ПРОБЛЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

И. А. Меньшенина (Витебск, ВГМУ)

ФОРМАЛИЗАЦИЯ РАСПОЗНАВАНИЯ КОМПОНЕНТА ‘МЕТОДЫ’ В НАУЧНЫХ МЕДИЦИНСКИХ СТАТЬЯХ НА АНГЛИЙСКОМ ЯЗЫКЕ

В статье анализируются англоязычные научные тексты по медицине. Выделяются прототипические компоненты суперструктуры раздела «Методы» данного типа текстов. Для каждого из компонентов выявлены наиболее употребительные языковые маркеры. Предлагается алгоритм поиска семантических компонентов раздела «Методы» медицинского научного текста.

Исследования по созданию автоматических систем реферирования направлены на обеспечение быстрого отбора релевантной информации из огромного массива текстов, доступных в Интернете. Ускорение доступа к новейшим исследованиям особенно актуально для специалистов медицинского профиля, поскольку от принятия быстрых и верных решений зависят здоровье и жизнь пациентов.

Учитывая тот факт, что основной обмен новыми знаниями в научной среде, и в медицинской – в частности, происходит на английском языке, актуальными представляются исследования проблем смыслового свертывания текстов научных статей по медицине на английском языке.

Как основной жанр научной коммуникации научная статья являлась объектом многих лингвистических исследований, в которых, наряду с лексическими, грамматическими и стилистическими особенностями, отмечалась ее четкая композиционная организация. Композиция научной статьи является формально-структурной формой сегментирования текста на структурно-семантические блоки, характеризующиеся относительной завершенностью и автосемантией в пределах целого текста. Научная статья отличается строго логической последовательностью изложения по определенной модели, которая обнаруживает типовые черты в различных научных областях [1, с. 84].

Анализ корпуса текстов подъязыка медицины показал, что композиция статей включает такие разделы, как: *заголовок* ‘title’, *резюме* ‘abstract’, *введение* ‘introduction’, *методы* ‘methods’, *результаты* ‘results’ и *обсуждение* ‘discussion’.

В настоящем исследовании рассматриваются вопросы формализации компонента «Методы» научной статьи по медицине на английском языке. Материалом для исследования послужил корпус из 100 статей, опубликованных в ведущих англоязычных медицинских журналах «The Lancet», «The British Medical Journal (BMJ)», «The New England Journal of Medicine», «The Journal of the American Medical Association» и др.

Поскольку тексты компонента «Методы» достаточно объемны и могут иметь различную научную ценность, необходимо найти способы определения именно тех фрагментов содержания данного раздела, которые позволят читателю быстро и четко определить для себя новизну и актуальность информации. Для этого важно выделить основные смысловые компоненты раздела.

Являясь одним из важнейших параметров научной статьи, композиция представляет собой самое общее членение текста. Для выявления важных компонентов смысловой структуры текста, которые впоследствии найдут отражение в реферате, необходимо выделить его суперструктуру. Согласно теории ситуативных моделей Т. ван Дейка, суперструктура состоит из прототипических компонентов содержательного характера, свойственных тексту того или иного жанра и являющихся для него обязательными, независимо от его содержания [2; 3]. Каждый компонент семантической структуры текста вербализуется с помощью определенных слов, словосочетаний, лексических клише и синтаксических конструкций, маркирующих тот или иной фрагмент текста [3].

Учитывая сложность анализа семантической структуры научных текстов подъязыка медицины, для выделения наиболее характерных семантических компонентов раздела «Методы» научных медицинских статей были приглашены 5 специалистов медицинского профиля, владеющих английским языком.

Анализ смысловой структуры данного раздела показал, что она может включать 4 основных компонента, почти все из которых, в свою очередь, состоят из подкомпонентов:

1. Характеристика типа исследования.
2. Описание материала исследования:
    А. Величина и характеристика выборки;
    Б. Критерии включения и исключения.
3. Описание процедур:
    А. Единица и метод рандомизации;
    Б. Описание способа вмешательства.
4. Представление анализа эксперимента:
    А. Статистические методы;
    Б. Программное обеспечение.

Некоторые из компонентов были отмечены специалистами как более важные: это те, к которым они обращаются в первую очередь (например, компонент ‘характеристика типа исследований’). В то же время другие были признаны наименее значимыми. Например, компонент ‘представление анализа эксперимента’ не вызывает интереса, поскольку анализ полученных результатов обычно извлекается из соответствующего раздела научной статьи.

Для обеспечения алгоритмического поиска фрагментов текста, соотносимых с компонентами суперструктуры раздела «Методы», необходимо определить способы их маркирования в тексте. Поскольку в разделе «Методы» должны быть приведены лишь сухие факты, выраженные в цифрах, автор статьи, как правило, исключает из данного раздела комментарии и интерпретации, что, соответственно, будет отражаться на выборе маркеров для вербальной реализации его компонентов.

Рассмотрим более подробно компонент ‘характеристика типа исследований’.

Основной коммуникативной целью данного семантического компонента является описание варианта организации исследования. Медицинские исследования могут быть продольными и поперечными. Из продольных наиболее известны дорогостоящие проспективные исследования. Более распространены ретроспективные исследования (исследования типа сравнения с контро-

лем), в которых сравниваемые группы пациентов (изучаемая и контрольная) выделяются по состоянию на текущий момент, а затем особенности группы в прошлом рассматриваются ретроспективно. Поперечные исследования (т.е. исследования в определенный момент времени) являются в основном описательными, но это не исключает попыток реконструкции причинно-следственных отношений по поперечным данным [4].

Соответственно, маркерами данного компонента суперструктуры будут выступать термины *prospective, comparative, retrospective, randomized, blinded, etc.*, а также лексические индикаторы *We conducted/ performed/ undertook/ etc.:*

To evaluate the results, ***we conducted a prospective and comparative*** *study from March 2008 to December 2014* [L. 2109, Vol. 3].

We undertook a prospective, randomised *study involving 100 consecutive patients undergoing primary total knee replacement* [L. 2108, Vol. 7].

This ***prospective randomized clinical*** *trial comprised patients presenting for primary THA at a major teaching hospital between February 1998 and September 1999* [AJ. 2109, Vol. 1].

Семантический компонент 'описание участников исследования' включает субкомпоненты 'величина' и 'характеристика выборки', 'критерии включения и исключения'. Коммуникативное назначение данного компонента состоит в предоставлении детальной информации о популяции, на которой проводилось исследование (пол, возраст, диагноз, индекс массы тела, анамнез и т.п.), указании численности больных в контрольных и опытных группах, перечислении критериев включения пациентов в исследование и исключения из него.

Для вербализации компонента 'описание участников исследования' широко используются маркеры *groups, male, female, aged, years of age, a total of, inclusion criteria, exclusion criteria* и др., а также лексические индикаторы *We selected/ separated/ excluded/ included, were eligible/ divided/ assigned* и др.

We ***selected*** *59 patients with a diagnosis of primary osteoarthritis of the knee who underwent elective surgeries for TKA, and* ***separated*** *them into two* ***groups*** *using permuted block randomization. Participants were* ***male*** *and* ***female*** *patients* ***aged*** *60 to 80 years, with grade 3 or above according to the Kellgren and Lawrence classification, 10 indicated for TKA with no bone defects requiring additional grafts or implants, and did not have pronounced angular deformities. We* ***excluded*** *patients with psychiatric disorders, dependence on alcohol or illegal drugs, allergies to morphine, dipyrone or any local anesthetic, previous infection in the knee or other joints, systemic inflammatory diseases, congenital deformities or neurological disorders, and arthroplasty revision* [L. 2109, Vol. 3].

Назначение семантического компонента 'описание процедур', состоящего из подкомпонентов 'единица и метод рандомизации' и 'описание способа вмешательства', – дать полное представление о методе диагностики/лечения (дозировка препарата, режим его введения, аппаратура, на которой проводилось исследование/лечение, объем операции, порядок применения лечебных/диагностических методов и т.д.).

Выделить семантический компонент 'описание процедур' позволяют такие языковые маркеры, как глаголы *to receive, to estimate, to calculate, to perform, to measure, to tolerate, to relate, to follow, to do, to use, to give* и др., которые в большинстве случаев употребляются в форме страдательного залога. Приведем примеры:

*All the **patients were selected** by simple randomization by using a closed envelope technique. The patients were asked to open the envelope just prior to the surgery. Well-written informed consent was taken from all the patients enrolled in the study. Prior Ethics Committee's approval was obtained for the study. Spinal, combined with epidural anesthesia **was used in the majority of cases**. The standard surgical **steps were followed**. All the **patients were given** three doses of secondgeneration cephalosporin (one within 30 min before the procedure and two doses at 12-hour interval post-operatively). Three doses of 1 g intravenous tranexamic acid (one pre-operatively and two post-operatively at 12-hour interval) **were given to** all the patients.*

Основная коммуникативная цель семантического компонента 'представление анализа эксперимента', включающего подкомпоненты 'статистические методы' и 'программное обеспечение', описать пакет программного обеспечения и методы, с помощью которых производился анализ полученных результатов.

Вербальная реализация семантического компонента 'представление анализа эксперимента' осуществляется посредством глаголов речевой и умственной деятельности *to assess, to define, to analyze, to find, etc.,* фамилий ученых, а также ссылок на их работы и разработанные ими методики *Fisher's exact test, The Mann-Whitney U test, etc.*

***Fisher**'s exact test was used **to assess differences** in the incidence of pyrexia, the extent of bruising and the rate of manipulation and of infection. The **Mann-Whitney** U test was used **to analyse** all other **results**. Values are given as means with 95 % confidence intervals (CI).*

На основе полученных данных был разработан алгоритм поиска семантических компонентов раздела «Методы» в научной медицинской статье на английском языке. Фрагмент алгоритма приведен ниже.

A1	Анализ в тексте раздела «Методы (methods)» фрагмента, расположенного непосредственно за статусом данной суперструктуры		
↓			
A2	Определение *характеристики типа исследования*: фрагмент содержит термины *prospective, comparative, retrospective, randomized, blinded,* etc., а также лексические индикаторы *We conducted/ performed/ undertook/* и др.	Да →	B1
↓			
A3	Определение компонента 'описание популяции обследования': фрагмент содержит маркеры *groups, male, female, aged, years of age, a total of, inclusion criteria, exclusion criteria* и др., а также лексические индикаторы *We selected/ separated/ excluded/ included/ were eligible, were divided, were assigned* и др.	Да →	B2

A4	Определение компонента ‘описание процедур’: фрагмент содержит языковые маркеры (глаголы *receive, estimate, calculate, perform, measure, tolerate, relate, follow, do, use, give* и др., которые в большинстве случаев употребляются в форме страдательного залога)	Да →	B3
A5	Определение компонента ‘представление анализа эксперимента’: фрагмент содержит глаголы *to assess, to define, to analyze, to find* и др., фамилии ученых, а также ссылки на их работы и разработанные ими методики *Fisher’s exact test*, *The Mann-Whitney U test* и др.	Да →	B4
A6	Обработка фрагментов B1, B2, B3, B4		
A7	Составление реферата раздела «Методы (methods)»		

Рассмотренный метод, основанный на закономерностях структурно-семантической организации научной статьи, в сочетании с методами маркирования информации делает возможным алгоритмизацию поиска семантических компонентов раздела «Методы» медицинского научного текста. Следующий этап работы предусматривает сокращение отобранных фрагментов исходного текста с последующим их преобразованием в целях построения связного текста аннотации.

ЛИТЕРАТУРА

1. *Карпилович, Т. П.* Алгоритмизация поиска компонента «Результаты» при смысловом свертывании научного текста/ Т. П. Карпилович // Актуальные проблемы прикладной лингвистики : сб. науч. ст.: в 2 ч. / МГЛУ; редкол.: А. В. Зубов (отв. ред) [и др.] – Минск, 2008. – Ч. 2 – С. 82–89.
2. *Дейк, Т. А. ван.* Язык. Познание. Коммуникация: сб. работ / Т. А. ван Дейк; сост. В. В. Петрова; пер. с англ. под ред. В. И. Герасимова. – М. : Прогресс, 1989. – 312 с.
3. *Карпилович, Т. П.* Моделирование процесса смысловой компрессии текста: когнитивно-дискурсивный подход / Т. П. Карпилович. – Минск : Изд-во МГЛУ, 2003. – 226 с.
4. *Власов, В. В.* Структуры медицинских исследований/ В. В. Власов // Русский медицинский журнал [Электронный ресурс]. – 1996. – № 7. – Режим доступа : https://www.rmj.ru/articles/-obshchie-stati/STRUKTURY_”MEDICINSKIH_ISSLEDOVANIY/. – Дата доступа : 26.06.2019.

In the article English scientific articles in medicine are analyzed. The prototypical components of the semantic structure are discussed and their linguistic markers are identified.