**https://doi.org/10.30853/filnauki.2020.7.62**

Яцко Вячеслав Александрович

**Принципы разработки лингвистической информационной системы**

Цель исследования - сформулировать принципы разработки лингвистической информационной системы, предназначенной для интегрированного представления лингвистических ресурсов. Научная новизна - впервые предлагаются принципы универсальности и комплементарности разработки лингвистической информационной системы. Под универсальностью понимается поддержка всех видов исследований, включая теоретические, прикладные и информационные. Комплементарность предполагает адекватное распределение электронных ресурсов, которые выступают в качестве компонентов системы. Выделяется семантическая, прагматическая и функциональная комплементарность. Полученные результаты показали, что структура системы должна включать в качестве центральных элементов лингвистическую онтологию и национальный корпус.

Адрес статьи: www.gramota.net/materials/2/2020/7/62.html

Источник

**Филологические науки. Вопросы теории и практики**

Тамбов: Грамота, 2020. Том 13. Выпуск 7. С. 313-316. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2020/7/

**https://doi.org/10.30853/filnauki.2020.7.62** Дата поступления рукописи: 13.05.2020

***Цель исследования*** *– сформулировать принципы разработки лингвистической информационной системы, предназначенной для интегрированного представления лингвистических ресурсов.* ***Научная новизна*** *– впервые предлагаются принципы универсальности и комплементарности разработки лингвистической информационной системы. Под универсальностью понимается поддержка всех видов исследований, включая теоретические, прикладные и информационные. Комплементарность предполагает адекватное распределение электронных ресурсов, которые выступают в качестве компонентов системы. Выделяется семантическая, прагматическая и функциональная комплементарность.* ***Полученные результаты*** *показали, что структура системы должна включать в качестве центральных элементов лингвистическую онтологию и национальный корпус.*

*Ключевые слова и фразы:* лингвистические технологии; лингвистическая информационная система; принципы разработки; лингвистическая онтология; текстовые корпуса.

**Яцко Вячеслав Александрович,** д. филол. н., проф.
*Хакасский государственный университет имени Н. Ф. Катанова, г. Абакан*
*iatsko@gmail.com*

# Принципы разработки лингвистической информационной системы

*Исследование выполнено при поддержке гранта РФФИ 20-07-00124.*

Одним из закономерных результатов развития современных информационных технологий является создание предметно-ориентированных информационных систем. В медицине разработаны больничные информационные системы (*hospital information systems*), центральной частью которых выступает база данных, в которой собираются информация о пациентах, истории болезней, амбулаторные карты, результаты обследований, операций, анализов биоматериалов, поступающие благодаря обмену данными с подсистемами, такими как радиологическая и лабораторная. Информационно-поисковые системы индексного типа обеспечивают врачам доступ ко всей необходимой информации для наибольшей эффективности лечения [6]. Географические информационные системы предоставляют информацию о местоположении, форме и взаимосвязях географических объектов в виде картографических данных (координаты, адреса, рельеф местности), фотографических данных и изображений, цифровых данных (например, полученные со спутников данные о землепользовании, качестве и видах почв), данных в табличном формате (например, демографические данные об этническом составе населения, возрастных и социальных группах). Разные виды данных формируются в виде отдельных слоёв, которые могут накладываться друг на друга. Географические информационные системы широко используются в быту с целью поиска объектов, навигации, построения маршрутов. Вместе с тем наложение друг на друга разнородных слоёв позволяет выявить ранее неизвестные зависимости, которые могут иметь большое значение для принятия управленческих решений, проведения политики органами государственного управления. Типичный пример – наложение слоя с данными о преступности на картографические и табличные данные, в результате чего становится возможным выявить факторы, влияющие на уровень преступности и виды преступлений, разрабатывать и осуществлять необходимые превентивные меры, принимать решения о соответствующем распределении ресурсов правоохранительных органов [7]. Таким образом, географические информационные системы могут использоваться и в качестве систем управления, обеспечивая поддержку принятия решений.

Наряду с предметно-ориентированными, можно выделить и универсальные информационные системы, предназначенные для широкого контингента пользователей, независимо от их социального статуса и профессиональной принадлежности. Универсальные информационные системы предназначены для бытового (резидентного) информационного обслуживания населения. Предметно-ориентированные системы, как мы полагаем, можно разделить на два вида: системы, предназначенные для повышения эффективности деятельности специалистов данной области; системы, предназначенные для поддержки принятия управленческих решений руководителями предприятий и организаций (информационные системы управления). Системы, предназначенные для поддержки деятельности специалистов, также можно разделить на системы, обеспечивающие информационное обслуживание специалистов, осуществляющих практическую деятельность, а также системы, предназначенные для поддержки научных исследований (специальные системы). Также можно выделить специализированные системы, предназначенные для поддержки решений, пользователями которых выступают руководящие работники и менеджмент предприятий и организаций. Данная классификация информационных систем представлена на Рисунке 1.

Можно выделить следующие основные функции, выполняемые предметно-ориентированными информационными системами. 1. Интеграция различных подсистем или слоёв и обеспечение обмена данными между ними с целью создания единого информационного пространства. 2. Представление данных пользователям в табличном, графическом, визуальном форматах. 3. Генерация статистических данных об объектах и их распределении. 4. Предоставление возможности пользователям создавать модели объектов на основе интерпретации полученных данных. Информационные системы управления также выполняют дополнительную функцию мониторинга и отслеживания объектов и событий.

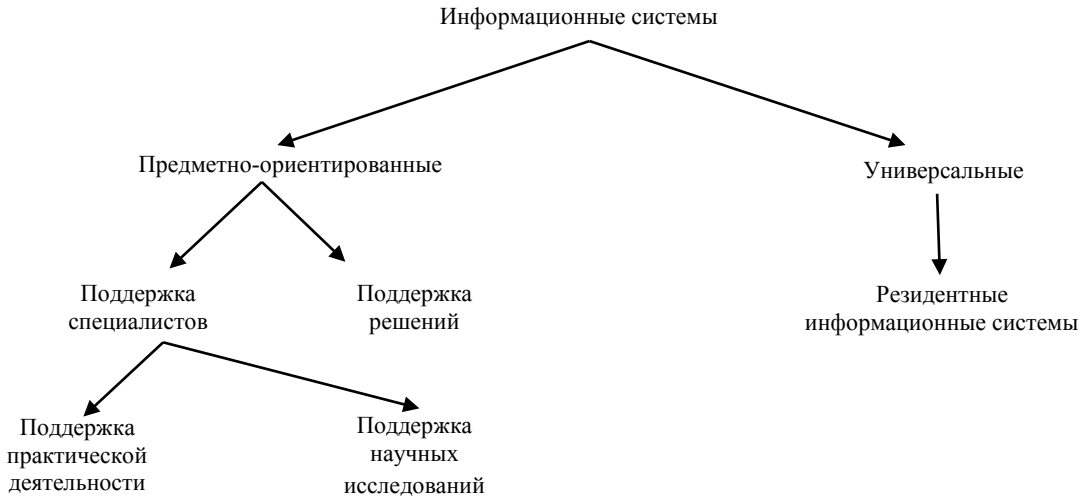


**Рисунок 1.** *Виды информационных систем*

***Задача*** данной статьи – сформулировать и обосновать основные принципы разработки информационной системы, предназначенной для интегрированного представления лингвистических технологий и поддержки как научных исследований, так и практической деятельности по языковому обучению. Вначале нами выделяются и описываются основные функции информационных систем, разработанных в других областях. Затем рассматриваются основные виды лингвистических технологий, которые соотносятся с различными группами пользователей. Далее описывается предполагаемая структура системы. Очевидно, что разработка такой системы является ***актуальной*** проблемой в связи большим количеством и разнообразием использующихся в настоящее время лингвистических технологий и систем. Ниже на основе таксономического метода исследования даётся характеристика различных видов технологий, выделяемых в зависимости от категории пользователей. Далее на основе семиотического метода исследования описывается структура лингвистической информационной системы. Семиотический метод предполагает выделение компонентов системы по семантическим, синтаксическим, прагматическим признакам. Создание такой системы имеет непосредственное теоретическое и ***практическое значение*** как для поддержки научных исследований в области языкознания с целью повышения их эффективности, так и совершенствования языкового обучения и двуязычного перевода.

Заметим, что идея создания лингвистической информационной системы предлагается нами впервые. Она была предложена на основе обобщения разработок в других предметных областях, указанных выше, которые и составляют теоретическую базу данного исследования. Ранее в языкознании ничего подобного не предлагалось ни в России, ни за рубежом. Для современных исследований в области языкознания характерны нацеленность специалистов на проведение конкретных разработок и исследований, недооценка общетеоретических проблем и недостаточное внимание фундаментальным исследованиям, к которым и относится предлагаемый проект. Достаточно сказать, что в настоящее время отсутствует общепринятый термин для обозначения раздела, связанного с разработкой информационных лингвистических технологий [2].

Разнообразные лингвистические технологии, разрабатываемые и применяемые в настоящее время, также могут быть идентифицированы по целевой категории пользователей и разделены на универсальные и специальные. Типичный пример универсальных лингвистических систем – информационно-поисковые системы индексного типа, такие, как «Яндекс» и «Гугл». Ими пользуются сотни миллионов людей, независимо от возраста, образования, социального положения. К универсальным относятся и системы голосового управления (голосовые ассистенты), предназначенные для обеспечения домашней автоматизации и использования электронных устройств в режиме “hands-free”, такие, как “Google Assistant”, “Cortana”, “Alexa”, “Siri” [8].

Специальные лингвистические технологии и системы представлены текстовыми корпусами, которые предназначены для поддержки научных исследований и практической деятельности по обучению языкам. Дидактические корпуса (learner corpora) содержат тексты, созданные студентами, изучающими иностранные языки. Данные тексты аннотируются тегами частей речи и тегами ошибок. Анализ больших массивов текстов позволяет выявить наиболее типичные ошибки, допускаемые студентами, а также объяснить причины этих ошибок, к которым относятся либо влияние родного языка, либо некоторые закономерности изучения данного языка как иностранного [11]. Сопоставление распределения частей речи в студенческих работах с их распределением в текстах, созданных носителями языка (эталонных текстах), позволяет выявить отклонения от стилистической нормы. Очевидно, что результаты анализа дидактических корпусов имеют непосредственное значение для практики обучения иностранным языкам.

Для поддержки научных исследований в области языкознания предназначены исследовательские корпуса разных видов [5]. Национальные корпуса представляют язык в данный момент его существования; они ориентированы на поддержку синхронных и сопоставительных исследований. Исторические корпуса предназначены для поддержки диахронных исследований. Параллельные корпуса включают два вида: трансляционные корпуса, содержащие эквивалентные тексты на двух или более языках, и контрастивные корпуса, содержащие тексты, относящиеся к одному жанру или типу. Трансляционные корпуса используются специалистами

в области теории и практики перевода, а контрастивные корпуса предназначены для специалистов в области автоматической классификации текстов.

К специальным относятся и приложения, предназначенные для статистического анализа текстовых документов, – конкордансы, которые на выходе выдают информацию о распределении единиц входного текста. Конкордансы генерируют три основных типа выходных данных – контекст ключевого слова, заданного пользователем; список слов с указанием частотностей; список коллокаций с указанием частотностей [10]. В обязательном порядке указывается количество уникальных слов и общее количество токенов. Некоторые конкордансы предоставляют пользователю дополнительные функции распознавания и статистического распределения н-грамм и кластеров, генерации ключевых терминов текста на основе вероятностно-статистического анализа, обработки аннотированных текстов. Эти выходные данные являются исходными для дальнейшей автоматической обработки текстов специалистами в области компьютерной лингвистики.

Специализированные технологии предназначены для поддержки принятия решений их пользователями и предполагают использование интеллектуального анализа текста, позволяющего выявить информацию, имплицитно присутствующую в тексте. К такой информации могут относиться, например, числовые коэффициенты, указывающие на интенсивность отрицательной или положительной оценки какого-то продукта покупателями. Основываясь на данной информации, менеджмент фирмы, производящей данный продукт, может принять решение о его модификации, продвижении или снятии с продажи. Отличие данных технологий состоит в том, что они являются онтологически-ориентированными. Для поддержки функционирования таких систем создаются лингвистические онтологии – многоуровневые таксономии, отражающие содержание соответствующей предметной области, а также формальные грамматики, в которых описываются правила распознавания единиц онтологии в речи (связном тексте).

Полагаем, что лингвистическая информационная система должна учитывать указанные виды технологий, а также потребности разных групп пользователей. Ядром системы должна стать онтология языкознания – иерархически структурированный информационно-справочный ресурс, отражающий структуру научной дисциплины. В основных узлах онтологии должны содержаться описания соответствующих разделов (направлений) и ссылки на работы, представляющие их содержание. Онтология предназначается для всех категорий специалистов и обучающихся. Другим центральным элементом должен стать национальный корпус, пользователями которого будет большое количество специалистов в области языкознания. К настоящему времени национальные корпуса созданы для многих языков, в том числе и для русского, который включает около 600 миллионов токенов [1].

Второй уровень системы образуют корпуса, предназначенные для специалистов более узких отраслей языкознания, в первую очередь это исторический корпус, параллельные корпуса и дидактический корпус. Оба этих вида корпусов включают подкорпуса, ориентированные на более узкие специализации. Исторический корпус может включать подкорпуса, отражающие определённый исторический период (например, XIX век), в то время как дидактический корпус – делиться на подразделы, соответствующие определённым языкам. Ко второму уровню также относятся программно-аппаратные платформы, предназначенные для дистанционного обучения языкам и лингвистическим дисциплинам, а также для поддержки разработок в области компьютерной лингвистики, включая конкордансы. К таким платформам можно отнести модули языков программирования и среды разработки, предназначенные для автоматической обработки текстовых документов. В качестве примеров можно привести *Natural Language Toolkit* для языка *Python* [4], а также пакет *Text mining* для языка *R* [9].

На периферии информационной системы находятся специализированные лингвистические ресурсы, находящиеся на стыке языкознания и других дисциплин. Это могут быть предметно-ориентированные лингвистические онтологии [3], которые применяются для интеллектуального анализа мнений, обнаружения экстремистского контента, обмена опытом, а также различные лексикографические ресурсы.

Считаем, что можно предложить следующие основные принципы разработки лингвистической информационной системы:

1) принцип универсальности;
2) принцип комплементарности.

Первый принцип предполагает, что информационная система предназначена для поддержки всех видов исследований, включая теоретические, прикладные и информационные. Второй принцип предполагает комплементарное распределение электронных ресурсов, которые выступают в качестве компонентов системы и обеспечивают её эффективное функционирование. Полагаем, что можно выделить три вида комплементарности: семантическую, прагматическую, функциональную. Семантическая комплементарность предполагает, что в лингвистическую информационную систему должны быть включены различные типы текстов, содержащиеся в разных видах корпусов. Национальный корпус должен дополняться историческим и дидактическим; моноязычный корпус – дву- и многоязычными корпусами; универсальные лексикографические ресурсы – специальными. Под функциональной комплементарностью мы понимаем сочетание онлайновых и оффлайновых режимов обработки документов, а также функции программного обеспечения. Независимо от того, насколько могут быть репрезентативны онлайновые корпуса, у исследователей-лингвистов возникают потребности в автоматическом анализе текстов, которые в них отсутствуют. Для этого могут применяться конкордансы, которые устанавливаются на локальные компьютеры и не требуют подключения к Интернету. Функциональность конкордансов должна дополнять функциональность онлайновых корпусов. В них должны быть как те функции, которые используются в корпусах, так и дополнительные функции, например, объединения, вычитания и пересечения текстов. Объединение текстов позволяет исследователю анализировать ряд текстов, относящихся к одному типу или жанру; вычитание позволяет удалять из текстов часть содержания, которое не интересует

исследователя, например, стоп-слова или наоборот – знаменательные слова; пересечение делает возможным выделить единицы, которые используются в разных текстах.

Было бы целесообразным включить в лингвистическую информационную систему в качестве отдельного приложения программное обеспечение, позволяющее получать на выходе весовые коэффициенты единиц текста, подсчитываемые с использованием различных метрик: TF*IDF, хи-квадрат, отношение шансов, прирост информации, формулы Байеса. Такое приложение могло бы быть полезным для решения целого ряда задач, связанных с автоматической классификацией текстовых документов, включая распознавание плагиата и фильтрацию спама.

Под прагматической комплементарностью мы понимаем удовлетворение с помощью компонентов системы информационных потребностей разных групп пользователей, указанных выше, в том числе специалистов в области теоретической и прикладной лингвистики, компьютерной лингвистики, лингводидактики, а также руководящего персонала предприятий и организаций, не входящих в сферу языкознания, но заинтересованных в использовании лингвистических технологий с целью принятия адекватных решений.

В данной статье были рассмотрены различные виды лингвистических технологий, соотносящиеся с различными категориями пользователей. На этой основе были сформулированы принципы разработки информационной системы, предполагающей интегрированное представление указанных технологий и ресурсов, а также описана структура такой системы. Очевидным является ***вывод*** о том, что разработка такой системы потребует значительных усилий со стороны научного сообщества, направленных на: 1) создание онтологии языкознания как основополагающего справочно-информационного ресурса, предназначенного для всех категорий пользователей, указанных выше; 2) модификацию существующих текстовых корпусов и создание новых. Особое внимание следует уделить переформатированию Национального корпуса русского языка, который в настоящее время не вполне соответствует требованиям, так как содержит тексты девятнадцатого и двадцатого веков. Оптимальным было бы их включить в исторический корпус, дополнив национальный корпус современными текстами; 3) стандартизацию терминологии и требований к разрабатываемым лингвистическим ресурсам, что позволит повысить эффективность научной коммуникации.

Решение таких масштабных задач потребует концентрации ресурсов, привлечения большого количества специалистов, создания соответствующего координационного центра, которым мог бы стать один из академических институтов.

*Список источников*


1. **Чеснокова И. Д., Маньшин М. Е.** Национальный корпус русского языка как основной инструмент поиска при лингвистических исследованиях (на примере поиска антонимов в публицистических текстах) // Известия Волгоградского государственного педагогического университета. 2018. № 5. С. 97-103.
2. **Яцко В. А.** Компьютерная лингвистика или лингвистическая информатика? // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2014. № 5. С. 1-10.
3. **Яцко В. А., Яцко Т. С.** Особенности структуры лингвистической онтологии // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2017. № 6. С. 16-25.
4. **Colton D.** Text classification using Python // Text mining and visualization: Case studies using open-source tools. Boca Raton, 2016. P. 199-220.
5. **Dash N. S., Arulmozi S.** Type and purpose of text // History, features, and typology of language corpora. Singapore, 2018. P. 67-83. DOI: 10.1007/978-981-10-7458-5_5.
6. **Farzandipour M., Meidani Z., Gilasi H., Dehghan R.** Evaluation of key capabilities for hospital information system: A milestone for meaningful use of information technology [Электронный ресурс] // Annals of Tropical Medicine and Public Health. 2017. Vol. 10. Iss. 6. URL: http://www.atmph.org/text.asp?2017/10/6/1579/222676 (дата обращения: 14.05.2020).
7. **GIS (Geographic Information System)** [Электронный ресурс]. URL: https://www.nationalgeographic.org/encyclopedia/geographic-information-system-gis/ (дата обращения: 14.05.2020).
8. **López G., Quesada L., Guerrero L. A.** Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces // Advances in Intelligent Systems and Computing. Cham: Springer, 2018. Vol. 592. P. 241-250.
9. **Oliveira N., Areal N.** Sentiment analysis of stock market behavior from Twitter using the R Tool // Text mining and visualization: Case studies using open-source tools. Boca Raton, 2016. P. 223-240.
10. **Wiechmann D., Fuhs S.** Concordancing software // Corpus Linguistics and Linguistic Theory. 2006. Vol. 2. Iss. 1. P. 107-127. DOI: 10.1515/CLLT.2006.006.
11. **Yu-Chun Lo, Jhih-Jie Chen, Ching-Yu Yang, Jason S.** Cool English: A grammatical error correction system based on large learner corpora // Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Santa Fe, 2018. P. 82-85.


## Principles of Linguistic Information System Design


**Yatsko Viatcheslav Alexandrovitch**, *Dr*
*Katanov State University of Khakasia, Abakan*
*iatsko@gmail.com*



The study aims to formulate principles for developing a linguistic information system meant to act as integrated representation of linguistic resources. Scientific originality of the research is in that principles of universality and complementarity of a linguistic information system development are suggested for the first time. Universality is understood as supporting all kinds of research including theoretical, applied and informational. Complementarity refers to adequate allocation of electronic resources functioning as system components. Such types of complementarity as semantic, pragmatic and functional are identified. The results of the study have demonstrated that the system structure must include linguistic ontology and a national corpus as its core elements.



*Key words and phrases:* linguistic technologies; linguistic information system; design principles; linguistic ontology; text corpora.