Acta Linguistica Petropolitana. 2020. Vol. 16.2. P. 160–180
DOI 10.30842/alp2306573716207

Автоматический метод языкового профилирования носителя диалекта (на материале восточносербского идиома села Берчиновац)*

А. Л. Макарова
Цюрихский университет (Швейцария); anastasia.makarova@uzh.ch

Д. В. Конёр
Институт лингвистических исследований РАН, Санкт-Петербург;
dsuetina@yandex.ru

Т. Вукович
Цюрихский университет (Швейцария); teodora.vukovic2@uzh.ch

А. Н. Соболев
Институт лингвистических исследований РАН, Санкт-Петербург;
sobolev@staff.uni-marburg.de

О. Винисторфер
Цюрихский университет (Швейцария); olivier-andreas.winistoerfer@uzh.ch

Аннотация. В настоящей статье представлен метод (полу)автоматического анализа фонетических и морфосинтаксических особенностей диалектного текста, который в перспективе может быть применен на большом объеме диалектных данных. Метод представлен на примере анализа индивидуального идиома носительницы тимокского говора села Берчиновац в районе города Княжевац Заечарского округа в Восточной Сербии. Приводится алгоритм поиска таких диалектных явлений, как наличие / отсутствие специфических (для данной диалектной зоны) фонем, удвоение прямого и косвенного объекта, способ выражения значений периферийных падежей, наличие постпозитивного артикля и т. д. Выявляются преимущества

* Публикация выполнена при поддержке следующих фондов: РФФИ 18-512-76002 ЭРА_а «Изучение дивергенции и конвергенции традиций Центральных Балкан: реализация и перцепция», EraNet Rus Plus grant/Swiss National Science Foundation IZRPZ0_177557/1 (TraCeBa project, https://traceba.net/); SNF100015_176378/1 (‘Ill-bred sons’, family and friends: tracing the multiple affiliations of Balkan Slavic).

и ограничения компьютерного анализа (по сравнению с «ручным») при попытке автоматизировать исторический и структурный лингвистический анализ.

Ключевые слова: автоматический анализ текста, языковое профилирование, носитель диалекта, балканославянские языки, сербские диалекты, тимокский диалект, идиолект носителя говора, село Берчиновац, Восточная Сербия.

Automatic language profiling of a dialect speaker: the case of the Timok variety spoken in the village of Berčinovac (Eastern Serbia)

A. L. Makarova
University of Zurich (Switzerland); anastasia.makarova@uzh.ch

D. V. Konior
Institute for Linguistic Studies, Russian Academy of Sciences, St. Petersburg; dsuetina@yandex.ru

T. Vuković
University of Zurich (Switzerland); teodora.vukovic2@uzh.ch

A. N. Sobolev
Institute for Linguistic Studies, Russian Academy of Sciences, St. Petersburg; sobolev@staff.uni-marburg.de

O. Winistörfer
University of Zurich (Switzerland); olivier-andreas.winistoerfer@uzh.ch

Abstract. In a previously published paper [Konior et al. 2019], which thematically led up to the present article, we explored the possibility of developing a quantitative tool for assessing the intrasystemic dialectal coherence and the degree of dialectal authenticity (preservation) for a particular variety of Slavic (and more broadly Balkan) dialectal speech. In order to do so, we analysed and manually counted all cases of presence or absence of specific phonemes, direct and indirect object reduplication, ways of expressing peripheral cases meaning, presence of a postpositive article, and some other language features. The data used for that purpose was extracted from "Linguistic Atlas of Eastern Serbia and Western Bulgaria" [SAOSWB]; an idiolect of a native speaker of the Timok dialect spoken in the village of Berčinovac (near the town of Knjaževac in the Zaječar district, Eastern Serbia) was chosen for analysis. Subsequently, the following question arose: how can the use of modern technologies for automatic text processing increase the efficiency of dialectologists' work, and what technical obstacles must be overcome in this regard? In the article, we present a method of (semi-)automatic analysis

of phonetic and morphosyntactic features in a dialect text with the use of morphological annotation (the tagger model is based on the ReLDI tagger [Ljubešić et al. 2016] and user Python scripts). An algorithm searching for some important dialect features is described and exemplified. Trying to imitate and automate historical and structural linguistic analysis, we open a discussion about the advantages and disadvantages of computer analysis of dialect data as compared with the manual analysis. In the future, the automatic method is expected to be helpful in managing larger amounts of dialect data.

Keywords: statistical methods in linguistics, machine text analysis, linguistic profiling, dialect speakers, Balkan Slavic languages, Serbian dialects, Timok dialect, idiolect of dialect speaker, village of Berčinovac, Eastern Serbia.

1. Введение

Данное исследование связано с двумя важными вопросами современной диалектологии — лингвистического свойства (классификация диалектов в свете глобальных языковых изменений, ведущих к «перерождению» традиционных местных идиомов в «региональные койне»)[1] и технического характера (перспективы машинной обработки диалектной речи).

Приступая к написанию статьи [Конёр и др. 2019], предваряющей настоящий текст, мы стремились понять, возможна ли разработка количественного инструмента для оценки внутрисистемной диалектной когерентности и «степени диалектной аутентичности» (или «сохранности») для того или иного варианта славянской (и шире — балканской) диалектной речи. Придя к выводу о том, что работа в этом направлении дает свои результаты [Конёр и др. 2019: 30–31], мы задались следующими вопросами: как применение современных технологий автоматической обработки текста может повысить эффективность работы диалектолога, а также какие препятствия технического характера необходимо преодолеть ради достижения этой цели? В настоящей статье мы постараемся приблизиться к ответам на эти вопросы. Структуру работы можно кратко охарактеризовать следующим образом: во *Введении* раскрывается проблематика и методы исследования, приводятся основные

[1] Об этих процессах в пограничных областях Восточной Сербии и Западной Болгарии подробнее см. в [Сикимич, Соболев 2020].

черты тимокских говоров, один из которых служит материалом для статьи. В *Разделе 2* обосновывается выбранный нами способ записи диалектных текстов, там же охарактеризована их аннотация; в *Разделе 3* речь идет об автоматизации извлечения данных из диалектного корпуса, а также о преимуществах и ограничениях компьютерного поиска примеров языковых черт в сравнении с ручным. В *Разделе 4*, посвященном языковому профилированию информанта, визуализированы и сопоставлены результаты автоматического и ручного анализа диалектных черт в идиолекте носителя говора с. Берчиновац. В *Заключении* обсуждаются результаты, перспективы и ограничения дальнейшей работы.

На конференции «Балканские языки и диалекты: корпусные и квантитативные исследования» (18–20 октября 2018 г., ИЛИ РАН) в докладе «Количественные методы исследования сербского тимокского диалектного текста» авторами настоящей публикации был представлен метод установления частотности и правил дистрибуции различительных диалектных признаков в речи носителя нестандартного южнославянского идиома. Данный метод был опробован на примере идиолекта Драгини Милкич (1906 г. р.), носительницы тимокского говора[2] с. Берчиновац в районе г. Княжевац Заечарского округа в Восточной Сербии [Конёр и др. 2019]. Нарратив Д. Милкич, записанный в 1990-е гг., был опубликован в третьем томе «Диалектологического атласа говоров Восточной Сербии и Западной Болгарии» [Соболев 1998] (далее — SAOSWB или Атлас). Поскольку на первом этапе исследования мы не располагали дигитализированным аннотированным текстом, извлечение данных для анализа осуществлялось вручную, что, впрочем, не повлекло за собой существенных временных затрат в связи со сравнительно небольшим объемом текста (4453 словоформы). Однако если перед исследователем будет поставлена задача провести подобный эксперимент на большем (от 20 тыс. словоформ) объеме данных, то неавтоматизированный поиск информации будет затруднен ограниченностью временного ресурса. Автоматизация этого процесса должна способствовать:

[2] Тимокский говор относится к торлакскому наречию — группе диалектов в Юго-Восточной Сербии, Западной Болгарии и северной части Северной Македонии. Эти диалекты принадлежат по своим историко-фонетическим характеристикам к сербскому диалектному континууму, а по морфосинтаксическим особенностям (приобретенным в результате контактного влияния балканизированных южнославянских диалектов и неславянских балканских языков) — к ареальной группе балканских языков.

1) обработке больших объемов данных; 2) снижению количества ошибок, вызванных «человеческим фактором»; 3) снижению влияния субъективной интерпретации языковых явлений при принятии решений о включении того или иного примера в выборку; 4) сбережению временного ресурса исследователей.

В связи с этим, вторым этапом нашей работы над методом языкового (диалектного) профилирования стала попытка его автоматизации с помощью морфологической аннотации [3] (модель таггера была разработана Теодорой Вукович [Vuković et al. 2019] на основе ReLDI-таггера [Ljubešić et al. 2016]; ниже аннотация будет описана более подробно) и пользовательских Python-скриптов. Тексты SAOSWB были аннотированы для включения в «Тимокский корпус» [4], изначально состоявший из диалектных текстов, записанных, в основном, в 2015–2017 гг.

Приведем в сокращенном виде список исследованных признаков — черт прототипического («идеального») тимокского говора [5]:

1) рефлекс прасл. **tj* реализован как *č*: **vętje* > *veče* ‘больше’, **větja* > *sveča* ‘свеча’, футуральная частица *če*;
2) рефлекс прасл. **dj* реализован исключительно как *ǯ*: **medja* > *meǯa* ‘межа’, **tjudje* > *čuǯo* ‘чужое’, **vidj-* > *viǯ-* ‘вид-’;
3) совпавшие рефлексы первичных и вторичных редуцированных реализованы как гласный среднего ряда среднего подъема *ə*: **sъn* > *sən* ‘сон’, **vъšъ* > *vəška* ‘вошь’, **dъsky* > *dəska* ‘доска’; в суффиксе *-*ъvъ* (*takəv* ‘так(ов)ой’); **dьnь* > *dən* ‘день’;
4) аналитическое маркирование:
   a. косвенного объекта (IO) и именного посессора (POSS): *i* ***na tuj ovcu*** (IO тот.OBL.SG овца.OBL.SG) *se toj dade prvo i venəc*

[3] Обращение только к морфологическому уровню продиктовано неразработанностью на сегодняшний день машинной аннотации южнославянской диалектной фонетики, синтаксиса и лексики.

[4] Описание корпуса, а также интерактивная карта пунктов с примерами аудио- или видеозаписей и фотоматериалами находятся на сайте Института балканистики Сербской академии наук и искусств: http://balksrv2012.sanu.ac.rs/webdict/timok/index.

[5] Разъяснение причин принятия тех или иных априорных решений в связи с грамматическими и прагматическими особенностями некоторых различительных диалектных признаков, а также итоговая лингвистическая характеристика исследуемого говора представлены в [Конёр и др. 2019].

‘и этой овце вначале надевают (букв. «дают») венок’; *təg u moje znaṕe i* ***na mojega tatu*** (IO мой.OBL.SG отец.OBL.SG) *odnela vodenicu* ‘тогда, насколько я знаю, [вода] унесла и мельницу моего отца’;

б. периферийных падежных отношений при имени существительном: ***ot sviṕu*** (от свинья.OBL.SG) *ostala samo glava* ‘от свиньи осталась только голова’, *orala sam səs pluk* (с плуг.NOM.SG) ‘я пахала плугом’, *on bil* ***u vojsku*** (в армия.OBL.SG) ‘он был в армии’;

в. косвенного объекта (IO) и местоименного посессора (POSS): *toj* ***na nas*** (IO 1PL.ACC) *pričala baba jedna* ‘это нам рассказывала одна пожилая женщина’;

г. периферийных падежных отношений при личном местоимении: *pokraj ńu* (рядом F. 3SG.ACC) ‘рядом с ней’, *səs ńega* (с M.3SG.ACC) ‘с ним’, *da pečemo leb* ***u ńu*** (в F. 3SG.ACC) ‘чтобы печь хлеб в ней’;

5) наличие постпозитивного артикля в именной группе: *dojde do nas voda-****ta*** (вода-F.SG.DEF) *do ovdeka* ‘дошла до нас вода досюда’;

6) аналитический компаратив прилагательных. Употребление частицы компаратива *po* и отсутствие суффикса компаратива: ***po****mlad* ‘моложе, более молодой, младший’;

7) местоименная редупликация:
   а. прямого объекта: *tebe* ***te****= stra* (2SG.ACC 2SG.ACC= страх.NOM.SG) ‘тебе страшно’;
   б. косвенного объекта: *tep* ***ti****= je dobro* (2SG.DAT 2SG.DAT= быть.PRS.3SG хорошо) ‘тебе хорошо’;

8) отсутствие частицы конъюнктива в конструкциях с модальными глаголами и в формах футура: *sad ču # pričam* (сейчас FUT.1SG # говорить.PRS.1SG) ‘сейчас расскажу’, ср. *sad ču* ***da*** *pričam* (сейчас FUT.1SG SBJV говорить.PRS.1SG) [Конёр и др. 2019: 21–22].

При попытке автоматического анализа был обнаружен ряд существенных ограничений, возникающих в результате автоматизации аннотирования и поиска данных. При этом было установлено, что автоматизация извлечения лингвистически релевантных данных может существенно повысить эффективность работы исследователя, но не может заменить экспертную деятельность.

2. Языковой материал. Транскрипция и аннотация

2.1. Обработка транскрибированных текстов

Оригинальная транскрипция диалектных текстов из Восточной Сербии и Западной Болгарии в SAOSWB выполнена максимально детально: обозначены места ударений и использованы специальные символы, отсутствующие в алфавитах соответствующих литературных языков, что позволяет анализировать тексты с точки зрения синхронной и исторической фонетики и фонологии. Такая транскрипция передает и внутреннюю вариативность: некоторые фонемы в одном и том же идиолекте могут реализовываться в одинаковых контекстах с помощью разных аллофонов (например, *dan* / *dəᵃn* / *daᵊn* / *dən* ‘день’). Печатный текст из Атласа был дигитализирован и переведен в формат .txt при помощи программы OCR Transkribus; транскрипция была несколько упрощена. Символы ударения были замещены верхним регистром, иные сложные (составные) знаки были заменены символами из стандартного сербского алфавита (например, палатальные согласные) или пропущены, если обозначаемые ими фонетические явления не имеют различительной функции (см. второй столбец *Таблицы 1*).

2.2. Аннотация

Анализируемый в настоящей статье материал из с. Берчиновац содержит диалектные тексты, снабженные дополнительной информацией о морфологических характеристиках каждого токена (морфологическая аннотация) и его основной форме (лемматизация). Лемматизация проводилась на базе стандартного сербского языка. В частности, в качестве «начальной формы» любого глагола был реконструирован инфинитив, несмотря на то что в южных сербских диалектах эта форма утрачена.

Данная информация была добавлена автоматически при помощи модели таггера ReLDI, специально обученной на образце диалектного материала[6]. Этот таггер использует аннотацию, разработанную на базе морфосинтаксического кодирования MULTEXT-East V5, в которой каждый символ в ряду обозначений кодирует часть речи токена

[6] https://github.com/bravethea/Torlak-ReLDI-Tagger-2019

и граммемы различных грамматических категорий: например, слово *žena* ‘женщина, жена’ аннотировано как Ncfsny: Noun, common, feminine, singular, nominative, animate (yes). Оригинальная кодировка была расширена позицией для постпозитивного артикля, который может быть кодирован как -v, -t или -n (в зависимости от облика самого артикля). Например, токен *ženata* ‘[эта] женщина’ будет аннотирован как Ncfsny-t, где последний символ (-t) отражает присутствие постпозитивного артикля *-ta*. В *Таблице 1* ниже приведены примеры исходной и упрощенной транскрипции.

Таблица 1. Пример всех слоев транскрипции и дополнительной аннотации

Table 1. Examples of the transcription and additional annotation

Исходная транскрипция	**Упрощенная транскрипция**	**Аннотация**	**Лемма**
pa	pa	Cc	*pa*
dójde	dOjde	Vma3s	*doći*
na	na	Sa	*ma*
nás	nAs	Pp1-pa	*mi*
vodáta	vodAta	Npmsa-t	*voda*
doovdéka	doovdEka	Rgp	*dovde*

3. Автоматизация извлечения данных. Преимущества и ограничения

Поиск языковых явлений может быть в значительной степени автоматизирован, если предметом этого поиска является видимый облик слова и дополнительная (лингвистическая) информация, аннотированная в данном тексте. В случае «Тимокского корпуса», частью которого должны стать тексты Атласа, такой информацией является морфосинтаксическая аннотация и лемматизация. Проблемной для автоматизированного поиска является задача найти информацию, которая не выражена эксплицитно в тексте. Особенно это касается фонетических исследований: известно, что корпусы в принципе редко используются для подобных целей, равно как и для поиска / изучения редких явлений [Birkner 2015; Dash 2018]. Рассмотрим процесс обработки каждого из выбранных для анализа диалектных признаков.

Диалектные различия 1 (рефлекс прасл. **tj*), 2 (рефлекс прасл. **dj*), 3 (совпадение рефлексов первичных и вторичных редуцированных в гласном среднего ряда среднего подъема *ə*). Данные признаки объединены в одну группу, поскольку с ними связано одно существенное ограничение автоматизированного поиска. Все они являются результатом исторического развития фонем, и для их ручного поиска в тексте исследователь должен обладать знаниями в области славянской этимологии. Для полной автоматизации поиска слов с тем или иным рефлексом какой-либо праславянской фонемы или сочетания фонем необходимо, чтобы корпус, в который входит данный текст, располагал не только морфологической аннотацией, но также аннотацией для каждого токена в виде некой предполагаемой праформы или формы в языке-доноре в случае заимствования. «Тимокский корпус» диалектных текстов содержит слои морфосинтаксической аннотации на базе системы MULTEXT-east для сербского языка [Erjavec et al. 2003] [7]. Отметим, что создание слоя этимологической аннотации не является невыполнимой задачей [Dash, Hussain 2013], однако это потребовало бы намного больше ресурсов и времени, поэтому не входило в изначальный план исследования.

Без этимологической аннотации, как в нашем случае, автоматизация извлечения данных заключается в поиске по тексту всех единиц, содержащих интересующие нас фонетические элементы *č*, *ć*, *dž*, *đ*, что в конечном итоге возвращает нас к необходимости их последующего этимологического анализа и установления, происходят ли эти элементы от праславянских **tj* и **dj* в случае каждого токена, а также удаления тех единиц, которые не подходят для анализа. Приведем в пример поиск лексем с рефлексом праславянского **dj*. Известно, что в данной диалектной области возможны два варианта: *đ* (стандартноязыковой вариант) и *dž* (прототипический тимокский вариант). Ниже, в *Таблице 2*, содержится небольшой фрагмент (несколько показательных примеров) вывода данных. Доли диалектной и стандартноязыковой реализации признака рассчитываются от общего числа полученных примеров обоих типов.

На материале небольшого текста из Берчиновца и таких нечастотных фонем, как *đ*/*dž*, подобный поиск может в некоторой степени облегчить процесс извлечения данных. Однако уже в случае рефлексов **tj* автоматизированный поиск (без наличия автоматической аннотации)

[7] http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html

Таблица 2. Рефлексы **dj*: диалектная и стандартноязыковая реализация

Table 2. Reflexes of **dj*: dialect vs standard realization

Диалектная реализация	**Стандартноязыковая реализация**
**dj* > *dž*	**dj* > *đ*
lédža ‘спина’ *prédžu* ‘прядут’	*róđena* ‘рожденная’ *govéđe* ‘говяжье’
80 %	20 %

незначительно оптимизировал бы ручную работу. Фонемы *ć*/*č* (ожидаемые рефлексы **tj*) намного более частотны в данном диалекте, а значит, последующая обработка результатов автоматизированного поиска заняла бы значительно больше времени.

Кроме того, если бы наш текст был в несколько раз больше, можно было бы применить другой подход[8]: не искать абсолютно все лексемы с этимологическим **tj*, **dj* или любым другим диалектным признаком, а выбирать из ряда заведомо наиболее частотных в данных текстах и анализировать их варьирование в корпусе. В случае SAOSWB такими частотными лексемами могли бы оказаться все производные (в том числе клитики футура) от глагола **xъtěti* для анализа рефлекса **tj* или **sъnъ* и **dьnь* для проверки признака «совпадение рефлекса редуцированных» (для проверки вторичных редуцированных — **došьlъ* и **jedьnъ*[9]).

В *Таблице 3* представлены результаты применения данного подхода к третьему диалектному признаку «совпадение рефлексов редуцированных»: мы проанализировали лексему со значением «день» (**dьnь*, лемма *dan*) и — для проверки рефлексов вторичных редуцированных — глагол движения «идти» (лемма *ići*) в форме причастия на *-l* м. р. ед. ч., с использованием лемматизации и морфологической аннотации.

---

[8] Подобный подход для анализа частотности слов с диалектной (vs стандартноязыковой) акцентуацией был реализован в [Vuković et al. 2020].

[9] Следует отметить, что случай с данным признаком — намного сложнее, чем с первыми двумя, следовательно, автоматический поиск по тексту не сможет оптимизировать работу исследователя, которому пришлось бы искать все возможные варианты рефлексов редуцированных *a*, *o*, *y* и *ə*, ведь эти гласные есть практически в каждом слове.

Таблица 3. Рефлексы редуцированных:
диалектная и стандартноязыковая реализация

Table 3. Reflexes of the reduced vowels:
dialect vs standard realization

Диалектная реализация		**Стандартноязыковая реализация**	
Корневой редуцированный	**Вторичный редуцированный**	**Корневой редуцированный**	**Вторичный редуцированный**
('vəzdƏn', 'RGP', 'vəzdan') ('dəna', 'Ncmsa', 'dan')	('dOšo', 'Vmp-sm', 'doći') ('dOšo', 'Vmp-sm', 'doći') ('dOšəl', 'Vmp-sm', 'doći')	('dAn', 'Ncmsn', 'dan')	('Išal', 'Vmp-sm', 'ići')
88 %		12 %	

В итоге по выбранным лексемам стандартноязыковая реализация составила 12 %, диалектная — 88 %. Это отличается от результата подсчета абсолютно всех слов с рефлексами редуцированных (37 % и 63 % соответственно), однако, как нам представляется, описанный выше подход — единственное решение для действительно большого объема данных.

Извлечение морфосинтаксических признаков было значительно оптимизировано (по сравнению с ручным поиском) после аннотации и лемматизации текста. В *Таблице 4* приведены результаты автоматизированного поиска в сравнении с результатами, полученными на предыдущем этапе исследования (до аннотации).

Диалектное различие 4а. Аналитическое маркирование косвенного объекта[10] и именного посессора.

Диалектное различие 4б. Аналитическое маркирование периферийных падежных отношений при имени существительном. Рассматриваются только те предлоги, которые в стандартном языке употребляются с генитивом (но не одновременно с аккузативом), локативом и инструменталисом.

[10] «Любой косвенный объект, выраженный именем существительным, будет считаться потенциально маркируемым аналитическим показателем *па* ‘на, к’ (все такие обнаруженные в тексте объекты принимаются за 100 % реализации в „каноническом“ говоре)» [Конёр и др. 2019: 22].

Таблица 4. Аналитическое маркирование

Table 4. Analytic case-marking

	Диалектная реализация	**Стандартноязыковая реализация**
4а.	tƏg u mojE znAnje i **na mojEga tAtu** (IO мой.OBL отец.OBL) odnEla vodenIcu ‘Тогда, насколько я знаю, [вода] унесла и мельницу моего отца’.	В тексте отсутствуют синтетические формы косвенного объекта и посессора, выраженного существительным.
	100 %	0 %
4б.	boluvAla sam nEšto **u glAvu** (в голова.OBL.SG) ‘У меня что-то болела голова’. a mI smo si **u selO** (в село.NOM.SG) ‘Мы живем себе в селе’.	ne znAm **kOje gOdine** (какой.F.GEN.SG год.GEN.SG) bIlo ‘Не знаю, в каком году это было’.
	99 %	1 %
4в.	pa dOjde **na nAs** (IO 1PL.ACC) vodAta doovdEka ‘и вода поднялась к нам досюда’ tOj **na nAs** (IO 1PL.ACC) pričAla bAba jednA ‘Нам это рассказывала одна пожилая женщина’.	pə nIšta, On dadE mEne [11] parU, jA **njEmu** (M.3SG.DAT) nEšto ‘и ничего, он дал мне монетку, и я ему что-то’
	80 %	20 %
4г.	s njU / səs njU (с F. 3SG.ACC) ‘с ней’ pOkraj njU (рядом F. 3SG.ACC) ‘рядом с ней’ s njEga / səs njEga (с M.3SG.ACC) ‘с ним’ s nAs (с 1PL.ACC) ‘с нами’ səs njI (с 3PL.ACC) ‘с ними’	В тексте отсутствуют синтетические формы местоимений в генитиве, инструменталисе и локативе.
	100 %	0 %

[11] Форма *mEne* в данном случае должна быть идентифицирована как датив (экавский рефлекс **menĕ* > *mene* в противоположность новоштокавскому исключению из общего экавизма **menĕ* > *meni*), однако таггер интерпретировал эту форму как аккузатив: *mEne* — Pp1-s**a** 'ja'.

Диалектное различие 4в. Аналитическое маркирование косвенного объекта (IO) и местоименного посессора (POSS)[12].

Диалектное различие 4г. Аналитическое маркирование периферийных падежных отношений при местоимении. Рассматриваются только те предлоги, которые в стандартном языке употребляются с генитивом (но не одновременно с аккузативом), локативом и инструменталисом.

Диалектное различие 5. Наличие постпозитивного артикля в именной группе. Поиск примеров, отличающихся этим признаком, — несложная задача, если текст снабжен морфосинтаксической аннотацией. Однако одним из ограничений автоматизированного поиска на данном этапе является нахождение тех примеров, которые открывают позицию для определенного артикля. Ручной анализ выявил, что в тексте имеется 115 таких именных групп (ИГ). Однако без дополнительного слоя специальной синтаксической аннотации («разметка кореференции», англ. coreference annotation [Deemter, Kibble 1999]), сообщающей о том, является та или иная группа определенной или нет, подобный поиск невозможен.

Диалектное различие 6. Аналитический компаратив (наличие частицы компаратива *po* и отсутствие суффиксов компаратива).

В тексте встретилась только одна форма компаратива: *'pOmladu'*. Синтетический (суффиксальный) компаратив не обнаружен.

Диалектное различие 7а. Местоименная редупликация прямого объекта (рассматриваются случаи именно удвоения объекта, а не любого иного употребления кратких местоимений при глаголе, например, в анафорической функции).

Этот и следующие признаки также вызывают некоторые трудности при применении нашего метода лингвистического профилирования информанта. Анализ примеров удвоения объекта в репрезентативных тимокских текстах [Escher 2021] показал, что этот феномен в принципе нерегулярен в данном идиоме и всегда связан с прагматическими характеристиками высказывания. Очевидно, что мы не можем назвать «потенциально удваиваемыми» абсолютно все прямые / косвенные объекты. Удваиваться, как правило, могут (но не должны) топикализированные объекты. Более того, значительная часть примеров местоименного удвоения в репрезентативных источниках представляет

[12] В данном различии учитываются только полные (полноударные) формы местоимений.

собой скорее аккузатив / датив темы, а не грамматическое средство указания на синтаксическую роль объекта (индексирование). В связи с этим задача найти в тексте все потенциально удвоенные объекты становится исключительно сложной не только для машинного, но и для экспертного поиска. Следует не только найти все прямые / косвенные объекты, но и проанализировать широкий контекст и прагматические особенности их употребления и принять решение, может ли объект в данном высказывании быть удвоенным. Единственный пример местоименного удвоения прямого объекта, выраженного именем существительным в анализируемом тексте — *jA **ga**= vIdim tUj mOjega čIču* (1SG.NOM M.SG.ACC= видеть.PRS.1SG здесь мой.M.ACC.SG дядя.ACC.SG) ‘я увидела (досл. «вижу») здесь моего дядю’ — далеко не очевидный случай. Нельзя быть полностью уверенными в том, что данное удвоение представляет собой именно индексирование, а не аккузатив темы.

Диалектное различие 7б. Местоименная редупликация косвенного объекта.

В тексте не встречается удвоение косвенного объекта.

Диалектное различие 8. Отсутствие частицы конъюнктива при модальных глаголах и в формах будущего времени (*Таблица 5*).

Таблица 5. Частица конъюнктива: диалектная и стандартноязыковая реализация

Table 5. Conjuntive particle: dialect vs. standard realization

Диалектная реализация	**Стандартноязыковая реализация**
čEkaj sAd **ču prIčam** (сейчас FUT.1SG говорить.PRS.1SG) ‘Подожди, сейчас буду говорить’.	on če da poglEda (M.3SG.NOM FUT.3SG SBJV смотреть.PRS.3SG) ‘он посмотрит’
37 %	63 %

4. Языковой профиль информанта

В данном разделе мы визуализируем итог исследования в виде графика, отображающего языковой профиль информанта (соотношение диалектной и инодиалектной реализации различительных признаков) в сравнении с профилем, полученным на предыдущем этапе исследования, т. е. при ручной обработке текста. Напомним, что для построения графика по итогам автоматического анализа соотношение различных рефлексов редуцированных рассчитывалось исходя из поиска

Рисунок 1. Языковой профиль информанта
по результатам автоматического анализа

Figure 1. Dialect profile of the informant
based on the results of the automatic analysis

нескольких частотных слов, а в случае с признаком «наличия постпозитивного артикля в референтной ИГ» мы использовали результаты ручного подсчета всех референтных ИГ в группе, потенциально открывающих позицию для определенного артикля.

Самые значительные различия в результатах анализа относятся к признаку «рефлексы редуцированных». Использование иного метода и отказ от подсчета абсолютно всех слов с данным рефлексом в тексте несколько исказили реальную картину. Однако, как уже упоминалось в *Разделе 3*, ввиду отсутствия этимологической аннотации, поиск абсолютно всех лексем с интересующим нас историко-фонетическим феноменом невозможен. Для больших объемов данных неизбежно придется принять описанный выше метод.

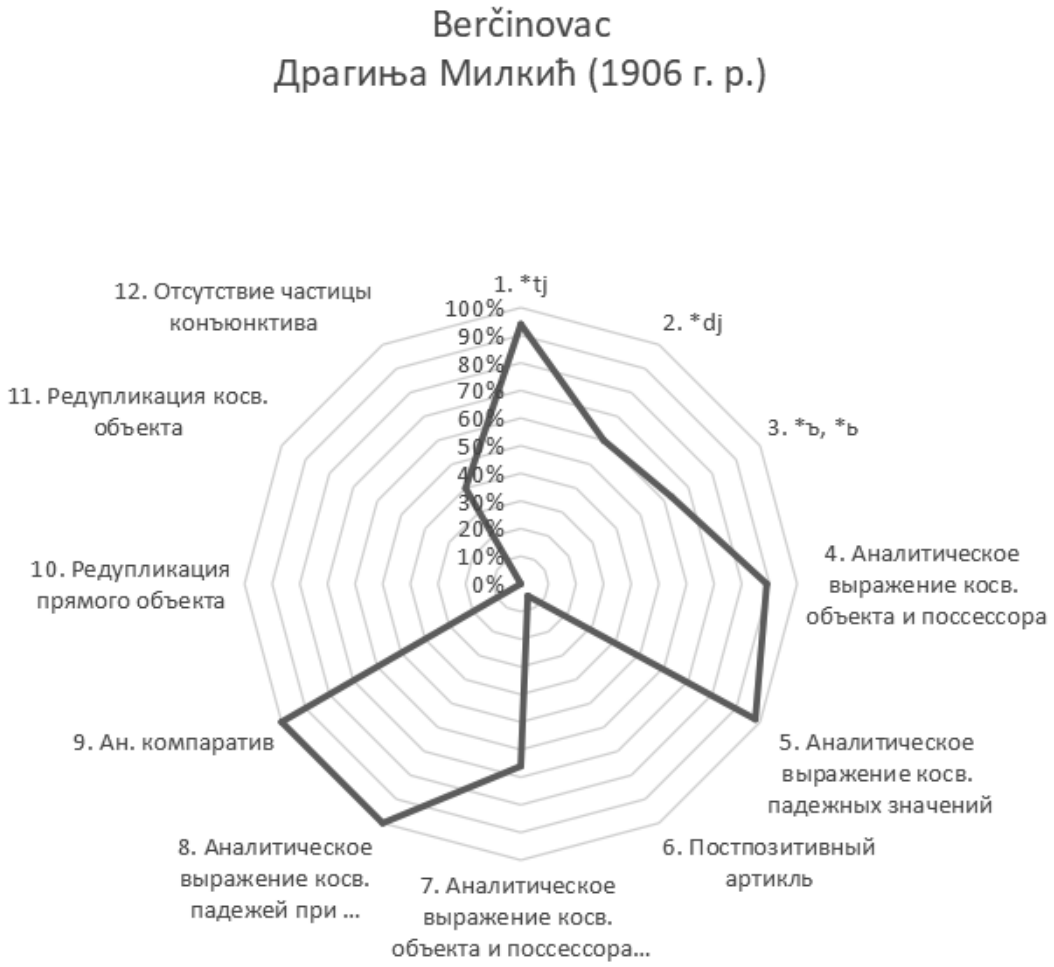

Рисунок 2. Языковой профиль информанта по результатам ручного анализа

Figure 2. Dialect profile of the informant based on the results of the manual analysis

5. Заключение

Известно, что корпусная лингвистика и автоматическая обработка / автоматический поиск по тексту имеют свои преимущества и недостатки, которые следует иметь в виду, используя предлагаемые ими методы. Как правило, корпусы представляют собой большие собрания текстов. Их ценность основывается на ясно сформулированных принципах: относительно большой объем включенных текстов, репрезентативность и наличие аннотации (дополнительной лингвистической и/или металингвистической информации, которая оптимизирует поиск). Объем корпуса чрезвычайно важен: чем он больше, тем

выше вероятность появления интересующих исследователя языковых явлений и тем правомернее будут выводы об их частотности и контекстных условиях употребления тех или иных форм. Репрезентативность, помимо того что она тесно связана с объемом корпуса, подразумевает наличие в нем аутентичных образцов текста на данном языке. Что касается аннотации, то корпусы, как правило, содержат информацию о частеречной принадлежности токенов и лемматизацию. Некоторые корпусы также содержат информацию о синтаксической структуре и (существенно реже) о других языковых уровнях — фонологии или прагматике.

Изначально «Тимокский корпус» [Vuković et al. 2019], частью которого стали и тексты из SAOSWB, разрабатывался с целью морфосинтаксического анализа отдельных языковых явлений («балканизмов»), характерных для южных диалектов Сербии, и не подразумевал наличие синтаксической и прагматической аннотации. Этим объясняются отдельные ограничения машинного анализа, показанные нами на примере текста из Берчиновца (например, невозможность автоматически найти все ИГ, открывающие позицию для определенного артикля).

Ситуация с историко-фонетическими языковыми признаками усложняется следующими обстоятельствами. Абсолютное большинство корпусов — это собрания письменных текстов, которые легко доступны в дигитализированном виде. Разработка диалектных корпусов и корпусов разговорного языка подразумевает намного больше усилий по сравнению с письменными корпусами, и именно поэтому они сравнительно немногочисленны [13]. В то же время они представляют собой исключительно ценный ресурс для исследования реального и естественного употребления языка. Корпусы разговорной речи базируются на фонетических или орфографических транскрипциях аудиозаписей, расшифровка которых редко бывает выполнена с максимальной детальностью, поскольку это требует высокой квалификации и значительных временных затрат транскрибирующего. Существует метод полу-орфографической (полуфонетической) транскрипции [Goedertier et al. 2000], который часто применяется для корпусов диалектной или разговорной речи, где нестандартные фонетические явления отображаются при помощи знаков стандартного алфавита. По причине упрощенной транскрипции диалектные корпусы редко используются для глубокого исследования

[13] Разнообразные корпусы устной речи представлены, например, на портале https://www.clarin.eu/resource-families/spoken-corpora.

фонетики и фонологии. Если это и происходит, то необходимым является принятие определенных допущений, как было показано на примере признака «рефлексы редуцированных».

Список условных сокращений

1, 2, 3 — лицо у глаголов и местоимений, ACC — аккузатив, DAT — датив, F — женский род, FUT — будущее время, GEN — генитив, IO — косвенный объект, M — мужской род, NOM — номинатив, OBL — косвенный падеж, PL — множественное число, POSS — местоименный посессор, PRS — настоящее время, SAOSWB — Диалектологический атлас Восточной Сербии и Западной Болгарии, SG — единственное число, SBJV — показатель зависимой глагольной формы, ИГ — именная группа.

Литература

Конёр и др. 2019 — Д. В. Конёр, А. Л. Макарова, А. Н. Соболев. Статистический метод языкового профилирования носителя диалекта (на материале восточносербского идиома села Берчиновац) // Вестник Томского государственного университета. Филология. 2019. № 58. С. 17–33. DOI: 10.17223/19986645/58/2.

Сикимич, Соболев 2020 — Б. Сикимич, А. Н. Соболев. Процессы дивергенции в разделенном государственной границей западноюжнославянском диалекте (на материале современной диалектной речи Восточной Сербии и Западной Болгарии) // Вестник Томского государственного университета. Филология. 2020. № 66. С. 158–176. DOI: 10.17223/19986645/66/9.

Соболев 1998 — А. Н. Соболев. О диалектологическом атласе Восточной Сербии и Западной Болгарии // Г. П. Клепикова (отв. ред.). Исследования по славянской диалектологии. Вып. 5. М.: Институт славяноведения РАН, 1998. С. 106–167.

Birkner 2015 — V. Birkner. The advantages and disadvantages of employing corpus evidence in sociolinguistic studies // The Teacher Magazine. 2015. Vol. 2. P. 11–17.

Dash 2012 — N. S. Dash. Etymological Annotation: a New Concept of Corpus Annotation // Proceedings of the 34th All India Conference of Linguists (34-AICL). Shillong, India, 2012. P. 100–104.

Dash, Arulmozi 2018 — N. S. Dash, S. Arulmozi. Limitations of language corpora // N. Dash, S. Arulmozi. History, features, and typology of language corpora. Singapore: Springer Singapore, 2018. P. 259–272.

Dash, Hussain 2013 — N. S. Dash, M. M. Hussain. Designing a Generic Scheme for Etymological Annotation: a New Type of Language Corpora Annotation // P. Bhattacharayya, K.-S. Choi (eds.). Proceedings of the 11th Workshop on Asian Language Resources. Nagoya: Asian Federation of Natural Language Processing, 2013. P. 64–71.

Deemter, Kibble 1999 — K. van Deemter, R. Kibble. What is coreference, and what should coreference annotation be? // A. Bagga, B. Baldwin, S. Shelton (eds.).

Proceedings of the Workshop on Coreference and Its Applications. Stroudsburg, PA: Association for Computational Linguistics, 1999. P. 90–96.

Erjavec et al. 2003 — T. Erjavec, C. Krstev, V. Petkevic, K. Simov, M. Tadic, D. Vitas. The MULTEXT-east morphosyntactic specifications for Slavic languages // T. Erjavec, D. Vitas (eds.). Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003. Stroudsburg, PA: Association for Computational Linguistics, 2003. P. 25–32.

Escher 2021 — A. L. Escher. Double argument marking in Timok dialect texts (in Balkan Slavic context). Zeitschrift für Slawistik. Forthcoming.

Goedertier et al. 2000 — W. Goedertier, S. Goddijn, J.-P. Martens. Orthographic transcription of the spoken Dutch corpus // M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhouer (eds.). Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece. Athens: National Technical University of Athens Press, 2000. P. 909–914.

Ljubešić et al. 2016 — N. Ljubešić, F. Klubička, Ž. Agić, I.-P. Jazbec. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian // N. Calzolari, Kh. Choukri, Th. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.). Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris: European Language Resources Association, 2016. P. 4264–4270.

Vuković et al. 2019 — T. Vuković, N. Muheim, O. Winistörfer, I. Simko, A. Makarova, S. Bradjan. Corpora and Processing Tools for Non-Standard Contemporary and Diachronic Balkan Slavic // I. Temnikova, I. Nikolova, N. Konstantinova (eds.). Proceedings of the Student Research Workshop associated with The 12th International Conference on Recent Advances in Natural Language Processing (RANLP 2019). Shoumen: Incoma, 2019. P. 62–68.

Vuković et al. 2020 — T. Vuković, B. Sonnenhauser, A. Escher. Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus. Manuscript.

Источники

SAOSWB — A. N. Sobolev. Sprachatlas Ostserbiens und Westbulgariens. Bd. I. Problemstellung, Materialen und Kommentare, Kartenanalyse. Bd. II. Sprachkarten. Bd. III. Texte. Marburg; Lahn: Biblion Verlag, 1998.

References

Birkner 2015 — V. Birkner. The advantages and disadvantages of employing corpus evidence in sociolinguistic studies. *The Teacher Magazine*. 2015. Vol. 2. P. 11–17.

Dash 2012 — N. S. Dash. Etymological Annotation: a New Concept of Corpus Annotation. *Proceedings of the 34th All India Conference of Linguists (34-AICL)*. Shillong, India, 2012. P. 100–104.

Dash, Arulmozi 2018 — N. S. Dash, S. Arulmozi. Limitations of language corpora. N. Dash, S. Arulmozi. *History, features, and typology of language corpora*. Singapore: Springer Singapore, 2018. P. 259–272.

Dash, Hussain 2013 — N. S. Dash, M. M. Hussain. Designing a Generic Scheme for Etymological Annotation: a New Type of Language Corpora Annotation. P. Bhattacharayya, K.-S. Choi (eds.). *Proceedings of the 11th Workshop on Asian Language Resources*. Nagoya: Asian Federation of Natural Language Processing, 2013. P. 64–71.

Deemter, Kibble 1999 — K. van Deemter, R. Kibble. What is coreference, and what should coreference annotation be? A. Bagga, B. Baldwin, S. Shelton (eds.). *Proceedings of the Workshop on Coreference and Its Applications*. Stroudsburg, PA: Association for Computational Linguistics, 1999. P. 90–96.

Erjavec et al. 2003 — T. Erjavec, C. Krstev, V. Petkevic, K. Simov, M. Tadic, D. Vitas. The MULTEXT-east morphosyntactic specifications for Slavic languages. T. Erjavec, D. Vitas (eds.). *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*. Stroudsburg, PA: Association for Computational Linguistics, 2003. P. 25–32.

Escher 2021 — A. L. Escher. Double argument marking in Timok dialect texts (in Balkan Slavic context). *Zeitschrift für Slawistik.* Forthcoming.

Goedertier et al. 2000 — W. Goedertier, S. Goddijn, J.-P. Martens. Orthographic transcription of the spoken Dutch corpus. M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhouer (eds.). *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000), Athens, Greece. Athens: National Technical University of Athens Press, 2000. P. 909–914.

Konior et al. 2019 — D. V. Konior, A. L. Makarova, A. N. Sobolev. Statisticheskiy metod yazykovogo profilirovaniya nositelya dialekta (na materiale vostochnoserbskogo idioma sela Berchinovats) [Quantitative method of language profiling of a dialect speaker (based on the material of the East Serbian idiom of the village of Bercinovac)]. *Tomsk State University Journal of Philology*. 2019. No. 58. P. 17–33.

Ljubešić et al. 2016 — N. Ljubešić, F. Klubička, Ž. Agić, I.-P. Jazbec. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. N. Calzolari, Kh. Choukri, Th. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Paris: European Language Resources Association, 2016. P. 4264–4270.

Sikimić, Sobolev 2020 — B. Sikimić, A. N. Sobolev. Processy divergentcii v razdelennom gosudarstvennoy granitcey zapadnoyuzhnoslavyanskom dialekte (na materiale sovremennoy dialektnoy rechi Vostochnoy Serbii i Zapadnoy Bolgarii) [Divergence Processes in the West South Slavic Dialect Divided by the State Border (Based on the Modern Dialect Speech of Eastern Serbia and Western Bulgaria)]. *Tomsk State University Journal of Philology*. 2020. No. 66. P. 158–176. DOI: 10.17223/19986645/66/9.

Sobolev 1998 — A. N. Sobolev. O dialektologicheskom atlase Vostochnoy Serbii i Zapadnoy Bolgarii [On the dialectological atlas of Eastern Serbia and Western

Bulgaria]. G. P. Klepikova (ed.). *Issledovaniya po slavyanskoy dialektologii* [Studies in Slavic Dialectology]. Iss. 5. Moscow: Institute of Slavic Studies RAS, 1998. P. 106–167.

Vuković et al. 2019 — T. Vuković, N. Muheim, O. Winistörfer, I. Simko, A. Makarova, S. Bradjan. Corpora and Processing Tools for Non-Standard Contemporary and Diachronic Balkan Slavic. I. Temnikova, I. Nikolova, N. Konstantinova (eds.). *Proceedings of the Student Research Workshop associated with The 12th International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Shoumen: Incoma, 2019. P. 62–68.

Vuković et al. 2020 — T. Vuković, B. Sonnenhauser, A. Escher. Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus. Manuscript.

Sources

SAOSWB — A. N. Sobolev. Sprachatlas Ostserbiens und Westbulgariens. Bd. I. Problemstellung, Materialen und Kommentare, Kartenanalyse. Bd. II. Sprachkarten. Bd. III. Texte. Marburg; Lahn: Biblion Verlag, 1998.