CENTRE D'ÉTUDES FRANCO-RUSSE
ЦЕНТР ФРАНКО-РОССИЙСКИХ ИССЛЕДОВАНИЙ

New Russian resources for the linguistic software Nooj

Improving linguistics resources and adding semantic tags for the Russian language

for Max Silberztein's Nooj software

Zagreb
Nooj Conference
June 2020

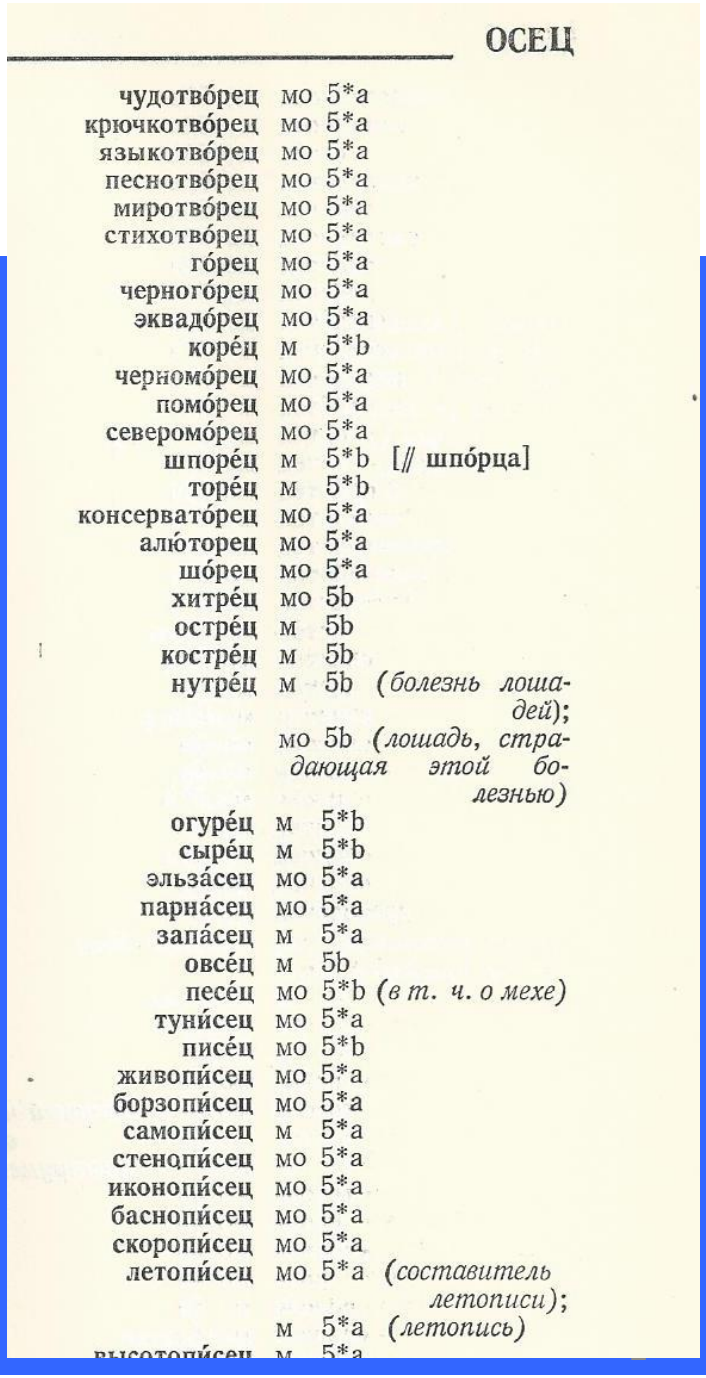
Vincent BÉNET
INALCO
CEFR-Moscow

Brief summary about Russian resources for Nooj

Zaliznyak dictionary

inverted dictionary

most useful dictionary for Russian grammar



ОСЕЦ

чудотво́рец мо 5*a
крючкотво́рец мо 5*a
языкотво́рец мо 5*a
песнотво́рец мо 5*a
миротво́рец мо 5*a
стихотво́рец мо 5*a
го́рец мо 5*a
черногóрец мо 5*a
эквадóрец мо 5*a
коре́ц м 5*b
черномóрец мо 5*a
помóрец мо 5*a
северомóрец мо 5*a
шпоре́ц м 5*b [// шпóрца]
торе́ц м 5*b
консерватóрец мо 5*a
алю́торец мо 5*a
шóрец мо 5*a
хитре́ц мо 5b
остре́ц м 5b
костре́ц м 5b
нутре́ц м 5b (*болезнь лошадей*);
мо 5b (*лошадь, страдающая этой болезнью*)
огуре́ц м 5*b
сыре́ц м 5*b
эльза́сец мо 5*a
парна́сец мо 5*a
запа́сец м 5*a
овсе́ц м 5b
песе́ц мо 5*b (*в т. ч. о мехе*)
туни́сец мо 5*a
писе́ц мо 5*b
живопи́сец мо 5*a
борзопи́сец мо 5*a
самопи́сец м 5*a
стенопи́сец мо 5*a
иконопи́сец мо 5*a
баснопи́сец мо 5*a
скоропи́сец мо 5*a
летопи́сец мо 5*a (*составитель летописи*);
м 5*a (*летопись*)
высотопи́сец м 5*a

Zagreb Conference Nooj
June 2020

Brief summary about Russian resources for Nooj

Zaliznyak electronic version → Nooj

покурол<есить	V	PF	INTR	sv-np	4a	#	
покус<ать	V	PF		sv	1a	#	
покус<иться	V	PF		sv	4b $II	#	
пок<усывать	V	IPF		nsv	1a?	#	
покут<ить	V	PF	INTR	sv-np	4c	#	
пок<учивать	V	IPF	INTR	nsv-np	1a	#	
пок<ушать	V	PF		sv	1a	#	
покуш<аться	V	IPF		nsv	$4(-с<иться	#	
покуш<ение	N	NEUT	INANI	s	7a	#	
пол	N	MASC	INANI	m	1c +n2	#	(настил) % д<о полу [//до п<олу,до п<ола];д
(пол)						#	п<од пол (в подвал,в подпол)
пол	N	MASC	INANI	m	1e	#	(биол.)
пол-лист<а				(?)	1	#	(дополняется формами от *полулист м 1в)
пол-л<итра				(?)	1	#	(дополняется формами от *полул<итр м 1а
пол-обор<ота				(?)	1	#	% завест<и с пол-обор<ота
п>ол-од<иннадцатого				(?)	2	#	
пол-очк<а				(?)	1	#	(дополняется формами от *полуочк<о с 3*
пол<а	N	FEM	INANI	j	1d	#	
полав<ировать	V	PF	INTR	sv-np	2a	#	
полаг<ать	V	IPF		nsv	1a?	#	(думать, считать)
полаг<ать	V	IPF		nsv	положит	#	(в сочетании с нач<ало, кон<ец, пред<ел ч
полаг<аться	V	IPF		nsv	(положи	#	
полаг<аться	V	IPF		nsv	1a	#	(быть установленным, причитаться)
пол<адить	V	PF		sv	4a	#	

мучить,V+Ipf+FLX=тешить
мутузить,V+Ipf+FLX=морозить
мучиться,V+sja+Ipf+FLX=тешиться
мшиться,V+sja+Ipf+FLX=дорожиться
мшить,V+Ipf+FLX=дорожить
мылиться,V+sja+Ipf+FLX=жариться
мылить,V+Ipf+FLX=жарить
мыслить,V+Ipf+FLX=медлить
мытариться,V+sja+Ipf+FLX=жариться
мытарить,V+Ipf+FLX=жарить
мягчить,V+Ipf+FLX=дорожить
мямлить,V+Ipf+intr+FLX=медлить
набавить,V+Pf+FLX=исправить
набагрить,V+Pf+FLX=замедлить
набаламутить,V+Pf+FLX=потратить
набатрачить,V+Pf+FLX=потешить
набедокурить,V+Pf+intr+FLX=поверить
набезобразить,V+Pf+intr+FLX=заморозить
набелиться,V+sja+Pf+Color+FLX=селиться
набелить,V+Pf+Color+FLX=похоронить
наблажить,V+Pf+intr+FLX=нагрешить

Brief summary about Russian resources for Nooj

- Linguistic resources for Russian language Zaliznyak's dictionary

-Very complete grammar information with some unformal unclassified added annotations

-All forms are included in the dictionary,
BUT NO information about derivation (prefixes or suffixes)

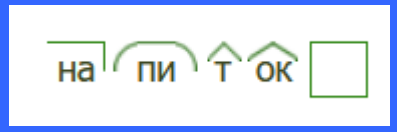
Vincent BÉNET
INALCO
CEFR-MOSCOW

Brief summary about Russian resources for Nooj

- Structure of Russian language :

PREF (s) + ROOT + SUFF (s) + DES + SUF2

	pi		t’	
na	pi	t ok		
na	pi		t’	sja
na	pi	va	t’	sja



How can we improve Nooj Russian resources?

PREFIX + ROOT + SUFFIX + DESINENCE

Verbal prefixation and derivation

Выпить *pf* – выпивать *ipf*
vy-pi-t’ → vy-pi-va-t’ *to finish drinking*

пить
pi-t’
To drink

напиться *pf* – напиваться *ipf*
na-pi-t’-sja → na-pi-va-t’-sja *to get /be drunk*

пропить *pf* – пропивать *ipf*
pro-pi-t’ → pro-pi-va-t’ *to spend sthg in drinking*

How can we improve Nooj Russian resources?

Implementing verbal prefixes and suffixes

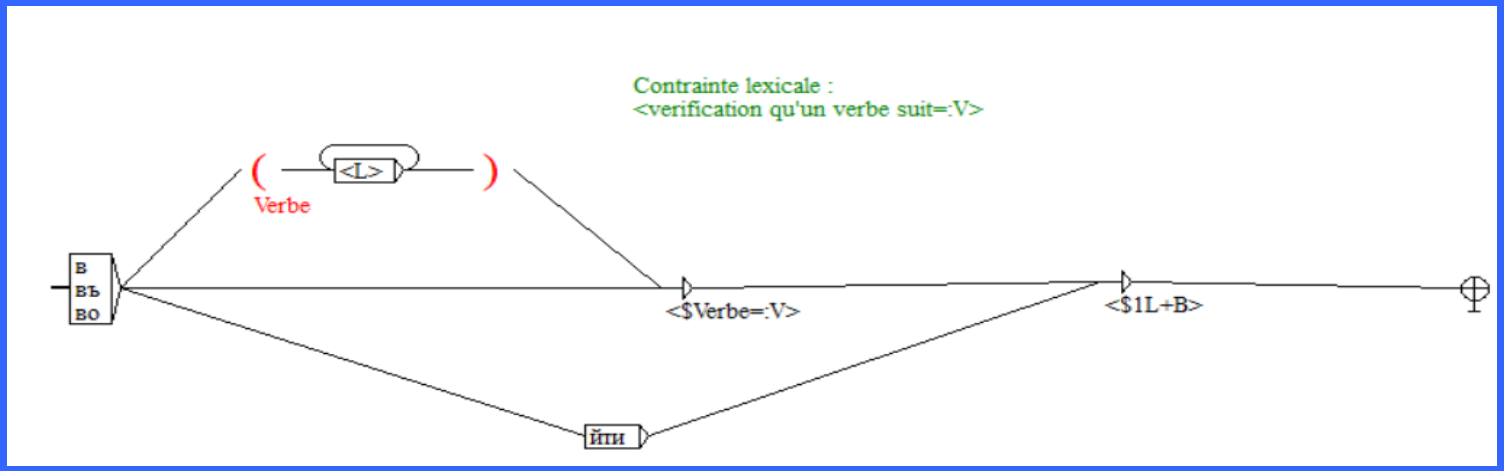
More information in the dictionary?

More derivational grammars .nom

пить
pi-t’

Пить… + DRV= «ва» +DRV=«ся» + DRV= «вы» + DRV = «на» + DRV=«за» + DRV=«про»

Grammar to recognize the verbal prefix В (meaning = into)



Improving lexical resources
Adding semantic resources to the main dictionary

Simple classification « A word and an idea »
House, sports, body, …

Semantic tags / annotation
of the Russian national Corpus

Semantic dictionnary of Tuzov
(160 000 words)

Improving lexical resources
Adding semantic resources to the main dictionary

Nooj properties.def

A_Sem = Animal | Color | (App

N_Sem = Hum | Prof | Parents | Body
Conc | Abstr | Org | Text |
Animal | Food | Health | Arts | Lit | Music | Sports
Topo | Country | River | City | Mount| Lake |
Posit | Time | Color ;
ADV_Sem = Time |Topo | Modal;
V_Sem = Color | Topo | Posit |Modal;

Improving lexical resources
Semantic resources added to the main dictionary

Prof = 900
Parent/ Kinship = 160 items
Forenames = 2280
Animal = 370
Food = 280 (Liquid = 25)
Body = 285
Health = 175
Arts = 65
Lit = 40
Music = 155
Sport = 65

Topo = 40
Country = 180
River = 15
City = 175
Mount = 5
Lake = 5

Posit = 25
Time = 135
Modal = 15
Color = 275

Improving lexical resources

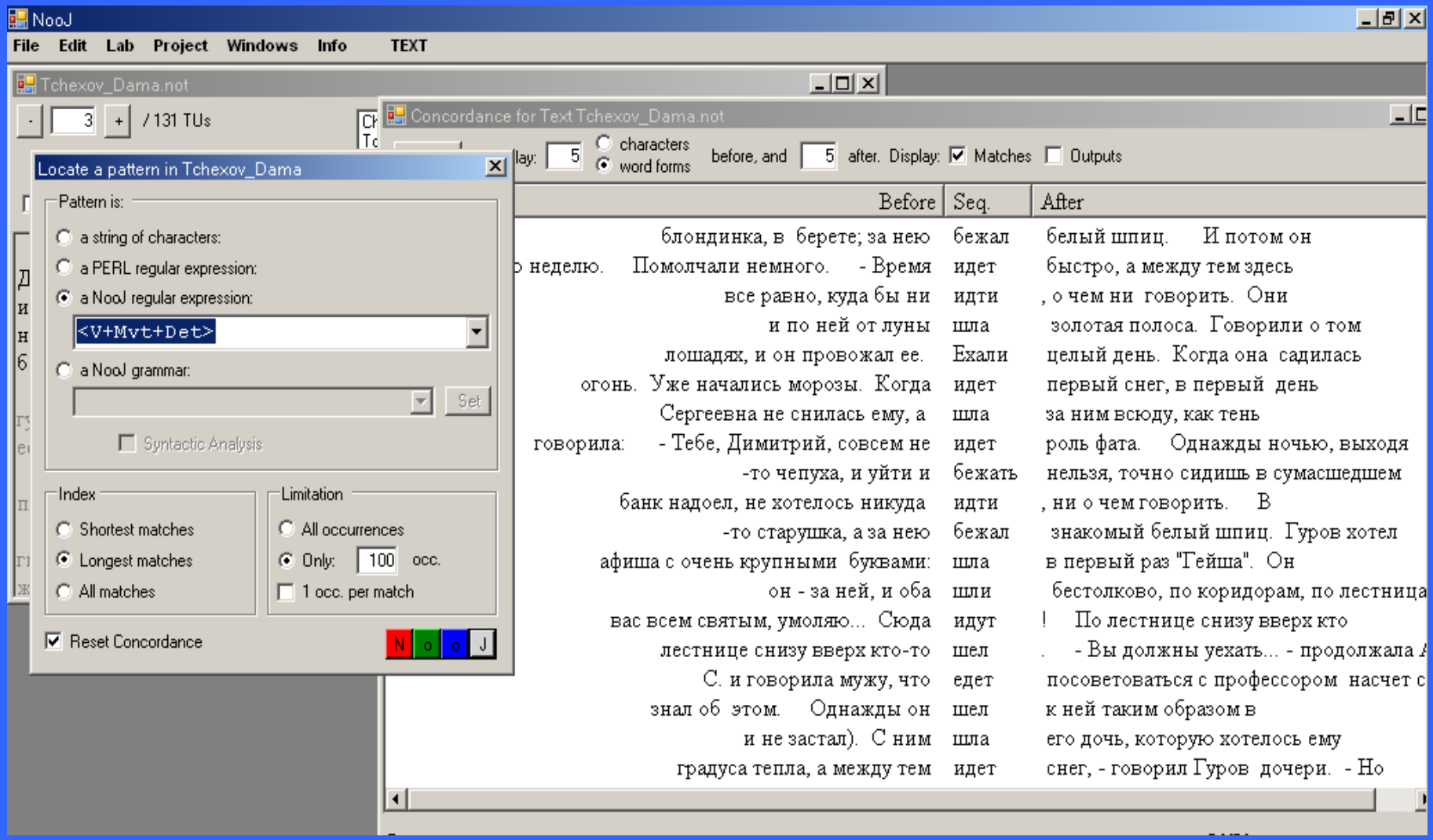
Added semantic resources to the main dictionary

Morphology or semantics ?

ходить,V+Mvt+Indet+ipf+intrLX=ходить
идти,V+Mvt+Det+ipf+intrLX=идти

входить,V+Mvt+Pvb+ipf+intrLX=ходить
войти,V+Mvt+Pvb+pf+intrLX=идти
выходить,V+Mvt+Pvb+ipf+intrLX=ходить
приеззжать,V+Mvt+Pvb+ipf+intrLX=акать

Working with Nooj Searching for « verbs of motion »



Improving lexical resources
Semantic resources added to grammars

Grammar for body parts

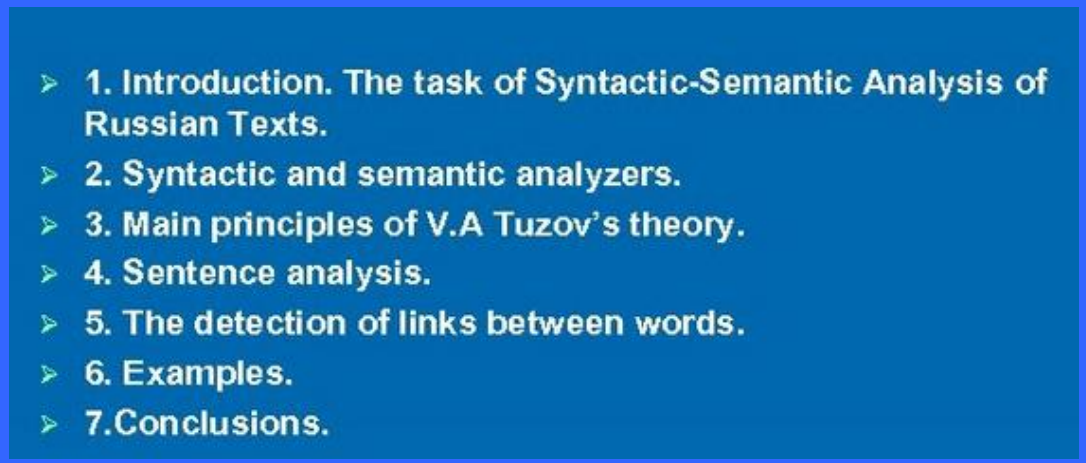
```
# Syntactic grammar
#
# Input Language is: ru
# Output Language is: ru
#
# Special Start Rule: Main
#
# Special Characters: '=' '<' '>' '\' '"' ':' '|' '+' '-' '/' '$' '_' ';'
#

Main = :Noms | :Adjectifs;
Noms = <брюшная полость> <подмышка> <аорта> <артерия> |
<борода> | <таз> | <бицепс> | <бицепс> | <рот> | <рука> | <лампочка> |
<волосы> | <лодыжка> | <ресница> | <ключица> | <клитор> | <копчик> | <сер
 <локоть> |  <бедро> | <зуб> | <диафрагма> | <палец> | <спина> | <двенади
 <мозг> | <плечо> | <эпидермис> | <желудок> | <надгортанник> | <печень> |
<колено> | <железа> | <таз> | <спираль> | <эритроциты> | <племянник> | <д
<резец> | <кишечник> | <радужка> | <нога> | <затылок> | <язык> | <язычок>
<костный мозг> | <мышца> | <миокард> | <ноздря> | <нерв> | <нос> | <пупок
<глаз> | <пищевод> | <лопатка> | <гвоздь> | <ухо> | <ротоглотка> | <палец
<поджелудочная железа> | <Вкусовые рецепторы> | <Паратироид> | <Ладонь> |
| <сплетение> | <запястье> | <волосы> | <большой палец> | <легкие> | <пре
<Позвоночник> | <Радиус> | <Селезенка> | <Ректум> | <Почка> | <Пателла> |
<крестец> | <кровь> | <мошонка> | <грудь> | <синус> | <бровь> | <грудная
```

The Tuzov dictionary

Improving Nooj ressources

The semantic dictionary by V.**Tuzov** (Saint Petersburg university) is a most complete electronic dictionary of 190 000 entries, with full morphological and semantical annotation.



The Tuzov dictionary

Improving Nooj ressources

1st job (almost done) : make it complete in Excel format and ready for Nooj annotation

2nd make it compatible with the Zaliznyak's dictionary (morphological tagging)

Vincent BÉNET
CEFR

CENTRE D'ÉTUDES FRANCO-RUSSE
ЦЕНТР ФРАНКО-РОССИЙСКИХ ИССЛЕДОВАНИЙ

The Tuzov dictionary

Improving Nooj ressources

There are 1526 different morphogical types
(that are the same as Zaliznyak's categories !)

There are 7452 semantical different types that can be reduced to 1638, using the category table.
(the types have been translated to English)

Vincent BÉNET
CEFR

The Tuzov dictionary

Improving Nooj ressources

$100/	Life+/	47
$100/0	Life+Existence-Inexistence	240
$100/1	Life+Perception	56
$100/11	Life+Perception+Viewing	299
$100/111	Life+Perception+Viewing+Sighted_Blind	38
$100/112	Life+Perception+Viewing+Far-viewing_Myope	6
$100/113	Life+Perception+Viewing+Perceptible_Unperceptible	68
$100/114	Life+Perception+Viewing+Visible_Univisible	15
$100/12	Life+Perception+Hearing	100
$100/13	Life+Perception+Smelling	44
$100/14	Life+Perception+Touching	8
$100/2	Life+Nutrition	599
$100/3	Life+Respiration	42
$100/31	Life+Respiration+Breathe	13
$100/4	Life+Development	84
$100/5	Life+Activities	18
$101/1	Life+Food	239

How many tags?

Semantic tag Signification Number of words

The Tuzov dictionary

Improving Nooj ressources

кормильщик	кищьлимрок	n%~пища	$101/1	{м3о 96}
кормитель	ьлетимрок	n%~пища	$101/1	{м2о 316}
кормить	ьтимрок	n%~кормление	$100/2	{г4н 6391}
кормиться	ясьтимрок	n%~кормление	$100/2	{г4н 10252}
кормление	еинелмрок		$100/2	{с7 610}
кормленщик	кищнелмрок		$12413402	{м3о 96}
кормленый	йынелмрок	n%~кормление	$100/2	{п1 1387}
кормный	йынмрок	n%~кормление	$100/2	{п1 36}
кормовой	йовомрок	n%~корма	$121324/2	{п1 553}
кормовой	йовомрок	n%~кормление	$100/2	{п1 553}
кормодобывание	еинавыбодомрок	n%~добыча	$1527	{с7 610}
кормодобывающий	йищюавыбодомрок	n%~добыча	$1527	{п4 7665}
кормозаготовительный	йыньлетивотогазомрок	n%~заготовка	$15231	{п1 428}
кормозаготовка	аквотогазомрок	n%~заготовка	$15231	{ж3 168}

Entry Entry2 Commentary Semantic tag Grammatical tag

CENTRE D'ÉTUDES FRANCO-RUSSE
ЦЕНТР ФРАНКО-РОССИЙСКИХ ИССЛЕДОВАНИЙ

The Nooj semantic dictionary

Improving Nooj ressources

быть	n%~будущее	$160/010/055	{г16сн 7788}
быть	n%~время	$165	{г16сн 7788}
быть	n%~действие	$155	{г16сн 7788}
быть	n%~зависимость	$1241/4123/135	{г16сн 7788}
быть	n%~звание	$1241/125	{г16сн 7788}
быть	n%~количество	$12/135	{г16сн 7788}
быть	n%~местопребывание	$12/35	{г16сн 7788}
быть	n%~нахождение	$12/35	{г16сн 7788}
быть	n%~носка	$1525325	{г16сн 7788}
быть	n%~отношения	$1/42325	{г16сн 7788}
быть	n%~приобретение	$15310/0/055	{г16сн 7788}
быть	n%~происхождение	$1241/175	{г16сн 7788}
быть	n%~психика	$1241/45	{г16сн 7788}
быть	n%~свойство	$1/10/005	{г16сн 7788}
быть	n%~состояние	$11135	{г16сн 7788}
быть	n%~существование	$111015	{г16сн 7788}
быть	n%~свойство	$1/10/00	{г16сн 7788}

The Nooj semantic dictionary

Improving Nooj ressources

говор	$14402	Knowledge+Communication+Information+Speech
говорение	$14402	Knowledge+Communication+Information+Speech
говорильный	$14402	Knowledge+Communication+Information+Speech
говорильня	$14402	Knowledge+Communication+Information+Speech
говорить	$14402	Knowledge+Communication+Information+Speech
говорить	$1526031	Action+Work+Intellectual+Commandment+Order
говориться	$14402105	Knowledge+Communication+Information+Speech+Изложение+Высказывание
сказать	$14402105	Knowledge+Communication+Information+Speech+Изложение+Высказывание
сказать	$1526031	Action+Work+Intellectual+Commandment+Order
сказаться	$1/37	Reason

Improving semantic ressources for Nooj

All the semantics tag are translated into English

First step:

- one morphological dictionary (allready existing)
- one semantical dictionary (soon available), both at Nooj format (soon available)

Second step:

more derivational and syntactic grammars to provide a full linguistic analysis

Improving semantic ressources for Nooj

Third step :

These dictionaries can be compiled in one big dictionary. (morphology and semantics)

Problems : more ambiguities may occur, that can be less productive and less useful for users.

Improving Russian ressources for Nooj

Thank you for your attention

vincent.benet@cnrs.fr
vincent.benet@inalco.fr