А.В.Сиренко

АВТОМАТИЗИРОВАННАЯ СИСТЕМА НАУЧНЫХ ИССЛЕДОВАНИЙ ПСИХОЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ (АСНИ ПМ): РАБОЧАЯ ЭКСПЛУАТАЦИОННАЯ ВЕРСИЯ 2012 г.[1]

1. Описание системы

1.1. Назначение программного комплекса

«Автоматизированная система научных исследований психолингвистических моделей» (АСНИ ПМ) — это программный комплекс, являющийся основной компонентой «Информационной системы когнитивных экспериментов» (ИСКЭ). Комплекс представляет собой набор компьютерных программ, выполняющих самостоятельно или автоматизирующих построение событийно-статистической модели психолингвистических экспериментов и ее функционирование. Событийно-статистическая модель формируется в виде взвешенной контекстно-зависимой грамматики, компоненты которой – множества символов и правил, создаются в процессе импорта исходных данных и в результате действий пользователя, после чего хранятся в СУБД и внешних файлах табулированных и бинарных форматов. В процессе моделирования необходимые данные загружаются в хешированные структуры, хранящиеся в ОЗУ.

1.2. Архитектура системы

Программный комплекс АСНИ ПМ имеет клиент-серверную архитектуру, включает в себя программу «Терминал АСНИ ПМ», поддерживающую взаимодействие с пользователем и выполнение бизнес-логики системы, а также СУБД PostgreSQL, ответственную за хранение данных.

1.3. Уровень взаимодействия с пользователем.

Модули первого уровня предоставляют пользовательский интерфейс, осуществляют простейшие операции проверки введенных данных, и обращения к СУБД. Емкие операции обработки выполняются посредством вызова модулей уровня бизнес-логики. Внешний вид экранных форм, а также их функциональные возможности представлены в руководстве пользователя.

1.4. Уровень бизнес-логики.

Модули уровня бизнес-логики не имеют пользовательского интерфейса и не хранят данных между запусками системы. Как правило, выполняются в отдельном программном потоке в рамках приложения ЛКТ. Часть функций этого уровня выполняются СУБД через исполнение хранимых процедур.

Модуль подготовки лингвистических данных. Выполняет загрузку в базу лингвистических данных дополнительных словарей из внешних текстовых файлов (синонимов и словоформ).

[1] В данном материале приводятся результаты исследований, выполненных при поддержке гранта РГНФ №12-04-12039в

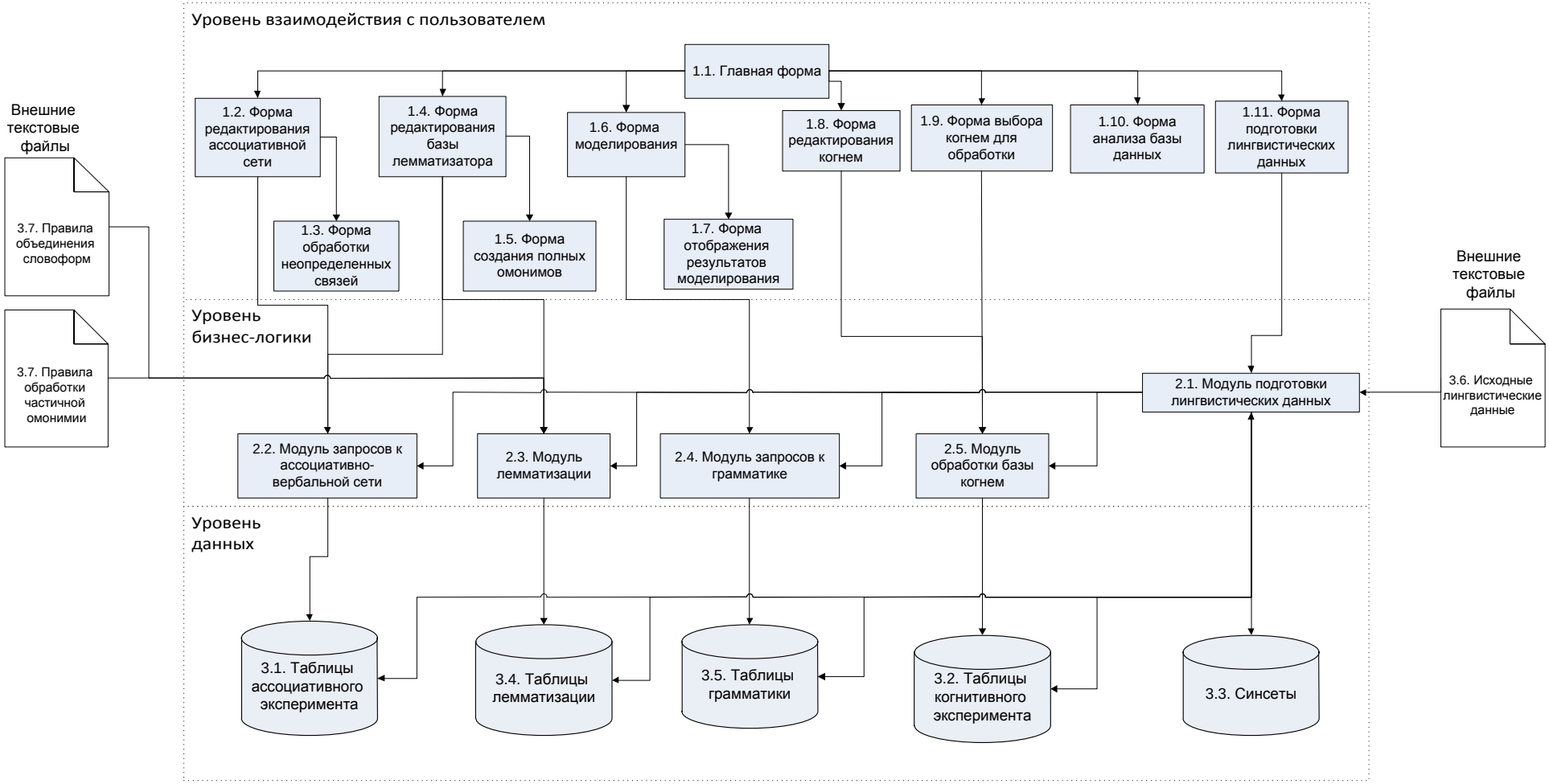

Рисунок 1. Архитектура терминала АСНИ ПМ.

Модуль генерации базы лемматизатора. Осуществляет поддержку действий «модуля лемматизации» уровня взаимодействия с пользователем. Функционирует в двух режимах:

- *Режим создания исходных данных* для лемматизации путем ручной обработки частичной омонимии и установки правил объединения словоформ. При этом создаются файлы последовательностей выполненных действий (скриптов) для их автоматического повторения.
- *Режим автоматической подготовки таблиц лемматизации*. Скрипты, созданные в ручном режиме, выполняются автоматически.

В ручном режиме входными данными являются управляющие команды пользователя, дополнительные лингвистические данные (исходные таблицы лемм и словоформ). Выходными данными являются сформированные окончательные таблицы лемматизации (дополнительные лингвистические данные), правила обработки частичной омонимии и правила объединения словоформ.

В автоматическом режиме входными данными являются дополнительные лингвистические данные (исходные таблицы лемм и словоформ), правила обработки частичной омонимии и правила объединения словоформ. Выходными данными являются сформированные окончательные таблицы лемматизации (дополнительные лингвистические данные).

Модуль подготовки ассоциативной сети. Осуществляет лемматизацию (применение готовых таблиц лемматизации, созданных модулем генерации базы лемматизатора), нормализацию ассоциативной сети.

Модуль поиска цепочек в ассоциативной сети. Выполняет поиск в ассоциативно-вербальной сети (далее АВС) по запросу модуля обработки фигур знания.

1.4. Уровень данных.

Уровень данных представлен как источниками информации, хранимыми в виде текстовых файлов (исходные лингвистические данные, правила обработки частичной омонимии, правила объединения словоформ), так и данными, работа с которыми обеспечивается реляционной СУБД.

2. Руководство пользователя Терминала «АСНИ ПМ» - ввод исходных данных, интеграция лексикографических объектов в базу данных.

Использование программного комплекса можно представить в виде четырех этапов:

1. ввод лингвистической информации из внешних файлов;
2. редактирование лингвистической информации;
3. построение событийно-статистической имитационной модели;
4. моделирование и просмотр результатов моделирования.

Руководство пользователя следует в своем содержании данному сценарию использования. Здесь мы рассмотрим этап подготовки ПО к использованию.

При запуске **«Терминала АСНИ ПМ»** отображается главная экранная форма приложения.

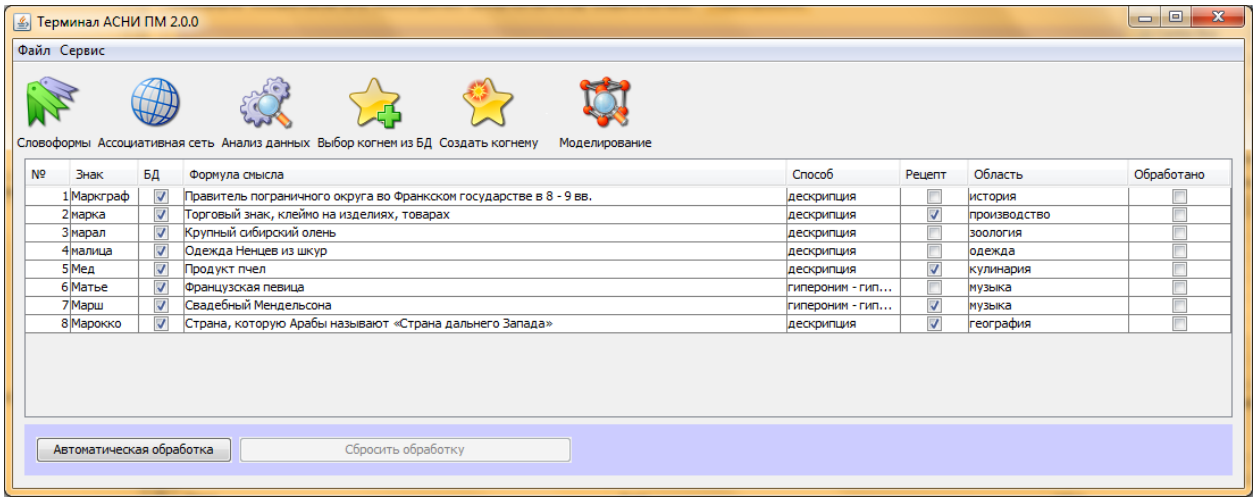

Рисунок 2. Главная форма.

Пятикомпонентная фигура знания (когнитивная единица) является ключевой структурой в программном комплексе, поскольку исходные данные моделирования представляются в форме когнем, даже если производится моделирование перехода от пропозиции к пропозиции. Главная форма отображает когнемы, выбранные для текущих операций. Когнемы могут быть представлены в базе данных – тогда они сохраняются между сеансами работы, а могут быть временными, «виртуальными» в терминах программного комплекса.
Флаг **«Обработано»** в таблице когнем показывает, переведены ли ЗНАК и ФОРМУЛА_СМЫСЛА когнемы в упорядоченное множество символов грамматики (обязательное требование для начала моделирования).
Кроме таблицы когнем главная форма содержит элементы управления:
Главное меню – ввод лингвистической информации, подготовка имитационной модели, выход из программы, анализ базы данных.
«Автоматическая обработка» - принудительный запуск токенизации фигуры знания.
«Словоформы» - редактирование базы лемматизатора.
«Ассоциативная сеть» - редактирование ассоциативной сети.
«Анализ данных» - просмотр содержимого базы данных, сохранение в текстовый файл.
«Выбор когнем из БД» - выбор из базы данных когнем для обработки.
«Создать когнему» - создание новой когнемы с возможностью сохранения в базе данных.
«Поиск с помощью грамматики» - ввод параметров моделирования.

2.1. Ввод лингвистической информации из внешних файлов.

Кратко ввод лингвистической информации был рассмотрен на этапе инсталляции программного обеспечения. В упрощенном виде он приемлем, если вы не планируете самостоятельно вносить изменения в исходные данные для построения модели, а желаете воспользоваться подготовленным материалом «как есть». В данной части рассмотрим процесс ввода более детально, учитывая возможность самостоятельной правки данных с помощью средств Терминала.
После запуска Терминала откройте форму **«Подготовка лингвистических данных».**
Вкладка **«Ассоциативная сеть»** содержит пункты меню:

- SR-пары;
- полная сеть;
- сократить сеть.

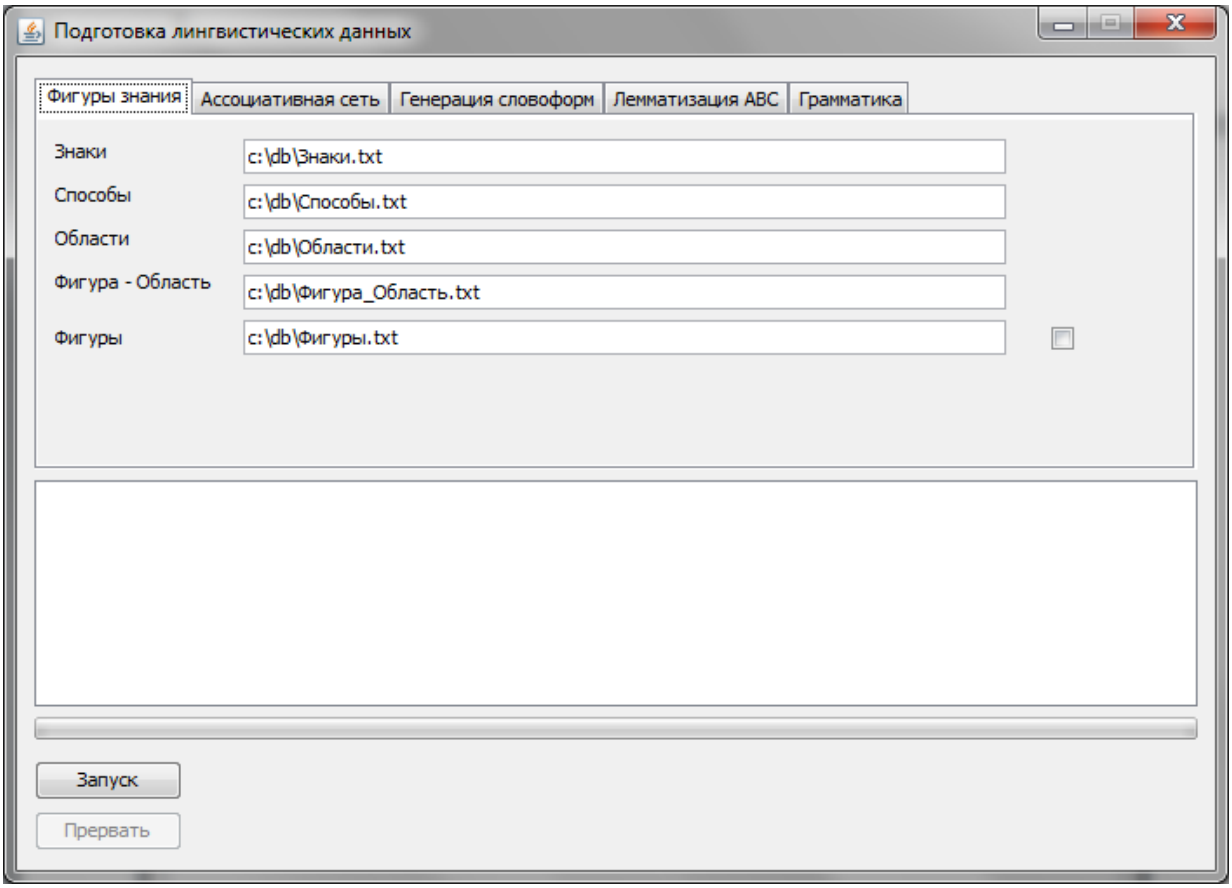

Рисунок 3. Форма подготовки лингвистических данных.

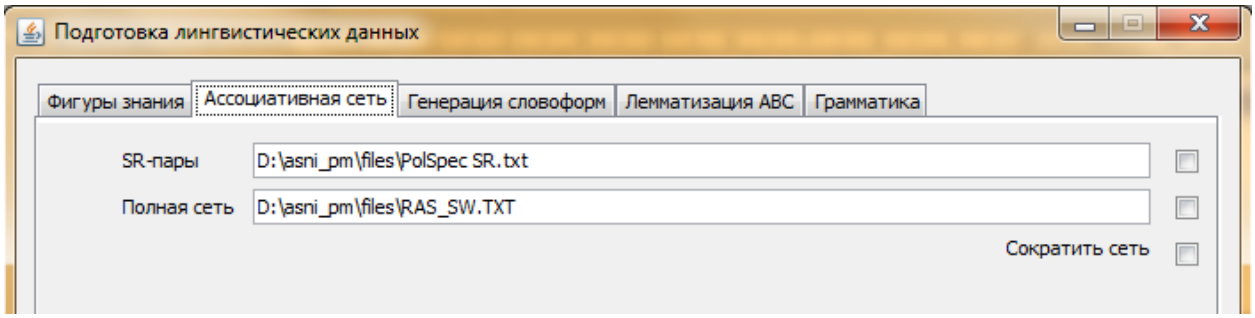

Рисунок 4. Ввод ассоциативной сети.

Выбор между **«SR-пары»** и **«Полная сеть»** является альтернативным. Импортируемые ассоциативные сети являются эквивалентными, незначительно отличающимися составом узлов и связей.

Пункт меню **«Сократить сеть»** позволяет убрать из сети единичные связи и те узлы, которые не имеют других связей, кроме единичных.

При построении сети из файла RAS_SW сокращенная сеть содержит 118 тыс. связей и 29 тыс. узлов (457 тыс. связей и 103 тыс. узлов в исходной сети). Работа с сокращенной сетью происходит быстрее и требует меньше аппаратных ресурсов, поэтому этот режим рекомендуется использовать при нехватке производительности.

Вкладка **«Генерация словоформ»** позволяет построить исходную базу словоформ лемматизатора на основе словаря iSpell. В программном комплексе используется база общеупотребительных слов словаря.

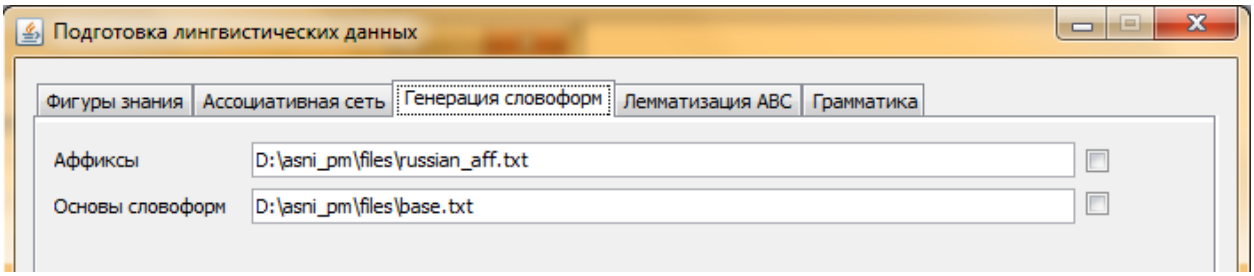

Рисунок 5 Ввод исходной базы словоформ

Пункт **«Аффиксы»** импортирует правила преобразования основ. Пункт «Основы словоформ» импортирует файл основ и производит их обработку на основе введенных ранее аффиксов. Обработка словоформ без введенных аффиксов приведет к остановке ввода данных по причине нехватки информации.
В результате генерации словоформ формируется: 1 298 673 словоформ, распределенных по 129332 леммам.
Вкладка **«Лемматизация АВС»** служит для обработки ассоциативной сети с помощью лемматизатора и построения двух ассоциативных сетей: лемматизированной и нелемматизированной.

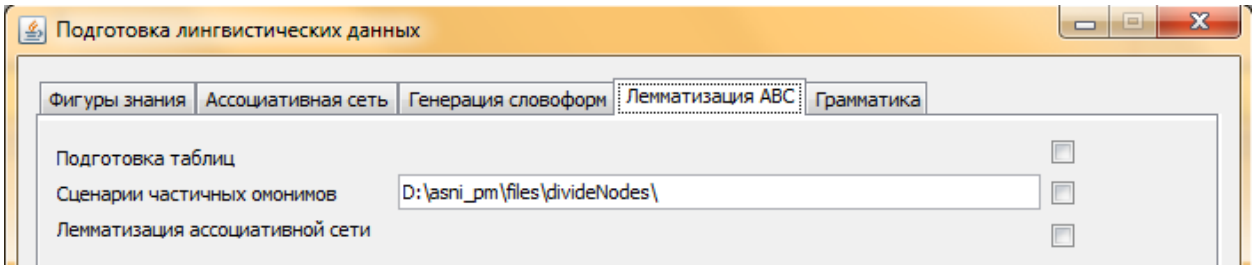

Рисунок 6. Лемматизация АВС

Пункт меню **«Подготовка таблиц»** выполняет: копирование ассоциативной сети, копирование исходных таблиц лемматизатора, назначение узлам ассоциативной сети словоформ. Копирование таблиц производится по причинам:

- структура ассоциативной сети для лемматизации отличается от исходной ассоциативной сети наличием полей, содержащих связи со словоформами лемматизатора;
- в процессе лемматизации автоматически, с помощью готовых скриптов, либо посредством диалоговых окон Терминала вносятся изменения в состав узлов и связей ассоциативной сети, а также таблиц лемматизатора.

Пункт **«Сценарии частичных омонимов»** выполняется при необходимости обработать частичную омонимию с помощью готовых скриптов. Набор скриптов может быть дополнен самостоятельно.
Пункт **«Лемматизация ассоциативной сети»** позволяет построить лемматизированную сеть по полной сети. Обязательным условием выполнения этой процедуры является однозначность определения лемм узлов сети. В случае неопределенности (например, согласно написанию, узлу соответствует несколько словоформ) необходимо обработать неоднозначные фрагменты сети.
После выполнения данных трех процедур над сокращенной ассоциативной сетью **RAS_SW** база данных программного комплекса содержит:

- Нелемматизированную сеть: **29 291** узлов, **132 456** связей.
- Лемматизированную сеть: **21 695** узлов, **123 719** связей.
- Множество лемм: **135 975** тыс. единиц.
- Множество словоформ: **1 306 134** единиц.

Вкладка **«Грамматика»** служит для построения вероятностной контекстно-зависимой грамматики на основе ассоциативной сети и дополнительных словарей.

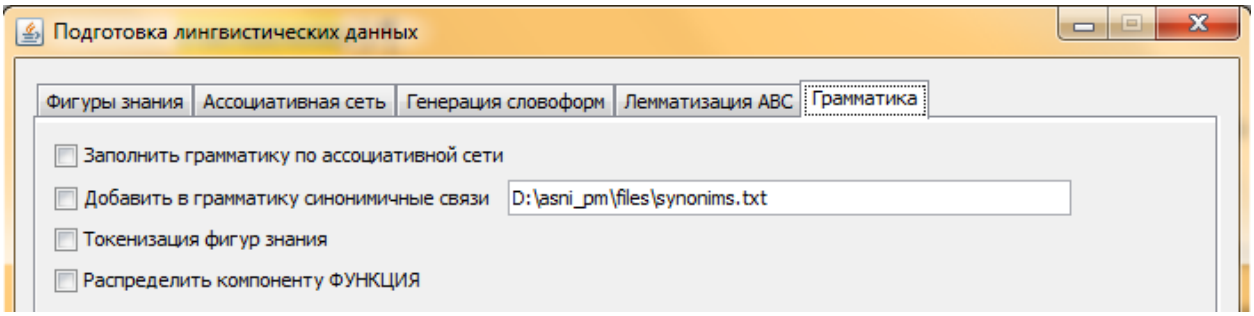

Рисунок 7. Построение грамматики

Пункт меню **«Заполнить грамматику по ассоциативной сети»** служит для первоначального построения таблиц символов и правил грамматики. При этом множество символов повторяет множество узлов ассоциативной сети, а ассоциативные пары формируют контекстно-свободные правила грамматики. Грамматика строится в двух вариантах: для лемматизированной и нелемматизированной сети.
Пункт меню **«Добавить в грамматику синонимичные связи»** позволяет дополнить грамматику правилами эквивалентных замен. Данные правила не имеют вероятностных характеристик. Для каждого синонимичного ряда (синсета) создается группа правил замены «от каждого к каждому». При необходимости, множество символов (а также таблицы лемматизатора) дополняются новыми элементами, входящими в синсет. Если в синсет входит лексема, которую нельзя однозначно сопоставить символу грамматики, в поле сообщений отображается запись о пропуске синсета. После обработки синсетов словаря синонимов Абрамова грамматика содержит 60 000 правил эквивалентных замен. В случае, если лемматизатору не удается однозначно идентифицировать словоформу, соответствующую элементу синсета, синсет игнорируется как неоднозначный, о чем сообщается в области уведомлений.
Пункт **«Токенизация фигур знания»** служит для назначения знакам и формулам смысла когнем в базе данных упорядоченных множеств символов грамматики. Токенизация позволяет оценить, какая часть естественно-языковой пропозиции может быть представлена грамматикой и используется при запуске моделирования. Главная форма терминала позволяет осуществить токенизацию активных фигур знания (задействованных в работе в текущий момент времени) перед процедурой моделирования. Вместе с тем, токенизация фигур знания необходима для ряда процедур построения грамматики, поэтому рекомендуется выполнять ее через форму подготовки лингвистических данных после построения грамматики и ввода синсетов.