирование Национального корпуса

5. Все работы, связанные с формированием, ведением и использованием Национального корпуса, проводятся в рамках единого национального подхода с обеспечением высокого уровня качества, безопасности и доступности данных.

6. Национальный корпус формируется в соответствии с настоящими Правилами за счет корпусов и подкорпусов.

7. Формирование Национального корпуса начинается с тщательного сбора и отбора текстов, представляющих все аспекты его функционирования – от письменных источников до устных материалов, которые отражают особенности языка, используемый в различных сферах и временных срезах.

8. Все материалы Национального корпуса соответствуют научным стандартам качества, имеют метаданные и источниковедческую информацию, проходят проверку

на грамматические и орфографические ошибки, а также редактируются и упорядочиваются в соответствии с единой системой кодирования и классификации.

9. Национальный корпус отражает разнообразие языковых практик, включая различные жанры, типы текстов, диалекты и стили, а также учитывает региональные и исторические особенности.

10. Все собранные тексты систематизируются и аннотируются по жанру, стилю, времени, источнику, по лексическим и грамматическим уровням.

11. После сбора и аннотирования текстов с учҰтом энциклопедического и источниковедческого анализа применяются методики лингвистической обработки и анализа данных, что позволяет создавать целевые эмпирические языковые базы для научных исследований и модели языка, используемые для разработки языковых технологий и лексикографических проектов.

12. Функции Национального корпуса:

1) Метаразметка предоставляет информацию, которая подробно описывает текст и его характеристики, включая автора, название, жанр, источник, дату создания, целевую аудиторию, стиль, размер, формат, структурные особенности;

2) Цифровая структура Национального корпуса обеспечивает систематизацию текстов в цифровом формате, что позволяет использовать современные инструменты автоматизированного анализа;

3) Многофункциональность Национального корпуса позволяет использовать его в различных областях, включая лингвистические исследования, лексикографию, обучение языкам, компьютерную лингвистику и нейролингвистическое программирование;

4) Прикладные задачи Национального корпуса позволяют автоматизировать обработку текстов, улучшать качество языковых моделей, разрабатывать эффективные инструменты для анализа текста, создания приложений для обучения языкам, машинного перевода и разработки искусственного интеллекта.

13. Структура Национального корпуса:

1) Национальный корпус состоит из отдельных корпусов и подкорпусов, каждый из которых охватывает конкретные направления;

2) Все корпусы и подкорпусы соответствуют техническим, функциональным и операционным характеристикам Национального корпуса, установленным Институтом с согласования Администратора, обеспечивая единообразие, совместимость и безошибочную интеграцию в процессе эксплуатации;

3) Все тексты в Национальном корпусе хранятся в унифицированных форматах, также обеспечивается интеграция API;

4) Текстовые данные охватывают широкий спектр жанров и стилей, включая научные работы, художественные произведения, публицистику, деловую документацию, разговорную речь и цифровые тексты;

5) Лексическая информация включает сведения о словах, их формах, значениях, сочетаемости и метаданные, такие как частотность, стилистика, синонимы, антонимы, с классификацией по частям речи и тематическим областям;

6) Грамматическая информация включает описание синтаксиса, морфологии и словообразования с разбором структуры предложений, частей речи, синтаксических связей и морфологических характеристик;

7) Стилистическая информация включает описание особенностей текста, различие между стилями, анализ языка в различных жанрах и классификацию текстов по категориям, таким как формальные и неформальные стили, жанры и их характерные черты;

8) В Национальном корпусе имеется кросс-языковая привязка, включающая параллельные тексты, которая позволяет проводить исследования в области машинного перевода, контрастивной лингвистики и изучения влияния языковых контактов;

9) Каждый текст снабжается метаданными, включая идентификационные данные, библиографические данные, жанрово-стилистические характеристики, лингвистические данные, контекстуальные данные, технические данные и аннотативные данные;

10) Медиафайлы, сопровождающие текстовые данные, служат для анализа различных аспектов языка в контексте мультимодальной информации;

11) Интерфейс обеспечивает удобное взаимодействие пользователя с Национальным корпусом для поиска, анализа и аннотирования данных;

12) Поисковая система позволяет пользователям искать и извлекать данные из базы Национального корпуса на основе различных критериев;

13) Официальная платформа Национального корпуса позволяет отслеживать количество словоупотреблений в реальном времени;

14) Структура данных позволяет добавлять в Национальный корпус новые категории и компоненты по мере необходимости.

## Глава 3. Ведение Национального корпуса

14. Национальный корпус непрерывно обновляется и расширяется за счет добавления новых материалов, отражающих изменения в языке и актуальные достижения в различных областях, с обеспечением доступа архивных версии.

15. Для повышения качества и актуальности базы данных обеспечивается взаимодействие с международными научными и лингвистическими организациями.

16. Для координации эффективного функционирования Национального корпуса создается рабочая группа, состоящая из профильных специалистов, которые обладают необходимыми знаниями и опытом.

17. Все материалы, загруженные в Национальный корпус, соответствуют требованиям по форматированию, орфографии, пунктуации, стилю и структуре.

18. Все участники, предоставляющие материалы для включения в Национальный корпус, обеспечивают их качество, точность, достоверность, актуальность.

19. Для обеспечения высокого качества данных регулярно проводится мониторинг.

20. Национальный корпус служит основой для разработки лексикографических и грамматических проектов, научных и образовательных материалов, а также для создания программных продуктов.

21. В случае сбоев или ошибок в процессе ведения Национального корпуса Институт принимает меры для их устранения, а также проводит обновления программного обеспечения Национального корпуса для обеспечения стабильной и бесперебойной работы.

22. Национальный корпус расширяется с учҰтом изменений в языке, новых тенденций, а также появления новых жанров, стилей и направлений, что гарантирует его актуальность и соответствие языковым инновациям.

23. В Национальный корпус могут быть включены материалы из открытых и доступных источников, если их использование не нарушает авторские права и соответствует нормам добросовестного использования.

24. Национальный корпус, корпусы и подкорпусы общедоступны и предоставляются для использования всем заинтересованным сторонам.

25. При ведении Национального корпуса учитываются этические принципы, включая исключение дискриминационного контента и обеспечение инклюзивности.