УДК 338.2

Рослякова Н.А., н.с., к.э.н., Институт проблем региональной экономики РАН, roslyakovaNA@gmail.com
Каневский Е.А., вед.н.с., ст.н.с., к.т.н., Институт проблем региональной экономики РАН, eak300@mail.ru
Боярский К.К., доцент, ст. преп., к. ф.-м.н., Университет ИТМО, boyarin9@yandex.ru

Roslyakova N.A. research fellow, Ph.D. in Economics, Institute for Regional Economic Studies RAS
Kanevsky E.A., lead res., senior res., Ph.D in Technical Sciences, Institute for Regional Economic Studies RAS
Boyarsky K.K., associate professor, Ph.D. in Physics, ITMO University

Исследование документов стратегического планирования анализаторами текста (на примере Стратегии социально-экономического развития Мурманской области)

Research of strategic planning documents by text analyzers (on the example of the Strategy of socio-economic development of the Murmansk region)

Ключевые слова: документы, стратегическое планирование, инструменты анализа, синтаксис предложения, семантика, омонимия.
Keywords: documents, strategic planning, analysis tools, sentence syntax, semantics, homonymy.

Аннотация. В работе ставится вопрос автоматического анализа текстов документов стратегического планирования. Причём анализатор текста должен позволять осуществлять содержательный анализ, а не ограничиваться сопоставлением значений плановых и фактических индикаторов документа. Поэтому использование контент-анализатора и парсера семантико-синтаксического типа позволяет как глубоко анализировать текст документа, так и обеспечивать различные формы представления результатов, что решает поставленные задачи.
Abstract. The paper raises the issue of automatic analysis of texts of strategic planning documents. Moreover, the text analyzer should allow for meaningful analysis, and not be limited to comparing the values of planned and actual indicators of the document. Therefore, the use of a content analyzer and a semantic-syntactic parser allows both deep analysis of the document text and providing various forms of presentation of results, which solves the tasks.

Анализ многочисленных документов стратегического планирования имеет большое значение как для адекватного и качественного исследовательского процесса, направленного на полное понимание и учёт аспектов, которые находятся в поле внимание региональных и федеральных властей, так и для выработки действенных и реальных (а не декларативных) рекомендаций по совершенствованию системы государственного стратегического планирования и системы регулирования и стимулирования социально-экономического развития регионов России.

В этом отношении большое значение имеет глубокий содержательный анализ документов стратегического планирования. Существует многочисленные статьи, посвященные этим вопросам. Так, например, на сайте Киберленинка [1], где агрегируются статьи из российских журналов, входящих в перечень ВАК, по запросу «анализ документов стратегического планирования» выдаётся свыше 11 тыс. статей. И это только статьи, вышедшие уже после 2020 года. Преимущественно эти статьи ориентированы на углублённый анализ документов того или иного региона или муниципального образования с ориентации на систему мер и индикаторов. Наиболее разработанным направлением является сравнительный (компаративный) анализ документов. Примерами таких работ являются [2, 3]. Однако следует подчеркнуть, что не существует статьей, где бы ставился вопрос углублённого анализа текстов машинными методами, с целью автоматизировать процесс анализа и постановки задач на поиск определённых смысловых конструкций в текстах стратегических документов. Примером работы, где ставятся вопросы анализа определённых содержательных конструкций, является [4]. Однако авторы оставляют в стороне вопрос разработки автоматизированных методов.

Таким образом, можно заключить, что содержательный анализ документов стратегического планирования представляет интерес и факторами, которые стимулируют поиск подходящего инструментария являются, во-первых, многообразие таких документов, во-вторых, наличие процесса их редактирования, доработки и переработки, в-третьих, развитие инструментария текстового и содержательного анализа, что позволяет ставить новые задачи для исследования текстов социально-экономического направления.

В рамках данной работы было сформулировано несколько запросов:

1. оценка частоты упоминания слов и степени их связанности в тексте;
2. отбор слов и более глубокий анализ контекста и окраски (положительной или отрицательной) при их упоминании;
3. фиксация некоторых слов, словосочетаний или смысловых конструкций и отслеживание как меняется контекст к ним с течением времени.

Для анализа была отобрана Стратегия социально-экономического развития Мурманской области до 2020 года и на период до 2025 года [5]. В общем смысле такие задачи были направлены на поиск в тексте указаний на определённые понятия:

1) **ценности**, декларируемые в стратегическом документе,
2) **мероприятия**, направленные на развитие и укрепление этих ценностей,
3) **эффекты**, которые отразят действенность мероприятий.

Простейший анализ текста может быть выполнен с помощью ДИСКАНТа – ДИалоговой Системы Классификации и АНализа Текста, прообразом которой явилась система ВЕГА [6,7]. Система предназначена для обработки как русскоязычных текстов, так и текстов, набранных латинскими буквами. Она обеспечивает составление словарей, пермутационный вывод информации, ее классификацию и кластерный анализ.

В качестве примера выберем упомянутую выше Стратегию социально-экономического развития Мурманской области. Первое впечатление об исследуемом тексте можно получить при обзоре алфавитного словаря (табл.1). Здесь хорошо видна вся группа слов, связанная с тем или иным понятием (термином) в тексте (в данном случае со словом СОЗДАНИЕ[1]).

Таблица 1. Фрагмент алфавитного словаря

Словоформа	Встречаемость
создаваемые	1
создаваемыми	2
Создавать	2
Создавая	1
Создает	5
Создана	2
Создание	78
Созданием	1
Создании	2
Созданию	6
Создания	15
Созданного	1
Созданных	1
Созданы	4
Создать	2
Создаются	3
Создающих	1

Более интересным является частотный словарь, позволяющий выделить те термины, которым в данном тексте уделяется наибольшее внимание (табл. 2).

Таблица 2. Верхушка частотного словаря

Лемма	**Встречаемость**
РАЗВИТИЕ	331
ГОД	294
ОБЛАСТЬ	261
МУРМАНСКИЙ	238
ГОСУДАРСТВЕННЫЙ	175
РЕГИОН	173
НАСЕЛЕНИЕ	153
ПОВЫШЕНИЕ	139
ЧИСЛО	126

[1] Здесь и далее леммы выделяются прописными буквами

ОБЕСПЕЧЕНИЕ	119
УСЛОВИЕ	106
СОЗДАНИЕ	102
РЕГИОНАЛЬНЫЙ	94
ДЕЯТЕЛЬНОСТЬ	93
РЕАЛИЗАЦИЯ	93

С учетом частоты встречаемости были подобраны термины, соответствующие указанным выше понятиям (приведены в порядке уменьшения частоты встречаемости).

Для понятия **ЦЕННОСТИ**: ОБРАЗОВАНИЕ (в смысле знаний), ПРИОРИТЕТ, БЕЗОПАСНОСТЬ, КУЛЬТУРА, ДОСТУПНОСТЬ, ТРУД, ЖИЗНЬ, КУЛЬТУРНЫЙ, ИНТЕРЕС, СЕМЬЯ, ДОХОД, РАБОТА (занятость), ОБРАЗ, ОБЩЕСТВО, ЖИЛЬЕ, ОБУЧЕНИЕ, ЗАНЯТОСТЬ.

Для понятия **МЕРОПРИЯТИЯ**: ОБЕСПЕЧЕНИЕ, РЕШЕНИЕ, ПОДДЕРЖКА, СОЗДАНИЕ, РЕАЛИЗАЦИЯ, ИСПОЛЬЗОВАНИЕ, СОВЕРШЕНСТВОВАНИЕ, ВКЛЮЧАТЬ, ОБЕСПЕЧИВАТЬ, МОДЕРНИЗАЦИЯ, СНИЖЕНИЕ, ВНЕДРЕНИЕ, УВЕЛИЧЕНИЕ, ОСВОЕНИЕ, РАСЩИРЕНИЕ, СТРОИТЕЛЬСТВО, СОХРАНЕНИЕ, ПРИВЛЕЧЕНИЕ, СТИМУЛИРОВАНИЕ, СОДЕЙСТВИЕ, ПОМОЩЬ, СОТРУДНИЧЕСТВО, УЛУЧШЕНИЕ, ОКАЗАНИЕ, УКРЕПЛЕНИЕ, УСИЛЕНИЕ, СОЗДАВАТЬ, СФОРМИРОВАТЬ, ОРИЕНТИРОВАТЬ, СОЗДАТЬ.

Для понятия **ЭФФЕКТ**: ПОВЫШЕНИЕ, СОЗДАНИЕ, РЕАЛИЗАЦИЯ, ОБРАЗОВАНИЕ (в смысле появление), РЕЗУЛЬТАТ, МОДЕРНИЗАЦИЯ, СНИЖЕНИЕ, ВНЕДРЕНИЕ, УВЕЛИЧЕНИЕ, ОСВОЕНИЕ, РАСШИРЕНИЕ, СТРОИТЕЛЬСТВО, СОХРАНЕНИЕ, ПРИВЛЕЧЕНИЕ, УВЕЛИЧИТЬСЯ, УЛУЧШЕНИЕ, ОПЕРЕЖАТЬ, ДОСТИЖЕНИЕ, УКРЕПЛЕНИЕ, ПРЕИМУЩЕСТВО, УСИЛЕНИЕ, ДИВЕРСИФИКАЦИЯ, ЗАНЯТОСТЬ, ОБЕСПЕЧЕННОСТЬ, ОТКРЫВАТЬСЯ, ПОСЛЕДСТВИЕ, СНИЗИТЬСЯ.

Некоторые выводы обо всей информации (не только текстовой) можно сделать, используя режим вывода по условию. Широкие возможности задания на поиск (по сути, на отбор) информации позволяют контролировать наличие в ней различных признаков (иногда весьма специфических). Так, например, можно отобрать абзацы, в которых содержатся начала слов «*численность*» и «*население*»[2]:

Начиная с 2000 года, ***численность*** *занятого* ***населения*** *Мурманской области имеет незначительнее колебания…*

С 2003 года также наметилось снижение доли ***численности населения*** *с денежными доходами ниже величины прожиточного минимума…*

[2] Последняя буква при задании в нашем случае должна быть опущена, так как здесь сравнение производится побуквенно начиная с начала слова (без использования морфологии).

*Основные показатели социально-экономического развития – **численность** занятых, реальные денежные доходы **населения**, темп роста ВРП будут иметь более позитивную динамику.*

Как видно, результаты поиска не всегда однозначны: если в двух первых предложениях речь идет о численности населения, то в третьем предложении это не так. Поэтому более удобный инструмент для анализа текста – пермутационный вывод текста по словарю, при котором выводится определенная окрестность целевого слова (18 символов слева и 39 символов справа на рис. 1). Такой способ вывода применяется как при поиске отдельных слов, так и при глобальном поиске по всему тексту.

```
коплению знаний и СОЗДАНИЮ современных научных и геоинформационых
к на юге региона, СОЗДАНИЕ порта Екатерининская гавань (сегодняшн
го порта, а также СОЗДАНИЕ Кольской военно-морской базы. Таким об
феры и перехода к СОЗДАНИЮ сферы социальных услуг как нового сект
огий, в том числе СОЗДАНИЮ в Мурманске международного центра тран
    Необходимость СОЗДАНИЯ прочной международно-правовой договорн
бежно потребуется СОЗДАНИЕ крупных транспортно-логистических узло
перспективы имеет СОЗДАНИЕ комплекса услуг круизного туризма по т
го пояса планеты. СОЗДАНИЕ такого инфраструктурного узла станет ж
вляет возможности СОЗДАНИЯ перспективного регионального кластера
аются перспективы СОЗДАНИЯ международных и трансграничных кластер
     Особенность СОЗДАНИЯ технологического кластера обеспечения
ра предполагается СОЗДАНИЕ производства диоксида титана, редких м
ера заключаются в СОЗДАНИИ на основе Мурманского транспортного уз
```

Рис. 1. Фрагмент пермутационного вывода по термину СОЗДАНИЕ

Такой вид вывода информации в принципе позволяет однозначно определять слова, расположенные контактно слева и справа от исследуемого нами целевого слова. Однако если перед нами стоит задача: установить, что именно планируется создать, т.е. определить слово в родительном падеже, стоящее правее термина (СОЗДАНИЕ), то далеко не всегда это можно определить по такому виду вывода информации. Иногда справа от целевого слова расположен предложный оборот времени или места действия (такой предложный оборот «прикрывает» искомый нами объект»). В частности, из рис. 1 видно, что только в четырех случаях из тринадцати зависимое слово стоит контактно, а в трех предложениях искомого объекта вообще не видно (фактически от исследуемого термина его отделяет более 4-х слов). Например, в последнем предложении «*…заключаются в **создании** на основе Мурманского транспортного узла сервисного **ядра** по обеспечению мореплавания по трассам Северного морского пути…*» разрыв составляет 6 слов. В этом случае ситуация осложняется еще тем, что справа от целевого слова *создании* имеется два слова в родительном падеже: *узла* и *ядра*, и если для человека понятно, что именно создается, то при автоматическом анализе текста возникают серьезные трудности.

Аналогичная ситуация может возникнуть и при обработке глагола СОЗДАТЬ; в этом случае искомым объектом будет являться прямое дополнение (винительный падеж). Примером является следующее предложение, в котором прямое дополнение отделено от глагола четырехсловным предложным оборотом и двумя прилагательными:

Эти и другие факторы предоставляют возможность ***создать*** *в течение ближайшего десятилетия международный транзитный* ***узел*** *в Мурманской области…*

Кроме изложенных выше методов анализа можно воспользоваться автоматической классификацией текста. Для любой автоматической классификации необходимо задать способы сравнения и установления степени близости. Эти способы не так однозначны, как кажется на первый взгляд. Даже человеку бывает нелегко решить: относить ли фрагменты из разных документов к одному и тому же классу или нет. Тем более компьютер со строгим алгоритмом работы может выдать нерелевантные результаты.

Особенностью системы ДИСКАНТ является возможность гибкой настройки механизма установления схожести высказываний. Прежде всего, следует определить, что, собственно, является единицей сравнения. Как правило, это слово или иная непрерывная последовательность знаков. Уже на этом этапе возможны различные подходы. Само понятие слова нечетко определено. Например, *Ростов-на-Дону* или *Кировский завод* – это неоднословные последовательности знаков, образующие единые лексические единицы. Как правило, на самом первом этапе анализа отбрасываются так называемые стоп-слова: предлоги, союзы и т.д. Здесь тоже не все однозначно: В предложениях: *Вижу я: на глазах у него будто слеза поблескивает* и *Ты на глазах у зрителя вершишь свой путь* в первом случае сочетание *на глазах у* это предлог-существительное-предлог, а во втором – сложный предлог, т. е. одна лексическая единица. Однако и предлоги удалять не всегда можно: *я за прививку* и *я против прививки* не одно и то же, несмотря на 100% совпадение знаменательной лексики. Анализ семантики предложных групп сейчас привлекает большое внимание [8, 9], однако вопрос о применимости этой семантики для целей классификации требует дополнительного исследования. Но с этой проблемой можно справиться с помощью относительно несложных правил, которые будут давать незначительный процент ошибок.

Как сравнивать слова? По словоформам, очевидно, малопродуктивно, поскольку русский язык относится к флективным синтетическим языкам со сложной морфологической изменяемостью слов. По леммам или по основе слова с помощью стемминга (отбрасывания «хвостов» слов по определенным алгоритмам)? Здесь уже нет однозначного ответа и исследователь должен решать, какой метод лучше подходит для данной задачи. Например, в документах встречаются слова *энергетикой* (отраслью) и *энергетиком* (работником этой отрасли). Лемматизатор определит их как различающиеся (*энергетика* и *энергетик*), а стеммер как совпадающие (основа у обоих *энергетик*). ДИСКАНТ позволяет использовать любой метод

по выбору исследователя. Лучшие результаты дает полный семантико-синтаксический анализ текста [10]. Такой анализ, проводимый с помощью парсера SemSin (подробнее см. ниже), позволяет в значительной степени снять омонимию, столь характерную для русского языка. Например, в предложениях *Необходимо развивать энергетику* и *Дать задание энергетику* в первом случае лемма определяется как ЭНЕРГЕТИКА (семантический класс «Отрасли»), а во втором как ЭНЕРГЕТИК (семантический класс «Профессия»).

Какие слова сравнивать? Как было показано в [11], для русского языка наилучшие результаты дает сравнение только существительных, в некоторых случаях существительных и прилагательных, а учет глаголов снижает экспертную оценку качества сравнения. ДИСКАНТ позволяет задавать частеречный состав сравниваемых слов и устанавливать весовые коэффициенты для различных частей речи. Кроме того, исследователь имеет возможность выделить некоторые слова как ключевые и приписать им повышенный вес при сравнении.

Особенностью системы ДИСКАНТ является возможность проводить сравнение по семантическим классам [12]. Важность этого способа сравнения иллюстрирует пример из [11]. Возьмем два предложения:

Ленточки бескозырок матросов кайзеровского флота имели надпись прописными печатными буквами, вышитыми золотой или серебряной канителью

и

Пилотка кроилась из темно-синего плотного сукна, обычно с черной или темно-синей подкладкой из искусственного шелка.

Эти предложения, несмотря на очевидную общность тематики, вообще не имеют одинаковых слов, поэтому при обработке «по словам» мера сходства равна нулю. Однако в соответствии с классификатором слова *бескозырка, пилотка, подкладка* относятся к классу одежды, а слова *ленточка, канитель, сукно, шелк* – к классу тканей. Соответственно, отбирая в качестве слов для сравнения существительные, при вычислении сходства «по классам» получаем меру сходства $\cos \varphi = 0{,}71$.

Конечно, любой классификатор в определенной степени субъективен и несет отпечаток предпочтений составителя. Для того, чтобы оптимизировать классификатор для целей конкретного исследования в ДИСКАНТе предусмотрена возможность работы с тремя степенями подробности классификатора: грубой, промежуточной и точной. Например, при использовании точного классификатора слова *православие* и *крещение* будут интерпретироваться как принадлежащие к разным классам, а в грубом классификаторе классы будут совпадающими.

Еще раз подчеркнем, что выбор способа сравнения, весовых коэффициентов и уровня отсечки определяется конкретной задачей, и выбирается исследователем.

Следует отметить также, что многие термины, входящие в приведенные выше понятия, имеют семантическую неоднозначность. Так, в частности, из отобранных терминов неоднозначными являются следующие[3]:

СОЗДАНИЕ Действительность Создание
СОЗДАНИЕ ФО Живой Человек Личность
СОЗДАНИЕ Действие Труд Дело

ОБРАЗОВАНИЕ Действительность Событие Начало
ОБРАЗОВАНИЕ ФО Неодуш. Материалы Породы
ОБРАЗОВАНИЕ Действие Занятие Воспитание Обучение

ДОСТИЖЕНИЕ ФО Живой Человек Успех-Неуспех
ДОСТИЖЕНИЕ Действие Труд Дело

УКРЕПЛЕНИЕ ФО Прочность
УКРЕПЛЕНИЕ Действие Борьба Нападение-Защита

ОТКРЫВАТЬСЯ Действительность Событие Начало
ОТКРЫВАТЬСЯ ФО Вид
ОТКРЫВАТЬСЯ Свойства Закрытость

КУЛЬТУРА ФО Природа Растения
КУЛЬТУРА ФО Живой Человек Цивилизация
КУЛЬТУРА Знания Науки Естествознание Биология

ТРУД Знания Литература Корресп. Статья
ТРУД Действие Труд

СЕМЬЯ ФО Неодуш. Множества
СЕМЬЯ ФО Живой Человек Личность Родня

РАБОТА Поселения Учреждения
РАБОТА Знания Сообщение Документы Запись
РАБОТА Действие Труд Работа

ОБРАЗ Способ
ОБРАЗ Образ
ОБРАЗ Знания Искусство Художество Живопись Картины

ВКЛЮЧАТЬ Действие Труд Работа
ВКЛЮЧАТЬ ФО Место Направление Внутри-Вне

ОБЕСПЕЧИВАТЬ Действие Труд Умственный Управление Исполнение
ОБЕСПЕЧИВАТЬ Действие Занятие Обладание Приобретение-Потеря

[3] Здесь приведены лексемы и их классы в упрощенном виде.

Очевидно, что для решения всех этих проблем необходимо привлечения методов синтаксического и семантического анализа. Воспользуемся семантико-синтаксическим анализатором SemSin, сочетающим в себе функции лемматизатора, синтаксического и семантического анализатора [13]. Парсер включает в свой состав словари, классификатор, блок морфологического анализа и набор продукционных правил.

Основной словарь, построенный на основе модифицированного словаря Тузова [12], содержит более 196 тысяч лексем. Для каждой лексемы указаны морфологические характеристики, а также номер (или номера) своего семантического класса и актанты или валентности (для подключения зависимых слов). Одной лексеме могут соответствовать несколько семантических омонимов (например, *коса* как волосы, *коса* как побережье и *коса* как утварь), которые относятся к разным классам.

Словарь фразем содержит около 5,5 тысяч айтемов, состоящих из двух или более слов. Это сложные предлоги или наречия, а также образные выражения или составные имена собственные. Словарь предлогов включает в себя более 2,5 тысяч айтемов. В каждом из них указаны сам предлог, падеж существительного требуемого предлогом, семантический класс этого существительного и тип полученной связи (например, "*Где*", "*Когда*" и т. д.). Классификатор содержит 1700 классов, образующих дерево, построенное по семантическому принципу. Одна лексема может относиться к нескольким классам.

Объем словаря обеспечивает распознавание около 96% слов при переходе к более современной партии новостных текстов. Примерно в половине случаев отсутствующие в словаре слова являются именами собственными, многие из которых система способна распознать автоматически.

Этот парсер анализирует текст по абзацам. Прежде всего, текст разбивается на токены и каждое слово обрабатывается морфологическим анализатором. Результат разбора выдается в виде одной или нескольких лемм с морфологическими характеристиками и классами (или набором классов) с указанием соответствующих актантов. После этого запускается предсинтаксический модуль, который делит абзац на предложения, уточняет написание и морфологические характеристики некоторых конструкций (слов с дефисами, составных и алфавитно-цифровых числительных), пытается решить проблему с неизвестными словами и осуществляет разбор фразем [14].

Затем подключается синтаксический модуль, использующий около 480 правил [15, 16]. В процессе анализа предложения одновременно выполняются снятие грамматической и частеречной омонимии, сегментация предложения и построение синтаксического дерева зависимостей. Во многих случаях разрешается и лексическая омонимия. Результат синтаксического анализа хранится в виде XML-файла, в котором для каждого слова представлены его лемма, часть речи, грамматические признаки (одушевленность, род, число, падеж, время и т.д.), основной номер клас-

са, идентификатор родительского узла и тип связи с ним, а также ссылки на слова, семантически тесно связанные с данным.

Полученное дерево содержит максимально полную информацию о предложении. Эта информация может в дальнейшем служить основой для решения самых разных задач: выявления терминов [17], классификации текстов [18] и т. д. В данной работе обсуждается вопросы углубленного анализа текста.

Важную роль в определении смысла высказывания играют именные группы, представляющие связку двух существительных, в которой подчиненное слово стоит в родительном падеже. При отсутствии такого слова-слуги смысл главного слова-хозяина остается недоопределенным. Одно дело *позиция делегации на переговорах*, другое – *позиция артиллерии*. Однако правильное подключение слов в родительном падеже при автоматическом анализе текста является непростой задачей. Ее сложность определяется, в основном, двумя факторами.

Во-первых, для русского языка характерны длинные цепочки из слов в родительном падеже. Может оказаться, что последовательность связей по родительному падежу такая же линейная, как и последовательность слов в предложении. Связываются контактно стоящие слова, как на рис. 2:

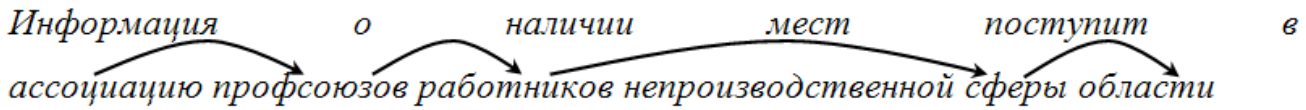

Рис. 2. Линейная связь слов в предложении

Достаточно часто по родительному падежу связываются удаленные слова, т.е. связь оказывается нелокальной [19]. И речь идет не о прилагательных, как в предыдущем примере, а о предложных группах и распространенных оборотах, например:

Открыта дорога импорту высоких технологий, в том числе ***созданию*** *в Мурманске международного* ***центра****…*

Интенсивное развитие регионального туристско-рекреационного комплекса в свою очередь будет способствовать ***созданию*** *в среднесрочной перспективе (на втором этапе реализации Стратегии)* ***кластера*** *северного дизайна…*

Особенно сложно правильно найти слово-слугу, если между ним и хозяином стоит предложная группа с существительным, которое само способно подсоединять слова в родительном падеже, как на рис. 3:

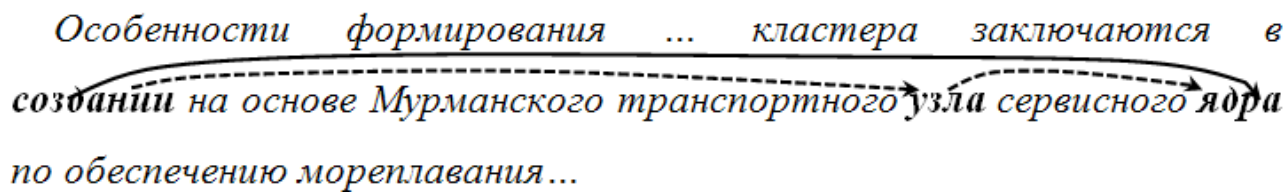

Рис. 3. Нелинейная связь слов в предложении

Формально вполне возможно построение связей, показанных пунктиром, хотя это очевидно неверно. Для правильного разбора таких конструкций парсером SemSin нами были разработаны специальные правила.

Описанные выше системы контентного и семантико-синтаксического анализа текста позволяют в значительной степени автоматизировать процесс поиска и классификации отобранных языковых конструкций. При этом встаёт вопрос представления результатов анализа для дальнейшей работы по сути документа стратегического планирования (или другого анализируемого текста). В этом отношении большой интерес представляют два вида выводов: дерево, отражающее взаимосвязи слов в предложении или абзаце, и облако слов, отражающее наиболее частотные слова и степень их связанности между собой.

На рис. 4 приведено дерево разбора предложения «*Особенности формирования производственного и транспортно-логистического кластера заключаются в* ***создании*** *на основе Мурманского транспортного узла сервисного* ***ядра*** *по обеспечению мореплавания по трассам Северного морского пути и привлечении средств на основе принципа государственно-частного партнерства*».

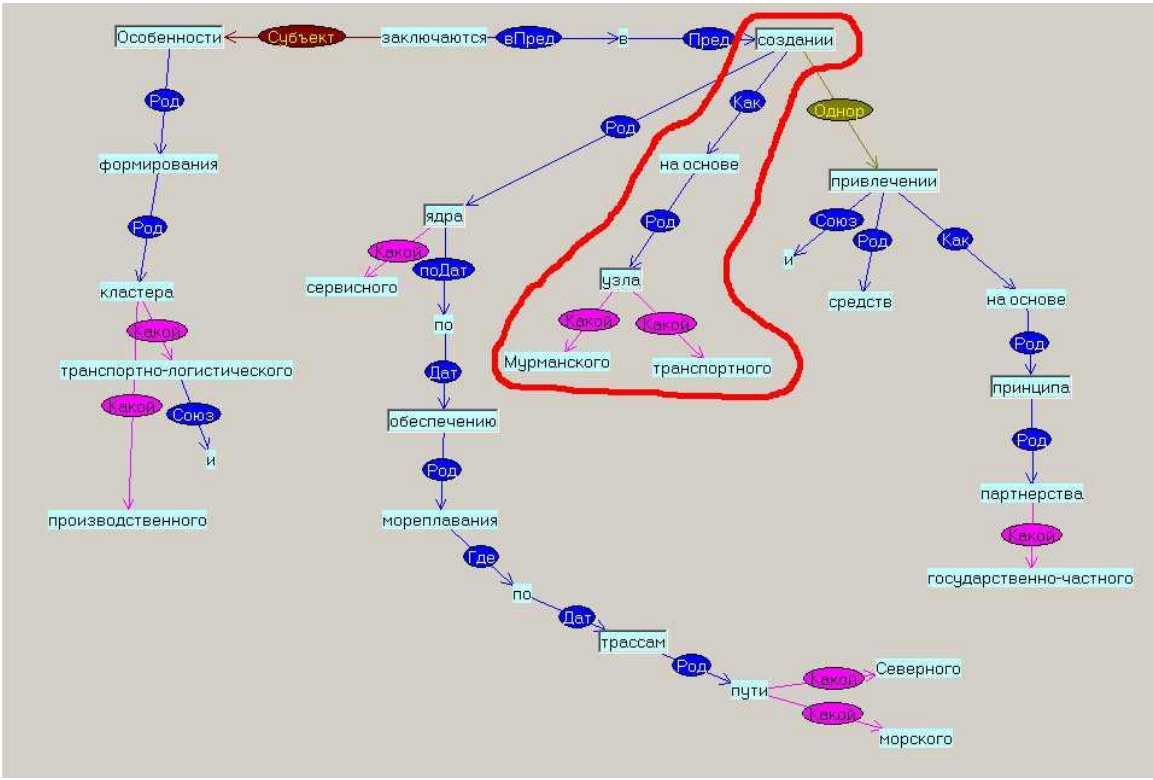

Рис. 4. Дерево разбора

Выделен блок (предложный оборот образа действия), характеризующий термин *СОЗДАНИЕ*. Используя графическое представление дерева, легко заметить, что в данном предложении речь идет не только о создании транспортного узла, но и о привлечении средств. Заметим, что в текстовом представлении этот блок маскирует предмет, который создается (транспортный узел).

Примеры облаков представлены ниже. Они были построены на основе пермутационных выводов, полученных на предыдущем этапе. Целью построения облаков слов на отдельных тематических фрагментах, а не на всём тексте целиком позволяет сконцентрироваться на исследование конкретных смыслов и более концентрированно увидеть связанные смыслы. Для построения использовался открытый ресурс Word in Out[4]. Здесь традиционно более интенсивным цветом отражается большая связанность категорий (слов) в облаке, размером – частота упоминания в исходном тексте. Взаимное расположение слов в облаке отражает тематические области и устойчивые смысловые конструкции (рис. 5 – 7).

Рис. 5. Облако слов «ценности»

[4] https://worditout.com/

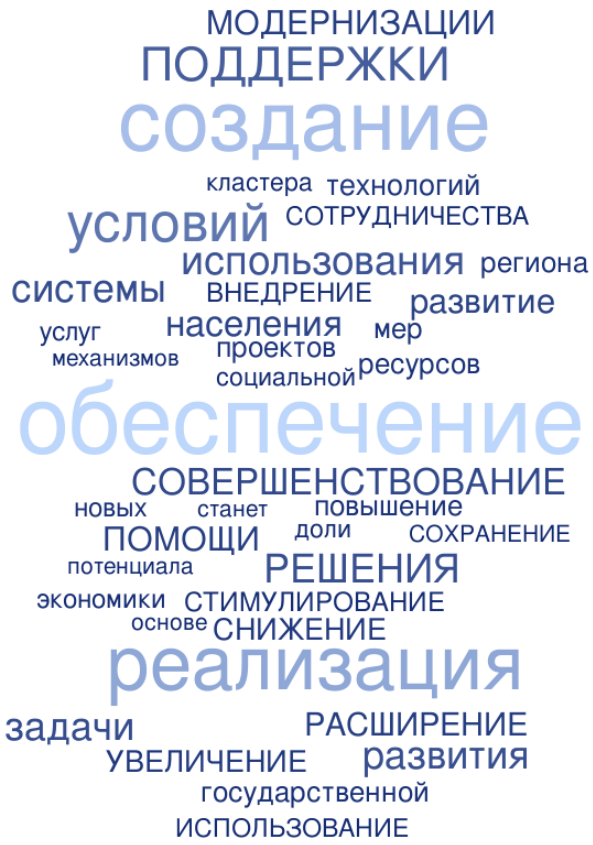

Рис. 6. Облако слов «мероприятия»

Рис. 7. Облако слов «эффекты»

Полученные результаты позволяют получить довольно наглядные конструкции, например, «безопасность семей», «поддержка занятости» в категории ценности,

«создание кластеров», «обеспечение ресурсов» в категории мероприятий, «создание технологий», «доступность услуг» в категории эффектов.

В целом можно заключить, что предложенные инструменты автоматизированного анализа текста позволяют достаточно быстро и качественно анализировать документы и представлять результаты в удобном для исследователя виде, что обеспечивает возможность дальнейшего анализа документов и разработки предложений по совершенствованию.

Литература

1. Научная электронная библиотека «КиберЛенинка», 2022 // «КиберЛенинка». URL: https://cyberleninka.ru/ (дата обращения 12.09.2022).

2. Лапенкова Н.В., Побываев С.А., Золотарев Е.В. Апробация разработанной методики компаративного анализа на примере зарубежных и отечественных документов стратегического планирования в области энергетической безопасности // Вопросы безопасности, 2021. № 4. С. 11-29.

3. Сургуладзе В.Ш. Идеологическое измерение стратегии национальной безопасности Российской Федерации: Сравнительный анализ документов 2015 и 2021 годов // Гуманитарные науки. Вестник Финансового университета, 2022. № 12 (1). С. 60-69.

4. Рубцов Г.Г, Литвиненко А.Н. Использование ценностно-ориентированного подхода в стратегическом планировании на примере реализации региональных стратегий развития субъектов Северо-Западного федерального округа // Вопросы управления, 2020. № 3 (64). С. 65-77.

5. Постановление Правительства Мурманской области от 25.12.2013 № 768-пп/20 «О стратегии социально-экономического развития Мурманской области до 2020 года и на период до 2025 года» (по состоянию на 10.07.2017).

6. Боярский К., Каневский Е. Вега – система классификации и анализа текста. – Deutschland, Saarbrűcken: LAP Lambert Academic Publishing GmbH & Co. KG, 2011. 148 с.

7. Каневский Е.А., Боярский К.К. ВЕГА – инструмент для лингвистических исследований // Прикладна лінгвістика та лінгвістичні технології:.MegaLing-2012. К.: Довіра, 2013. С. 113–123.

8. Azarova, I., Zakharov, V., Khokhlova, M., Petkevič, V.: Ontological description of Russian prepositions. In: CEUR Workshop Proceedings, 2552, pp. 245-257 (2020).

9. V. Zakharov, K. Boyarsky, A. Golovina, A. Kozlova.: Semantic Analysis of Russian Prepositional Constructions. RASLAN 2020. Recent Advances in Slavonic Natural Language Processing. Proceedings. Brno. 2020. P. 103-113.

10. K. Boyarsky, E. Kanevsky. Effect of Semantic Parsing Depth on the Identification of Paraphrases in Russian Texts. Filchenkov A., Pivovarova L., Žižka J. (eds) Artificial Intelligence and Natural Language. AINL 2017. Communications in Computer and Information Science, vol 789. Springer, Cham. 2018. P. 226-241.

11. Боярский К.К., Авдеева Н.А., Гусарова Н.Ф., Добренко Н.В. Каневский Е.А. Исследование специфики применения алгоритмов тематической сегментации для научных текстов. Аналитика и управление данными в областях с интенсивным использованием данных. XVII Международная конференция DAMDID/RCDL’2015. Обнинск, 13-16 октября 2015 г. М.: НИЯУ МИФИ. 2015. С. 181-189.

12. Тузов В.А. Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004. 400 с.

13. Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SEMSIN. // Научно-технический вестник информационных технологий, механики и оптики. 2015, т. 15, №5. С. 869–876

14. Боярский К.К., Каневский Е.А. Предсинтаксический модуль в анализаторе SemSin // Материалы XVI Всероссийской объединенной конференции "Интернет и современное общество". СПб. – СПб.: «Университетские Телекоммуникации», 2013. С. 280–286.

15. Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб.: ООО «МультиПроджектСистемСервис», 2011. С. 233–237.

16. Боярский К.К., Каневский Е.А. Система продукционных правил для построения синтаксического дерева предложения. Прикладна лінгвістика та лінгвістичні технології: MegaLing-2011. К.:Довіра, 2012. С. 73-80.

17. Боярский К.К., Арчакова Н.А., Каневский Е.А. Извлечение низкочастотных терминов из специализированных текстов. // Аналитика и управление данными в областях с интенсивным использованием данных. XVIII Международная конференция DAMDID/RCDL’2016. Ершово, 11-14 октября 2016 г. М: Торус Пресс. 2016. С. 211–216.

18. Боярский К.К., Арчакова Н.А., Каневский Е.А. Особенности кластеризации специальных текстов. // Интернет и современное общество. Сборник тезисов докладов. Труды XIX объединенной научной конференции «Интернет и современное общество», Санкт-Петербург, 22–22 июня 2016 г. – СПб: Университет ИТМО. 2016, С. 9–11.

19. Боярский К.К., Каневский Е.А. Нелокальные семантические связи в русскоязычных текстах» // Научно-технический вестник информационных технологий, механики и оптики, 2018. т 18, №5. С. 863–869.