**Результаты исследований**

**Формирование единого словаря строительной отрасли**

**Подготовительная работа**

Для проведения исследований по теме НИР был сформирован корпус, состоящий из 35 основных НПА. Перечень документов, вошедших в корпус, используемый в настоящей НИР приведен в таблице \_\_

Таблица \_\_\_. Корпус документов используемый в настоящей НИР

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | ID документа | Тип документа | Наименование НПА | Источник |
| 1 | K940001000\_ | Кодекс | Гражданский Кодекс РК (Общая часть) | <http://adilet.zan.kz/rus/docs/K940001000_> |
| 2 | K090000193\_ | Кодекс | О здоровье народа и системе здравоохранения | <http://adilet.zan.kz/rus/docs/K090000193_> |
| 3 | Z010000242\_ | Закон | Об архитектурной, градостроительной и строительной деятельности в Республике Казахстан | <http://adilet.zan.kz/rus/docs/Z010000242_> |
| 4 | Z040000603\_ | Закон | О техническом регулировании | <http://adilet.zan.kz/rus/docs/Z040000603_> |
| 5 | Z000000053\_ | Закон | Об обеспечении единства измерений | <http://adilet.zan.kz/rus/docs/Z000000053_> |
| 6 | Z1600000480 | Закон | О правовых актах | <http://adilet.zan.kz/rus/docs/Z1600000480> |
| 7 | K030000442\_ | Закон | Земельный кодекс РК | <http://adilet.zan.kz/rus/docs/K030000442_> |
| 8 | K070000212\_ | Закон | Экологический кодекс РК | <http://adilet.zan.kz/rus/docs/K070000212_> |
| 9 | K1700000125 | Кодекс | О недрах и недропользовании | <http://adilet.zan.kz/rus/docs/K1700000125> |
| 10 | Z020000332\_ | Закон | О геодезии и картографии | <http://adilet.zan.kz/rus/docs/Z020000332_> |
| 11 | Z1200000541 | Закон | Об энергосбережении и повышении энергоэффективности | <http://adilet.zan.kz/rus/docs/Z1200000541> |
| 12 | Z060000175\_ | Закон | Об особо охраняемых природных территориях | <http://adilet.zan.kz/rus/docs/Z060000175_> |
| 13 | Z040000567\_ | Закон | О связи | <http://adilet.zan.kz/rus/docs/Z040000567_> |
| 14 | Z1400000194 | Закон | О дорожном движении | <http://adilet.zan.kz/rus/docs/Z1400000194> |
| 15 | Z030000476\_ | Закон | Об автомобильном транспорте | <http://adilet.zan.kz/rus/docs/Z030000476_> |
| 16 | Z940007000\_ | Закон | О транспорте в Республике Казахстан | <http://adilet.zan.kz/rus/docs/Z940007000_> |
| 17 | Z010000245\_ | Закон | Об автомобильных дорогах | <http://adilet.zan.kz/rus/docs/Z010000245_> |
| 18 | Z1200000020 | Закон | О магистральном трубопроводе | <http://adilet.zan.kz/rus/docs/Z1200000020> |
| 19 | K030000481\_ | Закон | Водный кодекс РК | <http://adilet.zan.kz/rus/docs/K030000481_> |
| 20 | Z1200000532 | Закон | О газе и газоснабжении | <http://adilet.zan.kz/rus/docs/Z1200000532> |
| 21 | Z040000588\_ | Закон | Об электроэнергетике | <http://adilet.zan.kz/rus/docs/Z040000588_> |
| 22 | Z090000165\_ | Закон | О поддержке использования возобновляемых источников энергии | <http://adilet.zan.kz/rus/docs/Z090000165_> |
| 23 | Z1400000202 | Закон | О разрешениях и уведомлениях | <http://adilet.zan.kz/rus/docs/Z1400000202> |
| 24 | Z1600000486 | Закон | О долевом участии в жилищном строительстве | <http://adilet.zan.kz/rus/docs/Z1600000486> |
| 25 | K1400000235 | Кодекс | Об административных правонарушениях | <http://adilet.zan.kz/rus/docs/K1400000235> |
| 26 | K990000409\_ | Кодекс | Гражданский Кодекс РК (Особенная часть ) | <http://adilet.zan.kz/rus/docs/K990000409_> |
| 27 | Z1500000434 | Закон | О государственных закупках | <http://adilet.zan.kz/rus/docs/Z1500000434> |
| 28 | K1500000375 | Кодекс | Предпринимательский кодекс РК | <http://adilet.zan.kz/rus/docs/K1500000375> |
| 29 | K1500000414 | Кодекс | Трудовой кодекс РК | <http://adilet.zan.kz/rus/docs/K1500000414> |
| 30 | Z1100000469 | Закон | О специальных экономических зонах в Республике Казахстан | <http://adilet.zan.kz/rus/docs/Z1100000469> |
| 31 | Z960000006\_ | Закон | Об авторском праве и смежных правах | <http://adilet.zan.kz/rus/docs/Z960000006_> |
| 32 | Z920002900\_ | Закон | Об охране и использовании объектов историко-культурного наследия | <http://adilet.zan.kz/rus/docs/Z920002900_> |
| 33 | Z1400000188 | Закон | О гражданской защите | <http://adilet.zan.kz/rus/docs/Z1400000188> |
| 34 | Z1800000183 | Закон | О стандартизации | <http://adilet.zan.kz/rus/docs/Z1800000183> |
| 35 | P1000001202 | Технический регламент | Требования к безопасности зданий и сооружений, строительных материалов и изделий | <http://adilet.zan.kz/rus/docs/P1000001202> |

**Анализ списка терминов и их определений**

Результат формирования списка терминов и их определений приведён в Приложении \_\_\_ (**Список терминов и их определений**). Термины и определения из корпуса документов. В ходе формирования списка было выявлено следующее:

1. Не всегда в документах применяется буква «Ё»;
2. Порядок слов в терминах из 2-х и более слов может различаться. (Например: дом жилой – жилой дом);
3. Написание термина может быть дано неправильно. (Например: лифтовой холл – должно быть лифтовый холл);
4. Написание термина может быть приведено во множественном числе;
5. Термин приводится в полной и сокращенной форме, используемой далее по тексту документа. Например: “закладка материальных ценностей в государственный материальный резерв (далее – закладка)”;
6. В скобках приводится синоним термина. Например: “биоаналогичный лекарственный препарат (биоаналог, биоподобный лекарственный препарат, биосимиляр)”;
7. В среднем, один документ содержит определения 53-х терминов. На обработку одного термина уходит 5 минут.

Результаты проверки на дублирование терминов и определений приведены в Приложении \_\_ (**Дублирование терминов и определений**). Результаты проверки на коллизии в определениях терминов приведены в Приложении \_\_\_ (**Коллизии в определениях терминов**).

**Частотный анализ**

Список терминов с указанием количество упоминаний термина в тексте документа, выделенных в результате частотного анализа, приведён в приложении \_\_\_ (**Термины выделенные с помощью частотного анализа**). Встречаемость терминов в документах приведена в приложении \_\_ (**Встречаемость терминов**)

В среднем один документ содержит 307 терминов.

Далее был выполнен анализ покрытия терминов, выделенных ручным способом терминами, выделенными автоматизированным способом с помощью частотного анализа. Результат анализа приведён в приложении \_\_ (**Анализ покрытия**). Анализ показал, что **в документах присутствуют термины и их определения, которые в дальнейшем редко используются в документе** (Например: почвенный слой, временный землепользователь) или вообще не используются (Например: безопасность плотины, акт работодателя). Соответственно, такие слова и словосочетания не вошли в итоговый перечень терминов, выделенных с помощью частотного анализа.

**Анализ НТД**

Дополнительно к анализу НПА был проведён экспертный анализ 2-х документов НТД: В качестве объекта анализа были определены следующие документы:

1. СП РК 3.02-101-2012 ЗДАНИЯ ЖИЛЫЕ МНОГОКВАРТИРНЫЕ
2. СН РК 3.02–01–2018 ЗДАНИЯ ЖИЛЫЕ МНОГОКВАРТИРНЫЕ

Аналогично анализу НПА был проведён частотный анализ документов НТД. Результаты анализа приведены в приложениях \_\_\_\_ (Приложение. Частотный анализ НТД).

Дополнительно между выделенными терминами были установлены отношения SubClass\_Of (является подклассом). Основанием для установки отношений было существительное в именительном падеже, входящее в термин. Результат установки отношений оформлен в виде онтологии в формате OWL (Приложение. Пример онтологии SubClass\_Of.owl). В приложении \_\_\_ (**Снимок экрана Protege c примером онтологии**) приведен снимок экрана программного продукта Protégé 5.2.0 (http://protege.stanford.edu) с сформированной онтологией.

Полученный перечень терминов был использован в качестве основы для экспертного анализа корпуса документов и построения семантической сети ручным способом. Экспертный анализ проводился следующим образом:

* детальная вычитка документа;
* поиск терминов в документе на основе данных частотного анализа;
* определение связанных терминов и характера связи на основе экспертного анализа контекста предложений, в которых встречается искомый термин;
* отображение терминов и связей между ними в виде семантической сети.

Экспертный анализ НТД Здания жилые многоквартирные показал что:

* отсутствуют единые понятия - возникает необходимость в единой терминологии в НТД/НПА. В анализируемых документах встречаются множество слов, имеющих одно и то же либо близкое семантическое значение (например, жилой дом - жилое здание; квартира - помещение; инженерное обеспечение - инженерные коммуникации, инженерные системы, инженерное оборудование; возводятся - строятся и т.п.) либо один “объект” относится к множеству групп (например: балкон, лоджия относятся к нежилому помещению, также они относятся к летнему помещению);
* по тексту документов встречаются слова в скобках, сопутствующие к словам стоящим перед ними, которых понять можно по-разному (например, поверхностью (поверхностями) - в данном случае однозначно множественное число; несущих (капитальных) стен - вид стен, жилое помещение (квартира) - синонимы или как; жилой дом (жилое здание) - синонимы или как; целей (магазин, кафе, мастерская и ..) - уточнение и конкретизация; и т.п.);
* выявлено одинаковое определение к двум разным терминам в одном документе (подземный этаж, подвальный этаж)
* отсутствует единая структура и содержание документов - вышеперечисленные документы описывают требования к проектированию и строительству зданий жилых многоквартирных, из которых один устанавливает обязательные требования безопасности к зданиям жилым многоквартирным и (или) процессам их жизненного цикла в соответствии с требованиями технических регламентов, в другом - основные расчеты при проектировании зданий жилых многоквартирных. следовательно, в первую очередь, необходимо привести в порядок “Термины и определения” - данный структурный элемент документов содержать один и тот же перечень терминов с соответствующими определениями. “Нормативные ссылки” в нормативно-технических документов (т.е. СП РК) должны дополняться также и “Нормативными ссылками” нормативных правовых актов (т.е. СН РК). Вместе с тем, сама структура документов (главы/разделы/параграфы) должны соответствовать друг к другу  в целях получения полной и структурированной информации по проектированию и строительству зданий жилых многоквартирных.

Построенная ручным способом онтология (семантическая сеть?) приведена в приложении \_\_ (Приложение. Семантическая сеть НТД Здания жилые многоквартирные). Полный экспертный анализ НТД Здания жилые многоквартирные приведён в приложении \_\_\_ (Приложение. Отчет по анализу НТД Здания жилые многоквартирные).

Проведённый анализ также показал, что **в основном нормативные документы не содержат информации, позволяющей определить семантические отношения между терминами и необходимой для построения онтологий**. В основном **они содержат регуляторные требования**, задающие правила, которым должны соответствовать те или иные понятия, обозначенные в документах соответствующими терминами.

Для определения слов (глаголов), задающих уровень обязательности требований, был дополнительно проведен частотный анализ глаголов в корпусе документов. Список глаголов и частота их употребления приведены в приложении \_\_ (Приложение. Частотный анализ НПА глаголов). Анализ используемых глаголов показал что:

1. Для задания уровня обязательности требования используются разные глаголы, имеющие похожее семантическое значение. Например (допускается/рекомендуется/могут);
2. Практически отсутствуют глаголы, явно запрещающие то или иное действие проектировщика или устанавливаемую характеристику понятия;
3. Частица НЕ применяется к глаголам, что приводит к семантической неоднозначности (Например: рекомендуется – не рекомендуется).

В СТ РК 1.5-2013 “Общие требования к построению, изложению, оформлению и содержанию стандартов” в п. 5.1.5 задаются требования к использованию слов для задания обязательных нормативных требований и положений. Необходимо отметить, что приведённые в стандарте требования допускают очень широкую их трактовку.

В качестве примера стандарта, задающего очень строгие и не допускающие двойной трактовки требования к употреблению глаголов для задания уровня обязательности требований можно привести документ RFC 2119 (Request for Comments — документы из серии пронумерованных информационных документов Интернета, содержащих технические спецификации и стандарты, широко применяемые во всемирной сети). RFC 2119 “Ключевые слова для обозначения уровня требований в RFC” содержит только 5 (пять) глаголов и их синонимов, которые должны использоваться при описании стандартов и технических спецификаций применяемых в сети Интернет. Помимо указания самих глаголов, в отличие от СТ РК 1.5-2013, в документе также определяется уровень требований, задаваемых этими глаголами.

**Исследования контекста использования терминов**

Для корпуса документов НПА были построены модели Word2Vec. Для исследования выбраны следующие термины: Безопасность, Подрядчик, Заказчик, Градостроительный кадастр, Надзор, Кадастр, Жилой дом.

Списки семантически близких терминов и слов для отобранных терминов приведены в приложениях \_\_ (Приложения. Анализ Word2Vec.zip).

Результаты исследования показывают, что для отдельного документа присутствует свой практически уникальный контекст употребления того или иного термина. Это обусловлено тем, что НПА по своей природе является документом, содержащим обобщенное описание и нормы для одной конкретной предметной области. В практической плоскости можно зафиксировать контекст применения термина в отдельном НПА (или группе НПА) как образец применения термина в определенной предметной области и в последующем, при анализе других документов, имеющих отношение к предметной области, сравнивать контексты применения терминов.

**Методология формирования единого словаря строительной отрасли**