**Corpus driven research of semantic links in structure of hypertext (evidence from verb links)[[1]](#footnote-1)**

**Shulginov Valery A. (****shulginov.val@yandex.ru****)**

**Far Eastern Federal University, Vladivostok, Russia**

**Shulginov Vadim A. (****vadim-shulginov@yandex.ru****)**

**Rostelecom, Vladivostok, Russia**

**Abstract.** The paper presents electronic hypertext as a communicative-cognitive phenomenon, which has all the signs of textuality, and is also characterized by a complex structure and non-linear connection between chunks. We propose a corpus study of the semantics of hypertext transitions, based on the analysis of semantic proximity pairs link/target text, sentence/target text, source text/target text. The development of the corpus of hypertext fragments includes the following steps: development of a web scraping tool that extract required data from the list of sites, inventory and normalization of language units, linguistic and extra-linguistic markup of data. As a result, we have a database with the ability to search by the frequency of nomination links, POS-characteristic and semantic proximity. The paper analyzes the functioning of verb links as connectors by interdisciplinary and target text. We study lexical-semantic classes of verbs based on quantitative indicators of the semantics of the hypertext transition. This made it possible to identify typical strategies for arrange a hypertext. Key words: hypertext, semantic proximity, presupposition, cognitive linguistic, corpus, connector

Key words: hypertext, semantic proximity, presupposition, cognitive linguistic, corpus, connector

**Введение**

Сеть интернет представляет собой особую электронную среду, которая характеризуется нелинейностью, мультимедийностью, открытостью, бесконечностью и обеспечивает хранение и обмен информацией между пользователями. Весь массив данных в интернете, образует «вселенную электронных документов» (Ryazantseva, 2010), состоящую из узлов и разнонаправленных связей между ними. Линейный текст, погруженный в электронную среду, перенимает специфику ее функционирования: имплицитно заложенная в нем интертекстуальность материализуется с помощью гипертекстовых ссылок, которые обеспечивают возможность гипертекстового перехода между фрагментами. Под влиянием интернет-пространства текст трансформируется в коммуникативно-познавательную единицу нового типа, которая, с одной стороны, отвечает всем критериям текстуальности (целостность, связность, намеренность, приемлемость, информативность, ситуативность, интертекстуальность), с другой, характеризуется сложной структурой и нелинейными связями между фрагментами (Shulginov, 2016).

В электронной среде изменяются позиции адресанта и адресата: по мнению ряда исследователей (Derrida, Nelson, Shechtman и др.), в нем воплощается постструктуралистская концепция «смерти автора», согласно которой источник текста находится не в письме, а в чтении (Bart, 1994). Адресат перестает быть пассивным получателем информации, он конструирует сообщение таким образом, как это удобно ему, выбирая путь прочтения текста.

Однако подобная свобода читателя в восприятии и интерпретации электронного гипертекста оказывается весьма условной, так как именно автор определяет источник перехода и текст, который активируется с помощью ссылки. Основная стратегия автора (при условии, что его речевая деятельность направлена на кооперацию) заключается в оформлении тех гипертекстовых переходов, которые обеспечивают успешность коммуникативного акта. Ссылка становится способом прояснения референта высказывания, необходимого для адекватного восприятия текста читателем, что особенно актуально при толковании терминологической лексики в электронных энциклопедиях. Как отмечает Н.С Клочкова, «расширение тезауруса в целях установления верного референта высказывания или обеспечения правильного понимания слова … гарантирует полное смысловое восприятие текста» (Klochkova, 2015).

Таким образом, ключевым критерием описания электронного гипертекста становится наличие семантики гипертекстового перехода (Dedova, 2009), которая выражается в семантическом взаимодействии источника ссылки и его объекта.

Цель исследования заключается в описании формальных семантических связей в структуре электронного гипертекста с учетом корреляции данного параметра с рядом лингвистических и экстралингвистических факторов. В дальнейшем это позволит создать генеративную модель гипертекста на основе тематической кластеризации больших текстовых данных, а той авторской интенцией, которая характерна для определенного типа текстов. Необходимо отметить, что существует ряд исследований в области автоматического связывания документов (Cohn, Hofmann, 2001; Gruber и др. 2007), построенных на тематическом моделировании и кластеризации текстов. Наиболее близким оказывается исследование А. Грубера, М. Розена-Цви, Я. Вайса, в рамках которого была предложена скрытая тематическая модель электронного гипертекста (LTHM), построенная на коллекции статей из «Википедии». Она позволяет предсказывать вероятность актуализации ссылки, с учетом тематической характеристики лексемы, веса целевого текста, а также его тематической отнесенности. Существенный недостаток данной модели заключается в том, что в ней моделируются лишь сильные семантические связи, характерные для пар субстантивных однокомпонентных ссылок «термин/толкование» в публикациях научного стиля, что затрудняет процесс экстраполяции результатов на гипертекстовые структуры более свободной организации, в частности, концептуально-устные жанры, для которых характерно «сознательное затуманивание» семантики гипертекстового перехода (Dedova, 2009), а также использование глагольных ссылок, номинаций с несколькими компонентами.

Для составления универсальной типологии семантических связей в структуре электронного гипертекста мы выделяем трехчастные структуры диалогического типа, которые включают в себя исходный текст (текст-носитель ссылки), целевой текст (текст, к которому осуществляется переход при активации ссылки) и ссылку, обеспечивающую связность текстовых фрагментов. Причем, если рассматривать электронный гипертекст с позиции адресанта, то первичным является именно целевой текст, тогда как исходный текст и номинация ссылки содержат в себе интерпретацию целевого текста. Для номинации данных единиц мы используем термин «гипертекстема». Такой подход позволяет осуществить анализ системы личностного сознания как функционального базиса текстовой деятельности автора и, в конечном счёте, построить комплексную когнитивную модель электронного гипертекста.

1. **Принципы построения корпуса электронных гипертекстов**

На начальной стадии формирования корпуса[[2]](#footnote-2) было решено осуществлять сбор жанрово и стилистически однородных текстов. В состав основных источников вошли следующие электронные СМИ: «Коммерсант», «Известия», «РБК», «Новая газета», «ТАСС», «Дождь», «Новая газета», «Ведомости», «Интерфакс». В настоящий момент база данных включает в себя 31 тыс. текстов с общим словоупотреблением 12 млн. слов. Данные тексты входят в состав 51 тыс. гипертекстем. В настоящий момент распределение текстов по доменам выглядит следующим образом:

В связи с тем, что в основе исследовательского корпуса лежат трехчастные структуры с семантическим типом связи, необходимо выработать принципы квантитативного анализа на микро- (исходный текст, целевой текст и ссылка) и макроуровне (гипертекстема). Сбор данных осуществляется с помощью парсера, разработанного на основе библиотек Python: requests (обращение к веб-страницам), BeautifulSoup (чтение HTML-содержания), re и NLTK (выделения нужных элементов). Парсер формирует базу данных по следующему принципу:

* на первом этапе проводится обход электронных ресурсов, индексируются все ссылки на странице, собираются пары: исходная/целевая страница ссылки, а также адреса их доменов. Описанный процесс цикличен: в случае обнаружения ссылки на целевой странице, ей назначается статус исходной и алгоритм повторяется;
* на втором этапе парсер обходит полученные пары страниц для того, чтобы обнаружить полные текстовые фрагменты. При обнаружении нужных тегов все элементы исходного и целевого текста загружаются в базу данных.

Набор необходимых элементов включает в себя: название СМИ, текст статьи, заголовок, подзаголовок, автор, теги, дата публикации. Основная трудность связана с тем, что на каждом информационном ресурсе эти поля выполнены с различными элементами HTML-верстки. Так, например, элемент «заголовок» может обозначаться с помощью следующих тегов:

«РБК»: <span class="js-slide-title" itemprop="headline"> </span>

«Коммерсант»: <h2 class="article\_name" itemprop="alternativeHeadline"> </h2>

«ТАСС»: <span class="explainer\_\_title"> </span>

В связи с этим для каждого источника данных был создан отдельный файл конфигурации основных элементов.

На этапе инвентаризации гипертекстем особое внимание уделяется фрагментации исходного текста: помимо заданных полей, из текста извлекается номинация ссылки, а также предложение, в составе которого она употребляется. Предобработка гипертекстем включает в себя удаление стоп-слов, процесс токенизации и лемматизации исходных и целевых текстов.

Разметка полученных данных учитывает лингвистические и экстралингвистические признаки. В настоящий момент с помощью библиотеки Pymorphy2 выполнена полная морфологическая разметка номинаций ссылок.

Для определения количественных показателей семантической близости в парах «ссылка/текст», «текст/текст» был разработан лингвистический процессор, который выделяет ключевые слова в исходном и целевом текстах и строит для них векторные модели. Учитывая стилистическую и жанровую однородность состава корпуса, для выделения ключевых слов было решено применять меру TF- IDF, которая считается по формуле

$TFIDF\_{x,y}=tf\_{x,y}×log⁡(\frac{N}{df\_{x}}$),

где tf – частота слова в данном документе, N – общее число документов в корпусе, dfx  – количество документов, содержащих слово. Данная мера позволяет выявлять те лексемы, которые оказываются уникальны для конкретного текста относительно общего массива текстов. Векторное представление ключевых слов строится с использованием модели ruscorpora\_upos\_skipgram\_300\_5\_2018, обученной на Национальном корпусе русского языка.

Экстралингвистическая разметка ссылок включает в себя следующие параметры: переход осуществляется на главную/второстепенную страницу; содержит в себе адрес веб-страницы или текст; содержит название СМИ или нет; количество компонентов; переход на внутренний или внешний ресурс. Всё это позволяет оптимизировать поиск по корпусу.

**Результаты исследования и дискуссия**

Компонентный анализ номинаций ссылок показал, что с увеличением количества слов частотность их употребления убывает по экспоненциальному закону: чаще всего встречаются односложные ссылки (47%), двухкомпонентные встречаются примерно в 2 раза реже (17%), заданная регрессия продолжается и далее (доля трехкомпонентных ссылок 11%). Самые сложные номинации ссылок включают 9 компонентов (2%).

Большинство ссылок включает в свой состав лексемы знаменательных частей речи: глаголы (61%), существительные (38%) и прилагательные (15%), примерно такое же соотношение сохраняется и среди номинаций однокомпонентных ссылок.

Рассмотрим специфику функционирования однокомпонентных глагольных ссылок в структуре электронного гипертекста.

Среди глагольных номинаций наиболее частотными оказываются лексемы, которые относятся к лексико-семантической группе «информация, сообщение, речь»: (1) *заявить/заявлять,* (2) *сообщить/сообщать,* (3) *писать,* (4) *объявить,* (5) *рассказать/рассказывать,* (6) *назвать,* (7) *говорить,* (8) *анонсировать* и др. (табл.1).

|  |  |  |
| --- | --- | --- |
| **Номинация ссылки** | **Общее число вхождений** | **Семантическая близость к целевому тексту** |
| Заявить/заявлять | 1320/202 | 0,21/0,20 |
| Сообщить/сообщать | 1020/722  | 0,20/0,20 |
| Писать | 458 | 0,15 |
| Объявить | 366 | 0,19 |
| Рассказать/рассказывать | 359/130 | 0,16/0,16 |
| Назвать | 327 | 0,17 |
| Говорить | 297 | 0,18 |
| Утверждать | 76 | 0,20 |
| Анонсировать | 50 | 0,23 |
| Объяснить | 46 | 0,17 |
| Пояснить | 40 | 0,17 |

**Таблица 1. Распределение глагольных ссылок по частотности и семантической близости**

Как показал анализ косинусного расстояния между векторным представлением ссылки и ключевых слов целевого текста, данные ссылки вступают в слабую семантическую связь с целевым текстом (в диапазоне от 0,17 до 0,23). Это объясняется функциональной спецификой данного типа ссылок: с помощью них осуществляется гипертекстовый переход к тексту, который призван подтвердить эвиденциальность исходного текста. Таким образом, выбор номинации ссылки обуславливается установкой автора на подтверждение достоверности публикуемой информации, а не информирование о конкретном референте сообщения.

Кроме того, ближайший контекст ссылки также может включать в себя эпистемическую квалификацию целевого текста. Она выражается с помощью лексемы «ссылка» в форме творительного падежа, выполняющей функцию семантического предиката: (9) *сообщают «Ведомости» со ссылкой на гендиректора;* (10) *ранее Совет по правам человека при президенте (СПЧ) со ссылкой на ответ СКР сообщил;* (11) *как сообщила судья со ссылкой на материалы расследования;* (12) *цитирует Дунаева телеканал «360» со ссылкой на ТАСС.* Стремление подтвердить достоверность информации как в исходном, так и в целевом тексте является показателем того, что автор воспринимает гипертекстему как цельную неразрывную функциональную единицу.

В ходе исследования мы предположили, что при использовании глагольных ссылок лексико-семантической группы «сообщение» семантика гипертекстового перехода находит отражение во всей аргументно-предикатной конструкции. Такие конструкции включают следующие компоненты:

* агенс со значением «автор высказывания»: (13) *премьер-министр Тереза Мэй заявила;* (14) *сказала министр обороны Нидерландов Анк Бейлевелд;* (15) *глава ЦИК Элла Памфилова сказала;* (16) *Американский лидер также сказал;*
* агенс со значением «источник информации»: (17) *Как сообщало агентство Bloomberg;* (18) *сообщил ФОМ;* (19)  *ранее “Ъ-СПб” писал;* (20) *сообщают “Ведомости”;*
* пациенс: (21) *Росавтодор объявил конкурс;* (22) *Непосредственной причиной аварии в «Роскосмосе» назвали столкновение элементов при разделении первой и второй ступеней ракеты-носителя;*
* пациенс/место: (23) *Королев опубликовал на своем сайте заявление.*

Содержание целевого текста передается с помощью сложных предложений с изъяснительным типом придаточного предложения *(*(24) *пресс-секретарь президента России Дмитрий Песков сказал, что данные о падении на 20% «не соответствуют действительности»).*

Распределенный характер семантики гипертекстового перехода косвенно подтверждают показатели семантической близости в парах предложение/целевой текст. Предложения, предикативным центром которых являются глаголы сообщения, демонстрируют семантическую близость c целевым текстом около 0,36.

|  |  |  |
| --- | --- | --- |
| **Номинация ссылки** | **Семантическая близость в паре****ссылка/целевой текст** | **Семантическая близость в паре****предложение/целевой текст** |
| Заявить/заявлять | 0,21/0,20 | 0,36 |
| Сообщить/сообщать |  0,20/0,20 | 0,37/0,36 |
| Писать | 0,15 | 0,37 |
| Объявить | 0,19 | 0,35 |
| Рассказать/рассказывать | 0,16/0,16 | 0,34/0,43 |
| Назвать | 0,17 | 0.42 |
| Говорить | 0,18 | 0,33 |
| Утверждать | 0,20 | 0,39 |
| Анонсировать | 0,23 | 0,36 |
| Объяснить | 0,17 | 0,36 |
| Пояснить | 0,17 | 0,27 |

**Таблица 2. Показатель семантической близости в парах «ссылка/текст» и «предложение/текст»**

В настоящий момент осуществляется синтаксическая разметка ссылок, которая позволит выявить закономерности тем с между позицией ссылки в синтагме и тем показателем семантической близости, что она характеризуется.

Итак, фактическая пресуппозиция, подтверждающая достоверность информации в исходном тексте, маркируется с помощью глагольных лексико-семантической группы «сообщение». Семантическая связь непосредственно с содержанием текста осуществляет реализованная валентность предиката, которая значительно модифицируется: целевой текст незаконно занимает позицию объекта по отношению к ссылке-предикату. Таким образом, валентность реализуется и в линейном, и гипертекстовом пространстве.

Большая часть оставшихся глагольных ссылок образует поле номинаций, ориентированных на содержательную характеристику целевого текста. При этом, несмотря на потенциальное тематическое разнообразие целевых текстов, набор номинаций оказывается ограничен тематической спецификой публицистического дискурса. Мы выделяем следующие лексико-сематические группы (в скобках указана общее число вхождений и средняя семантическая близость в парах ссылка/текст ):

- судебно-административная деятельность: (25) *задержать* (297/0,26), *арестовать* (152/0,30), (26) *обвинить* (117/0,23), (27) *возбудить* (110/0,18), (28) *приговорить* (76/0,23), (29)подозревать (6/0,34), (30) *инкриминировать* (1/0,43);

- происшествия: (31) *пострадать* (93/0,18), (32) *погибнуть* (52/0,26), (33) *госпитализировать* (33/0,26), (34) *похоронить* (1/0,50);

- государственная деятельность: (35) *ратифицировать* (4/0,46), (36) *консолидировать* (2/0,44);

- деятельность в профессиональной сфере: (37) *подписать* (203/0,23), (38) *назначить* (63/0,26)

- финансово-экономическая деятельность: (39) *упасть* (33/0,20), (40) *подешеветь* (6/0,37), (41) *задолжать* (3/0,36), (42) *продать* (1/0,39),(43) *подорожать* (1/0,39),

- волеизъявление: (44) *одобрить* (91/0,20), (45) *поручить* (60/0,20), (46) *требовать* (66/0,20).

Данные глагольные ссылки характеризуются высоким показателем семантической близости по отношению к целевому тексту, вполне соотносимому с тем, что демонстрируют субстантивные ссылки близких лексико-семантических групп ((47) *пошлина* (0,46), (48) *санкция* (0,43), (49) *авария* (0,36)). Семантика гипертекстового перехода оказывается локализована в номинации ссылки, что, впрочем, не отменяет уточняющей функции актантов. Например, в предложении (50) *Ранее Басманный суд Москвы арестовал господина Нахлупина до 16 декабря* ссылка *арестовал* указывает на конкретную событие, в то время, как контекст уточняет субъектно-объектные отношения, а также дает пространственно-временную характеристику.

Глагольные ссылки, характеризующиеся сильной семантической связью с целевым текстом, могут выражать модальную оценку ситуации. Так, в предложении (51) *Мегин Келли заявила, что считает Владимира Путина очень умным человеком, которого не получится «перемудрить»* в качестве источника ссылки использована лексема *перемудрить*. Таким образом, автор статьи актуализирует оценочные компоненты в значении номинации ссылки, что, с одной стороны, выражает его субъективное отношение, с другой, стимулирует читателя к совершению гипертекстового перехода. Социальная направленность номинации ссылки выражается в лексемах с семантическим компонентом усиление ((52) *ужесточить*), репрессии ((53) *линчевать, (54) обезглавить*) или преодоления ((55) *прорвались*).

Итак, глагольные ссылки, указывающие непосредственно на содержание целевого текста, актуализирует те темы, которые послужили источником авторской рецепции при написании исходного текста. В этом случае ссылка выступает в качестве вторичного заголовка по отношению к целевому тексту, вступая с ним в референциальные отношения. Помимо этого, ссылка способна выполнять модальную и воздействующую функции.

**Заключение**

В нашем исследовании электронный гипертекст предстает как коммуникативно-познавательная единица, отвечающая критериям текстуальности, что позволяет применить корпусный подход для его изучения. При инвентаризации были выделены единицы диалогового типа, в составе которых целевой текст выступает в качестве стимула, исходный текст становится результатом авторской рецепции, а ссылка является средством маркировки их семантической связности. Данный корпус позволит описать формальные семантические связи в структуре электронного гипертекста с учетом лингвистических (морфологических, синтаксических) и экстралингвистических (жанр, автор, время создания и др.) параметров. Корпусное описание электронных гипертекстов позволит создать модель, способную генерировать электронный гипертекст с учетом жанровой характеристики текста и интенции автора электронного гипертекста.

Мы повели предварительный анализ функционирования глагольных ссылок в структуре электронного гипертекста и выделили три доминанты, определяющие специфику реализации семантики перехода в данных гипертекстемах. Во-первых, это гипертекстуальность как имманентный признак интернет-дискурса, определяющий процесс трансформации линейных синтаксических связей. Как показало исследование, в случае использования глагольных ссылок со значением «сообщение» видоизменяется валентность предиката, которая реализуется как в линейном, так и гипертекстовом пространстве.

Во-вторых, это тип текстов, погруженных в электронную среду. Корпусный анализ электронных СМИ показал, что ссылки, чьи номинации указывают на содержание целевого текста, распределяются по темам, соответствующие новостным разделам: судебно-административная деятельность, происшествия, государственная деятельность, деятельность в профессиональной сфере, финансово-экономическая деятельность. Таким образом, ссылка становится средством жанровой стратификации гипертекстов. Расширение жанрового разнообразия представленных в корпусе гипертекстов является закономерным шагом в развитии нашего проекта.

Наконец, это стратегия, которой автор придерживается при создании электронного гипертекста. Она отражается в степени локализации семантики гипертекстового перехода, а также функциональных особенностях использования ссылки и целевого текста.

**References**

1. Bart R. (1994), Selected Works by Bart R. Semiotics. Poetics [Izbrannye raboty. Semiotika. Poetika]. Moscow: Progress: Univers.
2. Beloedova A. V. (2017) Types of information sources in modern media discourse and problems of their reliability [Tipy istochnikov informatsii v sovremennom mediadiskurse i problemy ikh dostovernosti], Scientific journal of BelSU. Humanitarian sciences [Nauchnyye vedomosti BelGU. Ser. Gumanitarnyye nauki.], Vol.33, pp. 87 – 94.
3. Cherniavskaia V. E. (2013), Text Linguistics. Linguistics of discourse [Lingvistika teksta. Lingvistika diskursa]. Moscow: Flinta: Nauka.
4. Cohn D., Hofmann T. (2000), The missing link - A probabilistic model of document content and hypertext connectivity. In Advances in Neural Information Processing Systems 13, Neural Information Processing Systems (NIPS) 2000. pp 430 –436
5. Dedova O.V. (2005) Change of relations “author-reader” in the space of electronic hypertext [Izmeneniye otnosheniy “avtor-chitatel” v prostranstve elektronnogo giperteksta], Moscow University Bulletin. (Series 9. Philology) [Vestnik Moskovskogo universiteta. (Seriya 9. Filologiya)], Vol. 6, pp. 30 – 46.
6. Dedova O.V. (2008), Theory of Hypertext and Hypertext Practices in the Runet [Teoriya giperteksta i gipertekstovyye praktiki v Runete], Moscow: MAKS Press.
7. Gruber A., Rosen-Zvi M., Weiss Y. (2008), Latent Topic Models for Hypertext, available at: <https://www.semanticscholar.org/>
8. Klochkova E.S. (2015), Substantive hyperlinks as a means of expanding the thesaurus and creating an associative field [Substantivnyye giperssylki kak sredstvo rasshireniya tezaurusa i sozdaniya assotsiativnogo polya], Proceedings of the IV International Scientific Practical Conference “Hypertext as an object of linguistic research” [Materialy IV mezhdunarodnoy nauchno-prakticheskoy konferentsii “Gipertekst kak ob"yekt lingvisticheskogo issledovaniya”], Samara, pp. 52 – 59.
9. Kutuzov, A., Kuzmenko, E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham
10. Mangen A. (2008), Hypertext ﬁction reading: haptics and immersion. Journal of Research in Reading, Vol. 31. Issue 4, pp. 404 – 419.
11. Nelson T. (1974), Computer Lаb. Dream machines, Sausalito, CA: Mindful Press.
12. Patterson N (2000), Hypertext and the Changing Roles of Readers. G. English Journal,Vol. 90. № 2, pp. 74 – 80.
13. Paducheva E.V. (2008), The statement and its correlation with reality: Referential aspects of the semantics of pronouns [Vyskazyvaniye i yego sootnesennost' s deystvitel'nost'yu: Referentsial'nyye aspekty semantiki mestoimeniy], Moscow, LKI.
14. Ryazantseva T. I. (2010), Hypertext and electronic communication [Gipertekst i elektronnaya kommunikatsiya], Moscow: LKI.
15. Shechtman N.A. (2014), From narration to hypertext and narrative [Ot povestvovaniya k gipertekstu i narrativu], Orenburg: Publishing house of the OGPU.
16. Shulginov V. A. (2016), Cognitive model of hypertext [Kognitivnaia model' elektronnogo giperteksta], Bulletin of Kemerovo State University [Vestnik Kemerovskogo gosudarstvennogo universiteta], Vol.4, pp. 233 – 238.
1. 1 Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-312-00010 [↑](#footnote-ref-1)
2. Доступ к тестовой версии корпуса гипертекстем осуществляется по адресу http://hypercorpus.ru [↑](#footnote-ref-2)