ETHIOPIAN INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

QUESTIONS OF TRANSCRIPTION AND WORD TAGS WHEN CREATING A DIALECT CORPUS

Rustamova Zulfiya Khalmirzayeva
Tashkent State Pedagogical University named after Nizami
Associate Professor of the Department of Native Language
and its teaching methods in primary education

Abstract
Creating a dialect corpus - a linguistic database representing a specific dialect or regional variant of a language - is a complex task that requires solving a number of problems. Among these, two important issues were word transcription and tagging, important components of corpus construction. This article explores the complexities of transcribing and tagging words when creating a dialect corpus, highlighting the problems that arise during these processes, and discussing possible ways to overcome them.

Key words
text description, identification of time period, document selection.

Аннотация
Создание диалектного корпуса — лингвистической базы данных, представляющей конкретный диалект или региональный вариант языка, — сложная задача, требующая решения ряда задач. Среди них двумя важными проблемами были транскрипция слов и маркировка, важные компоненты построения корпуса. В данной статье исследуются сложности транскрипции и тегирования слов при создании диалектного корпуса, освещаются проблемы, возникающие в ходе этих процессов, и обсуждаются возможные пути их преодоления.

Ключевые слова
текстовое описание, идентификация периода времени, выбор документа.

Введение: Поскольку мы решили попробовать над этими текстами некоторые стандартные инструменты обработки текста (Reco, система тегов и инструмент фонетической транскрипции Phonet) а количество говорящих варьируется от носителей стандартного языка до говорящих только с анемофонией, использующих стандартный язык. Мы считаем, что обе транскрипции будут полезны: орфографическая и простая транскрипция. Аналогичному подходу последовали Коста и Таварес, которые расшифровали каждое слово своего диалектного корпуса с двумя уровнями детализации. Для автоматической обработки этих диалектных корпусов они используют только теги частей речи и морфологический анализ упрощенных транскрипций слов.

После тестирования многих существующих инструментов разметки диалектных текстов они решили пометить орфографически транскрибируемый корпус с помощью PALAVRAS, разработанного для автоматической обработки непрерывного португальского текста. Однако Phonet был разработан только для работы с орфографическими текстами на европейском португальском языке, поскольку модули ошибок и адаптации клавиатуры, необходимые для упрощенного процесса транскрипции, не были реализованы в Phonet.

Разработка больших и разнообразных корпусов письменных текстов имеет особое значение в лингвистических и компьютерных диалектных исследованиях. Однако, хотя ручная транскрипция диалектных вариаций (особенно в спонтанной речи) трудоемка, существует несколько стандартизированных способов. Использование тегов для классов слов и орфографически упрощенной транскрипции в диалектных корпусах может ускорить

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

создание больших корпусов и повысить производительность некоторых задач автоматической обработки диалектных текстов.

Однако данная ситуация порождает ряд теоретических и практических проблем: автоматическая обработка баз данных сложна и при необходимости производится вручную, поэтому занимает много времени. Следует решить вопрос о том, как сохранить и создать связи между литературными и диалектными словоформами, а также взаимоотношения между ними. С одной стороны, современные диалектические исследования поддерживаются цифровыми ресурсами, но с другой стороны, технологические усилия намного превосходят те выгоды, которые они приносят. Выбор содержания информаций в корпорации является следующей задачей. Однако основная задача состоит в том, чтобы определить, в какой степени корпус поддерживает тот или иной тип лингвистического исследования.

Цели исследования

- Организовать данные о разработанном корпусе для использования в качестве справочного материала для других исследователей в области лингвистики.
- Решение орфографических задач на синтез голоса, поэкспериментировать с инструментами, предназначенными для создания и выбора слов.
- Проектировать и создавать фонологические и морфологические инструменты с использованием машинного обучения.
- Создать вручную словарь слов, представляющих разные грамматики.

Транскрипция, процесс преобразования устной речи в письменную форму, является важным шагом в создании диалектного корпуса. Однако это не так просто, как кажется. Одной из основных проблем является отсутствие стандартизированной орфографии для диалектов, что может привести к несоответствиям в транскрипции. Разные транскрибаторы могут использовать разные варианты написания для обозначения одного и того же слова или фонемы, что приводит к несоответствиям в корпусе. Например, диалектное слово «gonna» (идти) может писаться как «gonna», «gona» или «gunna», что приводит к путанице и трудностям при разборе.

Еще одной трудностью транскрипции слов является отражение диалектных фонологических особенностей. Диалекты часто демонстрируют отчетливые фонологические особенности, такие как нестандартное произношение или слияние фонем, которые может быть трудно уловить в письменной форме. Например, в диалекте афроамериканского английского языка (AAVE) четко произносится гласный звук в словах «bit» и «bat», которые в письменной форме часто передаются как «пчела» и «но». Однако это представление может неточно отражать разговорный диалект, что вызывает опасения по поводу точности транскрипции.

Кроме того, транскрипция слов может оказаться трудоемким процессом, особенно для больших наборов данных. Ручная транскрибация одним транскрибатором или группой транскрибаторов может оказаться медленным и дорогостоящим процессом, что затрудняет достижение последовательности и точности. Чтобы решить эту проблему, были разработаны системы автоматического распознавания речи (ASR), преобразующие устную речь в письменную форму. Однако системы ASR не лишены ограничений, и их точность может варьироваться в зависимости от качества аудиозаписей, диалектных вариаций и фонового шума.

Помимо транскрипции слов, тегирование, процесс присвоения лингвистических аннотаций каждому слову корпусе, является еще одним важным компонентом создания диалектного корпуса. Маркировка включает в себя присвоение каждому слову тега части речи (POS), например существительного, глагола или прилагательного, а также определение

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

грамматических особенностей, таких как время, вид и наклонение. Однако маркировка может быть затруднена, особенно для диалектов с нестандартной грамматикой и синтаксисом.

Одна из проблем с маркировкой — сложность диалектной грамматики и синтаксиса. Диалекты часто имеют различные грамматические структуры, такие как время и вид глагола. Еще одна трудность в маркировке — отсутствие стандартизированных схем аннотации для диалектов. Разные исследователи могут использовать разные схемы аннотаций, что приводит к несогласованности маркировки в разных корпусах. Это может затруднить сравнение диалектных особенностей разных диалектов. Кроме того, отсутствие аннотированных корпусов диалектов может затруднить разработку и обучение моделей маркировки, что приведет к снижению точности и надежности.

Для преодоления этих проблем можно реализовать несколько решений. Во-первых, разработка стандартизированных схем орфографии и аннотаций для диалектов поможет обеспечить согласованность транскрипции и разметки. Этого можно достичь благодаря совместным усилиям исследователей и лингвистов по разработке руководящих принципов диалектной транскрипции и маркировки. Кроме того, использование инструментов автоматической транскрипции и маркировки, таких как системы ASR и алгоритмы машинного обучения, может помочь повысить эффективность и точность процесса.

Создание аннотированных диалектных корпусов может облегчить разработку моделей разметки для конкретных диалектов, что может повысить точность и надежность разметки. Этого можно достичь путем сбора и интерпретации больших наборов данных диалектной речи, которые можно использовать для обучения и оценки моделей тегирования. Кроме того, разработка руководств по маркировке для конкретных диалектов и схем аннотаций может помочь обеспечивать последовательность и точность маркировки.

Заключение.

Таким образом, создание диалектного корпуса — сложная задача, которая включает в себя несколько задач, включая транскрипцию слов и маркировку. Отсутствие стандартизированных схем орфографии и интерпретации, диалектных фонологических особенностей и грамматических сложностей могут привести к несоответствиям и неточностям в транскрипции и маркировке. Однако эти проблемы можно преодолеть путем разработки стандартизированных руководств, использования инструментов автоматической транскрипции и разметки, а также создания аннотированных диалектных корпусов. В конечном счете, создание точных и надежных диалектных корпусов имеет важное значение для лингвистических исследований, преподавания языка и развития языковых технологий.

Список использованных литератур:

1. Atkins, S., Clear, J. va Ostler, N. 1992. Korpus dizayn mezonlari. Adabiy va lingvistik hisoblash. 7(1): 1-16.
2. Biber, D. 1993. Korpus dizaynidagi vakillik. Adabiy va lingvistik hisoblash. 8(4): 243-257.
3. Cheng, W. 2011. Korpus tilshunosligini o'rganish: harakatdagi til. London: Routledge.
4. Hunston, S. 2002. Amaliy tilshunoslikdagi korpus. Kembrij: Kembrij universiteti nashriyoti.
5. Kennedi, G. 1998. Korpus tilshunosligiga kirish. Nyu-York: Addison-Wesley Longman Inc.
6. Leech, G. 1991. Korpus lingvistikasining zamonaviy holati. In: Aijmer K. va Altenberg, B. Eds. Ingliz korpus tilshunosligi: J. Svartvik sharafiga tadqiqotlar. London: Longman. Pp. 8-29.
7. McEnery, A., Xiao, R. and Tono, Y. 2005. Korpusga asoslangan til tadqiqotlari: ilg'or manbalar kitobi. London: Routledge.
8. Obidjon o'g'li, Y. O., & Xolmirzayevna, R. Z. (2023, April). KORPUS–TABIIY TILLARNING MILLIY XUSUSIYATINI SAQLAB QOLISHDA MUHIM INSTRUMENT.

eISSN: 2349-5715 pISSN: 2349-5707

Volume: 11, Issue 05, May-2024 SJIF 2019: 4.702 2020: 4.737 2021: 5.071 2022: 4.919 2023: 6.980 2024: 7,662

In INTERNATIONAL SCIENTIFIC CONFERENCES WITH HIGHER EDUCATIONAL INSTITUTIONS (Vol. 1, No. 14.04, pp. 61-64).

9. Rustamova, Z. (2022). LINGUISTIC BASIS OF CREATION OF DIALECTAL CORPUS OF UZBEK LANGUAGE. Science and Innovation, 1(8), 1255-1258.
10. Rustamova, Z. (2022). O'ZBEK TILI DIALEKTAL KORPUSINI YARATISHNING LINGIVISTIK ASOSLARI. Science and innovation, 1(B8), 1255-1258.