CORPUS PROCESSING: THE LINGUISTIC APPROACH DEVELOPING RESSOURCES FOR RUSSIAN, APPLICATIONS TO LINGUISTICS AND LANGUAGE TEACHING

M S Max Sylberztein, V B Vincent Bénet

CORPUS PROCESSING: THE LINGUISTIC APPROACH DEVELOPING RESSOURCES FOR RUSSIAN, APPLICATIONS TO LINGUISTICS AND LANGUAGE TEACHING.

M.S. Max Sylberztein
V.B. Vincent Bénet

Abstract

We present a set of linguistic resources developed for Russian with the Nooj platform: a grammatical dictionary associated with a set of grammars to describe the inflection, a semantic dictionary as well as a set of syntactic grammars that solve various types of ambiguities and recognize several named entities. We show then how these resources are used by NooJ to process corpora and create several pedagogical activities to teach the Russian language as a second language.

Key Words

Linguistics, Corpus linguistics, Natural Language Processing, Russian language.

Introduction

To study Russian corpora, most linguists make extensive and almost exclusive use of the "ruscorpora.ru" site of the Russian National Corpus, which contains texts with complete morphosyntactic and semantic tagging, i.e. completely disambiguated. Russian linguists can also use the "cfrl.ruslang.ru" collection site of the Computer Fund of Russian Language, which displays occurrences of wordform (and not by lemma) for a limited set of texts. In contrast, the NooJ software[1]

[1] Nooj is a linguistic development environment that can be used to formalize eight levels of linguistic phenomena: spelling and typography, inflectional, derivational and agglutinative

allows users to work with their own texts, including those that constitute the Computer Fund of Russian Language.

The INALCO institute has been collaborating with the Vinogradov Russian Language Institute of the Russian Academy of Sciences for over thirty years. The experience of working with J. Anoshkina's Unilex and A. Dialex programmes. Baranov's Unilex and Dialex programmes have made it possible to develop grammatical and morphosyntactic resources for NooJ in a relatively short period of time.

NooJ's dictionary was developed by converting and transforming Zalizniak's grammatical dictionary, which the computer-assisted research laboratory of Inalco had in electronic version under DOS. NooJ dictionaries are compiled versions of a list of words associated with their morphosyntactic and semantic description, formalized by corresponding paradigms.

The grammatical dictionary

There are three Russsian dictionaries in NooJ: a dictionary of common nouns; a dictionary of proper nouns, and a dictionary of substantive adjectives. The latter dictionary was constructed to avoid too long and tedious disambiguation between homographic forms of adjectives and nouns; if this dictionary is not activated, words like русский (russkij = russian) or новое (novoe = new) will be considered as adjectives (and not as potential substantives). Together, the three dictionaries represent about 3,500,000 wordforms, associated with over 95,000 different linguistic analyses.

---

morphology, local and structural and dependency syntax, transformational grammar and semantics. NooJ provides users with formal tools adapted to each type of phenomenon (regular, context-free, context-sensitive and unrestricted grammars) and makes it possible to develop resources with wide coverage, for over 20 languages. NooJ is used by linguists to describe natural languages, as a corpus processor in the digital humanities, and its engine has been used by several companies to build Natural Language Processing applications. NooJ is free and open source and runs on Windows, Mac OSX, LINUX and Unix, see: http://www.nooj-association.org.

Here is an extract from the dictionary "russe_morph.dic" and from the corresponding morphological grammar "russe_morph.nof".

волейбол,N+m+inan+Sport+FLX=завод
волк,N+m+an+Animal+FLX=волк
газировать,V+ipf+pf+FLX=интересовать

The series of codes following each entry indicate the category and linguistic properties of the word. N=Noun, m=Masculine, an=Animated, Sport is a semantic feature, and FLX specifies the morphological paradigm. Following are the paradigms:

завод = <E>/Im+s | <E>/Vi+s | а/Ro+s | у/Da+s | ом/Tv+s | е/Pr+s | ы/Im+p | ы/Vi+p | ов/Ro+p | ам/Da+p | ами/Tv+p | ах/Pr+p ;
волк = <E>/Im+s | а/Vi+s | а/Ro+s | у/Da+s | ом/Tv+s | е/Pr+s | и/Im+p | ов/Vi+p | ов/Ro+p | ам/Da+p | ами/Tv+p | ах/Pr+p ;
интересовать = <E>/Inf | <B5>ую/1+s+Pre | <B5>уешь/2+s+Pre | <B5>ует/3+s+Pre| <B5>уем/1+p+Pre | <B5>уете/2+p+Pre | <B5>уют/3+p+Pre | <B2>л/m+s+Pa | <B2>ла/f+s+Pa | <B2>ло/n+s+Pa | <B2>ли/p+Pa| <B5>уй/2+s+Imp | <B5>уйте/2+p+Imp | <B5>уя/Ger | <B2>в/Ger+Pf | <B5>ующий/Prtp+Pre+Act+m+s+Im | <B5>ующий/Prtp+ Pre+ Act+ m+s+ Vi | <B5>ующего/Prtp+Pre+Act+m+an+s+Vi | <B5>ующего/Prtp+Pre+Act+m+s+Ro | <B5>ующему/Prtp +Pre+Act+m+s+Da | <B5>ующим/Prtp+Pre+Act+mo+s+Tv | <B5>ующем/Prtp+Pre+Act+mo+s+Pr

Inflectional paradigms (FLX) associate each inflected wordform with several properties, such as Case (Im, Vi, Ro, Da, TV, Pr), as well as Number (s or p). Special operators such as <B> (delete current letter) are used to produce the corresponding inflected wordform.

As we can see, NooJ's codes are different from Zalizniak's ; they can be rapidly mastered though. The system of codes has been designed to be analytic so as to be better adapted to western slavist tradition.

All property codes have to be properly defined in a separate file (Properties.def). Because NooJ is used to formalize various languages, there is no unique standard: each language has its own system. This is

sometimes unfortunate for linguists who work on multiple languages of a given linguistic family (e.g. Russian, Belarusian and Ukrainian).

In the Russian NooJ module, the Properties.def file lists 11 categories : Adjectives (A), Adverbs (ADV), Nouns (N), Numerals (NUM), Pronouns (PRO), Verbs (V), Prepositions (PREP), Conjonctions (CONJ), Interjections (INTERJ), Particles (PART), and Parentheses or Parenthetic phrases (INTRO).

Some of the categories are further associated with a number of properties:

A_Forme = fc | fl | adv; (short, long and adverbial form)
A_Genre = m | f | n ; (masculine, feminine and neutral)
A_SGenr = an | inan ; (animated et non-animated)
A_Nombre = s | p; число (singular and plural)
A_Cas = Im | Vi | Ro | Da | Tv | Pr | Zv; (nominative, accusative, genitive, dative, instrumental, prepositionnal and vocative)
A_Deg = Comp | Sup ; (comparative et superlative)
A_Sem = App | Color | Body; (Semantic features)
ADV_Deg = Comp; (adverb comparatives)
ADV_Sem = Time | Topo | Modal; (Semantic features)

Nouns are associated with the same properties as Adjectives, plus a few extra:

N_Cas = Im | Vi | Ro | R2 | Da | Tv | Pr | P2 | Zv
N_Sem = Hum | Forename | Prof | Parent | Body | Conc | Abstr | Org | Text | Animal | Food | Arts | Lit | Music | Sports | Topo | Country | River | City | Mount| Lake | Posit | Time | Color

The following Categories (Numerals, Pronouns, Prepositions etc.) are closed sets with a limited set of properties, e.g.:

NUM_Cat = ord | card | coll ; (ordinal, cardinal, collective)
PRO_Cas = Im | Vi | Vip | Ro | Rop | Da | Dap | Tv | Tvp | Pr ;

The semantic description of prepositions and conjunctions is more complex. Indeed, the massive polysemy forced us to duplicate many entries (for instance: causality or origin for « из » et « от »).

By studying corpora, it will be possible to decide if it is better to keep these distinctions or not, taking into account the fact that they are often redundant with the semantic value of the following noun.

Semantic dictionary

We have associated over 5,000 entries with 33 semantic tags: Hum, Prof, Parent, Body, Conc, Abstr, Org, Text, Animal, Food, Arts, Lit, Music, Sports, Topo, Country, River, City, Mount| Lake, Posit, Time, Color. Here is a sample of NooJ's dictionary:

баран,N+m+an+Animal+FLX=артист
барсук,N+m+an+Animal+FLX=рыбак
барсучонок,N+m+an+Animal+FLX=волчонок
барсенок,N+m+an+Animal+FLX=утенок
барс,N+m+an+Animal+FLX=артист
(409 lexical entries are associated with the code Animal)

We are in the process of using these codes by using Tuzov's semantic dictionary. which contains over 145,000 entries; each entry is associated with a semantic tree. Presently, the dictionary (available as an Excel file)is in the following format:

Entry	*Semantic Code*	*Grammatical Code*
вагон	*$12132411*	*#{м1 12}*
вагонетка	*$12132411*	*#{ж3 168}*
вагонеточный	*$12132411*	*#{п1 36}*
вагонетчик	*$12413220*	*#{м3о 96}*
вагонетчица	*$12413220*	*#{ж5о 1304}*
вагонник	*$12413220*	*# {м3о 96}*
вагонный	*$12132411*	*# {п1 36}*

We are planning to convert this file into a clearer NooJ format:
wagon (вагон) Phys_Obj +Inanimate +Thing +Technics +Transport +Terrestrial +No_engine
wagonetka (вагонетка) Phys_Obj +Inanimate +Thing +Technics+ Transport+ Terrestrial+No_engine

wagonetochnyj (вагонеточный) Phys_Obj +Inanimate +Thing +Technics +Transport + Terrestrial +No_engine
wagonetchik (вагонетчик) Phys_Obj +Living +Human +Personality +Profession +Worker
wagonetchitsa (вагонетчица) Phys_Obj +Living +Human +Personality +Profession +Worker
wagonnik (вагонник) Phys_Obj +Living +Human +Personality +Profession +Worker
wagonnyj (вагонный) Phys_Obj +Inanimate +Thing +Technics + Transport + Terrestrial +No_engine

Current Limitations of the NooJ Russian dictionary

We had chosen to ignore the ë and the tonic accent because they are never indicated in written texts (except for pedagogical applications), but we are planning to add them. Note that Juras Hetsvich's team at the Institute of Informatics in Minsk has developed a Russian dictionary for NooJ that contains the ë and tonic accents. L. Verbitskaya's and V. Kasevich's dictionary does contain tonic accents and the ë as well.

In Zalizniak's dictionary, imperfective and perfective forms of a verb are represented as two independent lexical entries. Unfortunately, this means that NooJ cannot link them automatically. We are planning to use NooJ's DRV feature to link them.

Scientific and pedagogical applications

There are numerous applications of NooJ, including pedagogical ones. We give now an example of a lab session with NooJ.

After launching the NooJ software, select "ru" in the menu Info>Preferences, then load a text file. The window "Info > Preferences > Lexical Analysis" allows users to select lexical and morphological resources; the window "Info > Preferences > Syntactic Analysis" allows users to select syntactic and semantic grammars.

The first exercise consists of retrieving the list of all the wordforms associated with their linguistic analyses and their frequencies. as well as the list of all “unknown” wordforms, i.e. those that were not recognized by any of the linguistic resources selected in Info > Preferences.

Students can then see instantly the list of words that they don’t know. By double-clicking its occurrence in the concordance, they can see their wider contexts.

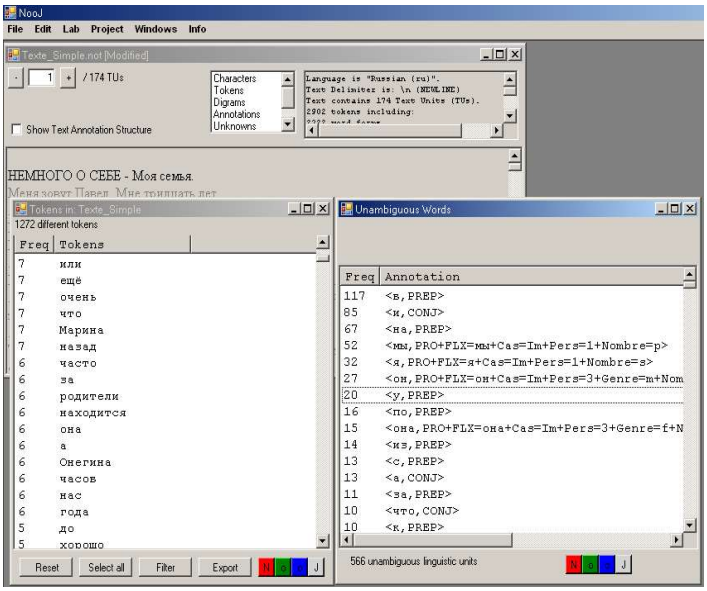

Fig. 1 : list of tokens and unambiguous words

NooJ’s lexical parser produces for a given sentence all its potential analyses, see figure 2. Because Nooj displays all the ambiguities, students must decide how to disambiguate them (e.g. is “мне” a dative or prepositional form?, does “лет” correspond to “лето” or “год”?).

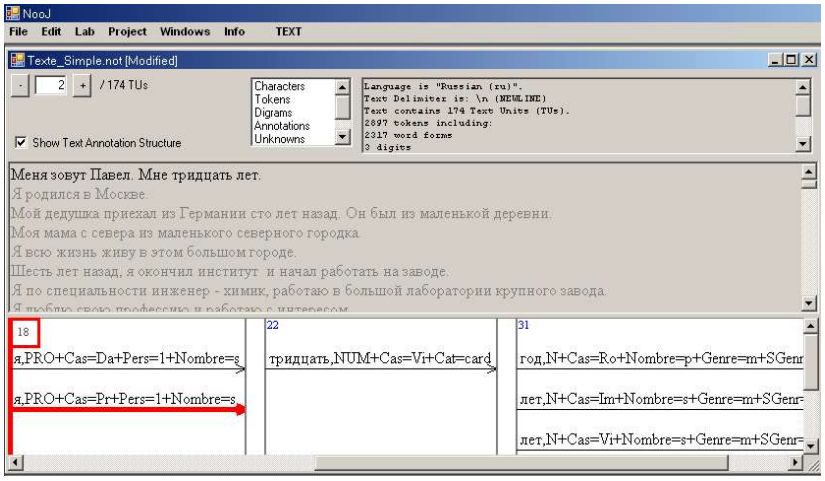

Fig.2: Lexical analysis

NooJ's syntactic grammars represent patterns that can be applied to texts to recognize certain syntactic structures and can therefore be used as queries.

For instance, grammar "Vb mvt" (movement verbs) recognizes simple movement verbs with no preverb. By applying this grammar to a text, one obtains the list of all movement verbs in the text. In the same way, grammar « Name.nog » can be used to extract from a text all the structures that correspond to "меня зовут" or "это называется".

These simple tools can thus be used as simple observation exercises that push the students to solve ambiguities.

Semantic features can also be used in class. Figure 3 below shows the occurrences of adjectives of color in Anton Chekhov's novel "The lady with the Dog", obtained by the simple query: <A+Color>. Semantic features can be used in literary studies. It is interesting to find out that the novel contains mostly occurrences of the black, grey and white colors, and that the red color is only mentioned once.

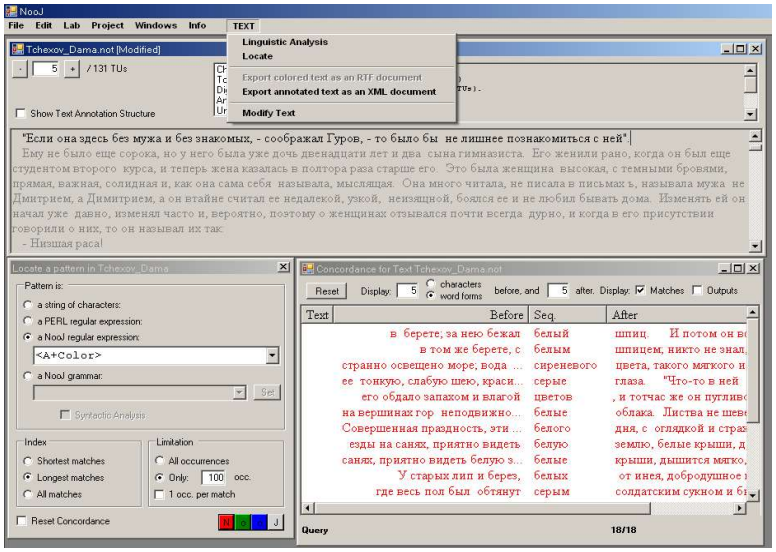

Fig. 3: Color Adjectives <A+Color>

Various thematic studies can be performed just by using the semantic codes previously mentioned. For instance, looking for “body parts” would show that hands and eyes are the ones that are the most frequent in the novel.

Finally, developing NooJ grammars (such as disambiguation rules) can constitute a proper goal in itself for the most advanced students. Figure 4 displays grammar « Prep_Na.nog » used to automatically disambiguate the wordform « на» (in “на столе” vs. “на, возьми!”).

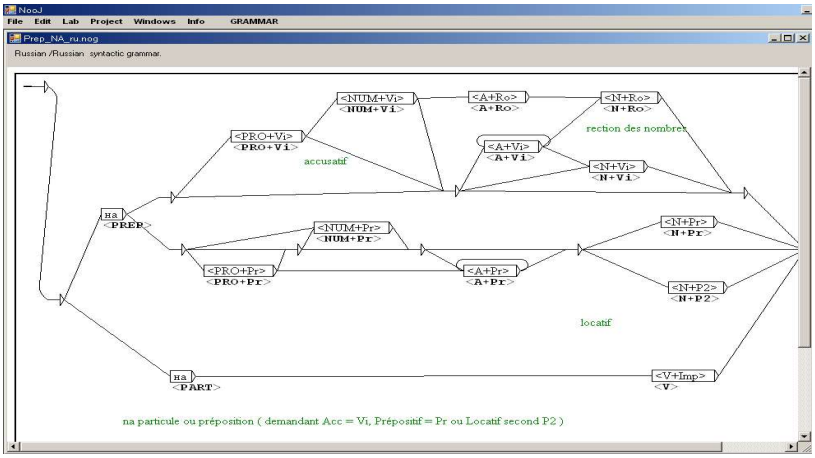

Fig. 4: Disambiguisating grammar for "HA"

Disambiguation grammars describe the minimal contexts needed to disambiguate certain wordforms. Figure 4 shows that the wordform "на" will be interpreted as a preposition if it is followed by a word in the accusative or prepositional form, whereas it will be a particle when it is followed by a verb in the imperative.

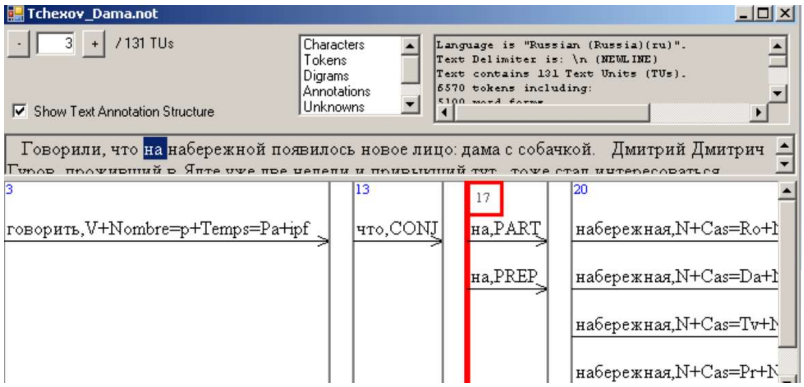

Fig.5: Annotated Text before the application of the disambiguation grammar for "HA"

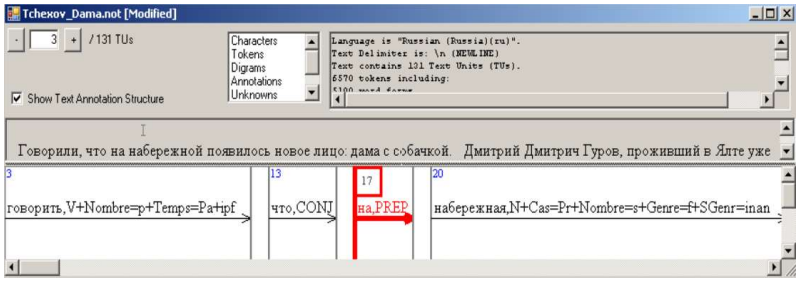

Fig. 6: Annotated Text after the application of the disambiguation grammar for "HA"

Finally, here is a grammar used to recognize the possessive structure in Russian:

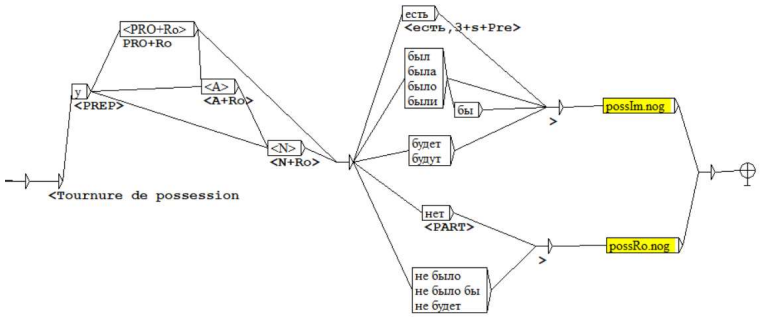

Fig.7: Grammar that represents the possessive structure in Russian

In Russian, the possessive structure uses the genitive and nominative cases. It is important (specifically in pedagogical applications), to recognize all its occurrences, at any tense and any person.

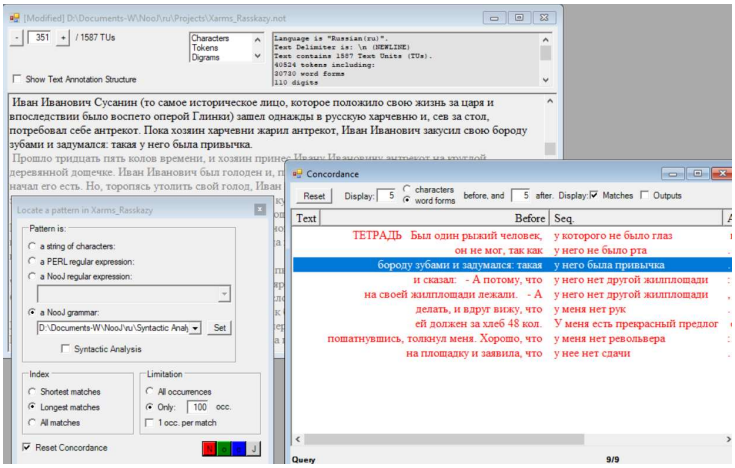

Fig. 8: Concordance produced by applying the Possessive grammar to the text

There are dozens of similar Nooj grammars that recognize Russian structures to express ages, dates, durations, locations, etc. that are of interest to all learners of Russian.

Conclusion,

As we have seen, linguistic resources are of interest both to linguists who can apply them to large texts as language teachers who can easily create exercises for students. NooJ's limitations are only those of the linguistic resources one is willing to develop, and new ones are developed for Russian as well as for other languages every day by linguists of the NooJ community.

References

1. *Zaliznyak A.A.* Grammaticheskij slovar' russkogo jazyka, ed. Russkij Jazyk, [Grammatical dictionary of Russian language] Moscow,1977

2. *Tuzov B.A.*, Komp'juternaja semantika russkogo jazyka, ed. Sankt Petersburg University, 2004 [Computer Semantics of Russian Language]

3. *Silberztein, Max.* 2016. Formalizing Natural Languages: the NooJ approach. Wiley Eds: Hoboken (NJ).

4. *Silberztein, Max.* 2003-. NooJ manual. available at the WEB site http://www.nooj4nlp.net/NooJManual.pdf .

5. *Bea Ehmann, László Balázs, Dmitry Shved, Vincent Bénet and Vadim Gushin,* The Russian linguistic resources of NooJ *in Space Psychological Research, in Nooj* (2012).

6. *Yury and Sviatlana Hetsevich* , Overview of Belarussian and Russiab Electronic Dictionnaries, *in Selected Papers from the 2011 International NooJ Conference* pp29-40.(2011).

Web Sites

Nooj available at: http:// www.nooj-association.org/
Russian National Corpus: http://ruscorpora.ru
Computer Fond of Russian Language: http://cfrl.ruslang.ru
Corpus of Standard Written Russian: http://www.narusco.ru

Max Silberztein
University of Franche-Comté (France).
E-mail: ***max.silberztein@gmail.com***

Vincent Bénet
INALCO, National Institute of Oriental Languages and Civilisations (France).
E-mail: ***vincent.benet@inalco.fr***