Automatic text processing and deeply annotated text corpora of Russian: interaction and mutual impact

Leonid Iomdin
A.A.Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences
iomdin@iitp.ru

Plan

1. Annotated text corpora of Russian: what are they?
2. SynTagRus treebank: basic facts
3. SynTagRus and ETAP parser: living together

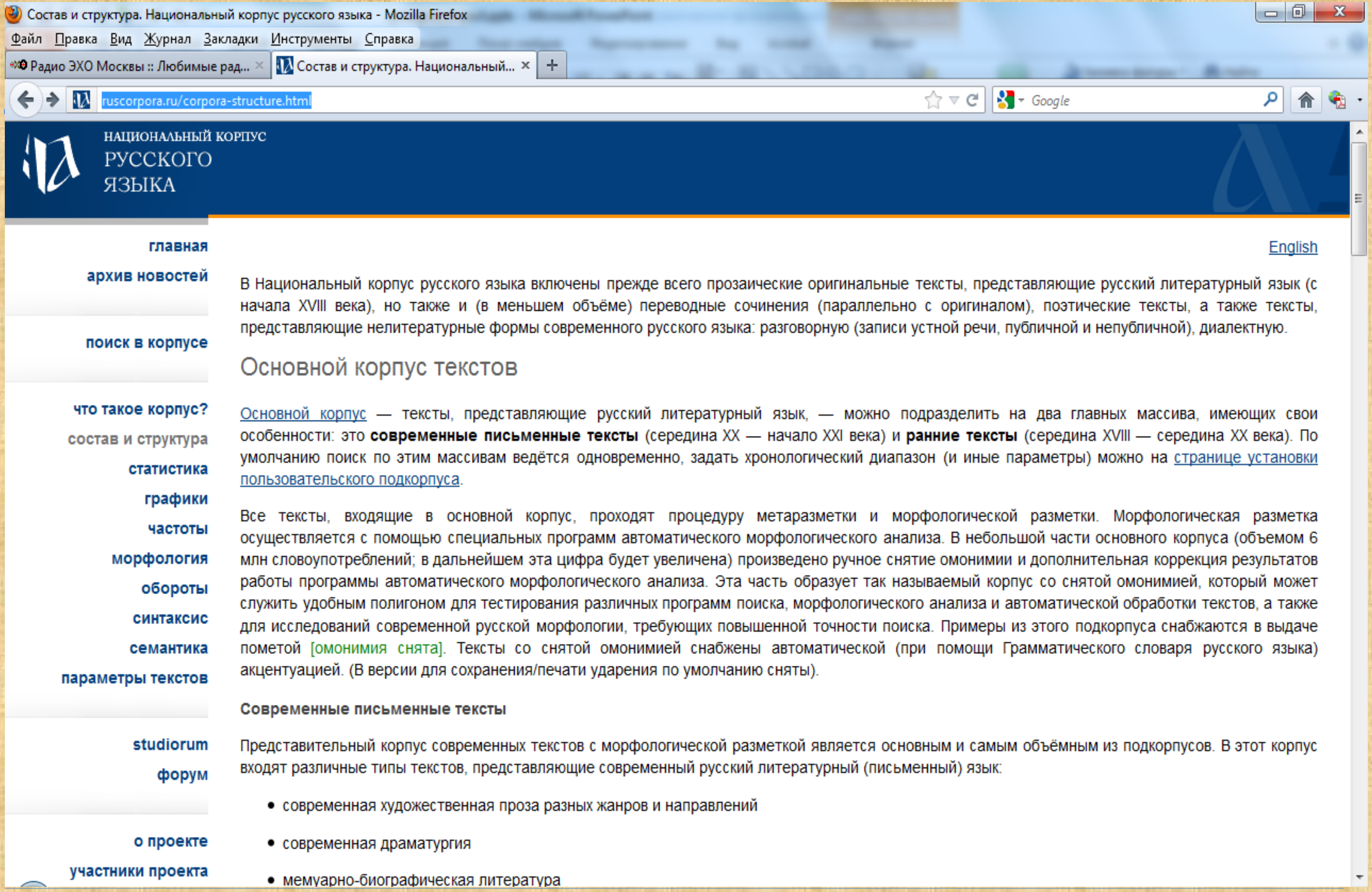
Summary

The main focus will be on SynTagRus, a corpus of Russian texts annotated with dependency-type syntactic structures, lexical meanings, and lexical functions. Statistical data collected from the corpus are used to improve lexical and syntactic disambiguation of automatic parsing. Other uses of SynTagRus include construction of dependency parsers by machine learning techniques and regression testing of the rule-based parser.

THE Russian corpus: National Corpus of Russian, НКРЯ

www.ruscorpora.ru
(hosted by Yandex, to be changed soon, non-commercial partnership to be founded)

National Corpus of Russian



В Национальный корпус русского языка включены прежде всего прозаические оригинальные тексты, представляющие русский литературный язык (с начала XVIII века), но также и (в меньшем объёме) переводные сочинения (параллельно с оригиналом), поэтические тексты, а также тексты, представляющие нелитературные формы современного русского языка: разговорную (записи устной речи, публичной и непубличной), диалектную.

Основной корпус текстов

Основной корпус — тексты, представляющие русский литературный язык, — можно подразделить на два главных массива, имеющих свои особенности: это **современные письменные тексты** (середина XX — начало XXI века) и **ранние тексты** (середина XVIII — середина XX века). По умолчанию поиск по этим массивам ведётся одновременно, задать хронологический диапазон (и иные параметры) можно на странице установки пользовательского подкорпуса.

Все тексты, входящие в основной корпус, проходят процедуру метаразметки и морфологической разметки. Морфологическая разметка осуществляется с помощью специальных программ автоматического морфологического анализа. В небольшой части основного корпуса (объемом 6 млн словоупотреблений; в дальнейшем эта цифра будет увеличена) произведено ручное снятие омонимии и дополнительная коррекция результатов работы программы автоматического морфологического анализа. Эта часть образует так называемый корпус со снятой омонимией, который может служить удобным полигоном для тестирования различных программ поиска, морфологического анализа и автоматической обработки текстов, а также для исследований современной русской морфологии, требующих повышенной точности поиска. Примеры из этого подкорпуса снабжаются в выдаче пометой [омонимия снята]. Тексты со снятой омонимией снабжены автоматической (при помощи Грамматического словаря русского языка) акцентуацией. (В версии для сохранения/печати ударения по умолчанию сняты).

Современные письменные тексты

Представительный корпус современных текстов с морфологической разметкой является основным и самым объёмным из подкорпусов. В этот корпус входят различные типы текстов, представляющие современный русский литературный (письменный) язык:

- современная художественная проза разных жанров и направлений
- современная драматургия
- мемуарно-биографическая литература

National Corpus of Russian

1. Main corpus (morphological annotation), over 300 million words. Subcorpus with resolved ambiguity: ("snjatnik"):
   7 million words
2. Syntactic corpus (morphological and syntactic annotation, lexical meanings and lexical functions): 770,000 words, over 52,000 sentences)

National Corpus of Russian

3. **Newspaper corpus**
4. **Parallel aligned corpora** (English-Russian, Russian-English, German-Russian, Ukrainian-Russian, Russian-Ukrainian, Belarussian-Russian, Russian-Belarussian, multilingual)

National Corpus of Russian

5. **Dialectal corpus**
6. **Poetic corpus**
7. **Learners’ corpus**
8. **Oral speech corpus**
9. **Accentological corpus**
10. **Multimedia corpus**
11. **Church Slavonic corpus**

SynTagRus

Syntactic Corpus of NRC and SynTagRus: the former is a subcorpus of the latter (is updated 3-4 times a a year, does not show lexical meanings or lexical functional annotation).

SynTagRus

The corpus is created semi-automatically: first, every sentence is processed by the ETAP parser and then manually corrected by at least two linguist experts.

SynTagRus

Currently the treebank contains over 52,000 sentences belonging to texts of a variety of genres (contemporary fiction, popular science, newspaper and journal articles dated between 1960 and 2012, texts of online news Wikipedia articles etc.) and is steadily growing.

SynTagRus

SynTagRus adopts a dependency-based annotation scheme, in a way parallel to the Prague Dependency Treebank but, in contrast, relying on the Meaning – Text theory by Igor Mel'čuk.

SynTagRus

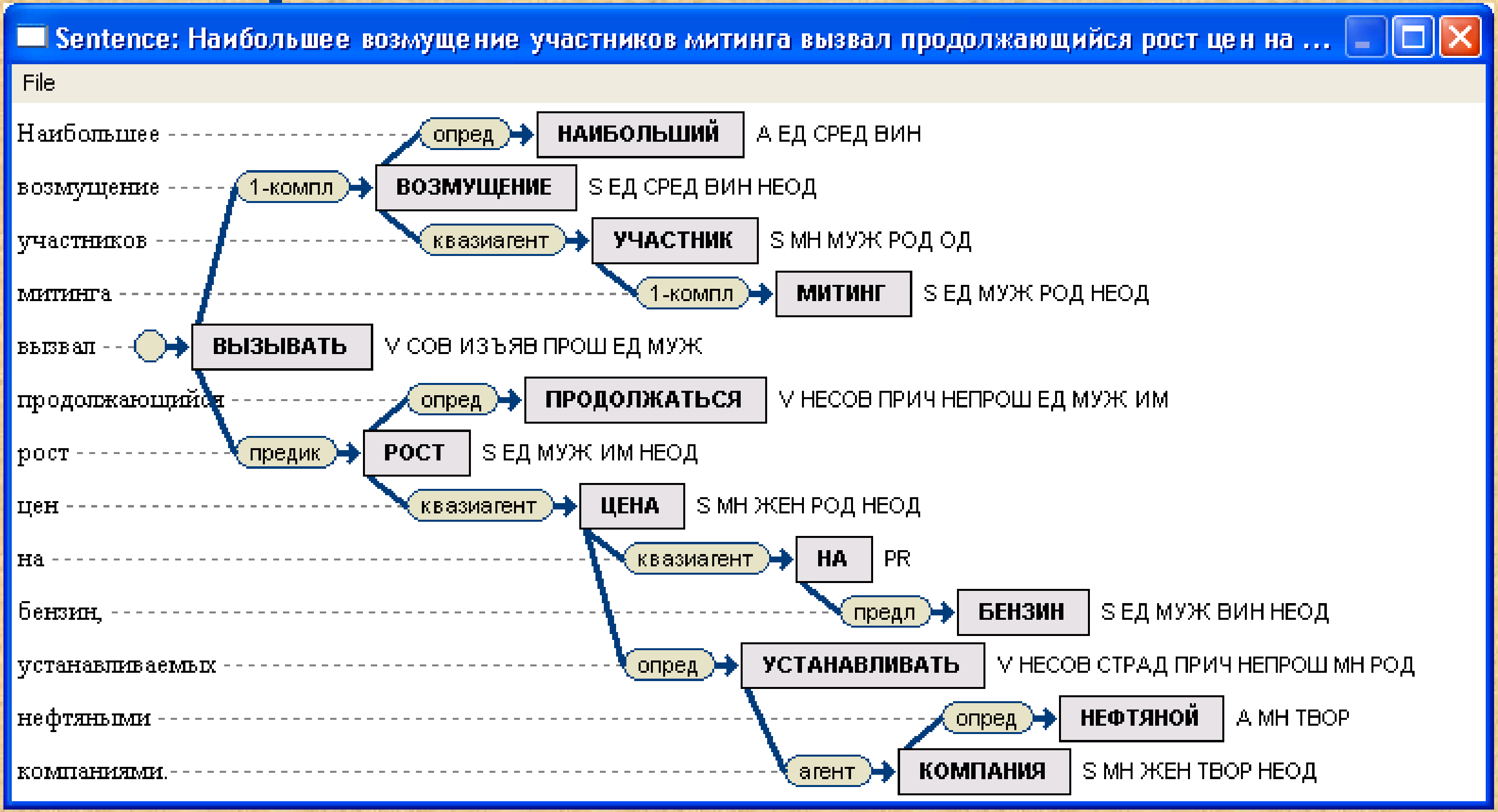
A sentence:

Naibol'šee vozmuščenie učastnikov mitinga vyzval prodolžajuščijsja rost cen na benzin, ustanavlivaemyx neftjanymi kompanijami

'It was the continuing growth of petrol prices set by oil companies that caused the greatest indignation of the participants of the meeting'.

SynTagRus

and its representation:



Sentence: Наибольшее возмущение участников митинга вызвал продолжающийся рост цен на ...

Наибольшее	опред	НАИБОЛЬШИЙ	A ЕД СРЕД ВИН
возмущение	1-компл	ВОЗМУЩЕНИЕ	S ЕД СРЕД ВИН НЕОД
участников	квазиагент	УЧАСТНИК	S МН МУЖ РОД ОД
митинга	1-компл	МИТИНГ	S ЕД МУЖ РОД НЕОД
вызвал		ВЫЗЫВАТЬ	V СОВ ИЗЪЯВ ПРОШ ЕД МУЖ
продолжающийся	опред	ПРОДОЛЖАТЬСЯ	V НЕСОВ ПРИЧ НЕПРОШ ЕД МУЖ ИМ
рост	предик	РОСТ	S ЕД МУЖ ИМ НЕОД
цен	квазиагент	ЦЕНА	S МН ЖЕН РОД НЕОД
на	квазиагент	НА	PR
бензин,	предл	БЕНЗИН	S ЕД МУЖ ВИН НЕОД
устанавливаемых	опред	УСТАНАВЛИВАТЬ	V НЕСОВ СТРАД ПРИЧ НЕПРОШ МН РОД
нефтяными	опред	НЕФТЯНОЙ	A МН ТВОР
компаниями.	агент	КОМПАНИЯ	S МН ЖЕН ТВОР НЕОД

SynTagRus

Nodes represent words (lemmas) assigned morphological and part-of-speech tags, whilst arcs are labeled with names of syntactic links. The tagging uses about 75 syntactic links, half of them proposed by Mel'čuk (1988).

SynTagRus

Normally, one token corresponds to one node in the dependency tree. There are certain exceptions:

- **composite words like *pjatidesjatiètažnyj* 'fifty-storeyed', where one token corresponds to two or more nodes;**
- **multiword expressions like *po krajnej mere* 'at least' where several tokens correspond to one node;**

SynTagRus

- so-called phantom nodes for the representation of hard cases of ellipsis, or gapping, which do not correspond to any particular token in the sentence (cf. *Ja kupil rubašku, a on galstuk* ‘I bought a shirt and he a tie’), which is expanded into *Ja kupil rubašku, a on kupilPHANTOM galstuk* ‘I bought a shirt and he boughtPHANTOM a tie’

SynTagRus and ETAP

Morphological Tagging of SynTagRus is based on a comprehensive morphological dictionary of Russian that counts about 130,000 entries (over 4 million word forms).

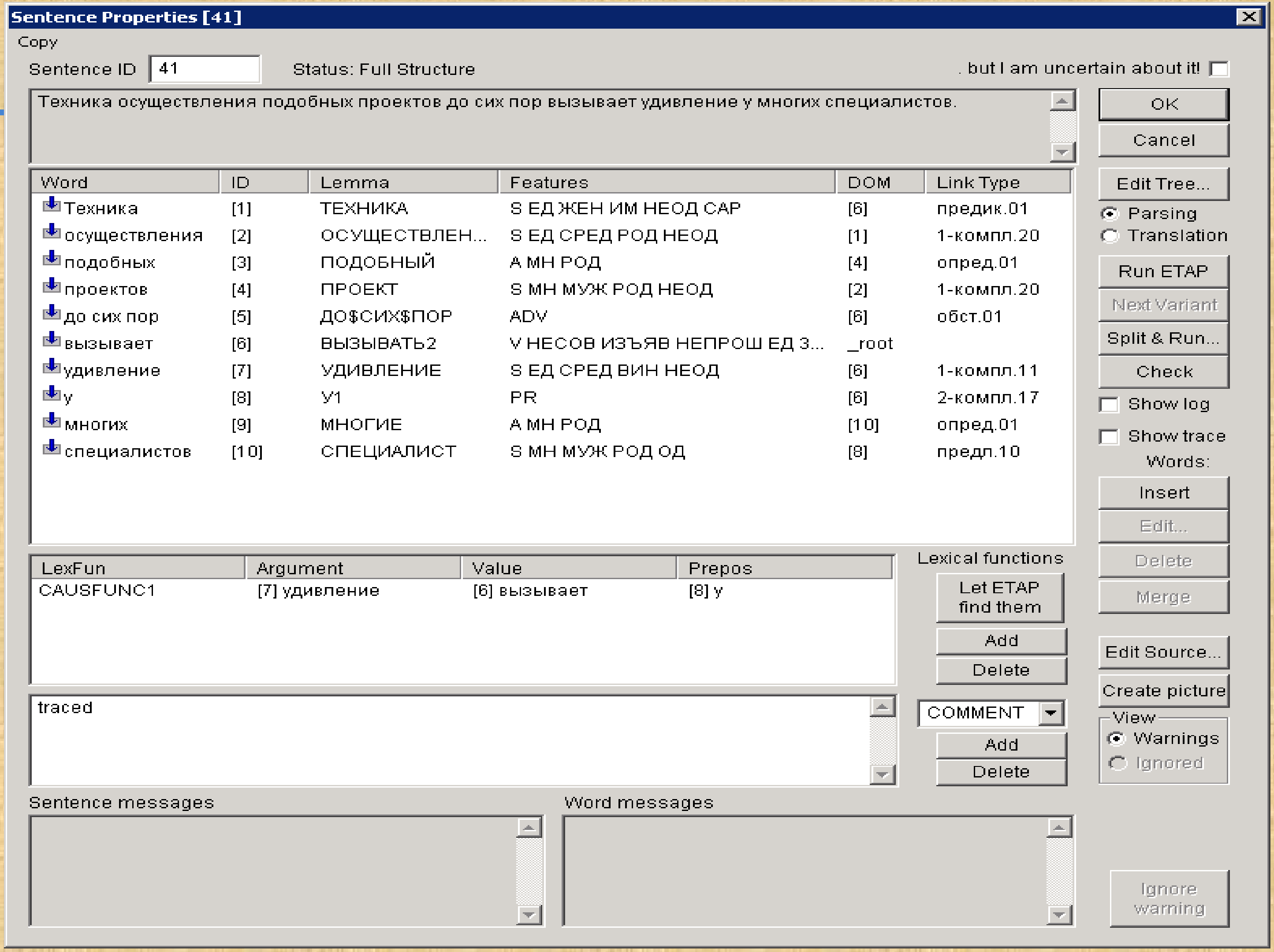
Recently, the dictionary was supplemented with full phonetic stress marking, which is used in a Russian text-to-speech synthesis system.

ETAP-3 morphological analyzer uses the dictionary to produce morphological annotation of words belonging to the corpus, which includes the lemma, POS tags, and, depending on POS, a set of morphological features.

SynTagRus and ETAP

The current version of SynTagRus contains partial lexical functional annotation. For collocations that could be presented with the apparatus of lexical functions, the tagging includes information on values and attributes of such lexical functions.

SynTagRus and ETAP



Sentence Properties [41]

Copy

Sentence ID: 41   Status: Full Structure   but I am uncertain about it! □

Техника осуществления подобных проектов до сих пор вызывает удивление у многих специалистов.

Word	ID	Lemma	Features	DOM	Link Type
Техника	[1]	ТЕХНИКА	S ЕД ЖЕН ИМ НЕОД САР	[6]	предик.01
осуществления	[2]	ОСУЩЕСТВЛЕН...	S ЕД СРЕД РОД НЕОД	[1]	1-компл.20
подобных	[3]	ПОДОБНЫЙ	A МН РОД	[4]	опред.01
проектов	[4]	ПРОЕКТ	S МН МУЖ РОД НЕОД	[2]	1-компл.20
до сих пор	[5]	ДО$СИХ$ПОР	ADV	[6]	обст.01
вызывает	[6]	ВЫЗЫВАТЬ2	V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3...	_root	
удивление	[7]	УДИВЛЕНИЕ	S ЕД СРЕД ВИН НЕОД	[6]	1-компл.11
у	[8]	У1	PR	[6]	2-компл.17
многих	[9]	МНОГИЕ	A МН РОД	[10]	опред.01
специалистов	[10]	СПЕЦИАЛИСТ	S МН МУЖ РОД ОД	[8]	предл.10

Lexical functions

LexFun	Argument	Value	Prepos
CAUSFUNC1	[7] удивление	[6] вызывает	[8] у

traced

COMMENT

● Parsing
○ Translation

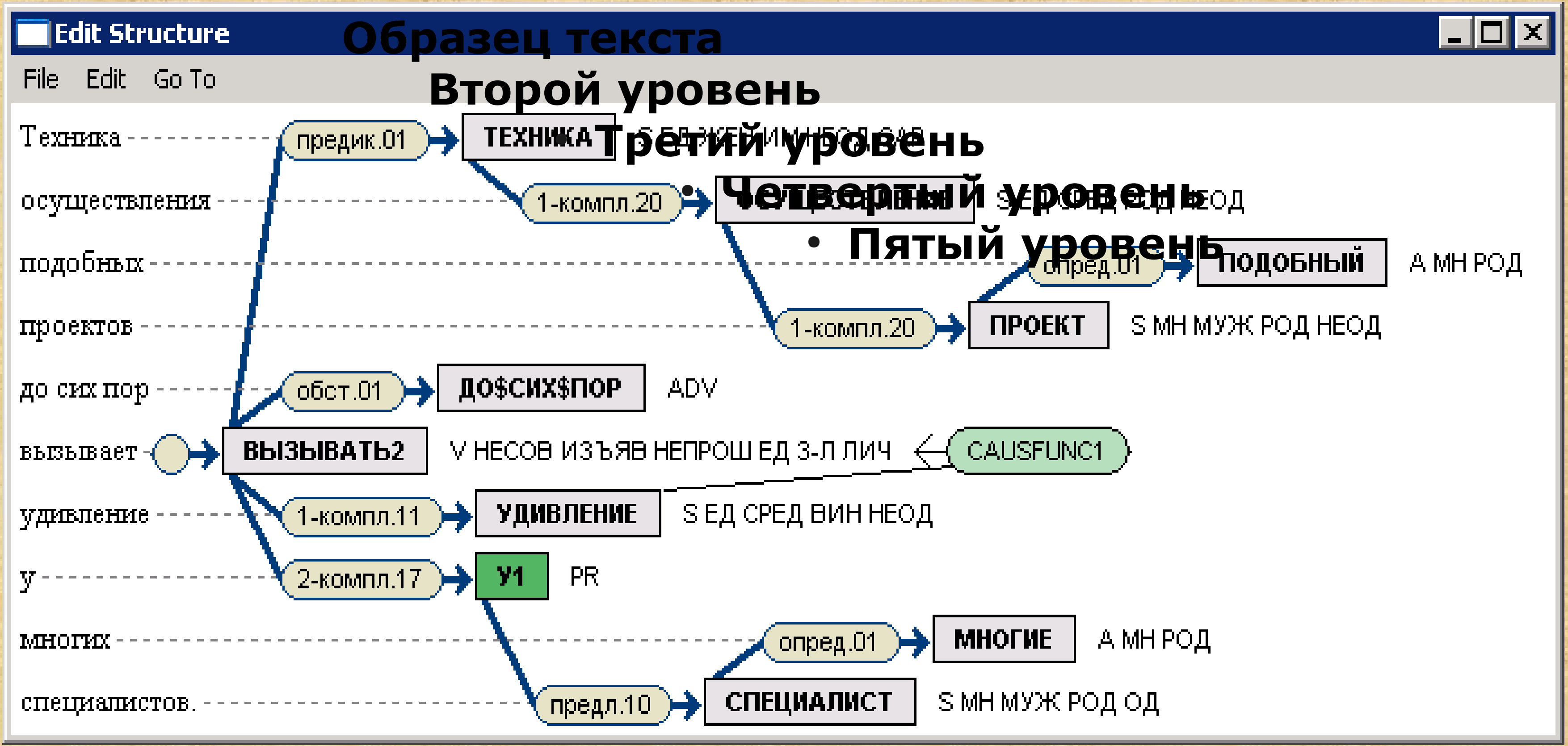
□ Show log
□ Show trace

View
● Warnings
○ Ignored

Sentence messages

Word messages

SynTagRus and ETAP



Образец текста

Второй уровень

Третий уровень

- Четвертый уровень
  - Пятый уровень

Техника	предик.01	ТЕХНИКА	S ЕД ЖЕН ИМ НЕОД [illegible]
осуществления	1-компл.20	ОСУЩЕСТВЛЕНИЕ	S ЕД СРЕД РОД НЕОД
подобных	опред.01	ПОДОБНЫЙ	A МН РОД
проектов	1-компл.20	ПРОЕКТ	S МН МУЖ РОД НЕОД
до сих пор	обст.01	ДО$СИХ$ПОР	ADV
вызывает		ВЫЗЫВАТЬ2	V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л ЛИЧ
удивление	1-компл.11	УДИВЛЕНИЕ	S ЕД СРЕД ВИН НЕОД
у	2-компл.17	У1	PR
многих	опред.01	МНОГИЕ	A МН РОД
специалистов.	предл.10	СПЕЦИАЛИСТ	S МН МУЖ РОД ОД

CAUSFUNC1

SynTagRus and ETAP

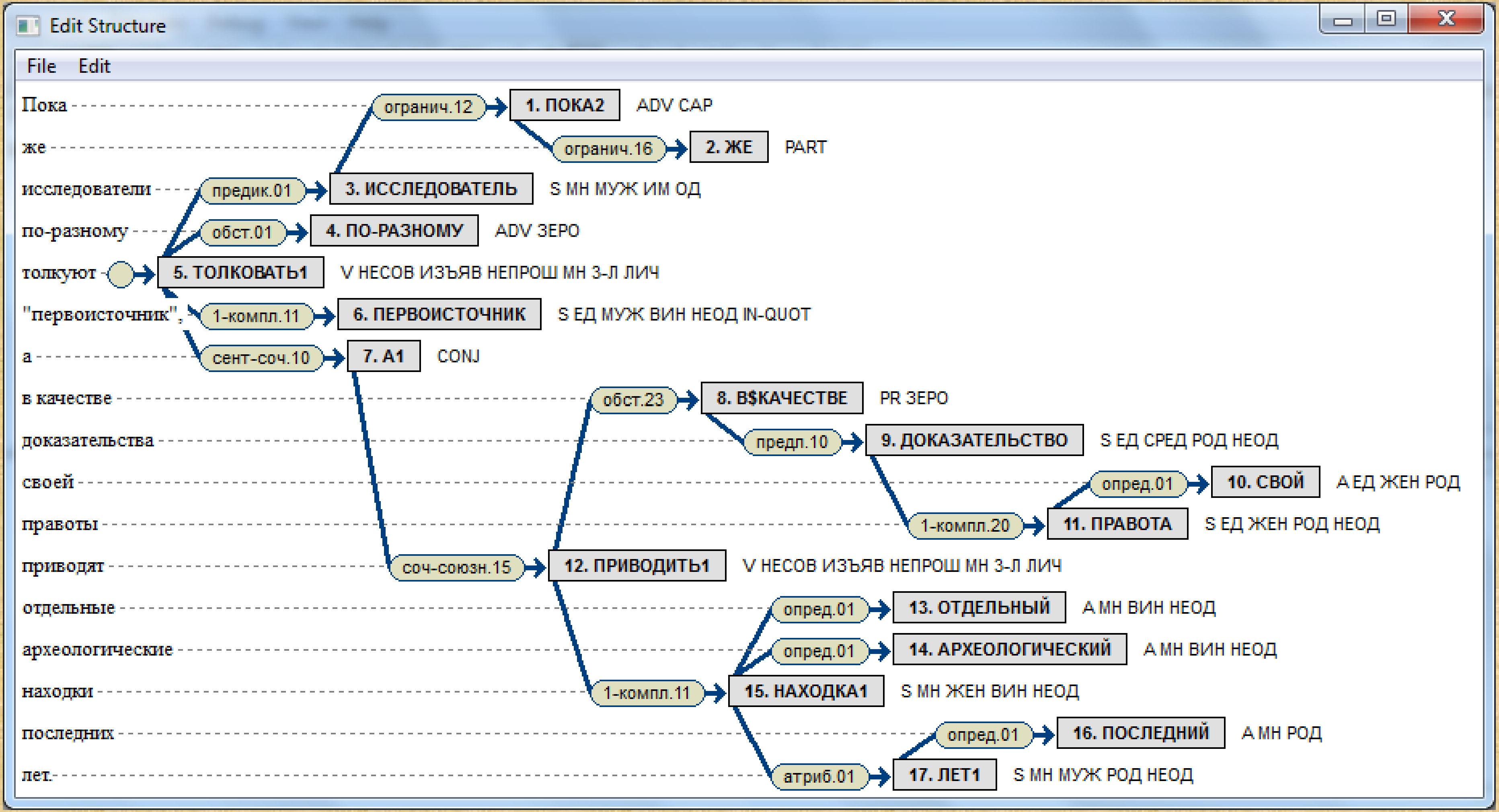
The current version of SynTagRus displays word senses as they are presented in the combinatorial dictionary of ETAP.

SynTagRus and ETAP

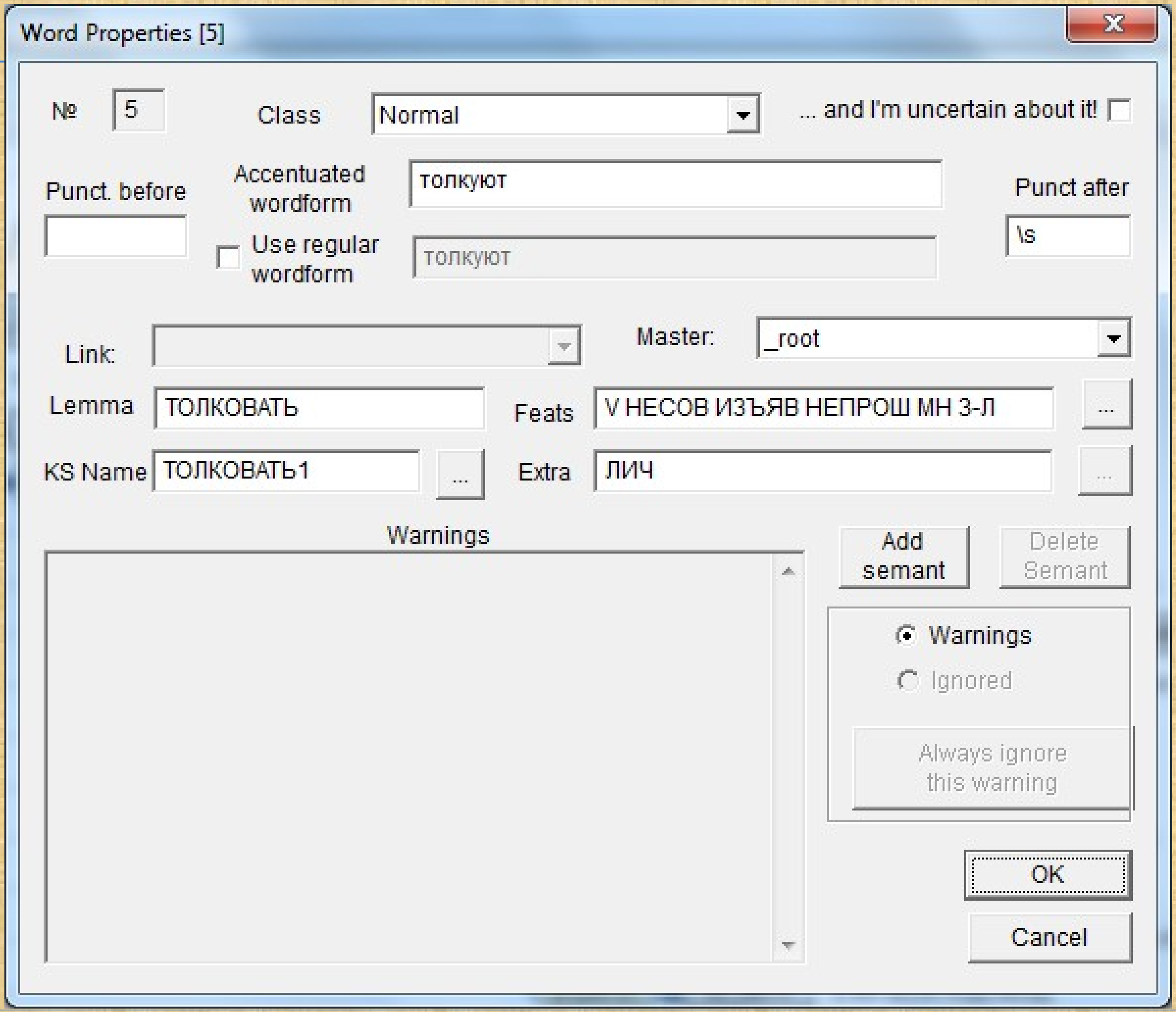
Пока же исследователи по-разному толкуют "первоисточник", а в качестве доказательства своей правоты приводят отдельные археологические находки последних лет.

‘So far, the researchers interpret differently the primary source, and as a proof of their being right they present isolated archeological findings of the recent years’

SynTagRus and ETAP



SynTagRus and ETAP



Word Properties [5]

№ 5

Class: Normal

... and I'm uncertain about it! □

Punct. before:

Accentuated wordform: толкуют

□ Use regular wordform: толкуют

Punct after: \s

Link:

Master: _root

Lemma: ТОЛКОВАТЬ

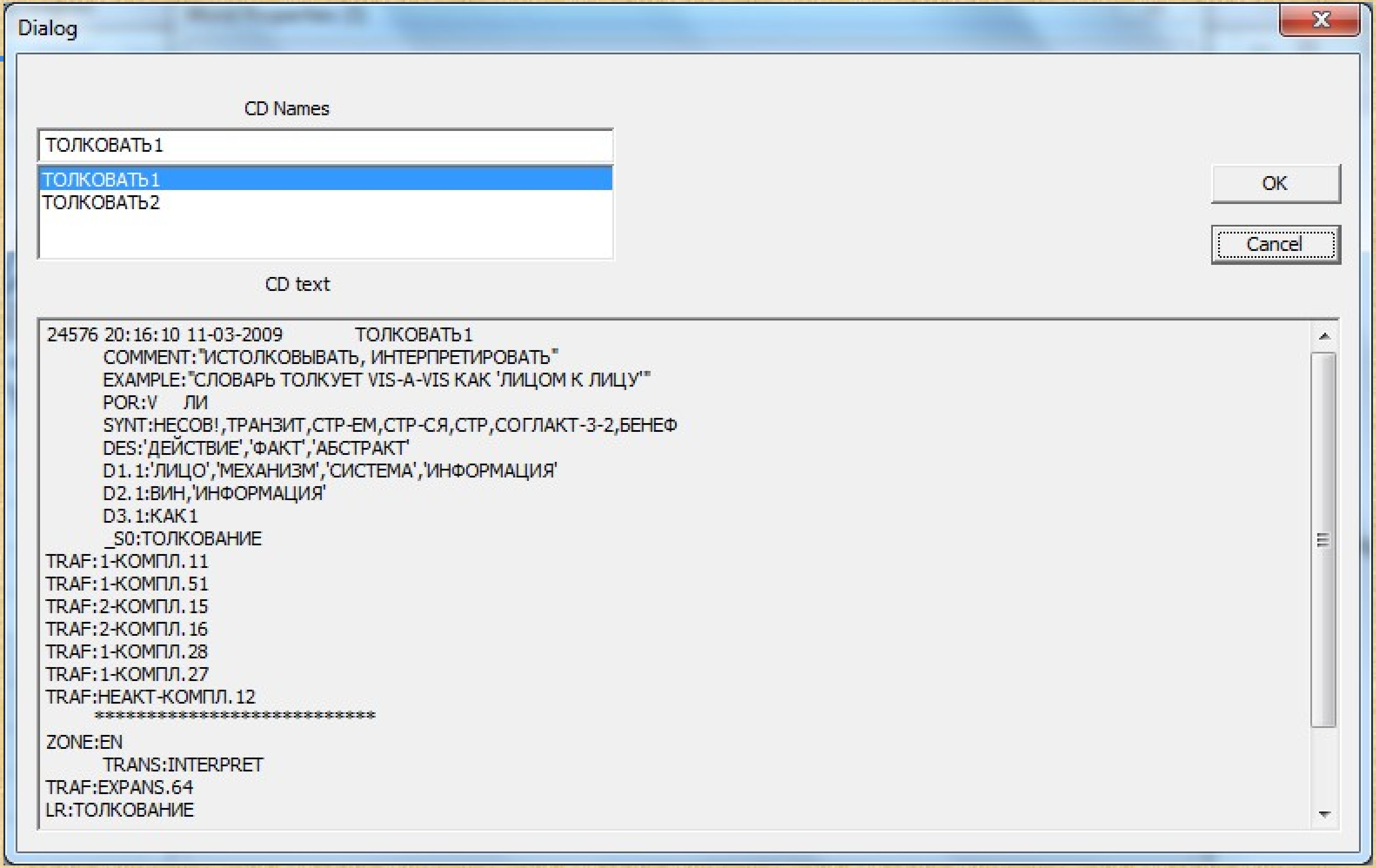
Feats: V НЕСОВ ИЗЪЯВ НЕПРОШ МН 3-Л

KS Name: ТОЛКОВАТЬ1

Extra: ЛИЧ

Warnings

SynTagRus and ETAP



ТОЛКОВАТЬ1

ТОЛКОВАТЬ1
ТОЛКОВАТЬ2

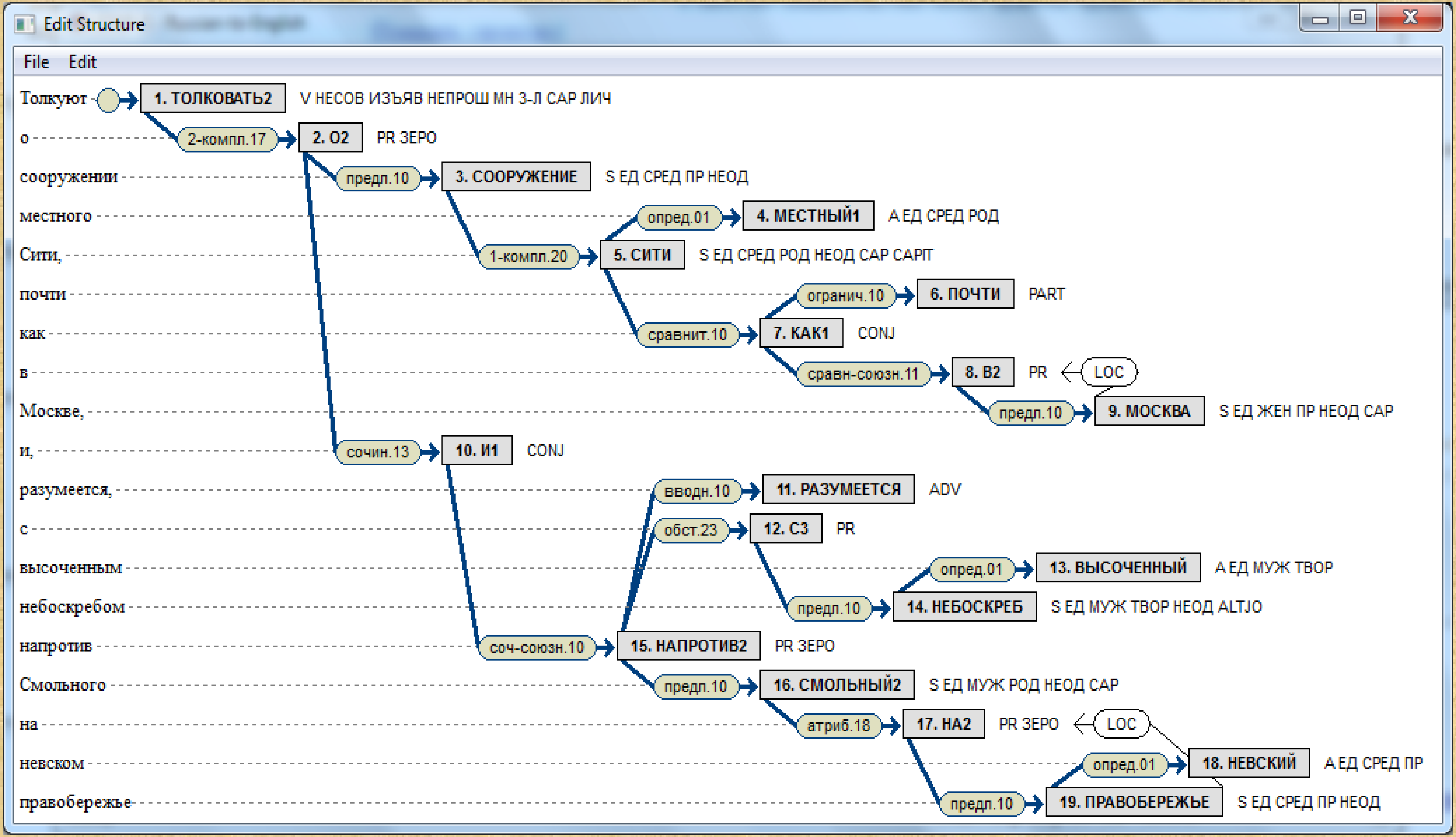
```
24576 20:16:10 11-03-2009          ТОЛКОВАТЬ1
        COMMENT:"ИСТОЛКОВЫВАТЬ, ИНТЕРПРЕТИРОВАТЬ"
        EXAMPLE:"СЛОВАРЬ ТОЛКУЕТ VIS-A-VIS КАК 'ЛИЦОМ К ЛИЦУ'"
        POR:V    ЛИ
        SYNT:НЕСОВ!,ТРАНЗИТ,СТР-ЕМ,СТР-СЯ,СТР,СОГЛАКТ-3-2,БЕНЕФ
        DES:'ДЕЙСТВИЕ','ФАКТ','АБСТРАКТ'
        D1.1:'ЛИЦО','МЕХАНИЗМ','СИСТЕМА','ИНФОРМАЦИЯ'
        D2.1:ВИН,'ИНФОРМАЦИЯ'
        D3.1:КАК1
        _S0:ТОЛКОВАНИЕ
TRAF:1-КОМПЛ.11
TRAF:1-КОМПЛ.51
TRAF:2-КОМПЛ.15
TRAF:2-КОМПЛ.16
TRAF:1-КОМПЛ.28
TRAF:1-КОМПЛ.27
TRAF:НЕАКТ-КОМПЛ.12
       ********************************
ZONE:EN
        TRANS:INTERPRET
TRAF:EXPANS.64
LR:ТОЛКОВАНИЕ
```

SynTagRus and ETAP

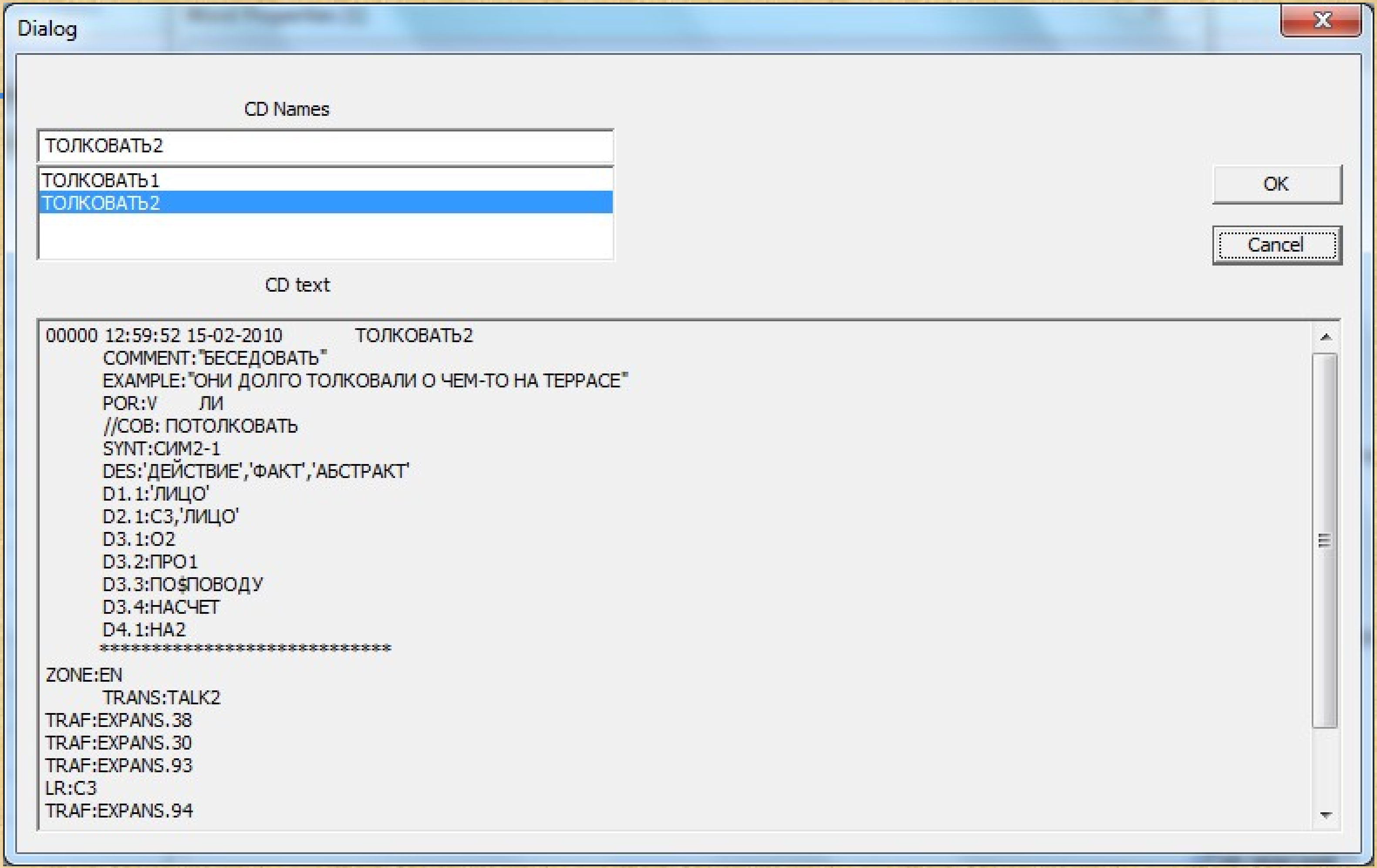
Толкуют о сооружении местного Сити, почти как в Москве, и, разумеется, с высоченным небоскребом напротив Смольного на невском правобережье.

‘They talk about the construction of a local City, almost like in Moscow and, naturally, with a very high skyscraper opposite Smolny on the Neva river bank.

SynTagRus and ETAP



SynTagRus and ETAP



CD Names

ТОЛКОВАТЬ2

ТОЛКОВАТЬ1
ТОЛКОВАТЬ2

CD text

```
00000 12:59:52 15-02-2010          ТОЛКОВАТЬ2
        COMMENT:"БЕСЕДОВАТЬ"
        EXAMPLE:"ОНИ ДОЛГО ТОЛКОВАЛИ О ЧЕМ-ТО НА ТЕРРАСЕ"
        POR:V      ЛИ
        //СОВ: ПОТОЛКОВАТЬ
        SYNT:СИМ2-1
        DES:'ДЕЙСТВИЕ','ФАКТ','АБСТРАКТ'
        D1.1:'ЛИЦО'
        D2.1:C3,'ЛИЦО'
        D3.1:O2
        D3.2:ПРО1
        D3.3:ПО$ПОВОДУ
        D3.4:НАСЧЕТ
        D4.1:НА2
        ********************************
ZONE:EN
        TRANS:TALK2
TRAF:EXPANS.38
TRAF:EXPANS.30
TRAF:EXPANS.93
LR:C3
TRAF:EXPANS.94
```

SynTagRus and ETAP

Syntagrus is not only a linguistic resource but also a computational resource which is used

- to collect various statistical data;
- to create training sets for machine learning;
- to develop automatic parsers (Nivre-Boguslavsky-Iomdin 2008)

SynTagRus beyond ETAP

- Over 30 free licenses provided to universities and academic institutions;
- 2 commercial licenses provided to big IT companies

Three main uses of SynTagRus within ETAP

1. It provides the statistics of the different syntactic constructions, lexical co-occurrences, patterns of ambiguities etc., which is used at several points of the algorithm if the statistical component is activated.

Three main uses of SynTagRus within ETAP

2. It serves as an efficient and accurate evaluation resource, which is used to evaluate the performance of ETAP parser and in this way find and resolve some of the system's bottlenecks.

Three main uses of SynTagRus within ETAP

3. It is used for regression testing of ETAP.

Regression Testing

- Periodically, ETAP is run on the whole corpus. Sentences that receive parses exactly equivalent to those stored in the corpus (between 30 and 35 percent of the bulk of the corpus) are selected as basis for regression testing.
- ETAP is then run on this test set to see if changes introduced in the dictionary, rules, or software affected the state of the test set.
- Regression testing has proven helpful in ensuring the stability of the parser and eventually improving it. Regression testing helps improve the SynTagRus itself: sometimes the discrepancies in parses detected by regression test runs point to erroneous annotation in the corpus, which is then corrected.