Технологии обработки языковых данных и документирование языков

План спецкурса

1. **Документирование малых языков как одна из главных задач полевой лингвистики (Week 1: Feb 12)**
   - 1.1. Что такое документирование языка
   - 1.2. Зачем нужна языковая документация?
     - 1.2.a. Срочность документирования языков, находящихся под угрозой.
     - 1.2.b. Какие языки спасать в первую очередь?
     - 1.2.c. Ценность первичного материала вне зависимости от теории.
   - 1.3. Типы языковых материалов, подлежащих документированию.
   - 1.4. Современные требования к языковому документированию.
   - 1.5. Организация процесса: участники, задачи, ресурсы, инструменты.
2. **Общие технологические проблемы документирования языков (Week 2: Feb 19)**
   - 2.1. Мультимедийные материалы.
     - 2.1.a. Необходимость мультимедийных материалов.
     - 2.1.b. Факторы, определяющие их качество: качество оборудования, условия и методика записи, протоколирование сеансов записи, разметка носителей.
     - 2.1.c. Форматы аудио- и видеофайлов, их объёмы и проблемы хранения.

♣ *ЛАБ: Рассчитать примерный объём корпуса*

   - 2.2. Текстовые (письменные) материалы. Проблемы систем транскрипции, шрифтов, кодировок, систем глоссирования.
   - 2.3. Общие вопросы стандартизации. Поддержка «смежных» стандартов.
   - 2.4. Интеграция ресурсов и инструментов.
3. **Компьютерное представление символов. Кодировки. Unicode. Форматирование текста (Week 3–4: Mar 5, 12)**
   - 3.1. [Повторение основ] Компьютерное представление информации. Принципиальное единство электронного представления программы и данных, данных разного вида (текст, изображение, звук, числа…).
   - 3.2. Представление букв. Кодировки ASCII, ISO-8859, КОИ-8. Совмещение разных кодировок (кодовые страницы).

♣ *ЛАБ: записать числами в ASCII строку «…»*

   - 3.3. Особые символы: управляющие символы; невидимые символы в MS Word.
   - 3.4. Нормализация текста: набор символов; пробелы и отступы; регистры. Использование стилей.

♣ *ЛАБ: Замена n пробелов/табуляций/пустых абзацев; удаление лишних пробелов в начале/конце абзаца; неразрывные пробелы. Макросы.*

♣ *ЛАБ: Написать библиотеку стилей.*

3.5. Проблемы со смешением кодировок. Средства борьбы (Кодировщик Лебедева, Штирлиц). Самодельные шрифты.

♣ *ЛАБ: Раскодировать фрагмент текста из багвалинской грамматики; e-mail.*

3.6. Unicode.

3.6.a. Идеология: «семантическое» кодирование (не начертание, а функция).

3.6.b. Количество символов, диапазоны Юникода. Проблемы Юникода.

3.6.c. Системы кодирования: UTF-8, 7, 16, 32; BE/LE.

3.7. Программы и шрифты, (не) поддерживающие Юникод.

3.7.a. Вставка символа: разные инструменты — разные результаты. Character Map, Insert Symbol, Alt + код (10), код (16) + Alt+X, назначение клавиш; BabelMap, Uniqoder, Keyman.

♣ *ЛАБ: записать числами в Юникоде строку «…»*

3.8. Шрифты для лингвистов. Самодельные шрифты. Старые (не-юникодовские) шрифты IPA. Современные шрифты.

3.9. Преобразования шрифтов. TECKit, ConsistentChanges; SILConverters.

4. Корпус глоссированных текстов (Week 5-6: Mar 19, 26)

4.1. Метаданные (сопутствующая информация). Стандарты OLAC, IMDI.

4.2. Разбиение текста на порции (предложения, ЭДЕ). Синхронизация с аудио и видео: необходимость отдельного слоя записи.

4.3. Представление глоссированных текстов. Обзор существующих практик. «Полужесткий» модульный стандарт.

4.3.a. Возможные и минимально необходимые слои информации.

4.3.b. Поморфемное глоссирование. Система разделителей. Унификация грамматических глосс.

4.3.c. Перевод. Различные версии перевода и их совмещение в одном поле. Типы комментариев.

4.3.d. Дискурсивная транскрипция (полная vs. упрощенная).

4.3.e. Особые проблемы: нулевые показатели, скрытые категории, аналитические словоформы, сложные слова, орфографические знаки.

♣ *ЛАБ: Отглоссировать фрагмент текста на изучаемом языке. Применить разные стили глоссирования (полный, сокращенный).*

4.4. Оформление глоссированного текста в текстовом редакторе.

4.5. Представление глоссированного текста в реляционной базе данных.

5. Toolbox: ведение словаря и автоматизация глоссирования (Week 7: Apr 2)

5.1. Назначение системы, история разработки Shoebox/Toolbox.

5.2. Toolbox как инструмент лексикографа. (Денис Паперно)

5.3. Автоматизация глоссирования.

5.3.a. Идеология глоссирования TB: глубинные формы в строке поморфемного членения. Проблемы с выведением поверхностных форм, способы борьбы.

5.3.b. Способы выбора правильного варианта разбора: задание контекста, принудительный выбор глоссы, Word Formulas. Проблемы.

♣ *ЛАБ: Отглоссировать фрагмент текста с применением формул.*

5.4. Экспорт данных из TB. Экспорт в RTF. Экспорт структурированных данных в XML. BoxReader/Writer.

6. **Фонетическая база данных (Week 8: Apr 9)**
   - 6.1. Назначение фонетической базы данных. Пример фонетической БД.
   - 6.2. Язык запросов.
   - 6.3. Подготовка материалов для БД.
7. **Интегрированная среда для документирования (Week 9-10: Apr 16, 23)**
   - 7.1. Идеология среды: независимость от ОС, Freeware, OpenSource, открытые форматы данных, обмен данными между приложениями.
   - 7.2. Компоненты среды, маршруты обмена данными.
     - 7.2.a. OpenOffice.org
     - 7.2.b. Java Runtime Environment
     - 7.2.c. Mozilla Firefox
     - 7.2.d. Toolbox
     - 7.2.e. BoxReader/Writer
     - 7.2.f. MannX, ELAN
     - 7.2.g. Tomcat
     - 7.2.h. MySQL
   - 7.3. OpenOffice.org против Microsoft Office
   - 7.4. Динамические веб-страницы. JSP, PHP.
8. **XML и преобразования структурированных данных (Week 11-12: May) (ADN / SA)**
   - 8.1. История форматов: HTML, SGML, XML, XHTML.
   - 8.2. XML как семейство языков. DTD, структура документов.
   - 8.3. Программы обработки XML: редакторы, парсеры, валидаторы.
   - 8.4. XML как формат обмена между приложениями.
   - 8.5. Преобразования XML-структур с помощью XSLT.
9. **Базы данных MySQL**