***V. Bénet & M. Silberztein***

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

**Abstract**

We present a set of linguistic resources developed for Russian: a morphological dictionary associated with a set of grammars to describe the inflection, a semantic component as well as a set of syntactic grammars that solve various types of ambiguities and recognize several named entities. We show how these resources can be used to automatically process corpora and how they can be used to teach Russian as a second language.

**Key Words**

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

**Introduction**

To study Russian corpora, most linguists and language teachers make extensive and almost exclusive use of the “ruscorpora.ru” site of the Russian National Corpus,[[1]](#footnote-1) which contains texts with complete morphosyntactic and semantic tagging, *i.e.* completely pre-disambiguated. Russian linguists can also use the “cfrl.ruslang.ru” collection site of the Computer Fund of Russian Language, which can display occurrences of wordforms (but not lemmas) for a verified but limited set of texts.

Most users of corpora (linguists, language teachers as well as many researchers in the humanities and the social sciences) need to explore and study other texts, such as those that constitute the Computer Fund of Russian Language, and more generally, any corpus of texts that they might have collected, from any source. Sketch Engine does offer access to Russian texts published on the Internet. Unfortunately, its linguistic functionalities, while satisfactory for users who are performing global statistical analyses on graphical wordforms, are not reliable enough for any precise linguistic analysis.[[2]](#footnote-2)

The NooJ software is a linguistic development environment platform that allows users to formalize eight levels of linguistic phenomena, for any written language: spelling and typography, inflectional, derivational and agglutinative morphology, local, 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), as well as software engineering tools that make it possible to develop, test, accumulate and share linguistic resources with wide coverage. NooJ linguistic resources at any level can then be combined automatically applied to large texts to perform various analyses, often in real time; hence, NooJ is being used as a corpus processor by many linguists, language teachers are more generally researchers in the digital humanities.[[3]](#footnote-3)

The INALCO institute[[4]](#footnote-4) has been collaborating with the Vinogradov Russian Language Institute of the Russian Academy of Sciences for over thirty years, working with J. Anoshkina's Unilex and A. Baranov’s Dialex software[[5]](#footnote-5), which have made it possible to develop grammatical and morphosyntactic resources for NooJ in a relatively short period of time, using Zalizniak's grammatical dictionary.[[6]](#footnote-6) However, these two software applications did not allow linguists to describe syntactic grammars and apply them to Russian texts.

**The grammatical dictionary**

We have developed three dictionaries: a dictionary of common nouns; a dictionary of proper nouns, and a dictionary of substantive adjectives. The latter dictionary was constructed to avoid the tedious description of homographic forms of adjectives and nouns; if this dictionary were not activated, words like русский [*Russian*] or новое [*new*] would always be processed as adjectives (and not as potential substantives). Together, the three dictionaries contain about 3,500,000 wordforms, associated with over 95,000 different linguistic analyses.

Following are extracts from the dictionary and the corresponding morphological grammar:

волейбол,N+m+inan+Sport+FLX=завод

волк,N+m+an+Animal+FLX=волк

газировать,V+ipf+pf+FLX=интересовать

The codes following each lexical entry indicate the category and linguistic properties of the word: N=Noun, m=Masculine, an=Animated; Sport is a semantic domain, etc. The FLX property specifies the entry’s morphological paradigm. For example, following is the definition of paradigm “завод”:

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 ;

Inflectional paradigms associate each inflected wordform with several properties, such as Case (Im, Vi, Ro, Da, TV, Pr) and Number (s or p). NooJ uses a dozen basic morphological operators such as <B> (delete current letter), as well as other specifically defined for a certain language; for instance, there is a special reduplication operator <D> specific to Amerindian languages, a special consonant finalization operator <F> for Semitic languages, etc.[[7]](#footnote-7)

NooJ’s operators are different from Zalizniak’s. They have been designed to be analytic so they are better adapted to the western Slavic tradition. In NooJ, all properties’ names and values must be defined in a separate property definition file.[[8]](#footnote-8)

The Russian property definition file defines 11 categories: adjectives (А), adverbs (ADV), nouns (N), numerals (NUM), pronouns (PRO), verbs (V), prepositions (PREP), conjunctions (CONJ), interjections (INTERJ), particles (PART), and pparenthetic phrases (INTRO). These categories are further associated with several properties. For instance, adjectives are associated with seven properties:

Genre = m | f | n ; *(masculine, feminine and neutral)*

SGenre = an | inan ; *(animated and non-animated)*

Number = s | p; *(singular and plural)*

Case = Im | Vi | Ro | Da | Tv | Pr | Zv; *(nominative, accusative, genitive, dative, instrumental, prepositional and vocative)*

Degree = Comp | Sup ; *(comparative and superlative)*

Sem = App | Color | Body ; *(Semantic features)*

Nouns are associated with the same properties as adjectives, plus a few extra, for example:

Sem = Human | Forename | Profession | Parent | Body | Concrete | Abstract | Organization | Text | Animal | Food | Arts | Literature | Music | Sports | Topography | Country | River | City | Mount | Lake | Posit | Time | Color ;

For instance, there are 409 lexical entries associated with the class “Animal”. Grammatical words (Numerals, Pronouns, Prepositions etc.) have a limited set of properties, e.g.:

Case = Im | Vi | Vip | Ro | Rop | Da | Dap | Tv | Tvp | Pr ;

The massive polysemy of grammatical words forced us to duplicate many entries (for instance: causality or origin for « из » et « от »). When we develop more and more complex syntactic grammars and apply them to corpora, we will be able to decide if it is better to process these words as highly ambiguous (*i.e.* associated with multiple different sets of properties), or not, depending on the fact that these properties might be redundant with the semantic value of the following noun or verb.

**Semantic dictionary**

We have associated over 5,000 lexical entries with 33 semantic features such as the “Sem” class (see above). Here is a sample of NooJ’s dictionary:

барсук,N+m+an+Animal+FLX=рыбак

барсучонок,N+m+an+Animal+FLX=волчонок

барсенок,N+m+an+Animal+FLX=утенок

The semantic classes are based on Tuzov’s semantic dictionary[[9]](#footnote-9), which is structured as an ontology (semantic tree), and contains over 145,000 entries, in a format similar to the following one:

*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}

We have converted this file into the clearer NooJ formalism. For example, $12132411 is now represented by the sequence of features: “+PhysObj +Inanimate +Thing +Technics +Transport +Terrestrial +Noengine”. We made sure that each class in Tuzov’s ontology corresponds to one and only one semantic feature in the resulting NooJ dictionary.

**Current limitations of our dictionary**

Initially, we had chosen to ignore the “ë” and the tonic accent because they are never indicated in written texts (except for pedagogical applications); we are planning to add them. Maximova and Gulyakova’s[[10]](#footnote-10) online Russian grammatical dictionary does not contain any written accent but ë. Note that Hetsevich’s team at the Institute of Informatics in Minsk has developed a Russian dictionary that contains the “ë” and tonic accents.[[11]](#footnote-11).

In Zalizniak’s dictionary,[[12]](#footnote-12) imperfective and perfective forms of a verb are represented as two independent lexical entries. Unfortunately, this means that NooJ cannot link them automatically and thus cannot perform transformations from one aspect to the other one. We are currently working on associating the perfective and imperfective forms of verbs (but only when they are linked semantically).

**Pedagogical applications**

There are numerous pedagogical applications of NooJ in the Language Teaching domain.[[13]](#footnote-13) Following is an example of a lab session:

After launching the NooJ software, select “ru” in the menu “Info⟶Preferences”, then load a text file. Select the lexical and morphological resources to be used via the command “Info⟶ Preferences⟶Lexical Analysis”; select the syntactic and semantic resources to be used via the command “Info⟶ Preferences⟶Syntactic Analysis”.

The first exercise consists of retrieving the list of all the wordforms associated with their linguistic analyses and their frequency, as well as the list of all “unknown” wordforms, *i.e.* those that were not recognized by any of the selected linguistic resources. Students can then see instantly the list of words that they don’t know. By double-clicking any occurrence in the concordance, they can then see their wider contexts in the text.

NooJ’s lexical parser displays for a given sentence all its potential analyses and ambiguities in the Text Annotation Structure (TAS), see figure 1. Annotations that are displayed in parallel correspond to linguistic ambiguities (at the lexical, morphological, syntactic or semantic levels). One exercise is therefore to ask students to solve all the ambiguities by deleting the wrong annotations in the TAS (*e.g.* is “мне” a dative or prepositional form? Does “лет” correspond to “лето” or “год”?).[[14]](#footnote-14)



*Fig. 1:* Lexical analysis

Students can enter queries in the form of syntactic grammars and apply them to the text, to locate certain sequences of interest. For instance, grammar “Vb mvt” (movement verbs) recognizes simple sequences that contain movement verbs with no preverb. By applying this grammar to a text, one obtains the list of all occurrences of 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 grammars are also be used by students to solve ambiguities.

Semantic features are also used in class. Figure 2 below shows the occurrences of adjectives of color in Anton Chekhov’s novel “The lady with the Dog”, obtained by entering the simple query: <A+Color>. It is interesting to find out that the novel contains mostly occurrences of the colors black, grey and white, and that the color red is mentioned only once.

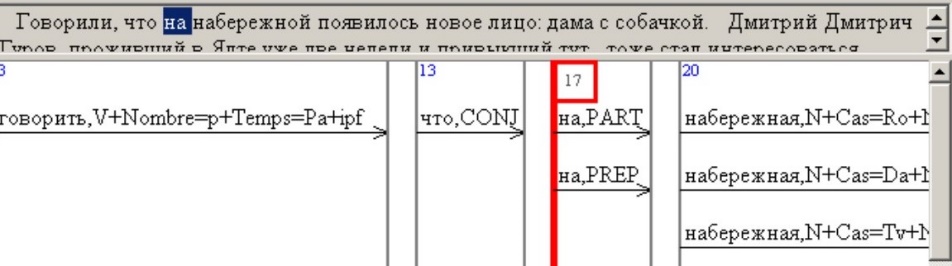
**

*Fig. 2:* Color Adjectives

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 more sophisticated grammars (such as disambiguation rules) constitute a goal for the most advanced students. Figure 3 displays a grammar used to automatically disambiguate the wordform “на” (in “на столе” vs. “на, возьми!”).

*Fig. 3:* **Disambiguisating grammar for “HA”**

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

****

*Fig. 4:* Initial Text Annotation Structure (TAS)

****

*Fig. 5:* **TAS after the disambiguation of “HA”**

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

**Conclusion**

We have seen that developing wide-coverage and precise linguistic resources is beneficial, to linguists who want to formalize a natural language, to users of corpora who want to apply sophisticated queries to their own texts, and to language teachers who want to demonstrate to their students how certain linguistic phenomena are occurring in real texts.

Our dictionary is unique because it is fully and freely available for download, and it is fully compatible with the NooJ platform, which guarantees that NooJ’s more sophisticated functionalities (such as the formalization of derivational morphology, of constituent and dependency syntax, as well as transformational engine, etc.) can be used to perform more and more sophisticated analyses.

The 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**

Anoshkina, J.G., 1993. Preparing frequency dictionaries and concordances on computer with Unilex-T.  Moscow, Institute of Russian Language, Computer Fund of Russian Language.

Baranov A.N., 2001. *Introduction to applied linguistics: Tutorial*. Moscow, Editions URSS. ISBN 5-8360-0196-0.

Frigière J. and Fuentes S., 2015. Pedagogical Use of NooJ dealing with French as a Foreign Language. In *Formalising Natural Languages with NooJ 2014*. Monti J., Silberztein M., Monteleone M., Di Buono M. Eds. Cambridge Scholars Publishing.

Hetsevich Y., Hetsevich S., 2012. Overview of Belarussian and Russian Electronic Dictionaries, *in Selected Papers from the 2011 International NooJ Conference* pp. 29-40.

Kilgariff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V., 2014. *The Sketch Engine: ten years on*. Published online: 10 July 2014. Springer.

Khokhlova M. 2009. *Applying Word Sketches to Russian*. In RASLAN, *Recent Advances in Slavonic Natural Language Processing*, pp 91-98.

Maksimova, E. A. “*Russian Grammatical Dictionary.”* *Seelrc.Org*, 2011

Nagel, S., 2002. Formenbildung im Russischen. *Formale Beschreibung und Automatisierung für das CISLEX-Wörterbuchsystem*.

Rodrigo A., Reyes S. and Bonino R., 2018. Some Aspects Concerning the Automatic Treatment of Adjectives and Adverbs in Spanish: A Pedagogical Application of the NooJ Platform. In *Formalising Natural Languages with NooJ 2017 and its Natural Language Processing Applications*. Springer CCIS Series #811.

Silberztein,M*.* 2003. NooJ manual. available at the WEB site <http://www.nooj4nlp.net/NooJManual.pdf>.

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

Silberztein M. 2021. Linguistic Resources for Corpus Processing: the ATISHS project. *Proceedings of the 56th Linguistic Colloquium*. Linguistik International: Peter Lang International Academic Publishers.

Tuzov *В.А.*, 2004. Komp’juternaja semantika russkogo jazyka, ed. Sankt Petersburg University, 2004 [Computer Semantics of Russian Language].

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Bénet Vincent**

INALCO, France.

***Email : vincent.benet@inalco.fr***

**Silberztein Max**

University of Franche-Comté (France).

***Email: max.silberztein@gmail.com***

1. See website: http://ruscorpora.ru [↑](#footnote-ref-1)
2. See (Kilgarriff et al. 2014) and (Khokhlova 20xx). (Silberztein 2021) shows that corpus processing tools that do not have access to precisely handcrafted dictionaries and grammars are not reliable enough for many linguistic analyses. [↑](#footnote-ref-2)
3. NooJ is a free software and runs on several platforms (Windows, Macos, Linux and Unix) and can be downloaded at: http://www.nooj-association.org. [↑](#footnote-ref-3)
4. The *Institut National des Langues et Civilisations Orienta*les (INALCO) is a French university specialized in the study and teaching of Oriental and Slavic Languages. [↑](#footnote-ref-4)
5. See (Anoshkina, 1993) and (Baranov, 2001). [↑](#footnote-ref-5)
6. See (Zalizniak, 1977). (Nagel 2002) describes a morphological Russian dictionary extracted from Dostoïevski’s *The Gambler* novel, but only 15% of its content is freely available. As it was developed on the INTEX/Unitex platform (see note 7), it contains very limited description (no syntax nor semantics), cannot be easily enhanced to link perfective and imperfective lexical entries and is not reversible and therefore cannot be used by a transformational engine such as NooJ’s. [↑](#footnote-ref-6)
7. For a comparison between NooJ, INTEX and Unitex, see http://www.nooj-association.org/intex-and-unitex.html. [↑](#footnote-ref-7)
8. The possibility for linguists to define their own set of properties is a crucial characteristic of NooJ; however, this freedom is sometimes unfortunate for linguists who work on several languages of a given linguistic family (*e.g.* Russian, Belarusian and Ukrainian), for which different property definition files have been designed by different teams of linguists. [↑](#footnote-ref-8)
9. See (Tuzov 2004). [↑](#footnote-ref-9)
10. See (Maximova & Gulyakova, 2011) [↑](#footnote-ref-10)
11. See (Hetsevich & Hetsevich, 2012). [↑](#footnote-ref-11)
12. See (Zaliznyak 1977). [↑](#footnote-ref-12)
13. See for instance (Frigière & Fuentes, 2015) and (Rodrigo 2018). [↑](#footnote-ref-13)
14. Sometimes, students have found problems in the coverage of some of our grammars: either an ambiguous word that should have been disambiguated, or a word that was wrongly disambiguated. With their feedback, we have been able to correct and enhance our disambiguation grammars. [↑](#footnote-ref-14)