### Общие положения

Редактор синтаксических структур (**stred.exe**) позволяет:

* Осуществлять ручную синтаксическую разметку предложений для текстов, записанных в специальном XML-формате;
* Разделять и объединять эти тексты,
* Вводить дополнительную информацию (комментарии) для всего текста или отдельных предложений.
* редактировать файл описания допустимых морфологических признаков и синтаксических отношений.

При ручной разметке осуществляется проверка допустимости имен синтаксических отношений и признаков, допустимости сочетания признаков.

Если StrEd-у доступна база лингвистических знаний ЭТАП и модуль etap.dll, то становятся доступны следующие функции:

* автоматическая разметка русских и английских текстов;
* проверка правильности размеченных текстов с диагностикой ошибок

### Дополнительные возможности: разбиение текстового документа на фразы и конвертация в формат tgt

Для разбиения текстового документа на фразы и конвертация в формат tgt в редакторе структур следует включить режим предварительной разметки. Этот режим включается автоматически при открытии текстового файла с расширением txt или snt. Можно также переключить режим вручную нажатием кнопки 1 (рис.1).

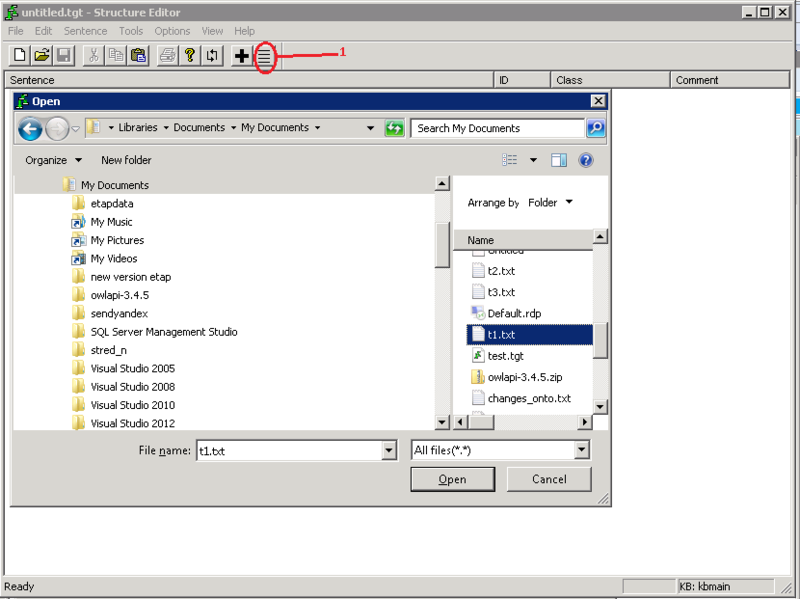


Рис. 1. Редактор структур в стандартном режиме

Документ при предварительном редактировании выводится в виде текста. Документ snt выводится без изменений, документ txt после считывания конвертируется. Редактирование текста производится стандартным образом, за исключением того, что начало предложения помечается специальным символом. Для разбиения предложения на части нужно поставить курсор в соответствующее место и нажать ”Enter”, для склейки нескольких предложений нужно удалить специальные символы. Для повторного запуска разбиения на предложения нужно нажать кнопку 1 (рис.2).

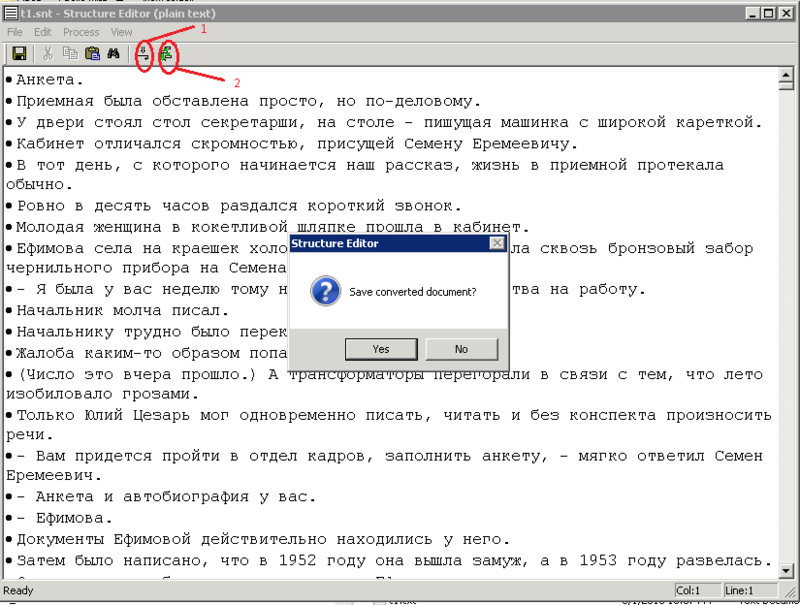


Рис. 2. Редактор структур в режиме предварительной разметки

Предварительная разметка документов включает в себя последовательность действий:

* Конвертирование размечаемого документа в формат snt (нормализация), включающее автоматическое разбиение его на фразы
* Исправление ошибок в размечаемом документе, в том числе ошибок разбиения на фразы
* Конвертирование полученного документа в формат tgt.

Документ snt является текстовым документом в кодировке UTF-8. Документ состоит из предложений, признаком конца предложения является символ перевода строки. Слова в предложениях разделяются пробелами или знаками препинания, несколько идущих подряд пробелов недопустимы. Символы – знаки препинания представляются единообразно: кавычки только символом “, но не « и т. д. Документ имеет расширение snt.

Конвертирование текстового документа в формат snt включает в себя:

* Замену символов перевода строки на пробелы
* Замену «нестандартных» символов — знаков препинания на стандартные
* Замену нескольких идущих подряд пробелов на один пробел
* Вставку символов перевода строки в местах, распознанных как конец предложения.

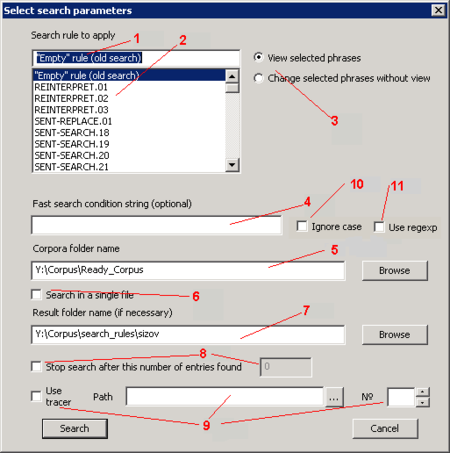
### Поиск по корпусу

Встроенная в Stred система поиска и модификации синтаксически размеченных фраз позволяет:

1. Использовать для поиска фраз по синтаксически размеченным документам запросы в старом формате
2. Использовать для поиска фраз запросы, записанные на языке FORET в специальных правилах обработки синтаксических структур. При таком поиске запрос в старом формате может быть использован для быстрого предварительного отсева неподходящих фраз.
3. Изменять с помощью соответствующих инструкций языка FORET найденные фразы либо в автоматическом, либо в интерактивном режиме.
4. Создавать из найденных (и, если нужно, измененных) фраз отдельный документ и выводить его в Stred для просмотра и редактирования. В этом случае к каждой фразе добавляется ссылка на ее исходное расположение. Возможно также сразу записывать изменения в исходные документы.
5. Вносить снабженные ссылками проверенные и отредактированные фразы в исходные документы.

Основной (правиловый) вариант поиска и замены устроен следующим образом – фразы из обрабатываемых документов поочередно загружаются в ЭТАП, потом к каждой фразе применяется одно или несколько выбранных правил. Если фраза удовлетворяет их условиям, то она либо изменяется, либо записывается в результаты поиска. В программе также реализован более быстрый но менее удобный и функциональный *старый”, поиск, то есть отбор размеченных предложений в результирующий файл по заданному напрямую шаблону без загрузки предложений в ЭТАП.*

Параметры поиска и замены задаются через диалог *Select search parameters*, вызываемый из *Главное меню -> Tools -> Corpus search and modify...*



В строке Search rule to apply (1) указывается, какими средствами будет осуществляться поиск и замена в документах. Это может быть

1. Одиночное правило из списка (2). Выбранное правило последовательно применится к каждой фразе из выбранных документов.
2. Правила из одноименной группы списка (2) - если в (1) указано только имя группы без номера. В этом случае к фразам документа применятся все правила, входящие в группу.
3. Поисковый запрос в старом формате – если в (1) выбрана зарезервированная команда “Empty” rule – old search.

Список (2) содержит правила из специализированного кортежа для правил обработки синтаксических структур. Эти правила создаются обычным способом – с помощью компилятора правил. Для написания поисковых правил в меню компилятора правил добавлен новый раздел *Corpus search and replace*, он находится в конце списка. После написания и компиляции правила готовы к использованию и появляются в данном списке

Радиокнопки (3) позволяют выбрать действия над «выбранными» фразами, к которым применились правила (или для которых выполнился поисковый запрос). При выборе пункта *View selected phrases* «выбранные» фразы запоминаются в специальном файле, причем для каждой фразы указывается документ – источник и порядковый номер в нем. После окончания процесса поиска этот файл автоматически открывается для редактирования, после которого окончательные варианты фраз вносятся в исходные документы с помощью функционала restore. При выборе пункта *Change selected phrases without view* фразы вносятся в исходные документы непосредственно после обработки правилами.

В строке *Fast search condition string* (4) записывается поисковый запрос в старом формате, который либо используется для предварительной фильтрации фраз, либо - если в (1) выбрана команда *“Empty” rule – old search* – осуществляет быстрый поиск. Например, в этой строке можно записать выражение *W/LEMMA=НОСИТЕЛЬ*, *W/KSNAME=НОСИТЕЛЬ1* и т.д.

В зависимости от значения чекбокса (6), в строке (5) задается либо папка, в которой непосредственно или опосредовано лежат синтаксически размеченные текстовые документы (галочка снята), либо имя одного документа, по котор(ым/ому) будет проводиться поиск и замена. По умолчанию в строке (5) указывается корневая папка корпуса «СинТагРус».

В строке (7) указывается имя папки, документами, куда записывается файл с отобранными фразами и файл с фразами, при обработке которых возникала ошибка.

Чекбокс (8) позволяет задать максимальное количество фраз, после обработки которых дальнейшая работа системы прекращается. Эта опция полезна для отладки правил и поисковых запросов, т.к. позволяет обнаружить в них ошибки, не прогоняя по всему массиву документов.

Чекбокс (9\*) позволяет протрассировать правила поиска и замены на конкретной фразе конкретного документа. При включенном чекбоксе в строке (9\*) задается имя документа а в (9\*) – номер фразы. Когда обработка дойдет до указанной фразы, будет вызван стандартный диалог установки трассировщика.

При включенном чекбоксе (10) поиск игнорирует разницу между заглавными и строчными буквами в значениях поисковых фильтров . При включенном чекбоксе (11) значения поисковых фильтров рассматриваются как регулярные выражения.

Поиск начинается после нажатия кнопки *Search*. В процессе обработки фраз в специальном окне интерактивно отображаются результаты обработки – имя обрабатываемого документа, процент просмотренных фраз, количество успешно обработанных фраз и количество фраз, при обработке которых возникли ошибки. После просмотра всех фраз в указанных документах в специальном диалоговом окне высвечиваются параметры, заданные для поиска и его результаты. Если в (3) была нажата радиокнопка *View selected phrases*, документ с отобранными фразами выводится в Stred для просмотра и редактирования, после которого входящие в него фразы фраз вносятся в исходные документы вызовом команды *Главное меню -> Tools ->Restore*. Также в отдельном документе сохраняются фразы, в которых возникли ошибки. Список этих фраз (если они есть) также выводится пользователю.

### Быстрый поиск

**«Быстрый поиск»** позволяет до загрузки предложения в ЭТАП предварительно отбирать предложения, к которым будет применяться правило. Стоит отметить, что на сегодняшний день на полном корпусе поиск работает крайне медленно, и использование «быстрого поиска» **существенно ускоряет** его работу. Мы настоятельно рекомендуем пользоваться предварительными фильтрами, когда это возможно. Синтаксис хорошо знаком тем, кто уже работал с поиском по корпусу. Например, чтобы работать только с предложениями, в которых есть лемма «ЧЕЛОВЕК», подходит строка быстрого поиска W/LEMMA=”ЧЕЛОВЕК”. Если нас интересует некоторое СинтО (например, 1 комплетивное), входящее в узел, то можно задать фильтр W/DOMLINK=1-компл. Аналогично, для указания СинтО, по которому к узлу присоединено зависимое слово, фильтр примет вид W/DEPLINK=1-компл. Подробнее см Tgtproc.exe.

### Написание правил поиска и замены

Правила могут описывать некоторый набор условий для отбора предложений. Условия могут включать проверку значений атрибутов слов и предложения в целом, которые пришли из разметки XML. Предложения, соответствующие описанным условиям, можно изменять с помощью инструкций, в том числе добавлять, удалять и модифицировать атрибуты, после чего можно сохранять изменения в предложении и/или записывать его в результирующий файл. Для этого добавлено несколько новых предикатов и инструкций. Одно из существенных отличий от привычного поиска состоит в том, что все действия с фразой определяются автором правила. То есть, если привычный поиск подсвечивает (устанавливает атрибут «STATUS» равным "WARNING") найденные узлы и предложения автоматически, то теперь это должно быть предусмотрено правилом. Также можно отметить, что:

1. большинство мест предикатов и инструкций для проверки разметки(за исключением контекстных переменных) заполняются строками, которые должны быть заключены в кавычки. В частности, имена и значения атрибутов задаются такими строками.
2. поскольку принудительная капитализация строк при трансляции словарей и правил отменена, имена и значения атрибутов нужно записывать в правильном реестре (“LEMMA” и т.п. – прописными буквами).
3. при сравнении атрибутов строчные и прописные буквы различаются. Исключение - специальный атрибут WORDFORM (значение - словоформа). Если при проверке значения словоформ требуется различать капитализацию, следует использовать атрибут CS\_WORDFORM.

### Зарезервированные атрибуты слова

Атрибуты, кодирующие характеристики слова, которые нельзя менять через **SETATTR:/SETATTRREG:**. При необходимости характеристики слова должны меняться соответствующими инструкциями, правильное значение атрибутов будет сгенерировано при записи структуры.

* ID - порядковый номер слова в предложении. Для изменения использовать инструкции PERGR...NOM и PERUZ...NOM
* DOM - номер слова-хозяина, LINK - имя связи. Для изменения использовать инструкции изменения связей
* FEAT - морфологические характеристики слова, EXTRAFEAT - остальные характеристики, используемые ЛП ЭТАП-3. Для изменения использовать инструкции изменения характеристик
* KSNAME - имя статьи КС. Для изменения использовать инструкции изменения ссылок на статьи КС
* PBEF и PAFT - строки, соответствующие знакам препинания перед (PBEF) и после (PAFT) слова во внутреннем представлении фразы. В общем случае (очень редко) не совпадает со знаками препинания во входном тексте. Для изменения использовать инструкции изменения знаков препинания.

Атрибуты, кодирующие характеристики слова, которые можно менять только через **SETATTR:/SETATTRREG:**.

* LEMMA - основная форма слова. Если не задана явно, то для опознанных слов, статьи которых есть в КС, совпадает с именем статьи КС без индекса, для слов, статьи которых есть в МС, но нет в КС, совпадает с именем статьи МС без индекса, для неопознанных слов совпадает с капитализированной словоформой.
* MSNAME - Если слова нет в КС, но оно есть в МС, атрибут равен имени статьи.
* WORDFORM - словоформа в исходном тексте без различения капитализации.
* CS\_WORDFORM - словоформа в исходном тексте c различением капитализации.
* WORDBEFORE - знаки препинания перед словом в исходном тексте.
* WORDAFTER - знаки препинания после слова в исходном тексте.