Лекция 4. Экстралингвистическая разметка. Метаданные.

В.П. Захаров

Санкт-Петербургский государственный университет

Метаразметка

- **Метаданные** – *структурированные данные о данных*:
  - помогают установить порядок среди хаоса,
  - позволяют осуществить автоматическое обнаружение и обработку данных.

Экстралингвистическая разметка

- "внешняя", "интеллектуальная" разметка
  - библиографические характеристики
  - типологические характеристики
  - тематические характеристики
  - социологические характеристики
  - .....
- "формальная" структурная разметка
  - текст, раздел, глава, часть, абзац, предложение ...
- технико-технологическая разметка
  - кодировка
  - даты обработки
  - исполнители
  - источник электронной версии
  - .....

"Внешняя", "интеллектуальная" разметка

Нужна:

- для выявления взаимосвязи языка и условий его существования;
- для изучения отдельных подмножеств языка.

Выделяют **два класса факторов**, влияющих на язык текстов:

- внешние, внеязыковые факторы (E - external);
- внутренние факторы (I - internal).

(См. Sinclair (1996). Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html)

"Внешняя", "интеллектуальная" разметка (продолжение)

Синклер выделяет:

три группы ***E**-факторов*:

1. E1 (origin) - факторы, относящиеся к созданию текста автором;
2. E2 (state) - факторы, относящиеся к внешним признакам текста (включая устную или письменную речь);
3. E3 (aims) - факторы, относящиеся к причинам создания текста и его влиянию на аудиторию.

и две группы ***I**-факторов*:

1. I1 (topic) - предметная область текста;
2. I2 (style) - стилистические особенности (стиль, жанр).

Набор метаданных в «Национальном корпусе русского языка»

Первый блок:

1) *автор текста*: имя, пол, дата рождения (или примерный возраст);
2) *название текста*;
3) *время и место создания текста* (может указываться точно или приблизительно);
4) *объем текста*: для художественных произведений принято, что обычная длина рассказа – менее 5 тыс. слов; обычная длина повести – от 5 до 15 тыс. слов; обычная длина романа – более 15 тыс. слов.

Второй блок:

параметры метаописания трех основных *массивов* текстов корпуса:

1) художественных текстов;
2) нехудожественных текстов;
3) драматургии.

Художественные тексты (в НКРЯ):

1. *жанр текста*

нежанровая проза, автобиографическая проза, детектив, детская литература, историческая проза, криминальная литература, приключения, фантастика, юмор и сатира

2. *тип текста*

автобиографическая проза, анекдот, ассоциативная проза, боевик, детектив, очерк, литературное письмо, повесть, притча, пьеса, рассказ, роман, сказка, триллер, эпопея, эссе и др.;

3. *хронотоп текста*

приблизительное указание на место и время описываемых в тексте событий

Реально предлагается следующее: древний Восток; Россия XVII в.; Россия XIX в.; Россия/СССР: советский период в целом; Россия, советский период – Германия 1920–1940-е годы; Россия/СССР – Европа 1960-1980-е годы; Россия/СССР: перестройка; Россия/СССР: советский и постсоветский период; Америка: современная жизнь; Израиль: современная жизнь; Средняя Азия: современная жизнь; ирреальный мир и др.

Также может быть «хронотоп не определен».

Нехудожественные тексты (в НКРЯ):

1. ***тип текста***

автобиография, акт, дневник, договор, документ, закон, заметка, заявление, инструкция, информационное сообщение, кодекс, комментарий, листовка, обзор, объявление, отзыв, отчет, очерк, письмо, постановление, проповедь, путеводитель, резюме, реклама, рекомендация, рецензия, рецепт, сочинение, справочник, статья, учебник, характеристика, хроника, эссе, юридический документ (включается также помета «тип не определен») и пр. (всего 62 параметра);

2. ***тематика текста***

открытый список в 5 подмножествах: бизнес, коммерция, экономика, финансы; война и вооруженные конфликты; дом и домашнее хозяйство; здоровье и медицина; зрелища и развлечения; искусство; криминал; наука (по разделам и отраслям); политика и общественная жизнь; право; производство; сельское хозяйство; спорт; природа; частная жизнь и т.п.

Служебная, или «имплицитная» метаразметка (в НКРЯ)

1. «текст-стиль», при этом выделяются академический, научно-популярный, официально-деловой, нейтральный, сниженный, сниженный с элементами грубого просторечия и жаргона, архаизованный, индивидуально-авторский, диалектный и пр. (всего 21);
2. аудитория-возраст;
3. аудитория-уровень образования;
4. аудитория-размер.

Программа метаразметки Systemic Coder

Systemic Coder - программа, облегчающая процесс метаописания корпуса текстов. Метаданные задаются на основе классификационной схемы.

Программа состоит из 5 интерфейсов.

Text Segmentation: разметка границ между сегментами текста;

Scheme Management: настройка классификационной схемы;

Coding: разметка текста;

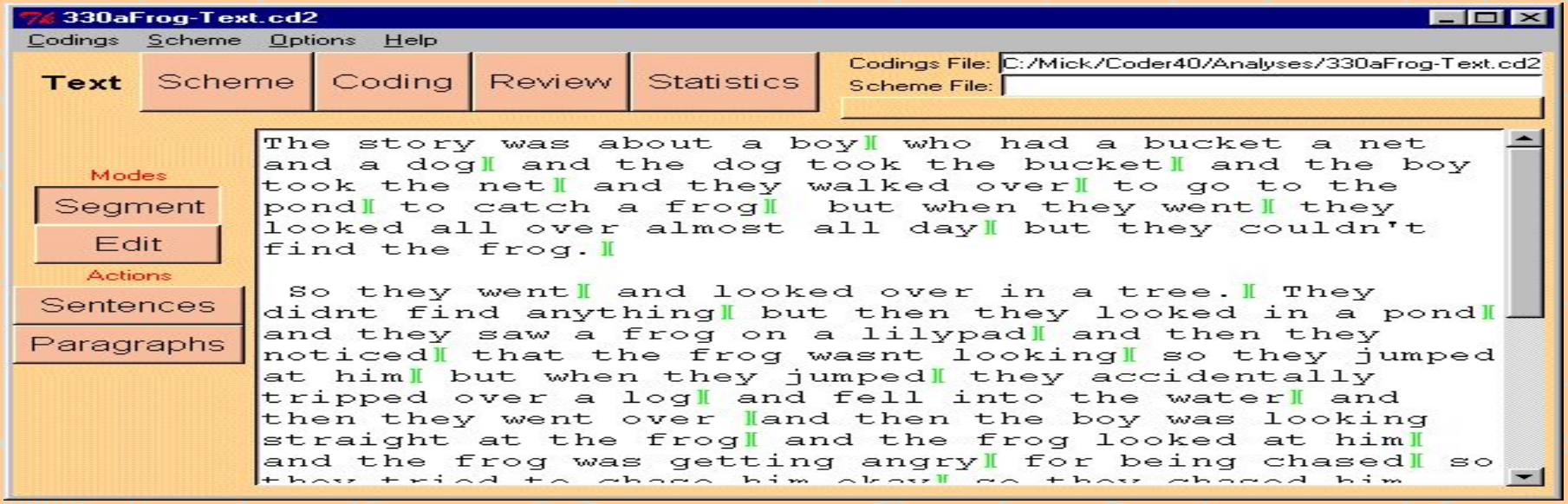
Review: просмотр размеченного текста;

Statistics: интерфейс, позволяющий получить описательную статистику о тексте, или разделить его на две или более совокупности и статистически их сравнить.

Деление текста на отдельные сегменты

Интерфейс *Разметки текста*. Текст, представлен в основном диалоговом окне - *текстовое окно*, слева расположен набор кнопок (*панель инструментов*).

Интерфейс метаразметки текста позволяет разделить загруженный текстовый файл на сегменты.

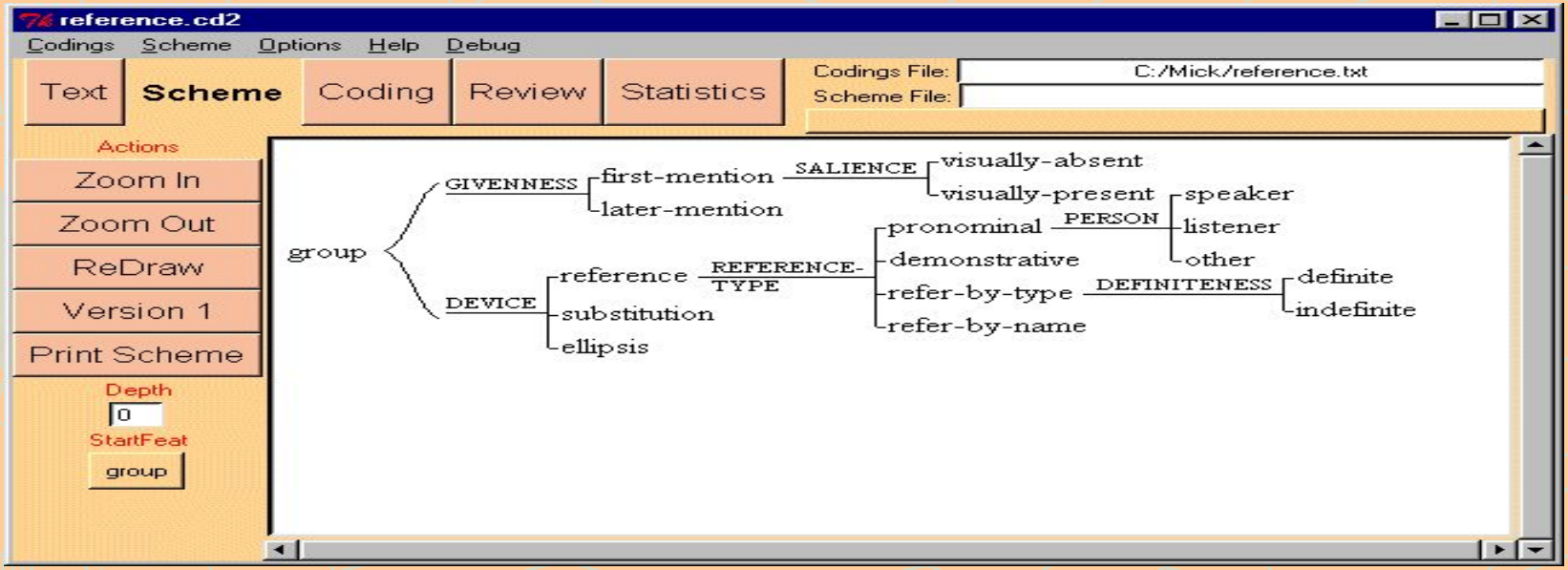


Классификационная схема

Классификация состоит из 3 частей:

- **имя (system name)**: идентификатор схемы;
- **признаки (features)**: варианты выбора;
- **условия ввода (entry-condition).**

Расширенная классификационная схема:



Создание и изменение классификационной схемы

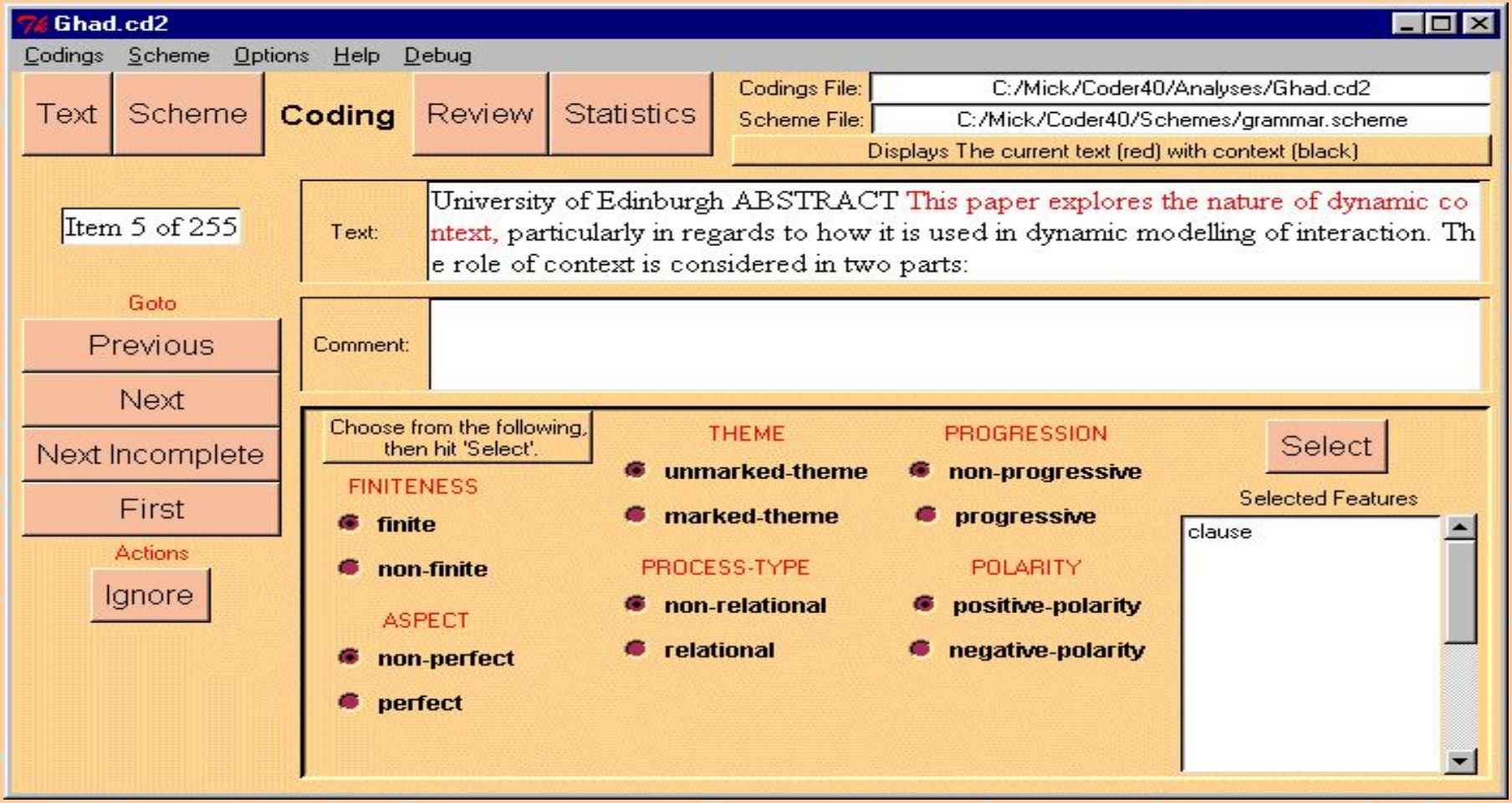
Управление классификационной схемой

- **Добавить признак (Add Feature)**: добавление нового признака в схему.
- **Переименование классификационной схемы (Rename System)**: изменение имени схемы.
- **Удаление классификации (Delete System)**: удаление классификации из схемы. Примечание: все признаки принадлежащие схеме и любая классификация, зависящая от нее будут также удалены. В настоящее время функция “Отменить” отсутствует.
- **Изменение условия ввода (Change Entry Condition)**: изменение условия ввода классификации с одного признака на другой.
- **Игнорировать/Не игнорировать подсхему (Ignore/Unignore Subnet)**: [New] Отключение классификации. Отключенная классификация выделена серым цветом. Она будет проигнорирована в кодировке и статистическом анализе.

Управление признаками

- **Добавить классификацию (Add System)**: создание макета классификации
- **Переименовать признак (Rename Feature)**: изменение имени признака
- **Удалить признак (Delete Feature)**: удаление признака. Примечание: все признаки, принадлежащие классификации и любая классификация, зависящая от нее будут также удалены. В настоящее время функция “Отменить” отсутствует
- **Редактирование примеров (Edit Realisations)**: [New] Вы можете добавить примеров, прикрепленных за признаками
- **Показать примеры (Show Examples)**: [New] Выбрав эту опцию вы перемещаетесь в интерфейс *Просмотра.*

Режим разметки



Интерфейс пользователя для поиска по метаданным:

Запросная форма НКРЯ для поиска по жанру текста:

- ✔ нежанровая проза
- ✔ автобиографическая проза
- ✔ детектив
- ✔ детская литература
- ✔ историческая проза
- ✔ криминальная литература
- ✔ приключения
- ✔ фантастика
- ✔ юмор и сатира

Интерфейс пользователя для поиска по метаданным:

Запросная формы НКРЯ для поиска по автору текста:

- Автор текста
- Пол:
  - ✔мужской
  - ✔женский
  - ✔любой
- Год рождения: от … до …