МОДЕЛИРОВАНИЕ ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ ПО НАУКАМ О ЗЕМЛЕ НА ОСНОВЕ ОНТОЛОГИИ[1)]

О.А. Курчавова
Институт проблем информатики РАН
e-mail: olga@170.ipi.ac.ru

Сообщение посвящено проблеме когнитивно-лингвистического моделирования полнотекстового научного документа на основе онтологий. Полнотекстовый научный документ рассматривается как единство линейного текста и вербально-графических компонентов в виде таблиц, графиков, диаграмм. Вербально-графические компоненты исследуются на примере таблиц. Рассматриваются вопросы вербально-графических компонентов, методика построения вербально-образных онтологий и средства предикации. В качестве материала для исследований используются статьи по наукам о Земле.

Ключевые слова: когнитивно-лингвистическое моделирование, вербально-графический компонент, изоконцептуальный объект, таблица, онтология.

1. Введение

В настоящее время задача извлечения научных знаний по вербально-графическим компонентам в виде таблиц и, в общем случае диаграмм, является нерешенной, хотя эти компоненты могут являться наиболее ценной информационной составляющей научного документа. Целью нашего исследования является разработка методики, которая позволит провести декомпозицию и структурно-семантический анализ полнотекстового научного документа и позволит индексировать документы с использованием вербально-образных онтологий [1].

Задачей исследования при построении вербально-образных онтологий является установление и описание лексических и семантических средств, использованных для представления смысла средствами линейного текста вербальных компонентов и вербально-графических коммуникативных компонентов научного дискурса по материалам статей по наукам о Земле. В качестве материалов исследования использовались как линейные текстовые фрагменты, так и вербально-графические коммуникативные компоненты, которые можно рассматривать как нелинейные текстовые фрагменты.

Для построения вербально-образных онтологий использовались систематические каталоги обозначений, принятых в науках о Земле (например, Стратиграфический кодекс) [2]. Кроме того, использовались условные обозначения, вводимые авторами статей. Систематические каталоги и авторские условные обозначения являются базовым лексическим ресурсом вербально-образных онтологий. Новизна метода состоит в целостном подходе к вербально-графическому компоненту, как целостному когнитивному объекту, аналогичному высказыванию на естественном языке в единстве его семантических и синтаксических средств [3].

Вербально-графические компоненты рассматриваются на примере таблиц. Для целей нашего исследования используется определение таблицы через визуальные составляющие графического конструкта в виде пересекающихся горизонтальных и вертикальных линий с текстовыми фрагментами, упорядоченными по вертикали и горизонтали. Таблицы выполняют важную текстообразующую функцию. По своей сущности таблица представляет собой особым образом организованный фрагмент текста, служащий для компактного представления материала, который предварительно подвергся структурной обработке [4]. Из этого определения можно сделать вывод, что в научных публикациях таблицы более удобны для представления структуры объекта исследования, его

[1)] Работа выполнена при финансовой поддержке РФФИ, проект N 00-06-8069

числовых параметров, отношений со сходными объектами, чем линейный текст. Обладая определенной графической структурой, таблицы данных позволяют наполнять их любым содержанием в кратком и структурированном виде.

2. Декомпозиция вербальных и вербально-графических коммуникативных компонентов

Использованный нами подход для декомпозиции вербального текста и вербально-образных компонентов основан на применении к ним лингвистических подходов. Процесс декомпозиции вербального и невербального текста происходит параллельно. При этом важным результатом исследований является тот факт, что средства синтаксического оформления вербально-графических высказываний являются изофункциональными по отношению к синтактическим средствам оформления линейного текста [5].

Выделяются следующие синтаксические единицы нелинейного текста:

1) *Сверхфразовое единство*. Это вербально-графический объект в целом. Средством номинации невербального сверхфразового единства является подрисуночная подпись. Например, "Синопсис валюхтинской микробиоты" (см. рис. 1).

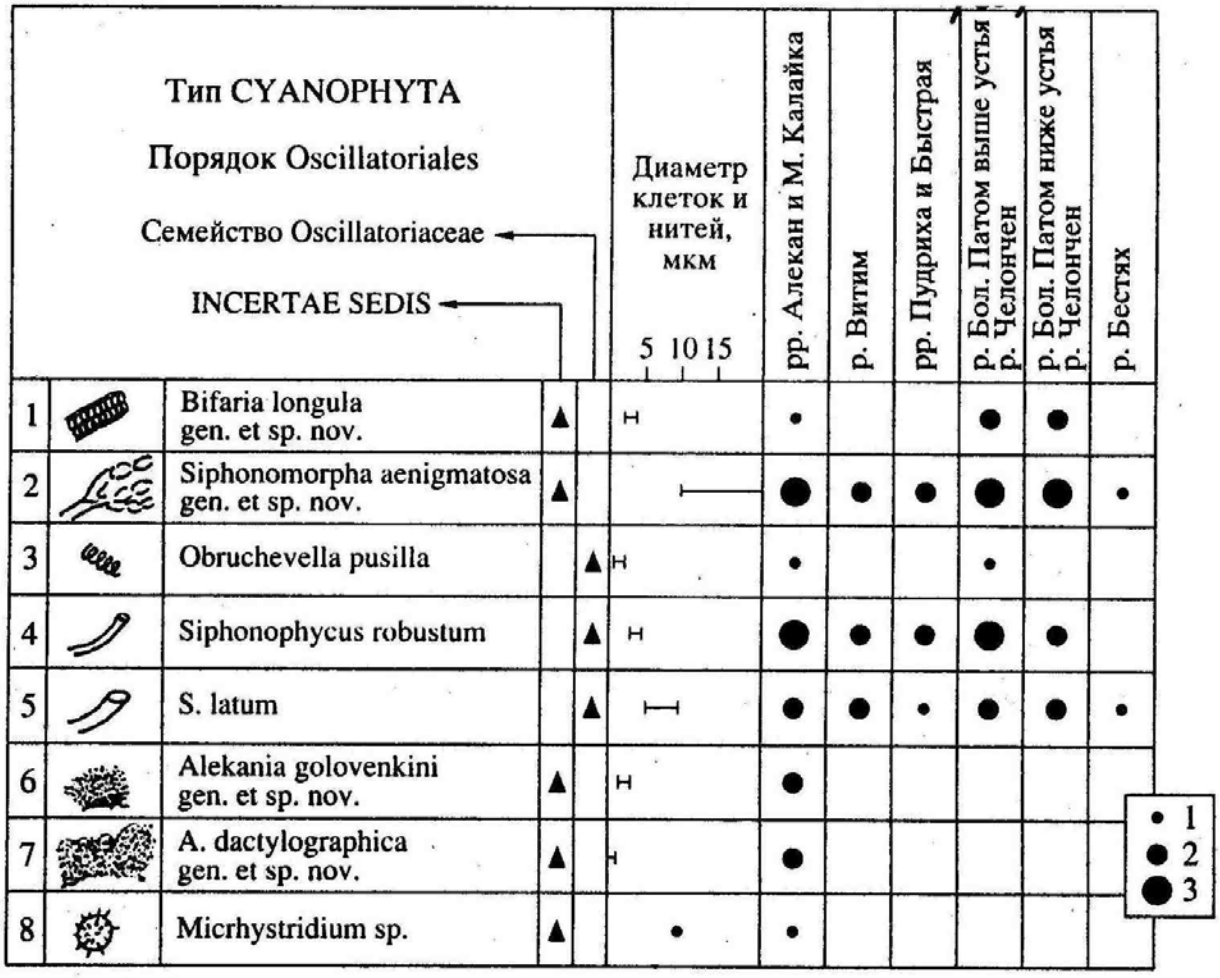

Рис. 1. Синопсис валюхтинской микробиоты:

1 – единичные экземпляры; 2 – более 10-ти экземпляров; 3 – десятки и первые сотни экземпляров.

2) *Невербальная клауза.* Аналог соответствует такой синтаксической единице как предложение. Например, *свита Сансонри выделяется в бассейне реки Рэсон*, или *свита Синичжу относится к периоду поздняя юра - ранний мел* (см. рис 2).

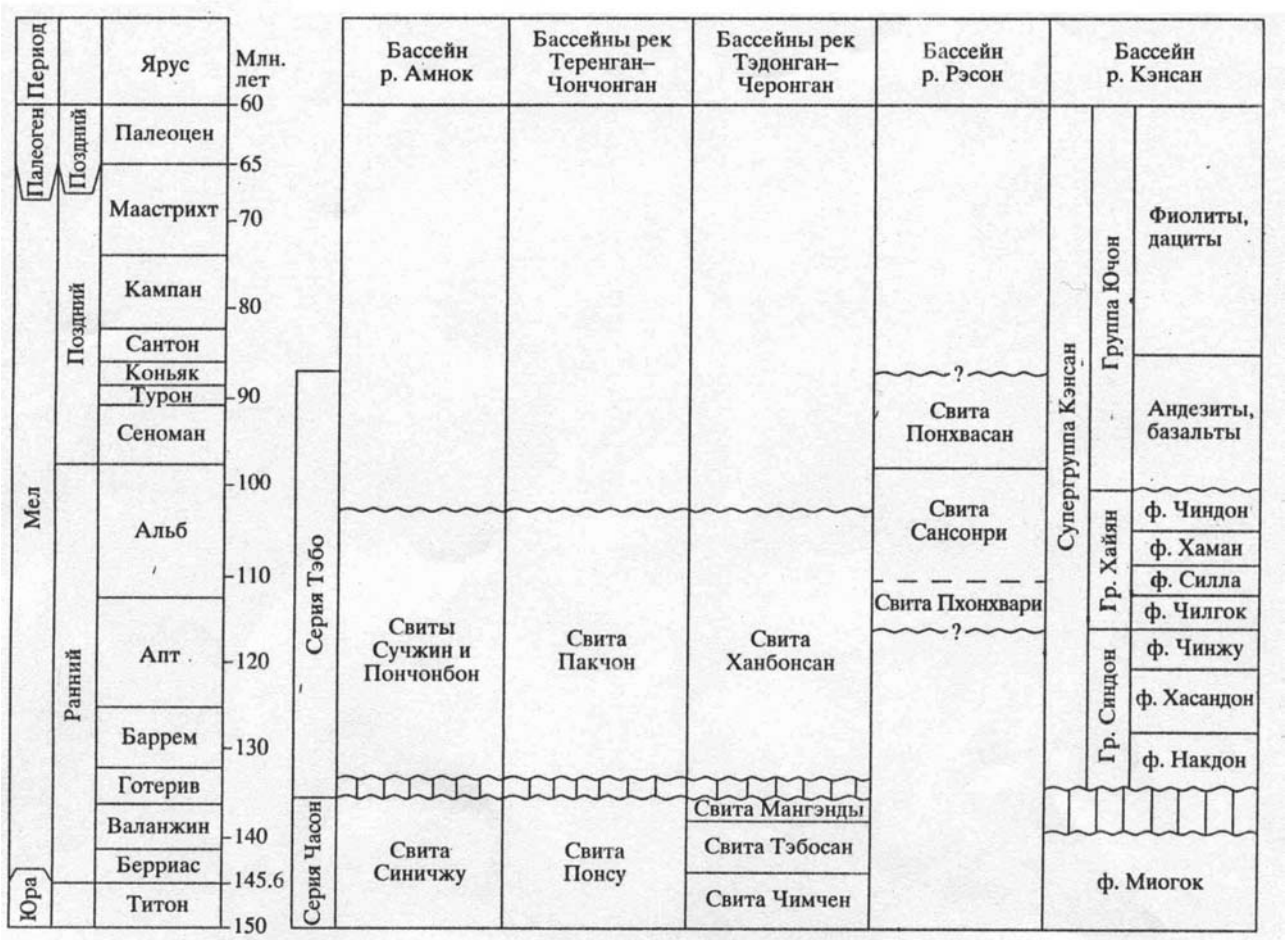

Рис. 2. Корреляция верхнемезозойских отложений северной и южной части п-ва Корея. Сокращения: ф – формация, гр – группа.

3) *Уровень синтагм*. Невербальные синтагмы - это строки и столбцы таблицы. Они соответствуют предикативным конструкциям конституирующим предложения линейного вербального текста. Например, бассейн реки Амнок, серия Тэбо.

4) *Уровень вербальных или невербальных лексем*. Невербальные лексемы соответствуют условным значкам (см. рис.3).

Для построения системы логико-семантических представлений вербальных и вербально-графических коммуникативных компонентов используется аппарат расширенных семантических сетей [6], применявшийся ранее для моделирования только вербальных коммуникативных компонентов. Модели, получаемые на базе этого аппарата, являются унифицируемыми декларативными представлениями - предикатно-актантными структурами, которые мы называем фрагментами семантической сети.

3. Разработка методики создания вербально-образных онтологий

При проектировании вербально-образных онтологий использовались аналогии с методами построения вербальных онтологий. При разработке онтологии ее верхним уровнем становятся термины высокой степени абстракции. Они представляют собой обобщенные концептуализации, действительные для всех предметных областей: такие как например, <предмет>, <явление>, <характеристика>, <параметр>. Нижний уровень онтологии отражает специфику конкретной области науки. Каркасом онтологий являются подрисуночные подписи. Например, "Корреляция верхнемезозойских отложений северной и южной частей п-ва Корея" (см. рис. 2). Слово "корреляция" представляет собой термин высокой степени абстракции со значением <отношение>. Это макроконцепт, задающий систему отношений в таблице. Такой тип отношений характерен для предметной области "Стратиграфия". Объектом корреляции являются верхнемезозойские отложения, сопоставляемые с системой шкал.

Рис. 3. Сводный литолого-стратиграфический разрез лапчанской и ботуобинской свит восточной окраины Тунгусской синеклизы

Шкалы могут быть лингвистическими, где значения представляются словами. Например, <Мел>: <Ранний >, <Поздний> (см. рис. 2). Лингвистическая шкала - это один из способов задания нечетких значений. Шкала может быть метрической, где значения задаются числами. Например, шкала абсолютного возраста (см. рис. 2). Встречаются также графические шкалы, где для их индексации необходимо найти в линейном тексте их лингвистическое или метрическое соответствие. (см. рис. 1, где графическую шкалу представляют кружки разного диаметра. Так, кружку с наименьшим диаметром соответствует лексема "единичные экземпляры").

Прототипической формой отношения является: *Отношение* (Шкала, Объекты отношения, место, время). Так для рис. 1 можно написать следующее: *Корреляция* (Шкала абсолютного возраста, Отложения, Корея, Верхнемезозойские).

Базовыми типами логико-семантических отношений в таблице являются родо-видовые, часть-целое, параметрические. Родо-видовые отношения иллюстрирует рис. 4.

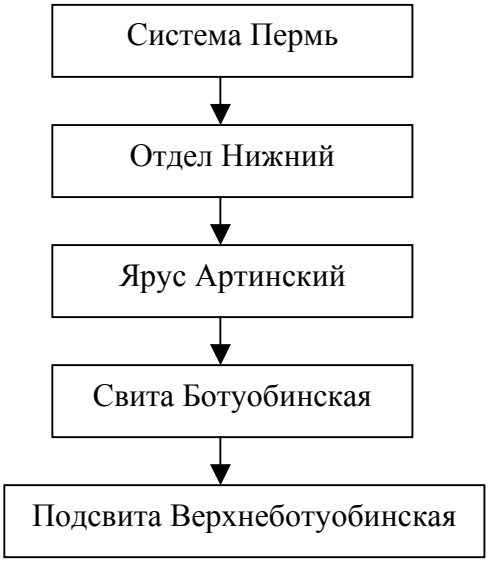

Рис. 4. Родо-видовые отношения (см. рис. 3).

Отношение часть - целое иллюстрирует следующий пример: *Состоять* (что, из чего), например, *Состоять* (Нижнеботуобинская подсвита, конгломераты, песчаники, алевролиты, угли) (см. рис. 5).

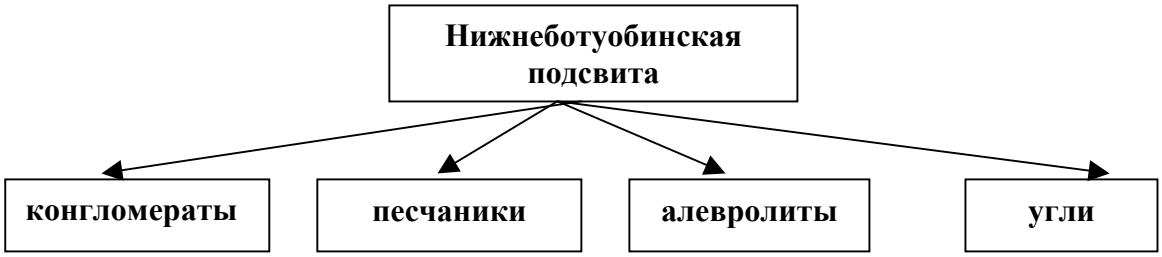

Рис. 5. Отношение часть – целое.

Примером параметрических отношений может служить: *Иметь возраст* (что, какой), например, *Иметь возраст* (свита Ханбонсан, 133-102 млн. лет) - см. рис. 2.

Тип логико-семантических отношений часто связан с топологией таблиц [7]. Следует отметить, что в простых таблицах с равным числом строк и столбцов без группировки присутствуют, как правило, только реляционные отношения. Отношения род-вид (см. рис. 6) возникают в сложных таблицах с группировкой строк и столбцов. Пример подобной таблицы см. на рис. 2.

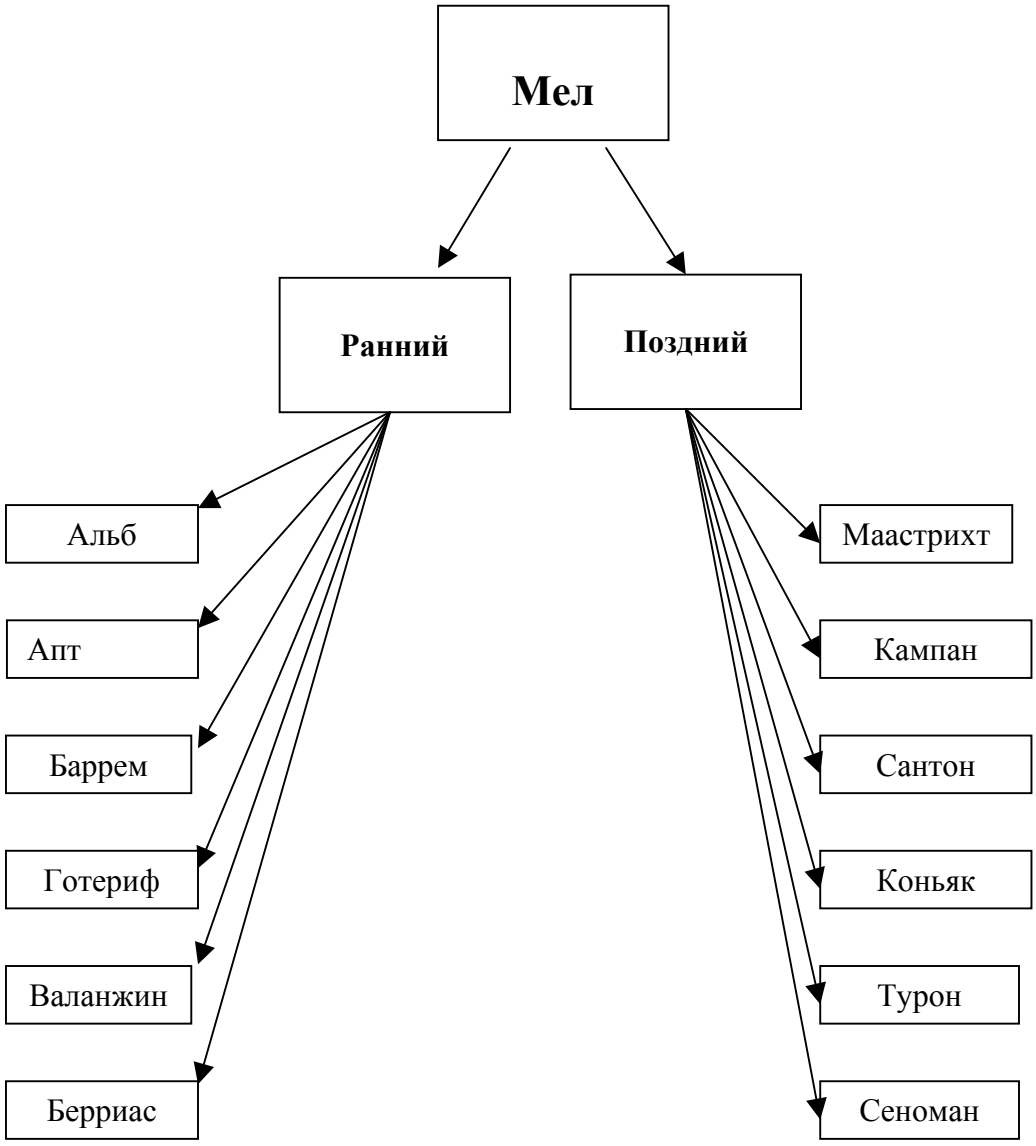

Рис. 6. Отношения род – вид, выражаемые в сложных таблицах за счет группировки строк и/или столбцов (на рис. 2 за счет столбцов).

4. Средства предикации

Любой вербально-графический коммуникативный компонент и его вербальные корреляты рассматриваются как единый объект, что в дальнейшем может служить основой для генерации вербально-графических представлений по естественно-языковому входу, и обратных преобразований невербальных представлений в конструкции естественного языка. Такого рода соответствия установлены для базовых синтаксических структур в пределах предложения; сверхфразового семантически единого комплекса, относящегося к одному вербально-графическому объекту.

Для текстов, относящихся к таким областям, как науки о Земле, выделено 2 широко распространенных типа вербально-графических объектов: специфицирующие и параметризующие. Основным средством задания предикации в объектах такого рода являются различные лингвистические и метрические шкалы. При этом, как правило, эти объекты задают динамику изменений некоторого параметра относительно заданной шкалы (системы шкал). Естественно-языковыми коррелятами такого рода диаграмм будут "зависимость некоторого параметра от другого параметра", "соотношение некоторого параметра с другим параметром", "распределение некоторого параметра относительно шкалы". Унифицированным представлением таких компонентов будут предикатно-актантные структуры, например:

зависеть (Параметр1, Параметр2),

распределяться (Что, по Чему),

соотноситься (Что, с Чем).

Для моделирования применен аппарат расширенных семантических сетей. Соотнесение объектов (ячеек таблицы) со шкалой является средством задания предикации на графическом метаязыке. Выделяются основные средства (глаголы, краткие прилагательные, отглагольные существительные) задания предикации в вербальном тексте. Устанавливаются семантические связи между основными средствами предикации в вербальном и вербально-графическом тексте. Вот некоторые из основных естественно-языковых средств задания отношений в тексте:

включать (что-свиты, что-формации);

выделять (что-свиту, что-формацию/серию);

иметь распространение (что, где, какое);

быть распространенным (что, где, как);

сменять(что, что, как, где);

характеризоваться (что, чем).

Все указанные предикаты являются естественно-языковыми аналогами способов организации ячеек таблицы. Одно и то же соотношение можно задать нарисовав таблицу или составить описание исследуемого материала на ЕЯ. Все выделенные предикаты являются характеризующими, специфицирующими средствами построения синтаксической структуры как вербального, так и невербального текста. Если отправной точкой при анализе полнотекстового документа является его вербальная часть, то присутствие в тексте таких предикатных выражений, как "включает", "относится" сигнализирует о возможности отсылки к некоторому вербально-графическому компоненту, который тоже присутствует в тексте. При этом, если предикат "включает" задает фокус внимания на некоторый титульный элемент, который представлен в таблице самой верхней ячейкой, то обратный ему предикат "относиться к" фокусирует внимание на входящий, включаемый элемент, который будет представлен в таблице любой ячейкой, отличной от самой верхней.

Предикатные выражения задают способы организации строк, столбцов и ячеек таблицы и их взаимное расположение. В свою очередь, табличные коммуникативные компоненты могут быть представлены совокупно множеством предикатных выражений. Например, свойство быть самой верхней ячейкой таблицы является синтаксическим средством вербально-графического языка, которое в качестве своих ЕЯ аналогов имеет предикатные выражения "включать", "состоять из".

Наши текстологические исследования позволяют сделать выводы, что как линейный текст, так и его вербально-графические компоненты порождаются на основе единообразного когнитивного аппарата, использующего компактный набор структур.

5. Заключение

1. По своей сущности таблицы представляют собой особым образом организованный фрагмент нелинейного текста, служащий для компактного представления материала, который предварительно подвергся структурной обработке.

2. Обладая определенной графической структурой, таблицы данных позволяют наполнять их вербальным и невербальным содержанием в кратком и структурированном виде.

3. Процесс декомпозиции вербального и невербального текста происходит параллельно.

4. Средства синтаксического оформления вербально-графических высказыванием являются изофункциональными по отношению к синтаксическим средствам оформления линейным текстом.

5. Выделяются следующие синтаксические единицы нелинейного текста: сверхфразовое единство, невербальная клауза, уровень синтагм, уровень невербальных или вербальных лексем.

6. При проектировании вербально-образных онтологий использовались аналогии с методами построении вербальных онтологий.

7. Каркасом онтологий являются подрисуночные подписи.

8. Важнейшим элементом семантических представлений являются шкалы.

9. Тип логико-семантических отношений часто связан с топологией таблиц.

10. Предикатные отношения задают способы организации строк, столбцов и ячеек таблицы и их взаимное расположение. Табличные коммуникативные компоненты могут представлены совокупно множеством предикатных выражений.

Литература

1. Гаврилова Т.А., Лещева И.А., Лещев Д.В. Использование онтологии в качестве дидактического средства // Труды международной конференции "Искусссственный интеллект 3'2000".- Донецк, 2000.
2. Стратиграфический кодекс.- Санкт-Петербург: Межведомственный-стратиграфический комитет, 1992.
3. Козеренко Е.Б. Когнитивно-лингвистическое моделирование полнотекстовых научных документов // Труды международной конференции "Искусственный интеллект 3'2000".- Донецк, 2000.
4. Аликаев Р.С. Язык науки в парадигме современной лингвистики.- Нальчик: Издательский центр «Эль-Фа», 1999.
5. Козеренко Е.Б. Унифицируемые категориально-функциональные представления для семантической разметки полнотекстового научного документа // Системы и средства информатики. Вып. 12.- М.: Наука, 2002.
6. Кузнецов И. П. Семантические представления.- М.: Наука, 1986.
7. Курчавова О.А. Таблицы как вербально-графические компоненты полнотекстовых научных документов // Труды международного семинара "Диалог-2001" по компьютерной лингвистике и ее приложениям. Т2.- Аксаково, 2001.
8. Филатова Н.И., Чанг Г.Х., Парк С.О. Корреляция верхнемезозойских осадочных и вулканических образований Кореи и обстановки их накопления // Стратиграфия. Геологическая корреляция.- 1999. Т. 7, N 4.
9. Белова М.Ю., Головенок В.К. Позднерифейские минерализованные микрофоссилии валюхтинской свиты Байкало-Патомского нагорья // Стратиграфия. Геоглогическая корреляция.-1999. Т. 7, N 4.
10. Ошуркова М.В. Возраст верхнемезозойских отложений восточного борта Тунгусской синеклизы по палинологическим данным // Стратиграфия. Геологическая корреляция. –1999. Т. 7, N 6.