Processing of Russian by the ETAP-3 Linguistic Processor

Haga clic para modificar el estilo de subtítulo del patrón

Igor Boguslavsky
Institute for Information Transmission Problems, Russian Academy of Sciences / Universidad Politécnica de Madrid

ETAP-3

- A multipurpose NLP environment developed in the Institute for Information Transmission Problems, Russian Academy of Sciences
- Theoretical foundations:
  - Igor Mel’čuk. Meaning □ Text Linguistic Theory
  - Jury Apresjan. Systematic Lexicography and the Theory of Integral Description of Language

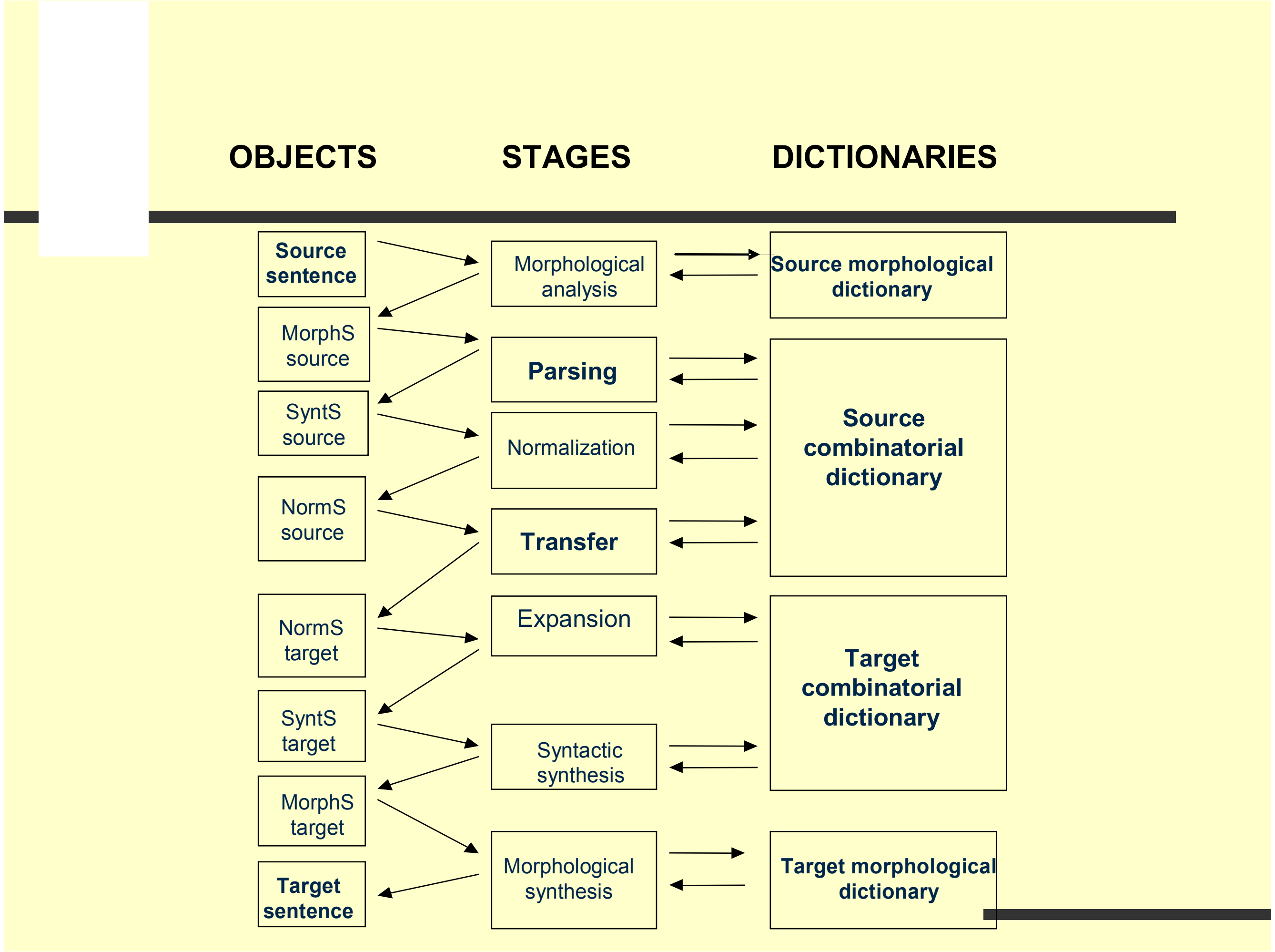
Motivation

- A non-commercial environment, primarily oriented at linguistic research, rather than the creation of a marketable software product.
- Main focus: computational modelling of natural languages.
  - NB: grammatical vs. non-grammatical
- Largest coverage: Russian and English

Major features

- Rule-based (with some statistical support of the parsing)
- Strict stratification
- Dependency syntactic structure
- Lexicalistic approach
- Self-tuning to the text processed
- Maximum reusability of modules and resources in different applications

General Architecture of Translation Process



OBJECTS

STAGES

DICTIONARIES

Source sentence

Morphological analysis

Source morphological dictionary

MorphS source

Parsing

SyntS source

Normalization

Source combinatorial dictionary

NormS source

Transfer

NormS target

Expansion

SyntS target

Syntactic synthesis

Target combinatorial dictionary

MorphS target

Morphological synthesis

Target sentence

Target morphological dictionary

Dictionaries

- Morphological dictionary (130,000+)
- Combinatorial dictionary
  - Lemma name
  - Syntactic features (out of 200+)
  - Semantic features (out of 40+)
  - Subcategorization frame (morphological, syntactic, semantic, lexical constraints)
  - Values of Lexical Functions
  - Rules of various types

Self-Tuning: Grammar vs. Dictionary

- General regularities: general rules applied to large lexical classes and resorted to in every sentence processing
  - Simple examples: Agreement Adj + N, N + V
- Less general regularities: Specific rules applied to closed lexical classes.
  - Simple example: composite numerals

ETAP options

- Machine Translation
- Ontology-based semantic analysis
- Interlingua-based analysis and generation (UNL)
- Paraphrasing on the basis of LFs
- Speech synthesis support
- Syntactic annotation of texts
- Grammar checking

Machine Translation

- **Russian □ English**
- Russian-French Prototype
- Russian-German Prototype
- Russian-Korean Prototype
- Spanish-English Prototype
- Arabic-English Prototype

Some salient features

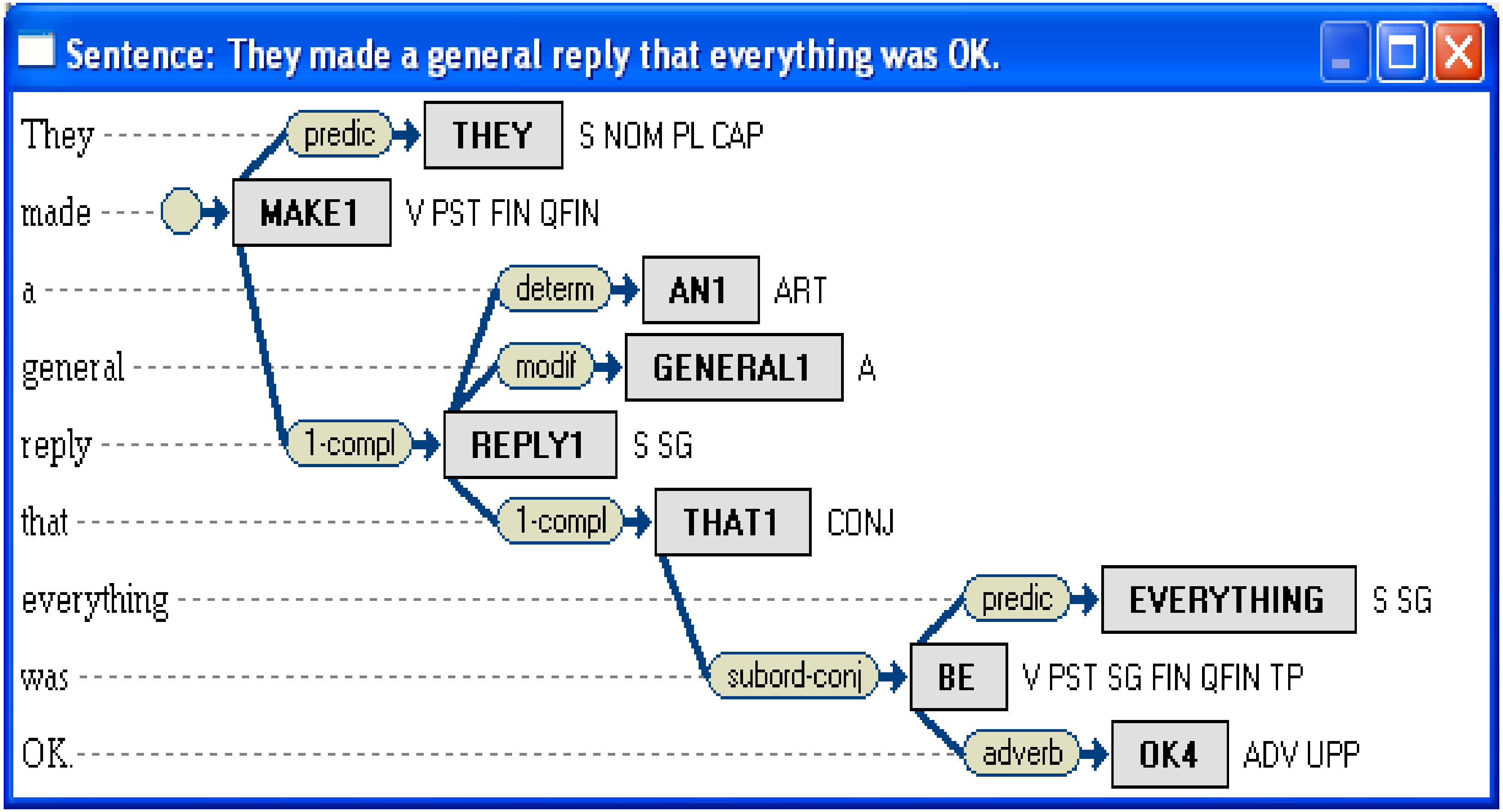
- Multiple parsing and translation facility
- Interactive disambiguation
- Extensive use of Lexical Functions for parsing, disambiguation, translation, paraphrasing.

Multiple Parsing/Translation

They made a general reply that …

- (a) ‘they replied in a general way that…’
- (b) ‘they forced a general to reply that…’

Syntactic Dependency Tree: First Option

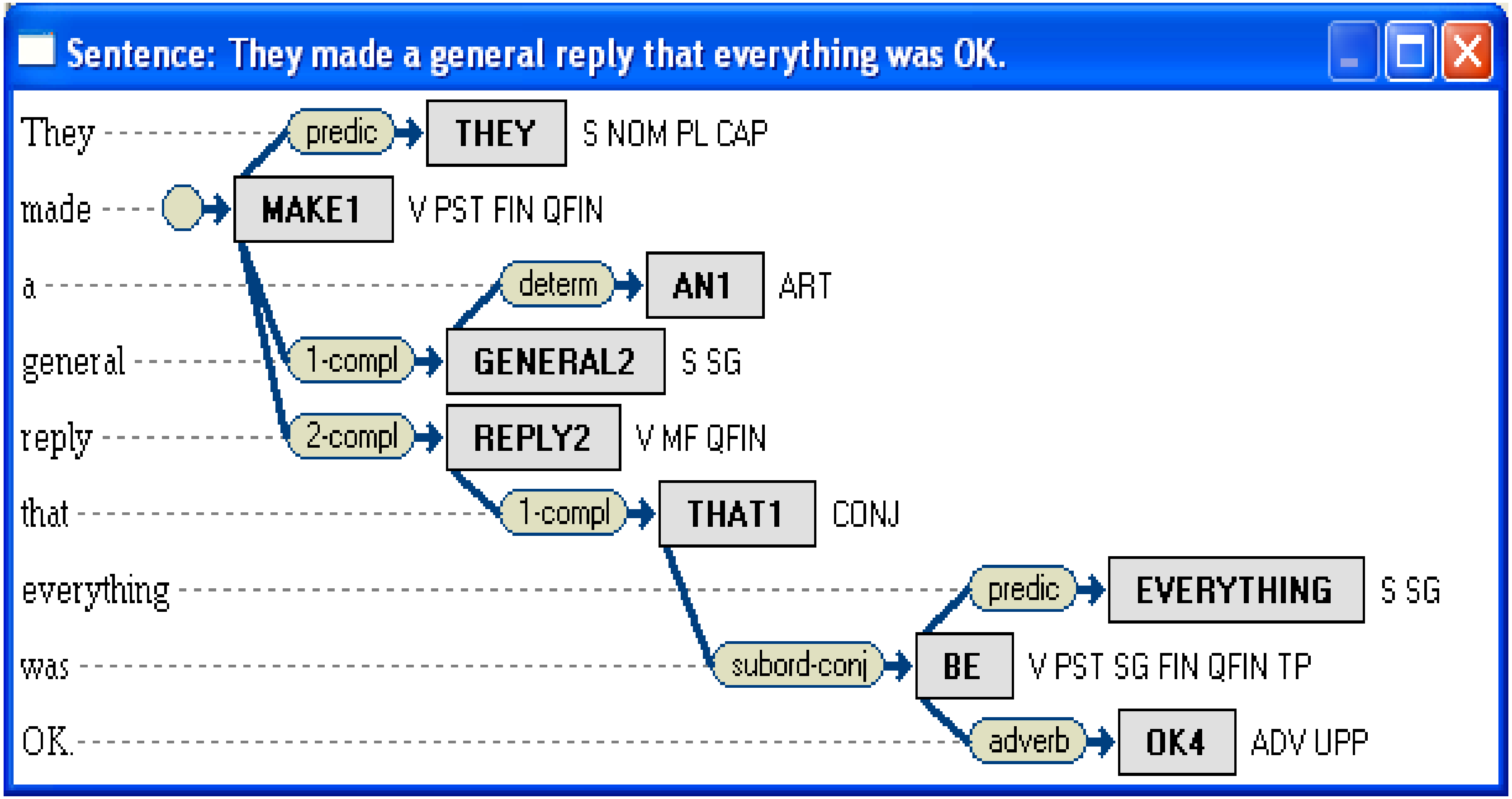


Sentence: They made a general reply that everything was OK.

Translation: First Option

Они дали общий ответ, что все было хорошо.

Syntactic Dependency Tree: Second Option



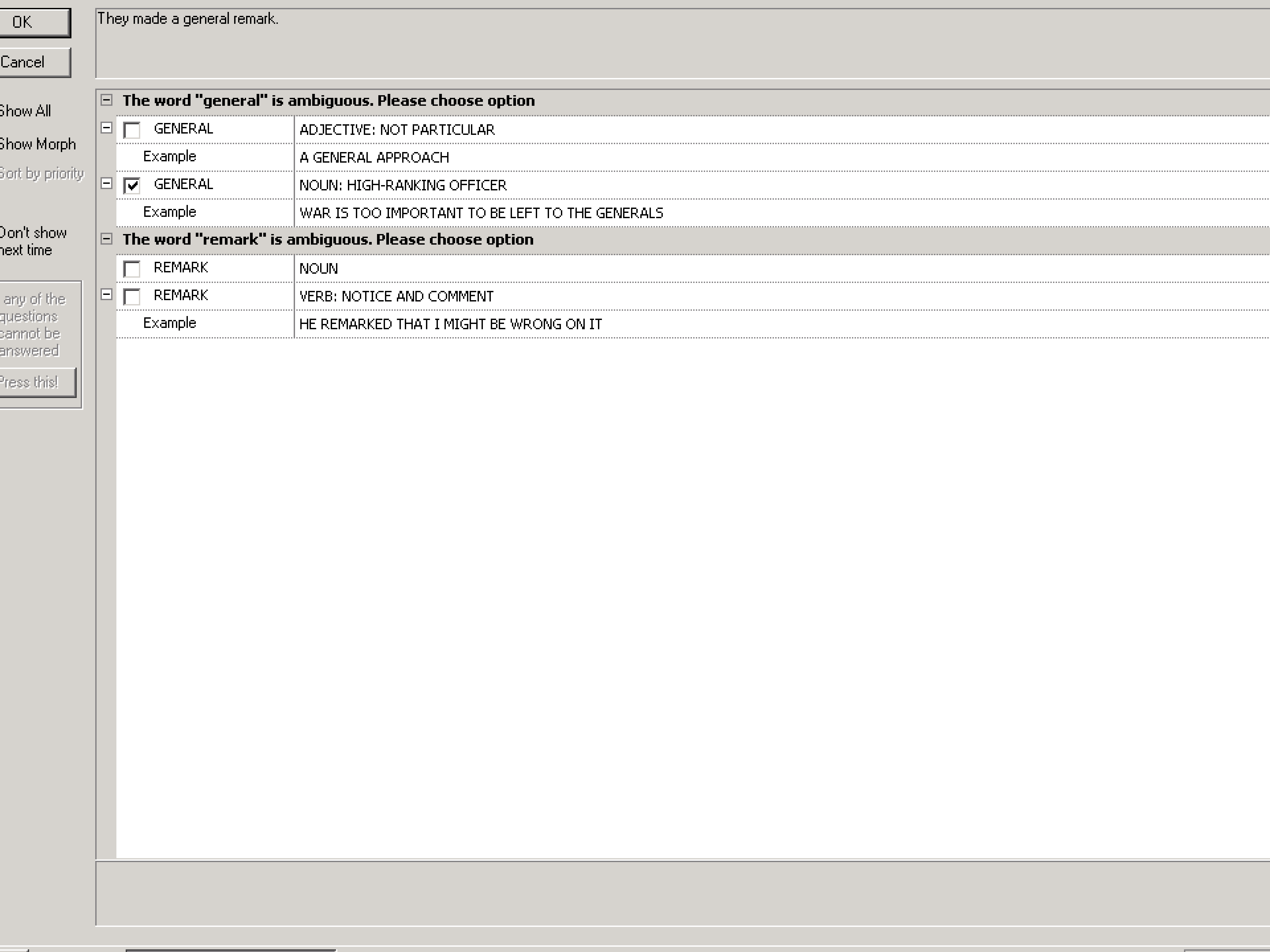
Sentence: They made a general reply that everything was OK.

Translation: Second Option

Они вынудили генерала отвечать, что все было хорошо.

Interactive disambiguation

Haga clic para modificar el estilo de subtítulo del patrón



They made a general remark.

The word "general" is ambiguous. Please choose option

□ GENERAL	ADJECTIVE: NOT PARTICULAR
Example	A GENERAL APPROACH
☑ GENERAL	NOUN: HIGH-RANKING OFFICER
Example	WAR IS TOO IMPORTANT TO BE LEFT TO THE GENERALS

The word "remark" is ambiguous. Please choose option

□ REMARK	NOUN
□ REMARK	VERB: NOTICE AND COMMENT
Example	HE REMARKED THAT I MIGHT BE WRONG ON IT

Lexical Functions

Haga clic para modificar el estilo de subtítulo del patrón

LFs {***R***, X, Y} are widely spread linguistically relevant meanings that are expresed differently in different languages.

- These meanings are language-independent.
- Their correlates are language-specific.
- MAGN (*disease*) = *grave*
- MAGN (*rain*) = *heavy*
- MAGN (*болезнь* ‘disease’) = *тяжелый,* lit. *heavy*
- MAGN (*дождь* ‘rain’) = *сильный,* lit. *strong*

Haga clic para modificar el estilo de subtítulo del patrón

Substitute LFs

- = those which replace the keyword in the given utterance without substantially changing its meaning or changing it in a strictly predictable way.
    - Synonyms, hypernyms, antonyms
    - Converse terms: *buy – sell, right – left, …*
    - Derivatives:
        - *encourage – encouragement,*
        - *to build – builder*
        - *nominate – nominee*
        - *teach – student*

Paraphrasing with converse terms

- *She **bought** a computer for 500 dollars from a retail dealer*
- *A retail dealer **sold** her a computer for 500 dollars*
- *She **paid** 500 dollars to the retail dealer for a computer*
- *The retail dealer **got** 500 dollars from her for a computer.*

Collocate LFs

- = those which appear in an utterance alongside the keyword.
- Adjectival LFs, such as MAGN
- Support verbs of the OPER / FUNC / LABOR family: play a leading role in paraphrasing

Paraphrasing based on collocates

- *He respects* [X] *his teachers*
- *He has* [Oper1(S0 (X))] *respect* [S0 (X)] *for his teachers*
- *He treats* [Labor1-2(S0 (X))] *his teachers with respect*
- *His teachers enjoy* [Oper2(S0(X))] *his respect*.

Rules for the previous example

- X □Oper1(X) + S0(X)
- X □Oper2(X) + S0(X)
- X □Labor12(X) + S0(X)

Some other rules

- X □Copul + S1(X)

He taught me at school – He was my teacher at school

- X □Func0 + S0(X)

They are arguing heatedly – A heated argument between them is on

- X □Func1 + S0(X)

He is afraid – Fear possesses him

- IncepOper1 + S0(X) □IncepOper2 + S0(X)

He conceived a dislike for her – She caused his dislike

- FinOper1 + S0(X) □FinOper2 + S0(X)

England lost control of this territory – This territory went out of England's control

Some more rules …

- LiquOper1 + S0(X) □LiquOper2 + S0(X)

The government deprived the monopolies of control over the prices – The government took the prices out of the monopolies’ control

- LiquOper1 + S0(X) □LiquFunc1 + S0(X)

We freed him of this burden – We lifted this burden from him

- X □IncepOper1 + Sres(X) □IncepFunc1 + Sres(X).

He learned physics – He acquired the knowledge of physics.

Some more…

- X □CausOper1 + Sres(X) etc.

He taught me physics – He gave me the knowledge of physics.

- LiquOper1 + Sinit (X) □LiquFunc1 + Sinit(X) etc.

A sudden bell woke him up – A sudden bell interrupted his sleep.

- CausFact0-M + X / CausFact1-M + X / CausReal1-M + X □IncepFact0-M + X / IncepReal1-M + X etc.

They sent him on leave for a few days – He went on leave for a few days.

Some more

- LiquFact0-M + X / LiquFact1-M + X / LiquReal1-M + X □ FinFact0-M + X / FinReal1-M + X etc.

He was deprived of his last chance to win in this event – He lost his last chance to win in this event

- Anti1Fact0-M(X) + X = *neg*Fact0-M(X) + X etc.

The plans of pacifying the aggressor failed – The plans of pacifying the aggressor did not succeed;

- Anti1Real1-M(X) + X □negReal1-M(X) + X etc.

The board of directors declined the compromise – The board of directors did not accept the compromise,

Some more…

- Anti1Real2-M(X) + X □negReal2-M(X) + X etc.

He swallowed up the insult – He did not avenge the insult

- Anti1Real3-M(X) + X □negReal3-M(X) + X etc.

The lecturer ignored the questions of the audience – The lecturer did not answer the questions of the audience, He neglected my advice – He did not follow my advice, Any soldier who violates the order is subject to court martial – Any soldier who does not obey the order is subject to court martial.

- X + Y □Anti1 + Anti2

He stopped violating the rules – He began observing the rules

Lexicographic support

- A paraphrasing system of this kind requires a good lexicographic source from which the appropriate LF values of words could be extracted. Such a source is provided by the combinatorial dictionary
- Combinatorial dictionaries of English and Russian: an inventory of 120+ LFs

LFs have a strong potential for NLP applications.

- LFs are used for:
  - Lexical and syntactic disambiguation
  - Adequate word selection in translation and text generation
  - Ontology construction
  - Reasoning
  - Synonymous paraphrasing of utterances
  - Anaphora resolution

Syntactic Disambiguation

- *support of the parliament*
  - ‘support by the parliament’
  - ‘support (given) to the parliament’

In lexical functional contexts, syntactic links are disambiguated:

- *The president had* [Y=OPER2(X)] *the support* [X] *of the parliament*
- *The president expressed* [Y=OPER1(X)] full *support* [X] *of the parliament*

LFs help establish a semantic relation

- Support verbs of the Oper-Func-Labor family attach one of the arguments of the noun
- Different LFs correspond to different arguments
- *Father gave me an **advice***
- *The proposal received much **attention**:*
  - In both cases the subject of the verb is an argument of the noun
  - Their roles are different:
    - *father* is the Agent of *advice*
    - *the proposal* is the Recipient of *attention*
- These verbs are LFs of different types.
  - *Give* = Oper1(*advice*). Its subject is the 1st argument of the noun, which is the Agent
  - *Receive* = Oper2(*proposal*). Its subject is the 2nd argument of the noun, which is the Recipient

Idiomatic translation of lexically restricted expressions: LF **Temp**

Temp (*March*) = *in : Temp*(*март*) = *в2*

Temp (*Tuesday*) = *on : Temp*(*вторник*) = *в1*

Temp (*dawn*) = *at : Temp*(*рассвет*) = *на2*

Temp (*moment*) = *at : Temp*(*момент*) = *в1*

Temp (*Easter*) = *at : Temp*(*пасха*) = *на1*

An LF corresponds to an ontological class

- LiquFunc0: ‘to cause to cease to exist or to be taking place’.
  - *to stop (the aggression), to lift (the blockade), to dispel (the clouds), to demolish (the building), to disperse (the crowd), to avert (the danger), to cure (the disease), to close (the dispute), to annul (the convention)…*
- LiquFact0: ‘to cause to cease functioning according to its destination’
  - *close (the eyes), stop (the car), land (the airplane), depose (the king), switch off (the lamp), neutralize (the poison), empty (the bucket), shut down (the factory).*

Anaphora resolution

- *The convention was signed by the United States, Honduras, El Salvador, Dominican Republic, Haiti, Argentina, Venezuela, Uruguay, Paraguay, Mexico, Panama, Bolivia, Guatemala, Brazil, Ecuador, Nicaragua, Colombia, Chile, Peru and Cuba but* ***it*** *was later annulled.*
- as many as 22 nouns that, theoretically, may be antecedents for the pronoun *it*.
- The correct antecedent, *convention*, appears furthest from the pronoun.
- The pronoun *it* occupies the object position of the lexical functional verb *annul*, and the argument of this LF (LiquFunc0) can only be the word *convention*.

LF-based inference

- *The blockade is lifted* (=LiquFunc0) □it does not exist any more.
- *He fulfilled* (= Real1) *the promise to buy a bicycle* □*He bought a bicycle.*

Paraphrasing and translation

- The source language word does not have a direct equivalent in the target language
  - *The congress was followed by a workshop*
  - *follow - sledovat'* (no passive)
  - Conv(*follow*) = *precede*
  - *The congress preceded the workshop*
  - *Surpass – prevosxodit'* (no passive)
  - Conv(*prevosxodit'*) = *ustupat'* ('be inferior')
  - *He is surpassed by nobody - On nikomu ne ustupaet* ('he is inferior to nobody')

Speech Synthesis Support

Haga clic para modificar el estilo de subtítulo del patrón

- In cooperation with Speech Synthesis Lab of the National Academy of Sciences of Belarus.
- Challenges:
  - Position of the word stress is highly variable.
  - Vowels are pronounced differently depending on their position wrt stress.
  - Phonetic ambiguity: same writing – different pronunciation.
    - Accentuated syllable: stoít (‘stands’) – stóit (‘costs’), Ivanóv (family name) - Ivánov (Gen.Pl. of *Ivan*)
    - Letter *e* stands for [e] and [jo]: vse – [vse] ‘everybody’ and [vsjo] ‘everything’, osel - [osel] ‘caved in’ and [osjol] ‘donkey’
  - Intonation.
- What is needed:
  - Lexico-grammatical information
  - Information on accentuation patterns (several hundred)
  - Automatic procedure of creation of the intonation contour, loudness, phoneme and pause length based on the analysis of definite properties of the input text and its prosodic annotation.

Each sentence should be supplied with prosodic annotation, which:

- divides the sentence into syntagms,
- marks accent units within syntagms,
- labels the intonational type of each syntagm.

This annotation cannot be performed sufficiently well if the syntactic structure of the sentence is not taken into account.

What has been done

We studied the interrelation between the syntactic structure of Russian sentences and their prosodic annotation.

As a result, a corpus of rules has been produced that discover prosodically significant elements in the syntactic structure of the sentence.

Corpus analyzed

Prosodically annotated texts of the television programmes contained in the database “Intonation of Russian news broadcasting”.

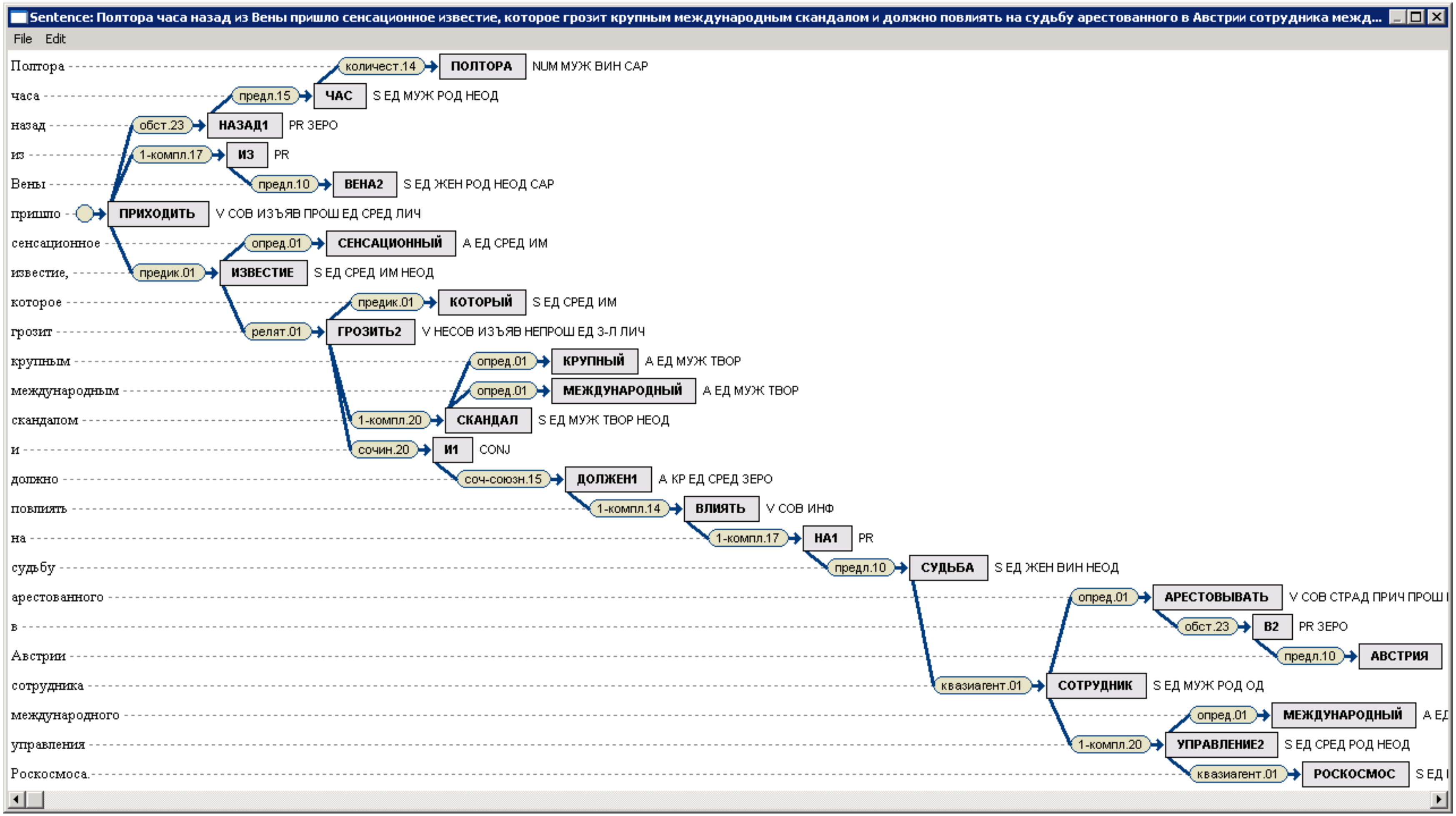
We ran these texts through the ETAP parser and found that there exist statistically significant syntactic correlates of prosodic accentuation of words in syntagms.

Syntactic experiment: annotated sentence

*Полтора часа *назад из *Вены пришло *сенсационное *известие, которое грозит *крупным международным *скандалом и должно *повлиять на судьбу арестованного в *Австрии сотрудника международного управления *РосКосмоса.*

‘An hour and a half ago a sensational message came from Vienna that threatens to provoke a large-scale international scandal and should affect the destiny of the staff member of the international department of RosKosmos arrested in Austria’

ETAP-produced tree



Sentence: Полтора часа назад из Вены пришло сенсационное известие, которое грозит крупным международным скандалом и должно повлиять на судьбу арестованного в Австрии сотрудника межд...

File Edit
Полтора количест.14 ПОЛТОРА NUM МУЖ ВИН САР
часа предл.15 ЧАС S ЕД МУЖ РОД НЕОД
назад обст.23 НАЗАД1 PR ЗЕРО
из 1-компл.17 ИЗ PR
Вены предл.10 ВЕНА2 S ЕД ЖЕН РОД НЕОД САР
пришло ПРИХОДИТЬ V СОВ ИЗЪЯВ ПРОШ ЕД СРЕД ЛИЧ
сенсационное опред.01 СЕНСАЦИОННЫЙ A ЕД СРЕД ИМ
известие, предик.01 ИЗВЕСТИЕ S ЕД СРЕД ИМ НЕОД
которое предик.01 КОТОРЫЙ S ЕД СРЕД ИМ
грозит релят.01 ГРОЗИТЬ2 V НЕСОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л ЛИЧ
крупным опред.01 КРУПНЫЙ A ЕД МУЖ ТВОР
международным опред.01 МЕЖДУНАРОДНЫЙ A ЕД МУЖ ТВОР
скандалом 1-компл.20 СКАНДАЛ S ЕД МУЖ ТВОР НЕОД
и сочин.20 И1 CONJ
должно соч-союзн.15 ДОЛЖЕН1 A КР ЕД СРЕД ЗЕРО
повлиять 1-компл.14 ВЛИЯТЬ V СОВ ИНФ
на 1-компл.17 НА1 PR
судьбу предл.10 СУДЬБА S ЕД ЖЕН ВИН НЕОД
арестованного опред.01 АРЕСТОВЫВАТЬ V СОВ СТРАД ПРИЧ ПРОШ I
в обст.23 В2 PR ЗЕРО
Австрии предл.10 АВСТРИЯ
сотрудника квазиагент.01 СОТРУДНИК S ЕД МУЖ РОД ОД
международного опред.01 МЕЖДУНАРОДНЫЙ A ЕД
управления 1-компл.20 УПРАВЛЕНИЕ2 S ЕД СРЕД РОД НЕОД
Роскосмоса. квазиагент.01 РОСКОСМОС S ЕД I

Matching

Prosodically marked words of this sentence:

- *Вены* – the rightmost element of the first complement phrase
- *назад* – the rightmost element of a circumstantial phrase
- *известие* – the right element of the subject phrase, from which a subordinate clause is cut off
- *скандалом* – the rightmost element of the first complement phrase;
- *Австрии* – the rightmost element of the participial phrase;
- *Роскосмоса* – the rightmost element of the first complement phrase.

Prosodic marking of other words of the sentence is syntactically non-motivated.

This experiment permitted to come up with several rules which identify prosodically marked words in the syntactic structure.

Conclusion

- ETAP processes Russian in a number of aspects:
  - Morphological analysis and generation
  - Dependency parsing (86%)
  - Translation from and into Russian
  - Semantic processing (interlingua, ontology-based)
- LFs are used for various purposes

Conclusion (cont)

- Speech synthesis support
  - Inclusion of accentuation patterns in the morphological dictionary helps improve accentuation.
  - Lexico-grammatical information helps disambiguate phonetically ambiguous word forms
  - The dependency SyntS helps determine prosodically marked elements of the sentence