

УДК 81.33+811.161.3

ЧАСТНАЯ КОМПЬЮТЕРНАЯ МЕТОДИКА ВЫЯВЛЕНИЯ БЕЛОРУССКИХ ОМОФОРМОВ В ЮРИДИЧЕСКИХ ТЕКСТАХ

Н. В. Супрунчук

*Минский государственный лингвистический университет,
ул. Захарова, 21Б-414, 220034, Минск, Беларусь, suprunchuk@mail.ru*

Анализируется новый пилотный корпус кодексов на белорусском и русском языках. С помощью процессора NooJ проведена автоматическая разметка и выявлены случаи множественного тэгирования (около 17 %). Предложены синтаксические правила для улучшения работы белорусского модуля NooJ, в частности для определения части речи, времени глаголов и падежа существительных. Результаты можно использовать для автоматического перевода и синтеза устной речи.

Ключевые слова: автоматическая обработка текста; время глагола; омонимия; полисемия; неопределённзначность; лингвистический процессор NooJ; корпус текстов.

A PRIVATE COMPUTER-BASED APPROACH FOR IDENTIFYING BELORUSSIAN HOMOPHORMES IN LEGAL TEXTS

N. V. Suprunchuk

*Minsk State Linguistic University,
Zakharova St., 21B-414, 220034, Minsk, Belarus, suprunchuk@mail.ru*

A new pilot corpus of law codes in Belarusian in Russian is analyzed. The NooJ processor was used to perform automatic tagging and identify cases of multiple tagging (about 17 %). Syntactic rules were proposed to improve the performance of the Belarusian NooJ module, in particular for determining part of speech, verb tenses and noun cases. The results can be used for automatic translation and speech synthesis.

Keywords: automatic text processing; verb tense; homonymy; polysemy; ambiguity; NooJ linguistic processor; text corpus.

Как известно, язык является знаковой системой, а знаки – двусторонними сущностями, то есть имеют план содержания и план выражения. Соотношение между двумя планами не всегда простое, симметричное. В связи с этим и выделяются синонимы, омонимы и многозначные слова. В качестве конструкторов, понятий омонимия и полисемия различаются, однако разграничить их на практике получается не всегда. Кроме того, известны факты частичного сходства планов выражения (ср. омоформы, омофоны, омографы) [1; 2]. В некоторых случаях полезно разде-

лять все эти виды совпадений, однако для автоматического анализа текстов не всегда важны указанные терминологические различия — главным оказывается более конкретный вопрос о том, можно ли однозначно понять ту либо иную форму или нет. Поэтому в компьютерной лингвистике довольно распространено обобщающее наименование «неоднозначность» или «неопределённость». И, соответственно, задача прояснения смысла называется «снятие, решение, ликвидация и т. д. неоднозначности». Так, создатели Национального корпуса русского языка отмечали: «Различия в разметке и в системах автоматического разрешения неоднозначности (WSD – word-sense disambiguation) определяются, прежде всего, тем, каковы потребности пользователей конечного продукта и каким способом (и с какими затратами) разработчики собираются добиться нужного результата» [3, с. 445], см. также [4]. Таким образом, в нашей статье анализируются случаи полисемии и омонимии, а также близкие к ним, без детализации различий в терминах.

Ряд законодательных актов нашей страны, начиная с Конституции, гарантирует равноправие двух государственных языков. В частности, в законе «О нормативных правовых актах» установлено следующее:

– «основными требованиями нормотворческой техники являются... ясность, простота и доступность языка изложения нормативного правового акта...» (ст. 28);

– «нормативный правовой акт принимается (издается) уполномоченным на то нормотворческим органом (должностным лицом) на белорусском и (или) русском языках» (ст. 54);

– «официальное опубликование нормативных правовых актов осуществляется на том государственном языке (языках), на котором они приняты (изданы), и с учетом требований, установленных законодательными актами» (ст. 60).

В Беларуси действует 27 кодексов, причем на белорусском языке был принят только один из них – Кодекс о культуре. Для исправления такой несимметричности и обеспечения равного доступа граждан к правовой информации был начат перевод законодательных актов на белорусский язык. В эту деятельность включились специалисты Академии наук, ряда вузов и других организаций, и некоторые из них в 2018 году объединились в рамках проекта Института проблем информатики НАН [5]. Кроме того, «в соответствии с поручением, данным Главой государства 1 марта 2019 г. в ходе “Большого разговора с Президентом”, при Национальном центре правовой информации был создан экспертный совет по переводу законодательных актов на белорусский язык. В состав экспертного совета вошли представители Администрации Президента Республики Беларусь, Национальной академии наук Беларуси, Палаты

представителей Национального собрания, Министерства юстиции, Министерства культуры, Национального центра правовой информации, Национального центра законодательства и правовых исследований, ведущих вузов страны и иных организаций, в том числе негосударственных, например общественного объединения «Товарищество белорусского языка имени Франциска Скорины»» [6].

В Институте проблем информатики белорусская версия кодексов готовилась с использованием современных компьютерных сервисов, которые помогают осуществлять перевод и вычитку текстов. В результате работы над правовыми актами была собрана коллекция кодексов. На их основе были подготовлены параллельный подкорпус «Белорусского Н-корпуса», а также, с помощью программы «Нудж», репрезентативный локальный корпус. Его характеристики: 52 463 словоупотребления, 40 647 словоформ, 1 796 цифр, 10 020 пробелов и знаков препинания, 83 788 тэгов. В ряде случаев автоматический модуль предложил несколько вариантов разметки, то есть выявил омонимы, омографы и подобные случаи, всего 3 172 (подробнее см. [7]). Самые частые из них, с частотой более 10, были классифицированы и анализируются далее.

Так, были обнаружены следующие неоднозначности:

1) грамматические (омоформы), когда одному плану выражения соответствует несколько граммов, например: *несла* – женский и средний род глагола; *закон, Беларусь* – именительный и винительный падеж существительного;

2) омоформы, связанные с категорией одушевленности: *асоба* – человек или организация; *эмбрыён* (одушевленное или нет); *новага, гэтага, нашага* и др. – род. падеж или вин. падеж прилагательных и местоимений: *няма новага (чалавека, дома) – бачу новага (чалавека)*;

3) предлоги, управляющие разными падежами: *у, на* и др.;

4) частичное совпадение парадигмы и синтаксической функции у слов разных частей речи: *рабочы* ‘1) существительное; 2) прилагательное’, *таксама* ‘наречие; союз; частица’;

5) лексико-грамматическая неоднозначность: *які* ‘1) местоимение; 2) мн. число существительного *як*’, *дзікі* ‘1) прилагательное; 2) животное «кабан»’, *варта* ‘1) охрана; 2) предикативное слово «стоит»’, *правы* ‘1) прилагательное «правый»; 2) существительное «право»’;

6) совпадение разрядов местоимений: *яго, гэта*.

Были выявлены и смешанные случаи; ср., например, *далей*: 1) сравнительная степень наречия *далеко*, т. е. «дальше», 2) наречие «в дальнейшем»; 3) родительный мн. ч. от слова *даль*: *гэтых далей* (наряду с *гэтых даляў*).

Для сокращения случаев неоднозначности был привлечен морфоло-

гический анализатор «Нудж». В ряде примеров он верно определяет разряд или грамматическую форму, но нередко выдает несколько вариантов разметки, что связано с недостаточной разработанностью формальных грамматик для белорусского языка. В связи с этим было принято решение подготовить частные формальные правила, которые можно было бы добавить к общему белорусскому модулю, созданному в 2012 году в лаборатории анализа и синтеза речи Национальной академии наук Беларуси.

Ниже представлено правило, позволяющее выбирать между местоимением *мае* и глаголом *мае* (рис. 1). Оно является одним из примеров автоматических синтаксических правил, помогающих прояснить случаи совпадения глагольных и именных форм (ср., также *складзе* (сущ.) – *складзЕ* (глагол), *разрываў* (сущ.) – *разрываў* (глагол) и др.).

Так, до его применения мы получали три варианта разметки словоформы в предложениях типа следующего:

пасажыр – фізічная асоба, якая мае праязны дакумент (білет), што сведчыць заключэнне дагавора перавозкі пасажыра ўнутраным водным транспартам... (Кодэкс унутранага воднага транспарту Рэспублікі Беларусь)

<мой, PRONOUN+FLX=МОЙ+Case=Accusative+Animation=Inanimate+Number=Plural+Group=Possessive>

<мой, PRONOUN+FLX=МОЙ+Case=Nominative+Number=Plural+Group=Possessive>

<мець, VERB+FLX=МЕЦЬ+Person=3+Aspect=Imperfective+Mood=Indicative+Tense=Present+Number=Singular>

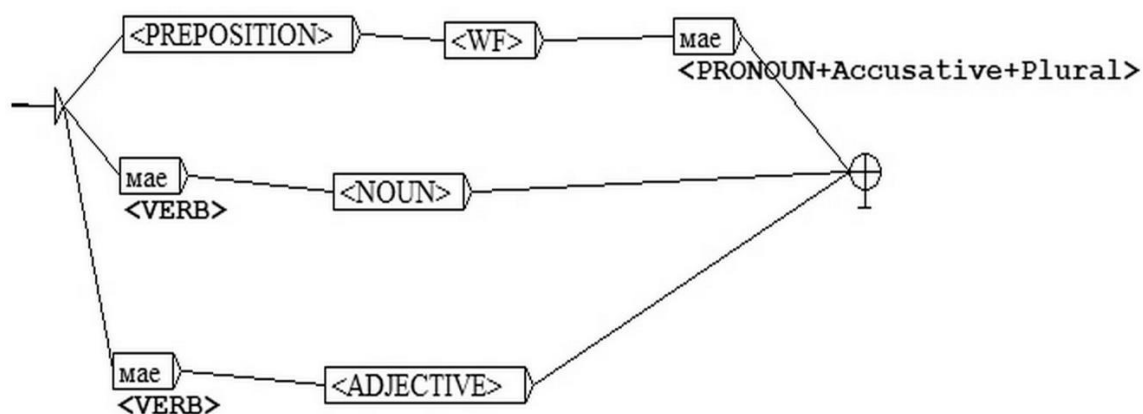


Рис. 1. Автоматическое правило для тэгирования формы *мае*

Второе и третье правило помогают различить родительный и да-

тельный падежи у существительных мужского рода на -у/-ю: *роду, краю* (рис. 2 и 3). Ср. неоднозначность разметки в примере ниже:

*У гэтым выпадку сума ў афіцыйнай грашовай адзінцы Рэспублікі Беларусь... вызначаецца па афіцыйным курсе адпаведнай валюты або ўмоўных (разліковых) грашовых адзінак на дзень **плацяжу**...* (Банкаўскі кодэкс)

<плацеж, NOUN+FLX=AAJIT+Meaning=Common+Case=Dative+
+Animation= Inanimate+Gender=Masculine>

<плацеж, NOUN+FLX=AAJIT+Meaning=Common+Case=Genitive+A
nimation= Inanimate+Gender=Masculine>

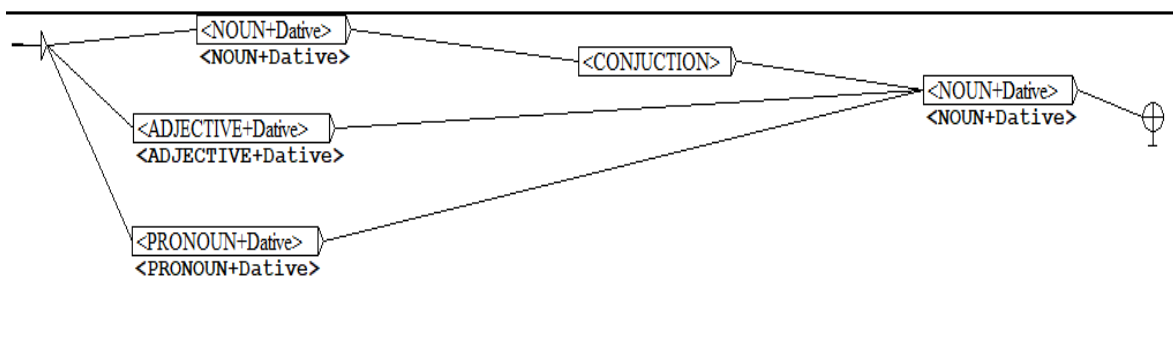


Рис. 2. Автоматическое правило для выявления форм дательного падежа существительных мужского рода

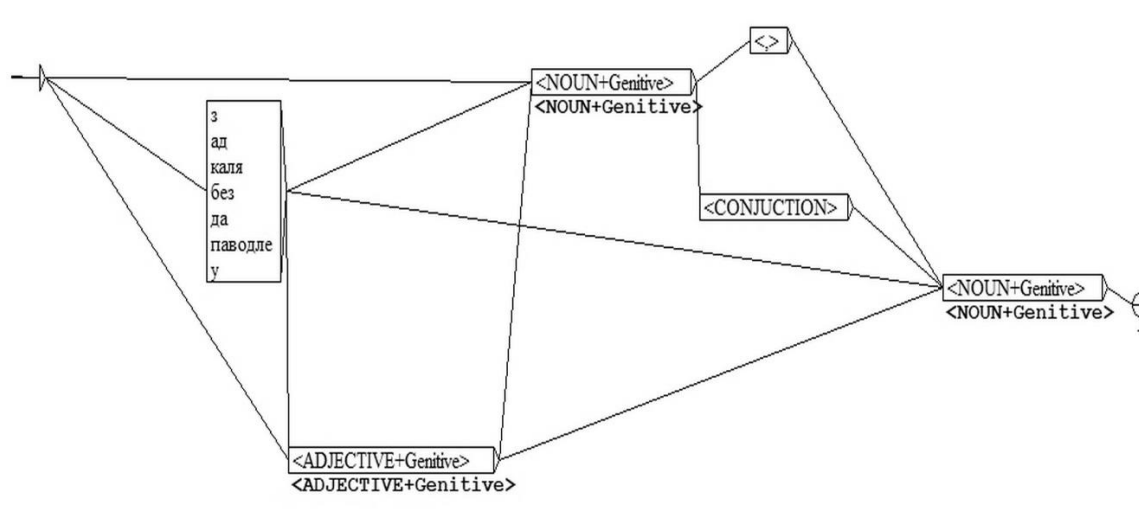


Рис. 3. Автоматическое правило для выявления форм родительного падежа существительных мужского рода (подготовили В. В. Воронович и Н. В. Супрунчук)

Таким образом, работа над параллельным белорусско-русским юридическим корпусом, подготовленным в рамках сотрудничества ряда специалистов по проекту лаборатории анализа и синтеза речи НАН Беларуси, позволяет сделать важные наблюдения над автоматической размет-

кой белорусских текстов. Были выявлены типы неоднозначности (лексическая, грамматическая, лексико-грамматическая омонимия и многозначность). В рамках экспериментального корпуса из 52 тысяч словоупотреблений обнаружено около 9 тысяч неоднозначных, т. е. около 17 % случаев требуют дополнительной разметки (автоматической или ручной). Предложены автоматические правила (NooJ grammars), позволяющие частично выполнять эту задачу. Выявлены показатели для разграничения формально идентичных форм глагола и существительного. Результаты исследования могут быть использованы при лингвистическом анализе текстов на белорусском языке, для компьютерного перевода и синтеза речи.

Библиографические ссылки

1. Зализняк А. А. Феномен многозначности и способы его описания // Вопросы языкознания. 2004. № 2. С. 20–45.
2. Старычонок В. Дз. Мнагазначнасць слова ў беларускай мове : у 3 кн. Мінск : Колорград, 2017.
3. Многозначность как прикладная проблема: лексико-семантическая разметка в Национальном корпусе русского языка / Е. В. Рахилина и др. // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции «Диалог 2006» / под ред. Н. И. Лауфер. М. : Изд-во РГГУ, 2006. С. 445–460.
4. Automatic Disambiguation of the Belarusian-Russian Legal Parallel Corpus in NooJ / V. Varanovich [et al.] // The 17th NooJ International Conference 2023 : book of abstracts / University of Zadar ; ed. L. Mijić [et al.]. Zadar : University of Zadar, 2023. P. 53–54.
5. Распрацоўка і развіццё лінгвістычнай базы ведаў юрыдычнай тэматыкі для сістэм машыннага перакладу і сінтэзу вуснага маўлення / Ю. С. Гецэвіч [і інш.] // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2021): доклады на XX Международной научно-технической конференции / отв. ред. А. В. Тузиков. Минск : ОИПИ НАН Беларуси, 2021. С. 184–188.
6. Одобрен перевод Избирательного кодекса на белорусский язык. 11 ноября 2019 г. [Электронный ресурс]. URL: https://ncpi.gov.by/about/history/2019/odobren-perevod-izbiratelnogo-kodeksa-na-beloruskiy-yazyk/?sphrase_id=311554 (дата обращения: 28.03.2024).
7. Creation of a legal domain corpus for the Belarusian module in NooJ: texts, dictionaries, grammars / V. Varanovich [et al.] // Communications in Computer and Information Science. 2022. Vol. 1758. Springer, Cham. P. 151–162. https://doi.org/10.1007/978-3-031-23317-3_18.