MASTER 1 & 2

Outils

Corpus de textes numériques russes : constitution et exploitation

Descriptif

Contenu des séminaires

Le traitement automatique de la langue

Objectifs et enjeux

Traitement automatique du russe

Corpus du CFRL

Corpus de RUSCORPORA

Etiquetage de RUSCORPORA

Recherche dans RUSCORPORA

Logiciel de Concordance et de traitement de corpus

NOOJ

MASTER 1 et 2

Etudes russes

Tous parcours,

parcours « sciences du langage » et parcours « enseignement »

M2 S9 (accessible également en M1 S7) :

Corpus de textes numériques russes

DESCRIPTIF

Objectif du cours : Utiliser les corpus de textes électroniques et leurs logiciels d'exploitation Ce cours peut être utile aux étudiants qui souhaitent travailler avec les nouvelles technologies, s'orienter vers les métiers de l'enseignement, de l'édition numérique, de la conception des méthodes de langues, ou de l'application des nouvelles technologies dans l'étude et l'enseignement de la langue ou de la culture

Pré-requis : niveau Informatique C2.

Contenu du cours: Présentation des différentes ressources comportant des textes russes électroniques (Internet, CD ROM, bibliothèques électroniques). Principes de la constitution de corpus. Rappels sur les polices, tables de caractères et transcodage. Utilisation de logiciels d'exploitation de corpus, initiation aux principes d'étiquetage de textes.

Evaluation : Examen théorique et pratique avec constitution de dossier (somme des travaux pratiques effectués pendant le semestre)

Ce cours est crédité de **3 ECTS** ou **6 ECTS** si un dossier est réalisé :

Code Inalco RUS4A02C (3 ECTS) et UE RUS4A12 (6 ECTS) RUS4A02C RUS4A02D

Horaire : ***13 semaines, soit 19h30 de cours***

PLAN DU SEMINAIRE

1-2. Rappels essentiels sur les codages / logiciels de conversion de codes, les bibliothèques en ligne, les types de textes disoponibles en ligne ou sur CD ROM

3-4. Définition et utilité des corpus de textes. Les corpus de textes en ligne / les corpus de textes constitués / méthodes de recherche dans les corpus. Consultation des différents corpus et constitution de son propre corpus

5-7. Maniement des logiciels d'exploitation de corpus (Nooj, Unitex, Simple Concordance etc.) Constitution de concordances.

8-10. Etude des lemmatiseurs, des normes d'étiquetage (parsage)

11-13 . Réalisation pratique d'un dossier

Traitement automatique des langues et Corpus de textes

Objectifs, enjeux , définitions, historique

Le **Traitement automatique du langage naturel** (abr. *TALN*) ou **Traitement automatique des langues** (abr. *TAL*) est une discipline à la frontière de la linguistique et de l'informatique, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. Parmi les applications les plus connues, on peut citer :

- la traduction automatique (historiquement la première application, dès les années 1950) ;
- la correction orthographique ;
- la recherche d'information et la recherche d'occurrence
- l'élaboration de lexiques
- les statistiques d'emploi
- le résumé automatique ;
- la génération automatique de textes ;
- la synthèse de la parole ;
- la reconnaissance vocale ;
- la reconnaissance de l'écriture manuscrite.

Le **traitement automatique de la langue** vise l'élaboration de programmes informatiques pour la manipulation automatique de données linguistiques, en vue de leur exploitation, de leur transformation. Cette manipulation automatique traite un objet d'entrée, un texte écrit, une séquence orale enregistrée, une liste de mots, et le transforme en objet de sortie, par exemple, une traduction, un texte écrit, un paradigme. Cette transformation est effectuée par un ordinateur au moyen d'un logiciel.

Pour être traitée, **la langue naturelle** doit évidemment être transformée en données linguistiques numérisées (textes écrits saisis à l'ordinateur, textes écrits au sens classique du terme : oeuvres littéraires ou extraits, notices, lettres, transcriptions de dialogues, etc. . .)

La difficulté de traitement réside dans la connaissance des principes de constitution interne de la langue naturelle. - décrire les règles de la langue, - les formaliser - les implémenter dans une machine afin de pouvoir les traiter.

Le traitement automatique des langues naturelles ne concerne que des données linguistiques composées de chaînes de caractères "finies".

Or la langue "naturelle" ne peut pas simplement se réduire à un simple codage d'information. Sont aussi importants que les données proprement dites les paramètres contextuels, situationnels et culturels qui composent le système de compréhension. Il faudra donc, à un moment ou à un autre, intégrer des données extra-textuelles qui permettront de prendre en compte un ensemble composé du lexique, de la syntaxe, de la sémantique et de la pragmatique. L'intégration de ces données extra-linguistiques se fera par un marquage ou un étiquetage des données linguistiques entrées dans la machine.

La langue naturelle devra donc, pour pouvoir être traitée automatiquement avoir été "étiquetée" suivant un certain nombre de critères propres au traitement voulu.

La langue est devenue ainsi un produit, et les divers traitements automatiques sont autant d'**enjeux** culturels, politiques, et économiques, ce que n'ont pas manqué de remarquer les industriels. C'est ainsi que ce sont développées ce qu'il est convenu d'appeler maintenant les industries de la langue.

le Corpus

Un **corpus** est un ensemble de documents, artistiques ou non (textes, images, vidéos, etc.), regroupés dans une optique précise. On peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifiques, etc.

Le corpus en linguistique
La branche de la linguistique qui se préoccupe plus spécifiquement des corpus s'appelle logiquement la *linguistique de corpus*.
On parle de corpus pour désigner l'aspect normatif de la langue : sa structure et son code en particulier.
Le corpus en littérature
Ensemble de textes vérifiés regroupant les œuvres complètes d'un auteur, ou bien un domaine particulier de son œuvre (correspondance, essais, prose, poésie, théâtre etc.)
Ensemble de textes choisis sur un domaine particulier, un sujet spécifique à une époque donnée, ou couvrant l'évolution du domaine sur plusiers époques. le corpus doit être établi en fonction de son étude.
Le corpus dans la science
Les corpus sont des outils indispensables et précieux en traitement automatique du langage naturel. Ils permettent en effet d'extraire un ensemble d'information utile pour des traitements statistiques.
D'un point de vue informatif, ils permettent d'extraire des tendances et notamment de construire des ensembles de n-grammes (théorie de probabilité statistique d'apparition d'une séquence éventuellement non – présente dans le corpus)
D'un point de vue méthodologique, ils apportent une objectivité nécessaire à la validation scientifique en traitement automatique du langage naturel. L'information n'est plus empirique, elle est vérifiée par le corpus. Il est donc possible de s'appuyer sur des corpus (à condition bien entendu qu'ils soient bien formés) pour formuler et vérifier des hypothèses scientifiques.

Corpus bien formé
Plusieurs caractéristiques sont à prendre en compte pour la création d'un corpus bien formé :

- la taille ;
- le langage du corpus ;
- le temps couvert par les textes du corpus ;
- le registre ;

Taille
Le corpus doit évidemment atteindre une taille critique pour permettre des traitements statistiques fiables. Il est impossible d'extraire des informations fiables à partir d'un corpus trop petit
Langage
Un corpus bien formé doit nécessairement couvrir un seul langage, et une seule déclinaison de ce langage. Il existe par exemple de subtiles différences entre le français de France et le français parlé en Belgique. Il ne sera donc pas possible de tirer des conclusions fiables à partir d'un corpus franco-belge sur le français de France, ni sur le français de Belgique.
Temps couvert par les textes du corpus
Le temps joue un rôle important dans l'évolution du langage : le français parlé aujourd'hui ne ressemble pas au français parlé il y a 200 ans ni, de façon plus subtile, au français parlé il y a 10 ans, à cause notamment des néologismes. C'est un phénomène à prendre en compte pour toutes les langues vivantes. Un corpus ne doit donc pas contenir de textes rédigés à des intervalles de temps trop larges.
Registre de langage
Il ne faut pas non plus mélanger des registres différents et le scientifique ne peut s'autoriser à extraire des informations d'un corpus destiné à un certain registre en les appliquant à un autre. Un corpus construit à partir de textes scientifiques ne peut être utilisé pour extraire des informations sur les textes vulgarisés, et un corpus mélangeant des textes scientifiques et vulgarisés ne permettra de tirer aucune conclusion sur ces deux registres.

Méthodologie du travail sur les corpus
Il serait maladroit d'un point de vue méthodologique d'appliquer des traitements statistiques sur le corpus qui a permis de faire ressortir un classement ou une modélisation du langage.
Lorsque l'on travaille avec des corpus, il convient donc de séparer un corpus initial en deux sous corpus:

- **le corpus d'apprentissage**, qui sert à retirer un modèle ou un classement à partir d'un nombre suffisant d'information ;
- **le corpus de test**, qui sert à vérifier la qualité de l'apprentissage à partir du corpus d'apprentissage.

Le calibrage des volumes des corpus se discute en fonction du problème, mais il est fréquent d'utiliser les 2/3 du corpus initial pour l'apprentissage et le tiers restant pour effectuer les tests.

Lorsque le volume du corpus initial n'est pas suffisant, il est possible de croiser les corpus de tests et d'apprentissage sur **plusieurs expérimentations**. Par exemple, si l'on découpe le corpus initial en 10 sous-corpus, numérotés de 1 à 10

- Expérience 1 : utilisation des corpus 1 à 8 en apprentissage, et 9 et 10 pour les tests;
- Expérience 2 : utilisation des corpus 1 à 6 et 9 et 10 en apprentissage, 7 et 8 pour les tests;
- ...

La mesure de qualité des résultats (précision ou rappel) est alors plus précise, mais **en aucun cas les corpus d'apprentissage et de tests n'ont été mélangé**.

Le projet de Fonds informatisé de la langue russe. (***http://cfrl.ru***)

La constitution du fonds informatisé de la langue russe

Le projet de constitution d'un fonds informatisé de la langue russe sous la direction et M. Andriouchtchenko a conduit à élargir le champ d'investigation de la linguistique russe, à commencer par ce que l'on entend par dictionnaire et langue littéraire.

La langue littéraire russe (= traduction du russe *russkij literaturnyj jazyk*. Ce concept de "langue littéraire", inventé par le linguiste Vinogradov, désigne la langue russe normée et attestée, et exclut le jargon, les dialectes, la langue vulgaire (*mat*)), regroupant traditionnellement la production littéraire artistique, la littérature de vulgarisation scientifique et la langue de la presse, inclut ce que l'on appelle couramment la langue commerciale, la langue scientifique et technique, la langue de documentation technique, qui sont des langues de communication à part entière.

Les dictionnaires académiques et grammaticaux, qui doivent être les sources principales des données pour les dictionnaires informatisés et les programmes informatiques doivent eux aussi être revus. En effet, ils ne satisfont pas aux exigences d'exhaustivité, de rigueur et de précision de description des faits de langue.

Objectifs de la constitution du fonds informatisé de la langue russe (1989) :

- l'élaboration et l'implémantation de systèmes d'automatisation de recherches linguistiques
- la mise sur support informatique de toute la richesse lexicale de la langue russe
- la constitution d'un fonds d'algorithmes et de programmes linguistiques,
- la constitution d'un fonds de systèmes d'analyse automatique et de synthèse du texte russe
- la constitution de systèmes automatiques d'interrogation en linguistique

Ceci impliquait que devaient être créés entre autres :

- un lexique général de la langue russe, contenant l'inventaire de tous les mots et des locutions figées mentionnées dans les dictionnaires, encyclopédies et autres sources.
- des fonds terminologiques englobant tout le lexique fixés par les systèmes d'informations sectoriels et les standards terminologiques
- un fonds dictionnairique et grammatical comportant tous les dictionnaires académiques de la langue russe et les données des grammaires académiques.

Il fallait donc transférer sur support informatique les dictionnaires existants, créer automatiquement des lexiques augmentés, unifier les données générales langagières et terminologiques, créer des programmes linguistiques (UNILEX, Dialex, Wordtab, MAK)

Comment transformer les données papier en données numériques? par la saisie manuelle et la saisie automatique par OCR.

sources : Thèse VB et Wikipédia

RUSCORPORA

What is the Corpus?

A corpus is a reference system based on an electronic collection of texts composed in a certain language. A national corpus represents that language at a stage (or several stages) of its development in all the variety of genres, styles, territorial and social variants of usage, etc.

A national corpus is created by linguists (specialists in corpus linguistics, a fast-developing discipline) for academic research and language teaching. Most of the major world languages have their own corpora. A well-recognized example is the British National Corpus, which is used as a model for many modern corpora. Among the Slavic languages, the Czech National Corpus (compiled at the Charles University of Prague) is notable.

A national corpus is distinguished by two features. Firstly, it is characterized by representative and well-balanced collections of texts. This means that such a corpus contains, if possible, all the types of written and oral texts present in the language (various genres of fiction, journalistic, academic, and business, as well as dialectal and sociolectal, texts). The proportion of text types in the corpus is based on their share in real-life usage at the time of composition. A representative corpus is necessarily a large one (containing up to several million tokens). The planned size of the Russian National Corpus is 200 million words.

Secondly, a corpus contains additional information on the properties of texts that are included. This is achieved by means of annotation. The annotation is a principal feature of the corpus, distinguishing the corpus from simple collections (also known as 'libraries') of texts on the Internet, such as, in Russian, the Maksim Moshkov library or the Russian Virtual Library. Such libraries are not well suited to academic work on the nature of language; they tend to focus on the content of texts rather than their language properties, while the creators of the Corpus recognize the importance of literary or scientific value of the texts, but see them as a secondary feature. Unlike an electronic library, the National Corpus is not a collection of texts which are deemed 'interesting' or 'useful' of themselves; the texts in the Corpus are interesting and useful for the study of language. Such texts might include not only great works of literature, but also works of a 'secondary' writer, or a transcription of an ordinary conversation.

The academic and teaching value of a corpus is dependent upon the variety of annotation. The Russian National Corpus currently uses four types of annotation: metatextual (information about the text), morphological, accentual and semantic; the introduction of syntactic annotation is planned for the near future. The system of annotation is constantly being improved.

Что такое Корпус?

Корпус — это информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т. п.

Национальный корпус создается лингвистами (специалистами по так называемой *корпусной лингвистике*, быстро развивающейся современной области языкознания) для научных исследований и обучения языку. Большинство крупных языков мира уже имеет свои национальные корпуса (различающиеся по полноте и уровню научной обработки текстов). Общепризнанным образцом является, в частности, Британский национальный корпус (BNC): на него ориентированы многие другие современные корпуса. Среди корпусов славянских языков выделяется Чешский национальный корпус, созданный в Карловом университете Праги.

Национальный корпус имеет две важные особенности. Во-первых, он характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Следует иметь в виду, что хорошая представительность достигается только при значительном объеме корпуса (десятки и сотни миллионов словоупотреблений). Планируемый составителями объем Национального корпуса русского языка — 200 млн. слов.

Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию). Разметка — главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представленных в современном интернете, в том числе и на русском языке (таких, как, по-видимому, наиболее известная «библиотека Максима Мошкова» или, например, «Русская виртуальная библиотека»). В настоящее время специалистами создана и пополняется также «Фундаментальная электронная библиотека» русской классической литературы, ориентированная на академический режим подачи текстов, максимально точное воспроизведение авторитетных печатных изданий. Однако такие библиотеки в необработанном виде для научных исследований языка пригодны очень ограниченно. Не следует забывать также, что библиотеки создаются теми, кому интересно в большей степени содержание текстов, чем их языковые качества. Для составителей Национального корпуса такие факторы, как увлекательность или полезность книги, ее высокие художественные или научные достоинства являются важными, но не первостепенными. Национальный корпус, в отличие от электронной библиотеки, — это не собрание «интересных» или «полезных» текстов; это собрание текстов, интересных или полезных для изучения языка. А такими могут оказаться и роман второстепенного писателя, и запись обычного телефонного разговора,

	и типовой договор аренды и т.п. — наряду, конечно, с классическими произведениями художественной литературы. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. В Национальном корпусе русского языка в настоящее время используется четыре типа разметки: метатекстовая, морфологическая, акцентная и семантическая; в ближайшее время планируется внедрение синтаксической разметки. Система разметки постоянно совершенствуется.
The need for the corpus The main purpose of the corpus is to facilitate academic research on the lexicon and grammar of a language, as well as the subtle but constant processes of language change within a relatively short period of time: from one to two centuries. The other purpose of the corpus is to serve as a reference point for lexical, grammatical, and accentological questions, and the history of the language. Modern IT-technologies make the processing of large volumes of text significantly simpler and faster, which creates the possibility for mass statistical analysis of texts. As a result, language research now yields results which could only be guessed at previously. Nowadays, truly scientific descriptions of grammars and academic dictionaries must be based on corpora of their respective languages. The use of corpus data is desirable (if not always strictly necessary) in other, more specialized language research. Therefore, the main users of national corpora are linguists of various profiles. Nevertheless, the corpus is useful for non-linguists too. Reliable statistical information on language use in a certain period or by a certain author may be of interest for researchers of literature, history and other humanitarian subjects. National corpora are also useful for language teachers, both native and foreign; language textbooks and teaching programs are increasingly oriented toward corpora. A corpus can be used for ascertaining the variants of usage of unknown words by foreigners, students, teachers, journalists, writers. Therefore, the corpus is aimed at people who are interested in the structure and usage of a language, be their interest professional or not.	**Зачем нужен национальный корпус?** Национальный корпус предназначен в первую очередь для обеспечения научных исследований лексики и грамматики языка, а также тонких, но непрерывных процессов языковых изменений, происходящих в языке на протяжении сравнительно небольших периодов — от одного до двух столетий. Другая задача корпуса — предоставление всевозможных справок, относящихся к указанным областям (лексика, грамматика, акцентология, история языка). Современные компьютерные технологии многократно упрощают и ускоряют процедуры лингвистической обработки больших массивов текстов. Раньше исследователь мог лишь просматривать тексты и вручную выписывать из них нужные примеры; эта предварительная (но абсолютно неизбежная) деятельность была очень трудоемкой и не позволяла обрабатывать большие массивы материала. Теперь ограничений на объем анализируемого материала и скорость поиска информации в нем по существу нет, а это означает, что в распоряжении исследователя оказываются колоссальные массивы текстов самого разного типа. Это не замедлило сказаться на развитии наших знаний о языке: возможность массовой — в том числе статистической — обработки текстов, недоступная прежде, позволила обнаружить в структуре и развитии языка такие закономерности, о существовании которых наука раньше или не подозревала, или лишь смутно догадывалась, но не могла строго обосновать. Теперь подлинно научные описания грамматического строя языков, а также авторитетные академические словари — практически все без исключений — должны составляться на основе корпусов этих языков. Учет корпусных данных оказывается крайне желательным (если не строго обязательным) и при многих других более специальных научных исследованиях.
The development of the National Corpus The Russian National Corpus covers primarily the period from the middle of the 18th to the early 21st centuries. This period represents the Russian language of both the past and the present in a wide range of sociolinguistic variants: literary, colloquial, vernacular, in part dialectal. The Corpus includes original (non-translated) works of fiction (prose, drama and poetry) of cultural importance which are interesting from a linguistic point of view. Apart from fiction, the Corpus includes a large volume of other sources of written (and, for the later period, spoken) language: memoirs, essays, journalistic works, scientific and popular scientific literature, public speeches, letters, diaries, documents, etc. The Russian National Corpus includes the following subcorpora: The Deeply Annotated corpus, containing sentences with full morphological and syntax structure markup,	**Как развивается Национальный корпус?** Национальный корпус русского языка охватывает прежде всего период от середины XVIII до начала XXI века: этот период представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах — литературном, разговорном, просторечном, отчасти диалектном. В корпус включаются оригинальные (непереводные) произведения художественной литературы (проза и драматургия, в дальнейшем также поэзия), имеющие культурную значимость, а также представляющие интерес с точки зрения языка. Но Национальный корпус ни в коей мере не является только корпусом языка художественной литературы. Помимо художественных текстов, в корпус в большом количестве включаются и другие образцы письменного (а для современного этапа — и устного) языка: мемуары, эссеистика, публицистика, научно-популярная и научная литература, публичные выступления, частная переписка, дневники, документы и т. п. Национальный корпус русского языка в настоящее время

The Parallel Russo-English corpus, which facilitates searches for all translations for a certain Russian or English word of phrase,

The Dialectal corpus, which includes recordings of dialectal speech from various regions of Russia and represents dialectal morphological variations,

The Poetry corpus, which facilitates searches not only by lexical and grammatical features but also by specifically poetical features, such as meter, rhyme types, etc,

The Educational corpus, a corpus of texts with disambiguated grammatical homonyms, which was adapted for the Russian school teaching program,

The Corpus of Spoken Russian which includes the recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies (1930-2007).

включает следующие подкорпуса:

глубоко аннотированный корпус, в котором для каждого предложения построена полная морфологическая и синтаксическая структура (дерево зависимостей);

параллельный русско-английский корпус текстов, в котором можно найти все переводы для определенного русского или английского слова или словосочетания;

корпус диалектных текстов, включающий запись диалектной речи различных регионов России с сохранением их грамматической специфики; предусмотрен специальный поиск с учётом диалектной морфологии;

корпус поэтических текстов, в котором возможен поиск не только по лексическим и грамматическим, но и по специфическим для стиха признакам (поиск определённого сочетания в сонетах, в эпиграммах, в стихотворениях, написанных амфибрахием, с определённым типом рифмовки и т. п.);

обучающий корпус русского языка — корпус со снятой омонимией, разметка которого ориентирована на школьную программу русского языка;

корпус устной речи - включает расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов 1930-2000-х годов.

Content and structure of the Corpus

The main corpus

The main corpus, which includes texts representing standard Russian, can be subdivided into 3 parts, each of which has its distinguishing features: modern written texts (from the 1950s to the present day), a subcorpus of real-life Russian speech (recordings of oral speech from the same period), and early texts (from the middle of the 18th to the middle of the 20th centuries). By default, the search is carried out in all the three sub-groups. It is possible to choose one of them and add search parameters on the “customize your corpus” page.

Every text included in the main corpus is subject to meta tagging and morphological tagging. Morphological tagging is carried out by computer programs for automated morphological analysis. In a small part of the main corpus (currently around 5 million tokens; this figure is set to increase with time) homonyms are disambiguated by hand and the results of automated morphological analysis corrected. This part is the model morphological corpus and serves as a testing ground for various search algorithms and programs of morphological analysis and automated processing. It can also be used for research on modern Russian morphology that requires particular preciseness. Examples of this subcorpus are annotated as “disambiguated” (“омонимия снята”). Disambiguated texts are automatically supplied with indicators of stress (from the *Grammatical dictionary of Russian*). Stress annotation may be turned off for printing or saving the search results.

Основной корпус текстов

Основной корпус — тексты, представляющие русский литературный язык, — можно подразделить на три главных массива, имеющих свои особенности: это **современные письменные тексты** (середина XX — начало XXI века), **корпус живой русской речи** (записи устных текстов того же периода) и **ранние тексты** (середина XVIII — середина XX века). По умолчанию поиск по этим трём массивам ведётся одновременно, выбрать один из них (и задать дополнительные 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 5 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 эталонный морфологический корпус, 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

Modern written texts

The representative corpus of morphologically tagged modern texts is the main and the largest of the subcorpora. The planned volume of the corpus is 100 million tokens. The corpus includes various types of texts representing modern standard (written) Russian:

- Modern fiction of various genres
- Modern drama
- Memoirs and biographies
- Journalism and literary criticism
- Scientific, popular scientific and teaching texts
- Religious and philosophical texts
- Technical texts
- Business and jurisprudence texts
- Day-to-day life texts, including texts not intended for publication (letters, diaries, etc.)

Texts are represented in proportion to their share in real-life usage. For example, the share of fiction (including drama and memoirs) does not exceed 40%.

The sources of book, magazine and newspaper texts included in the Corpus are usually proof-read electronic versions supplied by their respective publishers and the texts are used with publishers' permission.

Mid-18th to mid-20th century texts

Texts from the middle of the 18th century to the middle of the 20th century are also included in the Corpus and represent various genres (fiction, scientific texts, journalism, letters) but due to limited availability of such texts in electronic form or in modern reprints the proportion of fiction for this period is much higher than for the main corpus. Pre-1918 texts are given in modern orthography; peculiarities of their original orthography preserved in modern academic editions are also preserved in the Corpus.

Deeply Annotated Corpus

This subcorpus of the RNC contains texts augmented with morphosyntactic annotation. Besides the morphological information ascribed to each word in the text, every sentence has its syntax structure marked up.

The Deeply Annotated Corpus (DAC) uses dependency trees as its annotation formalism. Nodes in such a tree are words of the sentence, while its edges are labeled with names of syntax relationships. This way of representing the syntax structure originates from “Meaning ⇔ Text” linguistic model by Igor A. Mel’čuk and Alexander K. Zholkovsky. The repertory of syntactic relationships for the DAC, as well as other specific linguistic decisions on how to represent the syntax of Russian sentences, has been developed in the Laboratory for Computational Linguistics, Institute for Information Transmission, Russian Academy of Sciences that compiled the DAC.

Unlike the morphologically annotated portion of the RNC, the

сохранения/печати ударения могут быть сняты).

Современные письменные тексты

Представительный корпус современных текстов с морфологической разметкой является основным и самым объёмным из подкорпусов. Планируемый объем этого корпуса — 100 млн. словоупотреблений. 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
- журнальная публицистика и литературная критика
- газетная публицистика и новости
- научные, научно-популярные и учебные тексты
- религиозные и религиозно-философские тексты
- производственно-технические тексты
- официально-деловые и юридические тексты
- бытовые тексты (в том числе тексты, не предназначенные для публикации: личная переписка, дневники и т.п.)

Тексты представлены в определенной пропорции, отражающей их долю в общем массиве современных текстов. Так, доля художественных текстов (включая драматургию и мемуары) составляет не более 40%.

Источниками текстов, входящих в Корпус, для опубликованных книжных, журнальных и газетных текстов, как правило, являются выверенные электронные версии, предоставляемые издателями этих текстов (и используемые в Корпусе с разрешения издателей).

Тексты XVIII—середины XX вв. в Корпусе

Тексты XVIII—середины XX вв. в Корпусе представляют также различные жанры (художественная литература, научные тексты, частная переписка, публицистика), однако по причине доступности электронных версий и современных переизданий процент художественной литературы для этого периода гораздо выше, чем для второй половины XX в. Тексты, написанные до 1918 г., даются в послереформенной орфографии; те особенности оригинальной орфографии, которые сохраняются в научных переизданиях, сохраняются и в Корпусе.

Глубоко аннотированный корпус

Данный фрагмент Национального корпуса русского языка содержит тексты, снабженные морфо-синтаксической разметкой. Это значит, что помимо морфологической информации, приписанной каждому слову текста, для каждого предложения задана его синтаксическая структура.

Синтаксическая структура предложения, используемая в глубоко аннотированном корпусе (ГАК), представляет собой дерево зависимостей, в узлах которого стоят слова предложения, а ветви помечены именами синтаксических отношений. Такое представление о синтаксической структуре предложения восходит к лингвистической модели «Смысл ⇔ Текст» И.А.Мельчука и

DAC only contains fully disambiguiated annotations (i.e. both morphological and syntax ambiguity is resolved).

Parallel text corpus

The parallel text corpus is a special type of corpus where a text in Russian is complemented by its translation into a different language, and vice versa. The units of the original and the translated texts (usually, a unit is a sentence) are matched through a procedure known as “leveling”. A leveled parallel corpus is an important tool for various type of research, including studies on the theory of translation; it can also be used as a language teaching tool.

This site contains a small levelled Russo-English parallel text corpus.

Dialectal corpus

The dialectal corpus contains recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia. There is no intention to present the phonetic variation, but morphological, syntactic and lexical peculiarities of these texts are preserved. The subcorpus employs special tags for specifically dialectal morphological features (including those absent in standard language); moreover, purely dialectal lexemes are supplied with commentary.

Poetry corpus

At the moment the poetry corpus covers the time frame between 1750 and 1890s, but also includes some poets of the 20th century; currently, works of drama composed in poetry are not included. Apart from the usual morphological tagging (identical to that available for the non-disambiguated corpus), there is a number of tags adapted for poetry. For example, it is possible to search for texts written in various poetic meters such as amphibrach.

Educational corpus

The educational corpus is a small disambiguated corpus adapted for the Russian educational program, including works of fiction on the school reading list and several additional morphological features.

Corpus of Spoken Russian

The Corpus of Spoken Russian includes the recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies. To record the spoken specimens the standard spelling was used. The lexical, morphological and semantic queries are practicable. The building of the user's sub-corpora is available (for this purpose the usage of the sociological parameters is also possible). The corpus contains the patterns of different genres/types and of different geographic origins (Moscow, Sanct-Peterburg, Saratov, Ulyanovsk, Taganrog, Ekaterinburg, and so on). The corpus covers the time frame from 1930 to 2007.

А.К.Жолковского. Окончательный перечень синтаксических отношений, используемых в ГАК, а также целый ряд конкретных лингвистических решений, связанных с представлением синтаксической структуры предложения, был выработан в Лаборатории компьютерной лингвистики Института проблем передачи информации РАН. Силами коллектива этой Лаборатории и составлен ГАК.

В отличие от морфологически размеченного фрагмента Национального корпуса русского языка, ГАК целиком состоит из структур со снятой морфологической и синтаксической омонимией.

Корпус параллельных текстов

Особым типом корпуса является так называемый параллельный корпус, в котором тексту на русском языке сопоставлен перевод этого текста на другой язык или, наоборот, тексту на иностранном языке сопоставлен его перевод на русский язык. Между единицами оригинального и переводного текста (обычно — между предложениями) с помощью специальной процедуры устанавливается соответствие; эта процедура называется выравниванием, а тексты, соответственно, выровненными.

Выровненный параллельный корпус представляет собой важный инструмент для научных исследований (в том числе и для исследований по теории и практике перевода); он может также использоваться при обучении русскому и иностранным языкам.

В настоящее время на сайте Национального корпуса размещён небольшой выровненный параллельный русско-английский корпус.

Корпус диалектных текстов

Корпус диалектных текстов включает в себя записи диалектной речи (в орфографии, приближенной к стандартной) из различных регионов России. Задачи передать фонетическую информацию не ставится; при этом полностью сохранена морфологическая, синтаксическая и лексическая специфика текстов.

Корпус поэтических текстов

Корпус поэтических текстов включает стихотворные произведения. В настоящее время хронологический охват — примерно 1750-1850-е гг. (со включением нескольких авторов XX в.); в корпус пока не включены стихотворные драматические сочинения..

Обучающий корпус русского языка

Обучающий корпус русского языка — небольшой корпус со снятой омонимией, ориентированный на преподавание русского языка в школе (отобраны произведения из школьной программы, изучаемых в школьном курсе функциональных стилей, размечены словоизменительные типы и другие дополнительные морфологические признаки)

Корпус устной речи

Корпус устной речи включает в себя расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов. Использована русская стандартная орфография (при этом приводятся наиболее частотные и общепринятые стяженные формы). Возможен лексический, морфологический и семантический поиск, а также формирование пользовательских подкорпусов, в том числе и по социологическим параметрам. Включены тексты самых разных жанров и типов, разного происхождения с точки зрения географии (Москва, Санкт-Петербург, Саратов, Ульяновск, Таганрог, Екатеринбург, Норильск, Воронеж, Новосибирск и мн. др.). Хронологический охват корпуса 1930-2000-е гг.

Corpus Statistics

In January 2008, the Russian National Corpus contained 52 392 texts consisting of 149 357 020 tokens

Национальный корпус русского языка в январе 2008 г. содержал **52 392** текста общим объемом **149 357 020** словоупотреблений.

I. Подкорпус	**Число текстов**	**Число предложений**	**Число словоупотреблений**	**% словоупотреблений**
Основной корпус	59 489	16 205 733	193 915 626	56.8%
- в том числе со снятой омонимией	2 142	516 860	5 944 267	1.7%
Газетный корпус	181 175	8 553 495	113 292 003	33.2%
Диалектный корпус	197	20 273	194 283	0.1%
Обучающий корпус	229	65 666	664 751	0.2%
Параллельный корпус	326	1 257 224	17 570 179	5.1%
Поэтический корпус	37 869	574 567	6 225 404	1.8%
Устный корпус	2 830	1 536 190	9 606 442	2.8%
Всего:	**282 115**	**28 213 148**	**341 468 688**	**100%**

Вид текста	**Число текстов**	**Число предложений**	**Число словоупотреблений**	**% словоупотреблений**
Художественные письменные тексты	5 961	8 485 384	85 780 792	44.2%
Нехудожественные письменные тексты	53 528	7 720 349	108 134 834	55.8%
Всего:	**59 489**	**16 205 733**	**193 915 626**	**100**

II. Tokens by part of speech **Часть речи**	**Число словоупотреблений**	**% словоупотреблений**
существительное	1 693 312	28.5%
прилагательное	506 851	8.5%
числительное	102 039	1.7%
- в том числе записанное прописью	42 595	0.7%
- в том числе записанное цифрами	59 444	1.0%
числительное-прилагательное	24 535	0.4%
глагол	1 007 618	17.0%

наречие	246 213	4.1%
предикатив	42 280	0.7%
вводное слово	25 891	0.4%
местоимение-существительное	467 455	7.9%
местоимение-прилагательное	277 634	4.7%
местоимение-наречие	129 369	2.2%
местоимение-предикатив (*некого, нечего*)	680	0.0%
предлог	621 883	10.5%
союз	471 309	7.9%
частица	268 104	4.5%
междометие	8 375	0.1%
инициал	10 138	0.2%
прочие (иностранные слова, звукоподражания)	30 847	0.5%
Всего:	**5 934 533**	**100%**

Le standard de codage russe morphologique et sémantique

Морфологический стандарт Национального корпуса русского языка

Структура морфологической информации

Морфологическая информация, приписываемая произвольному слову в тексте, состоит из четырех «полей», или групп помет:

1. Лексема, которой принадлежит словоформа (указывается «словарная запись» данной лексемы и ее принадлежность к той или иной части речи).
2. Множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (например, род для существительного, переходность для глагола).
3. Множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола).
4. Информация о нестандартности грамматической формы, орфографических искажениях и т. п.

Части речи

S — существительное (*яблоня, лошадь, корпус, вечность*)
A — прилагательное (*коричневый, таинственный, морской*)
NUM — числительное (*четыре, десять, много*)
A-NUM — числительное-прилагательное (*один, седьмой, восьмидесятый*)
V — глагол (*пользоваться, обрабатывать*)
ADV — наречие (*сгоряча, очень*)
PRAEDIC — предикатив (*жаль, хорошо, пора*)
PARENTH — вводное слово (*кстати, по-моему*)
S-PRO — местоимение-существительное (*она, что*)
A-PRO — местоимение-прилагательное (*который, твой*)
ADV-PRO — местоименное наречие (*где, вот*)
PRAEDIC-PRO — местоимение-предикатив (*некого, нечего*)
PR — предлог (*под, напротив*)
CONJ — союз (*и, чтобы*)
PART — частица (*бы, же, пусть*)
INTJ — междометие (*увы, батюшки*)

Значения грамматических категорий

Род:

m — мужской род (*работник, стол*)
f — женский род (*работница, табуретка*)
m-f — «общий род» (*задира, пьяница*)
n — средний род (*животное, озеро*)

Одушевленность:

anim — одушевленность (*человек, ангел, утопленник*)
inan — неодушевленность (*рука, облако, культура*)

Число:

sg — единственное число (*яблоко, гордость*)
pl — множественное число (*яблоки, ножницы, детишки*)

Падеж:

nom — именительный падеж (*голова, сын, степь, сани, который*)
gen — родительный падеж (*головы, сына, степи, саней, которого*)

dat — дательный падеж (*голове, сыну, степи, саням, которому*)
acc — винительный падеж (*голову, сына, степь, сани, который/которого*)
ins — творительный падеж (*головой, сыном, степью, санями, которым*)
loc — предложный падеж (*[о] голове, сыне, степи, санях, котором*)
gen2 — второй родительный падеж (*чашка **чаю***)
acc2 — второй винительный падеж (*постричься **в монахи**; по **два** человека*)
loc2 — второй предложный падеж (*в **лесу**, на **осу́***)
voc — звательная форма (*Господи, Серёж, ребят*)
adnum — счётная форма (*два **часа́**, три **шара́***)

Краткая/полная форма:

brev — краткая форма (*высок, нежна, прочны, рад*)
plen — полная форма (*высокий, нежная, прочные, морской*)

Степень сравнения:

comp — сравнительная степень (*глубже*)
comp2 — форма «*по*+сравнительная степень» (*поглубже*)
supr — превосходная степень (*глубочайший*)

Вид:

pf — совершенный вид (*пошёл, встречу*)
ipf — несовершенный вид (*ходил, встречаю*)

Переходность:

intr — непереходность (*ходить, вариться*)
tran — переходность (*вести, варить*)

Залог:

act — действительный залог (*разрушил, разрушивший*)
pass — страдательный залог (только у причастий: *разрушаемый, разрушенный*)
med — медиальный, или средний залог (глагольные формы на *-ся*: *разрушился* и т.п.)

Форма (репрезентация) глагола:

inf — инфинитив (*украшать*)
partcp — причастие (*украшенный*)
ger — деепричастие (*украшая*)

Наклонение:

indic — изъявительное наклонение (*украшаю, украшал, украшу*)
imper — повелительное наклонение (*украшай*)
imper2 — форма повелительного наклонения 1 л. мн. ч. на *-те* (*идемте*)

Время:

praet — прошедшее время (*украшали, украшавший, украсив*)
praes — настоящее время (*украшаем, украшающий, украшая*)
fut — будущее время (*украсим*)

Лицо:

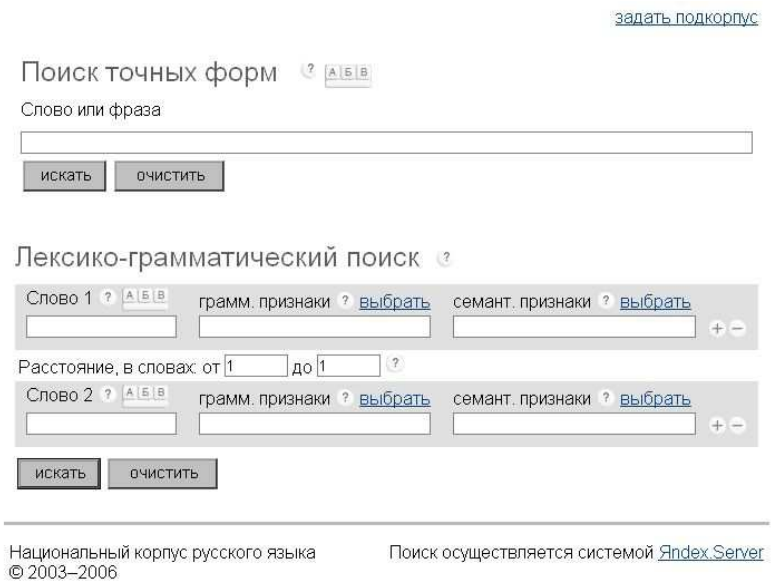
1p — первое лицо (*украшаю*)
2p — второе лицо (*украшаешь*)
3p — третье лицо (*украшает*)

Прочие признаки:

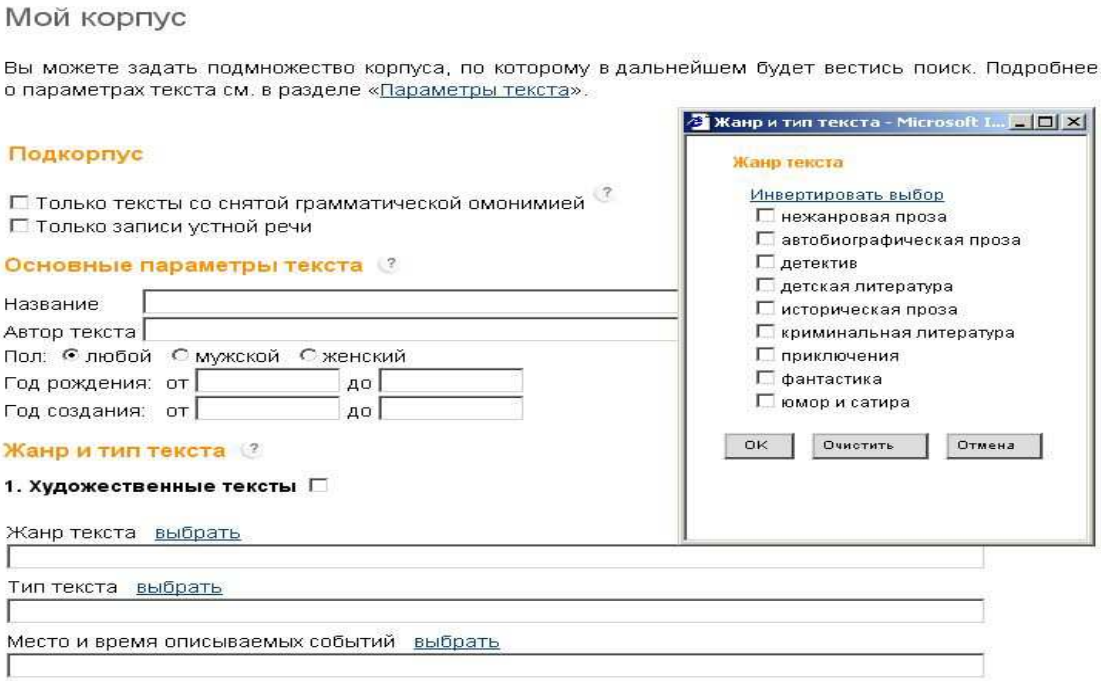
persn — личное имя (*Иван, Дарья, Леопольд, Эстер, Гомер, Маугли*)
patrn — отчество (*Иванович, Павловна*)
famn — фамилия (*Николаев, Волконская, Гумбольдт*)
0 — несклоняемое (*шоссе, Седых*)

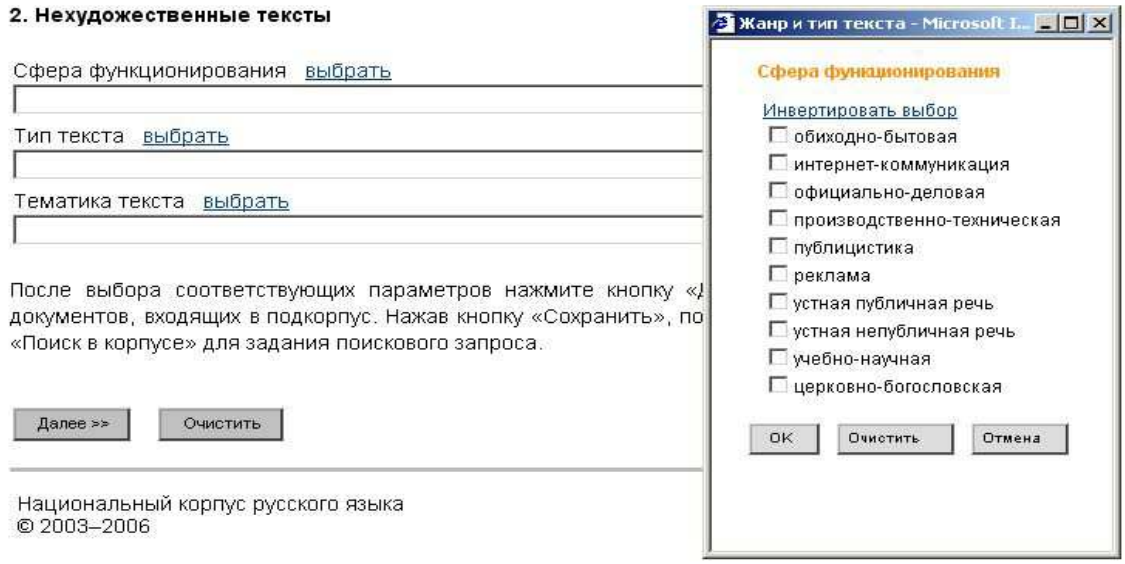
Часть указанных помет (а именно, второй винительный падеж, звательная форма, счётная форма, форма по+сравнительная степень, общий род, переходность, несклоняемость) присутствуют только в корпусе со снятой грамматической омонимией.

ПОИСК ФОРМ



подкорпус :





ÉTIQUETAGES MORPHOLOGIQUE, SYNTAXIQUE ET SEMANTIQUE (PARSING)

Etiquettes pour un marquage grammatical des unités lexicales pour le russe.

- partie du discours
- morphologie
- syntaxe
- étymologie (mot emprunté, langue étrangère)
- éventuellement des indications sémantiques, stylistiques et de domaine de langue.

codage simplifié de la version électronique du dictionnaire de Zalizniak

La structure de la base contient 6 champs : M, NO, AC, CAT, TY, DI:

structure ancienne (conservée)	structure nouvelle	signification
ID	**ID**	**numéro d'entrée**
M	**M**	**entrée**
NO	**NO**	
ACC	**SCH**	**schéma accentuel**
CAT	**CAT**	**partie du discours**
	SSCAT1	**genre /aspect**
	SSCAT2	**ssgenre : animé-inanimé** **/ss aspect : det-indet**
	P1	**particularités de déclinaison / conjugaison** à définir (locatif 2, génitif 2)
	P2	**particularités de déclinaison / conjugaison** à définir (pluriel irreg)
	P3	**autres particularités :** voyelle mobile
TY	**TY**	**type**
DI	**DI**	**remarques diverses**

```
м (= masculin)
ж (= féminin)
с (= neutre)
мо(= masculin animé)
жо (= féminin animé)
со (= neutre animé)
мн (= pluriel)
св (= perfectif)
св нп (= perfectif intransitif)
св безл (=perfectif impersonnel)
нсв (=imperfectif)
нсв нп (=imperfectif intransitif)
нсв безл (=imperfectif impersonnel)
нсв-св (= verbe à double aspect imperfectif/perfectif)
н (= adverbe)
предик (= adverbe prédicatif)
п (=adjectif)
союз (=conjonction)
межд (=interjection)
част (=particule)
предл(=préposition)
ввод (= mot introductif)
  срав (=comparatif))
```

TY = le type de déclinaison ou conjugaison correspondant aux numéros des tables du dictionnaire papier (1a, 1b, 3*a, 4a $I(-<a-)1, etc. .). Ces types ne sont évidemment pas très « parlants ».

DI = divers, c'est-à-dire du texte contenant des commentaires sur les entrées.

DICTIONNAIRES en VERSION ELECTRONIQUE

Application : les dictionnaires de Ожегов, Зализняк, Фасмер sur Internet
http://starling.rinet.ru/cgi-bin/main.cgi?flags=wygtmnl

О. Н. Ляшевская, С. А. *Шаров* НОВЫЙ ЧАСТОТНЫЙ СЛОВАРЬ

dict.ruslang.ru/freq.php

www.artint.ru/projects/frqlist.phP

NOOJ

à télécharger **sur** **http://www.nooj4nlp.net/pages/nooj.html**

C'est un logiciel de traitement de corpus puissant, auquel on peut greffer des données lexicales et sémantiques. L'utilisation la plus simple est la constitution d'un lexique à partir d'un ou plusieurs textes , la recherche d'occurrences et la constitution de concordances.

L'objet de cet exposé est pas de décrire la démarche d'écriture et de poser les problèmes ainsi que es solutions apportées, et ensuite une petite séance de travaux pratique pour l'installation des ressources et le maniement du logiciel..

PRINCIPES D'ECRITURE DU LEMMATISEUR

SOURCE :
Les 96000 entrées du dictionnaire Zaliznjak au format électronique avec leur codage tel qu'il est présent dans la version papier

ETIQUETTES

Un jeu d'étiquette mis au point pendant la redaction de ma thèse, fait à partir du codage de Zaliznjak.

Pourquoi avoir créé un autre jeu :
- Codage de Zaliznjak est très complet mais mal commode à utiliser Нсв нп1а о 5 еть
Мо 3*b – (….)

- **manque d'étiquettes** : pas de N, pas de V, mais des étiquettes groupées . Ceci était justifié lorsque chaque octet comptait ainsi on a 7 codes pour les substantifs мо, м, со, с, жо, ж, мн ; mais ces codes sont synthétiques et malcommodes à utiliser. Ceci a été décomposé et redistribué en N+m+f+n+s+p+an+inan, soit 8 codes, mais donne la possibilité simple de trier.
- **Etiquettes « accentuelles »** servent de base au dictionnaire (les schémas accentuels a b c) Assez peu utile pour le moment et pour les textes envisagés. Mais les modèles ont été fait pour garder l'accent.
- **Mots avec voyelle mobile non indiqués clairement** (codage * mal commode, pas d'indication automatique de la nature de la voyelle ni de sa place dans le mot)
- étiquettes superflues (explication du sens du mot ou autre schéma accentuel)
- **indications en clair inutilisables informatiquement** : les particularités sont données dans le cadre de notes ou directement en clair dans les colonnes du dictionnaire (par exemple les doublons ями / ьми de certains mots de 3ème déclinaison)
-

PROBLEMES
Techniques
Problèmes techniques
Excel traite 65000 lignes, le dictionnaire a 96000 entrées
Nécessité de composer des petits dictionnaires correspondants aux types
Ecrire une routine permettant le tri par odre alphabétique inverse / normal, afin de pouvoir trier sélectivement par la fin des mots

Accents
Les accents sont un problème que nous avons mis de côté.
- Problème technique liée à l'écriture de l'accent :
Pas de caractère propre pour le caractère accentué mais une combinaison de 2 caractères
U +caractère + U + accent 301 F008 etc. et le n° n'est pas fixe.
--> 1 caractère de plus que le nombre de lettres
Seules quelques polices proportionnelles affiche les accents
Gestion compliquée (surtout) en cas d'accent mobile nécessitant l'écriture de paradigmes non seulement en fonction du nombre de schémas accentuels et de type de déclinaison mais également en fonction du nombre de lettres dans le mot
-Nécessité d'écrire des routines informatique propres
De plus :
- Aucun correcteur orthographique n'inclut les accents
- Aucun texte susceptible d'être traité par Nooj (texte scanné ou en provenance d'internet) ne comporte les accents
- Il n'y a pas de standard de notation pour les accents
--> on a donc écrit des paradigmes souvent redondants (pour l'instant) tenant compte des schémas accentuels en vue d'un éventuel développement ultérieur.

La lettre ë
Absente de toute la typographie usuelle ; Dans un premier temps la lettre a été retirée
Des doublons ont été entrés quand le ë est initial (ёж & еж). Dans les flexions, il a été entré deux formes боев боёв pour бой

```
ёж,N+m+inan+FLX=нож
ёрник,N+m+an+FLX=бульдог
ёршик,N+m+an+FLX=бульдог
ёршик,N+m+inan+FLX=чайник
ёрш,N+m+an+FLX=богач
ёрш,N+m+inan+FLX=нож
ежик,N+m+an+FLX=бульдог
еж,N+m+an+FLX=богач
еж,N+m+inan+FLX=нож
ерник,N+m+an+FLX=бульдог
ершик,N+m+an+FLX=бульдог
ершик,N+m+inan+FLX=чайник
ерш,N+m+an+FLX=богач
ерш,N+m+inan+FLX=нож
```

Reconnaissance simple / génération / surgénération / quid des neologismes

Nooj fonctionne dans les deux sens :
En reconnaissance de formes & en génération de formes
Ceci a posé des problèmes qui ont fait l'objet de plusieurs discussions contradictoires

Pour l'auteur du logiciel Max Sylberztein, il faut éviter des générations du type je pleus, tu pleus, il pleut, nous pleuvons etc. Donc pas de вечерею, вечереешь etc.., pas non plus de читанный, de писав etc.
Nous voyons donc tout de suite la multiplication des paradigmes que cela va entraîner pour les verbes et la complexité d'attribution des types aux entités
Zalizniak utilise нсв св нп, codes qui ont été automatiquement repris dans notre version
On pourra évidemment définir automatiquement dans les paradigmes les gérondifs et participes convenables avec la correspondance suivante :
НСВ – GER IPF + PARTICIPE PRES ACTIF +PARTICIPE PASSE ACTIF + PARTICIPE PRES PASSIF
НСВ + НП – GER IPF + PARTICIPE PRES ACTIF +PARTICIPE PASSE ACTIF
СВ – GER PF + PARTICIPE PASSE ACTIF + PARTICIPE PASSE PASSIF
СВ + НП – GER PF + PARTICIPE PASSE ACTIF

Cela va multiplier par 4 les paradigmes pour un même type de verbe, on risque d'avoir des formes non reconnues, et il faudra systématiquement vérifier que les formes желанный битый сказуемое sont bien dans le dictionnaire…

PROCEDURE : DEFINITION DES PROPRIETES ET DES CARACTERISTIQUES AFFECTEES

A_Forme = fc | fl | adv;
A_Genre = m | f | n ;
A_SGenr = an | inan ;
A_Nombre = s | p;
A_Cas = Im | Vi | Ro | Da | Tv | Pr | Zv;
A_Deg = Comp | Sup ;

ADV_Deg = Comp;

N_Genre = m | f | n ;
N_SGenr = an | inan ;
N_Nombre = s | p;
N_Cas = Im | Vi | Ro | R2 | Da | Tv | Pr | P2 | Zv ;
N_Sem = Hum | Conc | Abstr | Org | Text |
ConcColl +
Cpmc | Immeub | Qual | Anim | Loc | Pdc |
Sent | Quant | Mat | Liq | Alim | Vehicl | Pr |
Tmp | Atm | Geom | CollHum | CollImmeub |
Mach ;

```
NUM_Cat = ord | card | coll
NUM_Genre = m |  f |  n ;
NUM_SGenr = an | inan ;
NUM_Nombre = s | p;
NUM_Cas = Im | Vi | Ro | Da | Tv | Pr ;

PRON_Genre = m | f | n;
PRON_SGenr = an | inan ;
PRON_Nombre = s | p;
PRON_Pers = 1 | 2 | 3;
PRON_Cas = Im | Vi | Vip | Ro | Rop | Da | Dap |
Tv | Tvp | Pr  ;

V_Pers = 1 | 2 | 3;
V_Asp = Ipf | Pf;
V_Temps = Pre | Pa | Fu;
V_Mode = Inf | Ind | Imp | Cond | Ger | Prtp ;
V_Voix = Act | Pss ;
V_Genre = m  |  f  |  n ;
V_Nombre = s | p ;
V_Sem = Intr | Tr | Refl ;
V_Cas =  Im | Vi | Ro | Da | Tv | Pr ;

PREP;
CONJ;
INTERJ;
PART;
INTRO;
```

ETABLISSEMENT DES TYPES

A partir du Zaliznjak

Substantifs :
1 type par code Zaliznjak + spécificités particulières (voyelle mobile, mouillure, yod etc.) non notées par un code
Exemple :
m1a=завод
#m1a=артист
#m2a=портфель
#mo2a=житель
#m3a=чайник
#m3a=бульдог
#m4a=марш
#m4a=товарищ
#m5a=месяц
#m5a=принц
#m6a=случай
#m6a=герой
#m7a=сценарий
#m7a=викарий
#m1b=топор
#m1b=бегун
#m2b=словарь
#m2b=карась
#m3b=пирог
#m3b=рыбак
#m4b=нож
#m4b=богач
#m5b=кострец
#m5b=кузнец
#m6b=лишай
#m6b=холуй
#m7b=кий
#m1c=дар
#m3c=плуг
#m4c=харч
#m6c=буй
#m1c=кол
#m3d=казак
#m1e=зуб
#m2e=голубь
#m3e=волк
#m4e=обруч
#m2f=конь

+ Spécificités non directement codées : palatalisations, pluriels irreguliers, voyelle mobile etc.

ECRITURE DES PARADIGMES : noms féminins, masculins

```
карта  = <E>/Im+s | <B>y/Vi+s | <B>ы/Ro+s |
<B>e/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s |
<B>e/Pr+s
 | <B>ы/Im+p | <B>ы/Vi+p | <B>/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ax/Pr+p ;
корова = <E>/Im+s | <B>y/Vi+s | <B>ы/Ro+s |
<B>e/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s |
```

<B>е/Pr+s | <B>ы/Im+p | <B>/Vi+p | <B>/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ах/Pr+p ;
неделя = <E>/Im+s | <B>ю/Vi+s | <B>и/Ro+s |
<B>е/Da+s | <B>ей/Tv+s | <B>ею/Tv2+s | <B>е/Pr+s
| <B>и/Im+p | <B>и/Vi+p | <B>ь/Ro+p | <B>ям/Da+p
| <B>ями/Tv+p | <B>ях/Pr+p ;
весна = <E>/Im+s | <B>у/Vi+s | <B>ы/Ro+s |
<B>е/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s | <B>е/Pr+s
| <B>ы/Im+p | <B>ы/Vi+p | <B2>ен/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ах/Pr+p ;
тюрьма = <E>/Im+s | <B>у/Vi+s | <B>ы/Ro+s |
<B>е/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s | <B>е/Pr+s
| <B>ы/Im+p | <B>ы/Vi+p | <B3>ем/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ах/Pr+p ;
судьба = <E>/Im+s | <B>у/Vi+s | <B>ы/Ro+s |
<B>е/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s | <B>е/Pr+s
| <B>ы/Im+p | <B>ы/Vi+p | <B3>еб/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ах/Pr+p ;
доска = <E>/Im+s | <B>у/Vi+s | <B>и/Ro+s |
<B>е/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s | <B>е/Pr+s
| <B>и/Im+p | <B>и/Vi+p | <B2>ок/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ах/Pr+p ;
серьга = <E>/Im+s | <B>у/Vi+s | <B>и/Ro+s |
<B>е/Da+s | <B>ой/Tv+s | <B>ою/Tv2+s | <B>е/Pr+s
| <B>и/Im+p | <B>и/Vi+p | <B3>ег/Ro+p |
<B>ам/Da+p | <B>ами/Tv+p | <B>ах/Pr+p ;
семья = <E>/Im+s | <B>ю/Vi+s | <B>и/Ro+s |
<B>е/Da+s | <B>ей/Tv+s | <B>ею/Tv2+s | <B>е/Pr+s
| <B>и/Im+p | <B2>ей/Vi+p | <B2>ей/Ro+p |
<B>ям/Da+p | <B>ями/Tv+p | <B>ях/Pr+p ;

случай = <E>/Im+s | <E>/Vi+s | <B>я/Ro+s |
<B>ю/Da+s | <B>ем/Tv+s | <B>е/Pr+s | <B>и/Im+p |
<B>и/Vi+p | <B>ев/Ro+p | <B>ям/Da+p |
<B>ями/Tv+p | <B>ях/Pr+p ;
герой = <E>/Im+s | <B>я/Vi+s | <B>я/Ro+s |
<B>ю/Da+s | <B>ем/Tv+s | <B>е/Pr+s | <B>и/Im+p |
<B>ев/Vi+p | <B>ев/Ro+p | <B>ям/Da+p |
<B>ями/Tv+p | <B>ях/Pr+p ;
зуй = <E>/Im+s | <B>я/Vi+s | <B>я/Ro+s | <B>ю/Da+s
| <B>ем/Tv+s | <B>е/Pr+s | <B>и/Im+p | <B>ев/Vi+p
| <B>ев/Ro+p | <B>ев/Vi2+p | <B>ев/Ro2+p |
<B>ям/Da+p | <B>ями/Tv+p | <B>ях/Pr+p ;
воробей = <E>/Im+s | <B2>ья/Vi+s | <B2>ья/Ro+s |
<B2>ью/Da+s | <B2>ьем/Tv+s | <B2>ьем/Tv2+s |
<B2>ье/Pr+s | <B2>ьи/Im+p | <B2>ьев/Vi+p |
<B2>ьев/Ro+p | <B2>ьям/Da+p | <B2>ьями/Tv+p |
<B2>ьях/Pr+p ;
сценарий = <E>/Im+s | <E>/Vi+s | <B>я/Ro+s |
<B>ю/Da+s | <B>ем/Tv+s | <B>и/Pr+s | <B>и/Im+p |
<B>и/Vi+p | <B>ев/Ro+p | <B>ям/Da+p |
<B>ями/Tv+p | <B>ях/Pr+p ;

ECRITURE DES PARADIGMES : adjectifs, numéraux

полный = <B2><L>о/fc+m+s | <B2>а/fc+f+s |
<B2>о/fc+n+s | <B2>ы/fc+p | <E>/fl+Im+m+s |
<E>/fl+Vi+m+s | <B2>ого/fl+Vi+m+an+s |
<B2>ого/fl+Ro+m+s | <B2>ому/fl+Da+m+s |
<B2>ым/fl+Tv+m+s | <B2>ом/fl+Pr+m+s |
<B2>ое/fl+Im+n+s | <B2>ое/fl+Vi+n+s |
<B2>ого/fl+Ro+n+s | <B2>ому/fl+Da+n+s |
<B2>ым/fl+Tv+n+s | <B2>ом/fl+Pr+n+s |
<B2>ая/fl+Im+f+s | <B2>ую/fl+Vi+f+s |
<B2>ой/fl+Ro+f+s | <B2>ой/fl+Da+f+s |
<B2>ой/fl+Tv+f+s | <B2>ою/fl+Tv2+f+s |
<B2>ой/fl+Pr+f+s | <B2>ые/fl+Im+p |
<B2>ые/fl+Vi+p | <B2>ых/fl+Vi+p | <B2>ых/fl+Ro+p
| <B2>ым/fl+Da+p | <B2>ыми/fl+Tv+p |
<B2>ых/fl+Pr+p | <B2>ее/Comp |
<LW>по<RW><B2>ее/Comp | <B2>ей/Comp |
<LW>по<RW><B2>ей/Comp;

пять = | <E>/Im | <E>/Vi | <B>и/Ro | <B>и/Da |
<B>ью/Tv | <B>и/Pr;
пятеро = <E>/Im | <B>ых/Vi+an | <B>ых/Ro |
<B>ым/Da | <B>ыми/Tv | <B>ых/Pr;
восемь = <E>/Im | <E>/Vi | <B3>ьми/Ro |
<B3>ьми/Da | ю/Tv | <B3>ьми/Pr;
пятьдесят = <E>/Im | <E>/Vi | <L5><B>и<R5>и/Ro |
<L5><B>и<R5>и/Da | <L5>ью<R5>ью/Tv |
<L5><B>и<R5>и/Pr;
восемьдесят = <E>/Im | <E>/Vi |
<L7><B>ьми<R5>и/Ro | <L7><B>ьми<R5>и/Da |
<L5>ью<R5>ью/Tv | <L7><B>и<R5>и/Pr;
сто = <E>/Im | <E>/Vi | <B>а/Ro | <B>а/Da |
<B>а/Tv | <B>а/Pr;
двести = <E>/Im | <E>/Vi | <B4>ухсот/Ro |
<B4>умстам/Da | <B4>умястами/Tv |
<B4>ухстах/Pr;

ECRITURE DES PARADIGMES : verbes

читать = <E>/Inf | <B2>ю/1+s+Pre |
<B2>ешь/2+s+Pre | <B2>ет/3+s+Pre |
<B2>ем/1+p+Pre | <B2>ете/2+p+Pre |
<B2>ют/3+p+Pre
| <B2>л/m+s+Pa | <B2>ла/f+s+Pa |
<B2>ло/n+s+Pa | <B2>ли/p+Pa
| <B2>й/2+s+Imp | <B2>йте/2+p+Imp
| <B2>я/Ger
| <B2>ющий/Part+Pre+Act+m+s+Im |
<B2>ющий/Part+Pre+Act+m+s+Vi |
<B2>ющего/Part+Pre+Act+m+an+s+Vi |
<B2>ющего/Part+Pre+Act+m+s+Ro |
<B2>ющему/Part+Pre+Act+m+s+Da |
<B2>ющим/Part+Pre+Act+mo+s+Tv |
<B2>ющем/Part+Pre+Act+mo+s+Pr |
<B2>ющая/Part+Pre+Act+f+s+Im |
<B2>ющую/Part+Pre+Act+f+s+Vi |
<B2>ющей/Part+Pre+Act+f+s+Ro |
<B2>ющей/Part+Pre+Act+f+s+Da |
<B2>ющей/Part+Pre+Act+f+s+Tv |
<B2>ющею/Part+Pre+Act+f+s+Tv |
<B2>ющей/Part+Pre+Act+f+s+Pr |
<B2>ющее/Part+Pre+Act+n+s+Im |
<B2>ющее/Part+Pre+Act+n+s+Vi |
<B2>ющего/Part+Pre+Act+n+s+Ro |
<B2>ющему/Part+Pre+Act+n+s+Da |
<B2>ющим/Part+Pre+Act+n+s+Tv |
<B2>ющем/Part+Pre+Act+n+s+Pr |
<B2>ющие/Part+Pre+Act+p+Im |
<B2>ющие/Part+Pre+Act+p+Vi |

<B2>ющих/Part+Pre+Act+an+p+Vi |
<B2>ющих/Part+Pre+Act+p+Ro |
<B2>ющим/Part+Pre+Act+p+Da |
<B2>ющими/Part+Pre+Act+p+Tv |
<B2>ющих/Part+Pre+Act+p+Pr |
<B2>вший/Part+Pa+Act+m+s+Im |
<B2>вший/Part+Pa+Act+m+s+Vi |
<B2>вшего/Part+Pa+Act+m+an+s+Vi |
<B2>вшего/Part+Pa+Act+m+s+Ro |
<B2>вшему/Part+Pa+Act+m+s+Da |
<B2>вшим/Part+Pa+Act+m+s+Tv |
<B2>вшем/Part+Pa+Act+m+s+Pr |
<B2>вшая/Part+Pa+Act+f+s+Im |
<B2>вшую/Part+Pa+Act+f+s+Vi |
<B2>вшую/Part+Pa+Act+f+s+Vi |
<B2>вшей/Part+Pa+Act+f+s+Ro |
<B2>вшей/Part+Pa+Act+f+s+Da |
<B2>вшей/Part+Pa+Act+f+s+Tv |
<B2>вшею/Part+Pa+Act+f+s+Tv |
<B2>вшей/Part+Pa+Act+f+s+Pr |
<B2>вшее/Part+Pa+Act+n+s+Im |
<B2>вшее/Part+Pa+Act+n+s+Vi |
<B2>вшего/Part+Pa+Act+n+s+Vi |
<B2>вшего/Part+Pa+Act+n+s+Ro |
<B2>вшему/Part+Pa+Act+n+s+Da |
<B2>вшим/Part+Pa+Act+n+s+Tv |
<B2>вшем/Part+Pa+Act+n+s+Pr |
<B2>вшие/Part+Pa+Act+p+Im |
<B2>вшие/Part+Pa+Act+p+Vi |
<B2>вших/Part+Pa+Act+an+p+Vi |
<B2>вших/Part+Pa+Act+p+Ro |
<B2>вшим/Part+Pa+Act+p+Da |
<B2>вшими/Part+Pa+Act+p+Tv |
<B2>вших/Part+Pa+Act+p+Pr |
<B2>емый/Part+Pre+Pss+m+s+Im |
<B2>емый/Part+Pre+Pss+m+s+Vi |
<B2>емого/Part+Pre+Pss+m+an+s+Vi |
<B2>емого/Part+Pre+Pss+m+s+Ro |
<B2>емому/Part+Pre+Pss+m+s+Da |
<B2>емым/Part+Pre+Pss+mo+s+Tv |
<B2>емом/Part+Pre+Pss+mo+s+Pr |
<B2>емая/Part+Pre+Pss+f+s+Im |
<B2>емую/Part+Pre+Pss+f+s+Vi |
<B2>емой/Part+Pre+Pss+f+s+Ro |
<B2>емой/Part+Pre+Pss+f+s+Da |
<B2>емой/Part+Pre+Pss+f+s+Tv |
<B2>емою/Part+Pre+Pss+f+s+Tv |
<B2>емой/Part+Pre+Pss+f+s+Pr |
<B2>емое/Part+Pre+Pss+n+s+Im |
<B2>емое/Part+Pre+Pss+n+s+Vi |
<B2>емого/Part+Pre+Pss+n+s+Ro |
<B2>емому/Part+Pre+Pss+n+s+Da |
<B2>емым/Part+Pre+Pss+n+s+Tv |
<B2>емом/Part+Pre+Pss+n+s+Pr |
<B2>емые/Part+Pre+Pss+p+Im |
<B2>емые/Part+Pre+Pss+p+Vi |
<B2>емых/Part+Pre+Pss+an+p+Vi |
<B2>емых/Part+Pre+Pss+p+Ro |
<B2>емым/Part+Pre+Pss+p+Da |
<B2>емыми/Part+Pre+Pss+p+Tv |
<B2>емых/Part+Pre+Pss+p+Pr |
<B2>ем/Part+Pre+Pss+m+fc |
<B2>ема/Part+Pre+Pss+f+fc |
<B2>емо/Part+Pre+Pss+n+fc |
<B2>емы/Part+Pre+Pss+p+fc;

ECRITURE DES PARADIGMES : verbes

прочитать = <E>/Inf | <B2>ю/1+s+Pre+Fu |
<B2>ешь/2+s+Pre+Fu | <B2>ет/3+s+Pre+Fu |
<B2>ем/1+p+Pre+Fu | <B2>ете/2+p+Pre+Fu |
<B2>ют/3+p+Pre+Fu
| <B2>л/m+s+Pa | <B2>ла/f+s+Pa |
<B2>ло/n+s+Pa | <B2>ли/p+Pa | <B2>й/2+s+Imp |
<B2>йте/2+p+Imp | <B2>в/Ger | <B2>вши/Ger |
<B2>вший/Part+Pa+Act+m+s+Im |
<B2>вший/Part+Pa+Act+m+s+Vi |
<B2>вшего/Part+Pa+Act+m+an+s+Vi |
<B2>вшего/Part+Pa+Act+m+s+Ro |
<B2>вшему/Part+Pa+Act+m+s+Da |
<B2>вшим/Part+Pa+Act+m+s+Tv |
<B2>вшем/Part+Pa+Act+m+s+Pr |
<B2>вшая/Part+Pa+Act+f+s+Im |
<B2>вшую/Part+Pa+Act+f+s+Vi |
<B2>вшую/Part+Pa+Act+f+s+Vi |
<B2>вшей/Part+Pa+Act+f+s+Ro |
<B2>вшей/Part+Pa+Act+f+s+Da |
<B2>вшей/Part+Pa+Act+f+s+Tv |
<B2>вшею/Part+Pa+Act+f+s+Tv |
<B2>вшей/Part+Pa+Act+f+s+Pr |
<B2>вшее/Part+Pa+Act+n+s+Im |
<B2>вшее/Part+Pa+Act+n+s+Vi |
<B2>вшего/Part+Pa+Act+n+s+Vi |
<B2>вшего/Part+Pa+Act+n+s+Ro |
<B2>вшему/Part+Pa+Act+n+s+Da |
<B2>вшим/Part+Pa+Act+n+s+Tv |
<B2>вшем/Part+Pa+Act+n+s+Pr |
<B2>вшие/Part+Pa+Act+p+Im |
<B2>вшие/Part+Pa+Act+p+Vi |
<B2>вших/Part+Pa+Act+an+p+Vi |
<B2>вших/Part+Pa+Act+p+Ro |
<B2>вшим/Part+Pa+Act+p+Da |
<B2>вшими/Part+Pa+Act+p+Tv |
<B2>вших/Part+Pa+Act+p+Pr |
<B2>нный/Part+Pa+Pss+m+s+Im |
<B2>нный/Part+Pa+Pss+m+s+Vi |
<B2>нного/Part+Pa+Pss+m+an+s+Vi |
<B2>нного/Part+Pa+Pss+m+s+Ro |
<B2>нному/Part+Pa+Pss+m+s+Da |
<B2>нным/Part+Pa+Pss+m+s+Tv |
<B2>нном/Part+Pa+Pss+m+s+Pr |
<B2>нная/Part+Pa+Pss+f+s+Im |
<B2>нную/Part+Pa+Pss+f+s+Vi |
<B2>нной/Part+Pa+Pss+f+s+Ro |
<B2>нной/Part+Pa+Pss+f+s+Da |
<B2>нной/Part+Pa+Pss+f+s+Tv |
<B2>нною/Part+Pa+Pss+f+s+Tv |
<B2>нной/Part+Pa+Pss+f+s+Pr |
<B2>нное/Part+Pa+Pss+n+s+Im |
<B2>нное/Part+Pa+Pss+n+s+Vi |
<B2>нного/Part+Pa+Pss+n+s+Ro |
<B2>нному/Part+Pa+Pss+n+s+Da |
<B2>нным/Part+Pa+Pss+n+s+Tv |
<B2>нном/Part+Pa+Pss+n+s+Pr |

<B2>нные/Part+Pa+Pss+p+Im |
<B2>нные/Part+Pa+Pss+p+Vi |
<B2>нных/Part+Pa+Pss+an+p+Vi |
<B2>нных/Part+Pa+Pss+p+Ro |
<B2>нным/Part+Pa+Pss+p+Da |
<B2>нными/Part+Pa+Pss+p+Tv |
<B2>нных/Part+Pa+Pss+p+Pr |
<B2>н/Part+Pa+Pss+m+s+fc |
<B2>на/Part+Pa+Pss+f+s+fc |
<B2>но/Part+Pa+Pss+n+s+fc |
<B2>ны/Part+Pa+Pss+p+fc;

ECRITURE DES DERIVATIONS

Dictionnaire dic

Григорий,N+Hum+m+an+FLX=Георгий+DRV=ьевич+DRV=ич
Гурий,N+Hum+m+an+FLX=Георгий+DRV=ьевич
Дементий,N+Hum+m+an+FLX=Георгий+DRV=ьевич
Дмитрий,N+Hum+m+an+FLX=Георгий+DRV=ьевич+DRV=ич
Димитрий,N+Hum+m+an+FLX=Георгий+DRV=евич+DRV=ич
Евгений,N+Hum+m+an+FLX=Георгий+DRV=ьевич+DRV=ич
Евсевий,N+Hum+m+an+FLX=Георгий+DRV=ьевич
Евстафий,N+Hum+m+an+FLX=Георгий+DRV=ьевич

Paradigmes nof

влович = <B2>лович/N+Im+m | <B2>ловича/N+Vi+m | <B2>ловича/N+Ro+m | <B2>ловичу/N+Da+m | <B2>ловичем/N+Tv+m | <B2>ловиче/N+Pr+m | <B2>ловна/N+Im+f | <B2>ловну/N+Vi+f | <B2>ловны/N+Ro+f | <B2>ловне/N+Da+f | <B2>ловной/N+Tv+f | <B2>ловною/N+Tv+f | <B2>ловне/N+Pr+f;
евич = <B>евич/N+Im | <B>евича/N+Vi | <B>евича/N+Ro | <B>евичу/N+Da | <B>евичем/N+Tv | <B>евиче/N+Pr | <B>евна/N+Im+f | <B>евну/N+f+Vi | <B>евны/N+f+Ro | <B>евне/N+f+Da | <B>евной/N+f+Tv | <B>евною/N+f+Tv | <B>евне/N+f+Pr;
ьевич = <B2>ьевич/N+Im | <B2>ьевича/N+Vi | <B2>ьевича/N+Ro | <B2>ьевичу/N+Da | <B2>ьевичем/N+Tv | <B2>ьевиче/N+Pr | <B2>ьевна/N+Im+f | <B2>ьевну/N+f+Vi | <B2>ьевны/N+f+Ro | <B2>ьевне/N+f+Da | <B2>ьевной/N+f+Tv | <B2>ьевною/N+f+Tv | <B2>ьевне/N+f+Pr; йлович = <B2>йлович/N+Im+m | <B2>йловича/N+Vi+m | <B2>йловича/N+Ro+m | <B2>йловичу/N+Da+m | <B2>йловичем/N+Tv+m | <B2>йловиче/N+Pr+m | <B2>йловна/N+Im+f | <B2>йловну/N+Vi+f | <B2>йловны/N+Ro+f | <B2>йловне/N+Da+f | <B2>йловной/N+Tv+f | <B2>йловною/N+Tv+f | <B2>йловне/N+Pr+f;
ьич = <B>ич/N+Im | <B>ич/N+Zv | <B>ича/N+Vi | <B>ича/N+Ro | <B>ичу/N+Da | <B>ичем/N+Tv | <B>иче/N+Pr | <B>ична/N+Im+f | <B>ичну/N+Vi+f | <B>ичны/N+Ro+f | <B>ичне/N+Da+f | <B>ичной/N+Tv+f | <B>ичною/N+Tv+f | <B>ичне/N+Pr+f;
ыч = ыч/N+Im | ыч/N+Zv | ыча/N+Vi | ыча/N+Ro | ычу/N+Da | ычем/N+Tv | ыче/N+Pr ;
ич = <B>ч/N+Im | <B>ч/N+Zv | <B>ча/N+Vi | <B>ча/N+Ro | <B>чу/N+Da | <B>чем/N+Tv | <B>че/N+Pr ;
йч = <B>ич/N+Im | <B>ич/N+Zv | <B>ича/N+Vi | <B>ича/N+Ro | <B>ичу/N+Da | <B>ичем/N+Tv | <B>иче/N+Pr ;
йлыч = <B2>йлыч/N+Im | <B2>йлыч/N+Zv | <B2>йлыча/N+Vi | <B2>йлыча/N+Ro | <B2>йлычу/N+Da | <B2>йлычем/N+Tv | <B2>йлыче/N+Pr ;

AFFECTATION DES PARADGIMES AUX ENTREES

adjectifs
адамов,A+FLX=отцов
адов,A+FLX=отцов
акулий,A+FLX=лисий
давящий,A+FLX=свежий
далекий,A+FLX=далекий
дальнейший,A+Sup+FLX=умнейший+DRV=наи
движущий,A+FLX=свежий
дворницкий,A+FLX=химический
легкий,A+FLX=яркий
легконогий,A+FLX=строгий
легонький,A+FLX=беленький

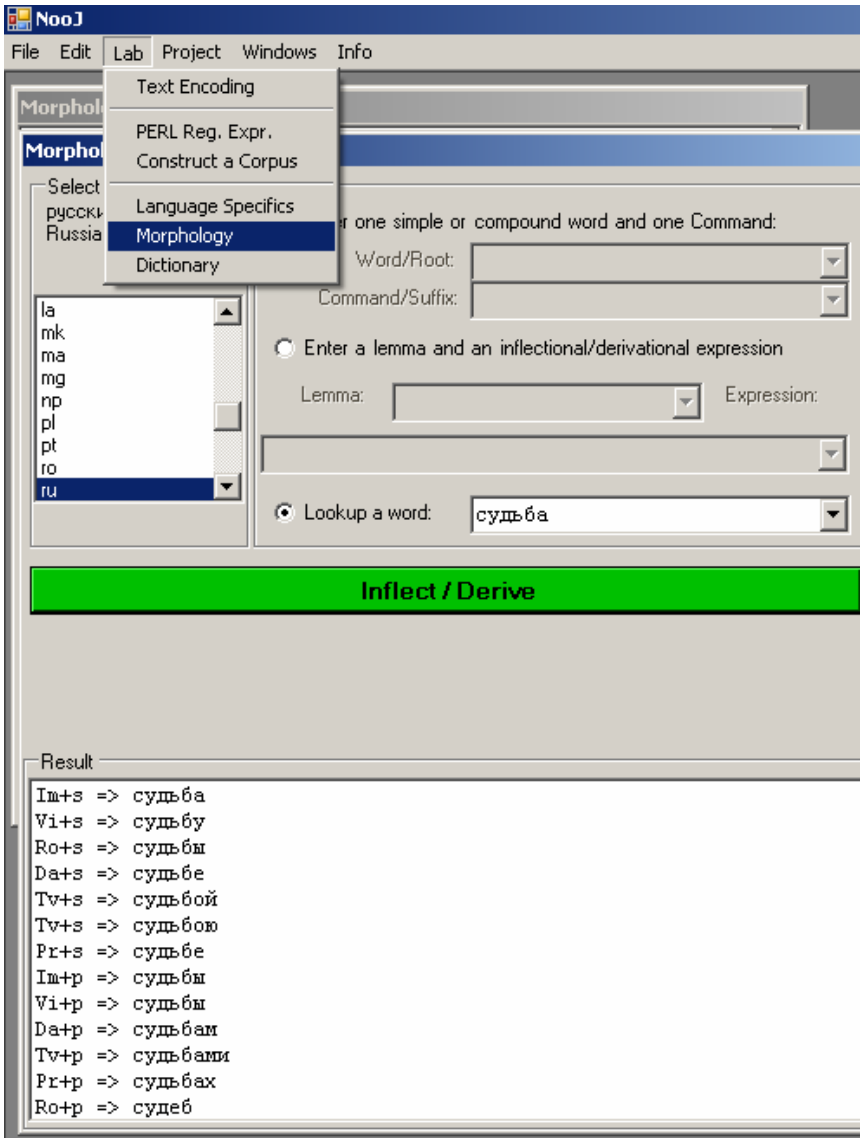
noms
аба,N+f+inan+FLX=j1b
абазинка,N+f+an+FLX=jo3oa
абака,N+f+inan+FLX=j3b
абака,N+f+inan+FLX=j3a
аббатиса,N+f+an+FLX=jo1a
аббревиатура,N+f+inan+FLX=j1a
аббревиация,N+f+inan+FLX=j7a
абдикация,N+f+inan+FLX=j7a
абдукция,N+f+inan+FLX=j7a

абажур,N+m+inan+FLX=завод
абазинец,N+m+an+FLX=украинец
абазин,N+m+an+FLX=артист
абаз,N+m+inan+FLX=завод
абак,N+m+inan+FLX=чайник
аббат,N+m+an+FLX=артист
абдомен,N+m+inan+FLX=завод
абдуктор,N+m+inan+FLX=завод
абелит,N+m+inan+FLX=завод
абзац,N+m+inan+FLX=месяц
абиетин,N+m+inan+FLX=завод
абиссинец,N+m+an+FLX=украинец
абитуриент,N+m+an+FLX=артист
verbes
предстательствовать,V+ipf+intr+FLX=интересовать
предчувствоваться,V+sja+ipf+FLX=интересоваться
предчувствовать,V+ipf+FLX=интересовать

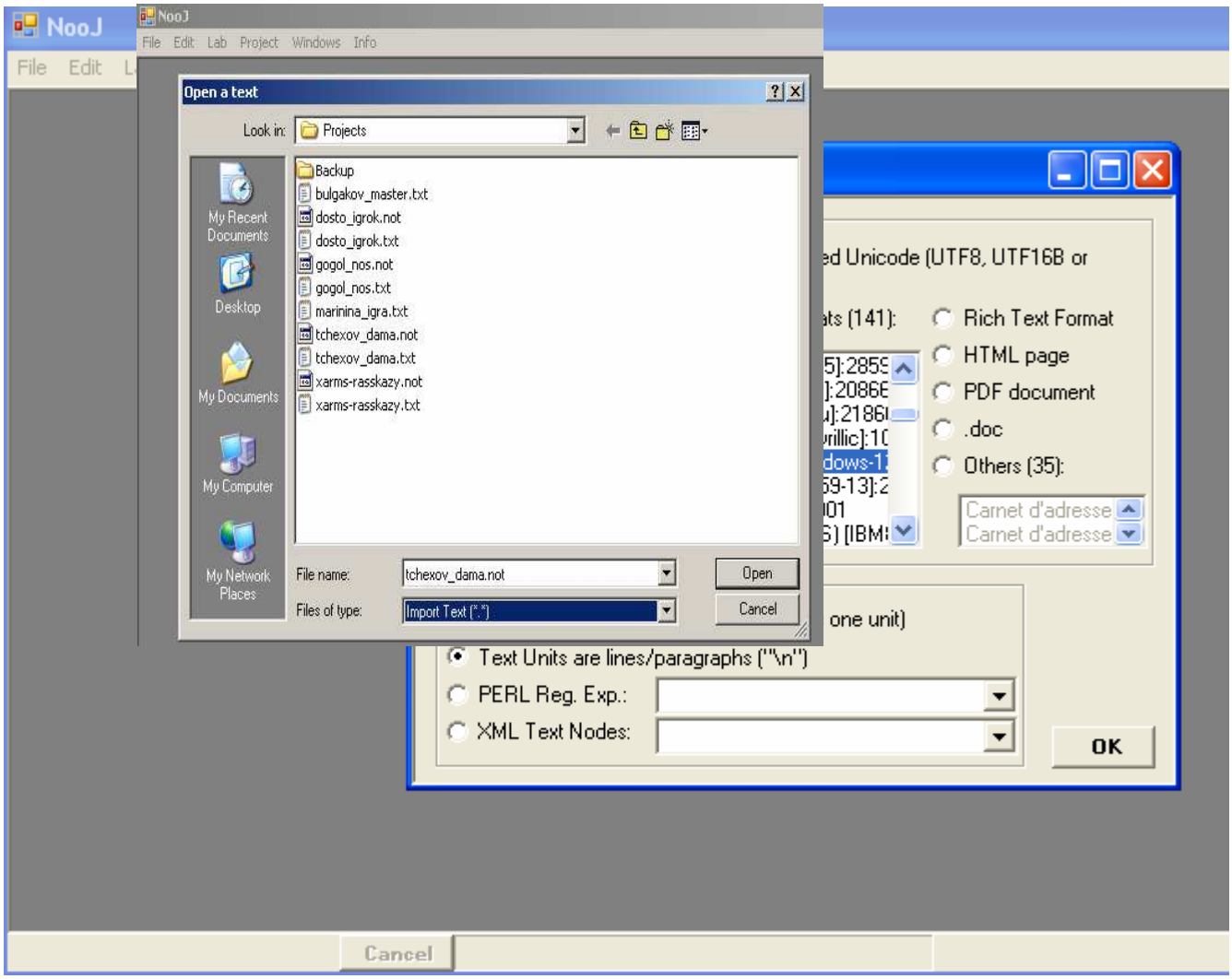
предшествовать,V+ipf+intr+FLX=интересовать
презентовать,V+ipf+pf+FLX=интересовать
президентствовать,V+ipf+intr+FLX=интересовать
преимуществовать,V+ipf+intr+FLX=интересовать
прелюбодействовать,V+ipf+intr+FLX=интересовать
премировать,V+ipf+pf+FLX=интересовать
застигнуть,V+pf+FLX=свергнуть
застынуть,V+pf+intr+FLX=привыкнуть
засунуть,V+pf+FLX=двинуть
затерпнуть,V+pf+intr+FLX=привыкнуть
затиснуться,V+sja+pf+FLX=тукнуться
затиснуть,V+pf+FLX=брызнуть
затихнуть,V+pf+intr+FLX=привыкнуть
заткнуться,V+sja+pf+FLX=вернуться
заткнуть,V+pf+FLX=толкнуть
затолкнуть,V+pf+FLX=толкнуть
затонуть,V+pf+intr+FLX=потянуть
затронуть,V+pf+FLX=двинуть
затухнуть,V+pf+intr+FLX=привыкнуть
затянуться,V+sja+pf+FLX=натянуться

COMPILATIONS / VERIFICATIONS / CORRECTIONS
ajout et suppression d'entrées ou de paradigmes

Vérification des paradigmes :



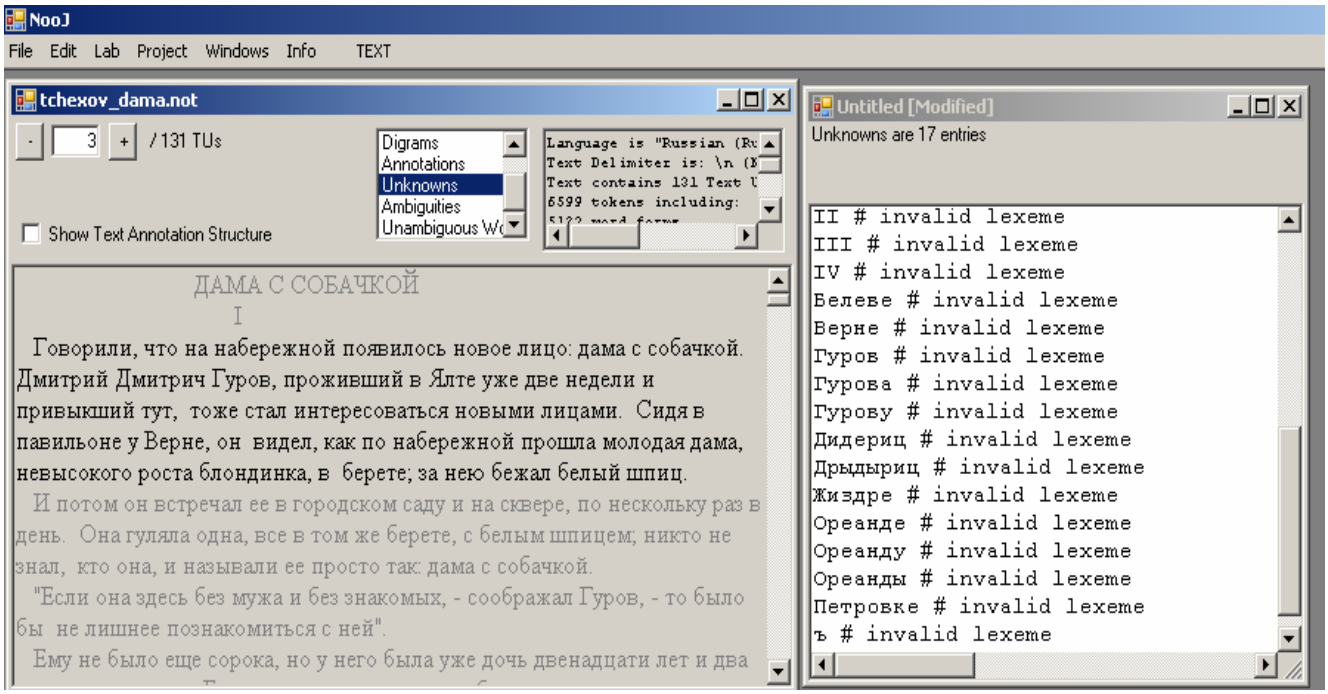
UTILISATION DES RESSOURCES LEXICALES AVEC UN TEXTE
Paramétrage du codage :
File / Open text / import text



UTILISATION DES RESSOURCES LEXICALES AVEC UN TEXTE

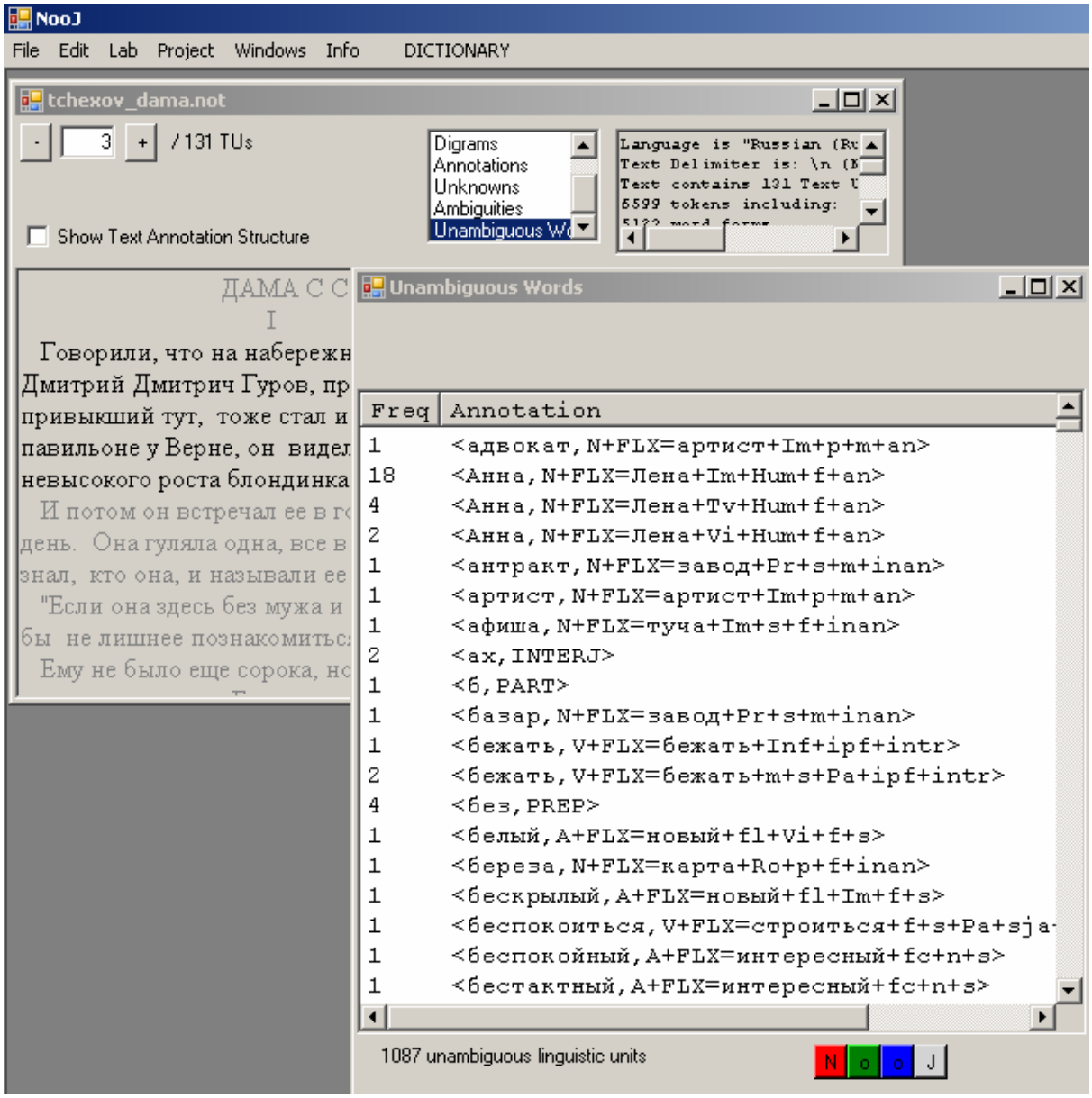
Choix du texte :

Travail sur le texte :
1) repérage des mots inconnus :



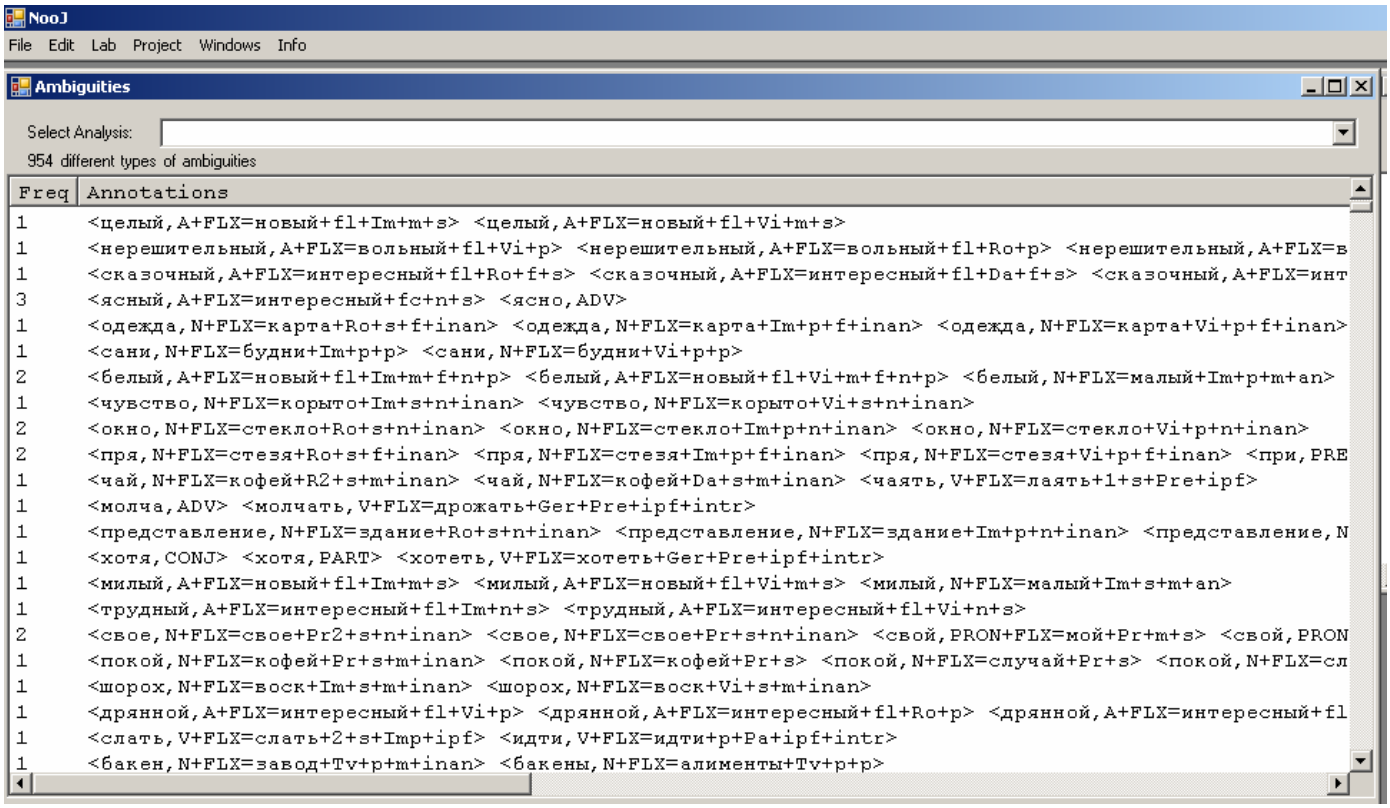
Travail sur le texte :

2) repérage des mots reconnus (nonambigus)



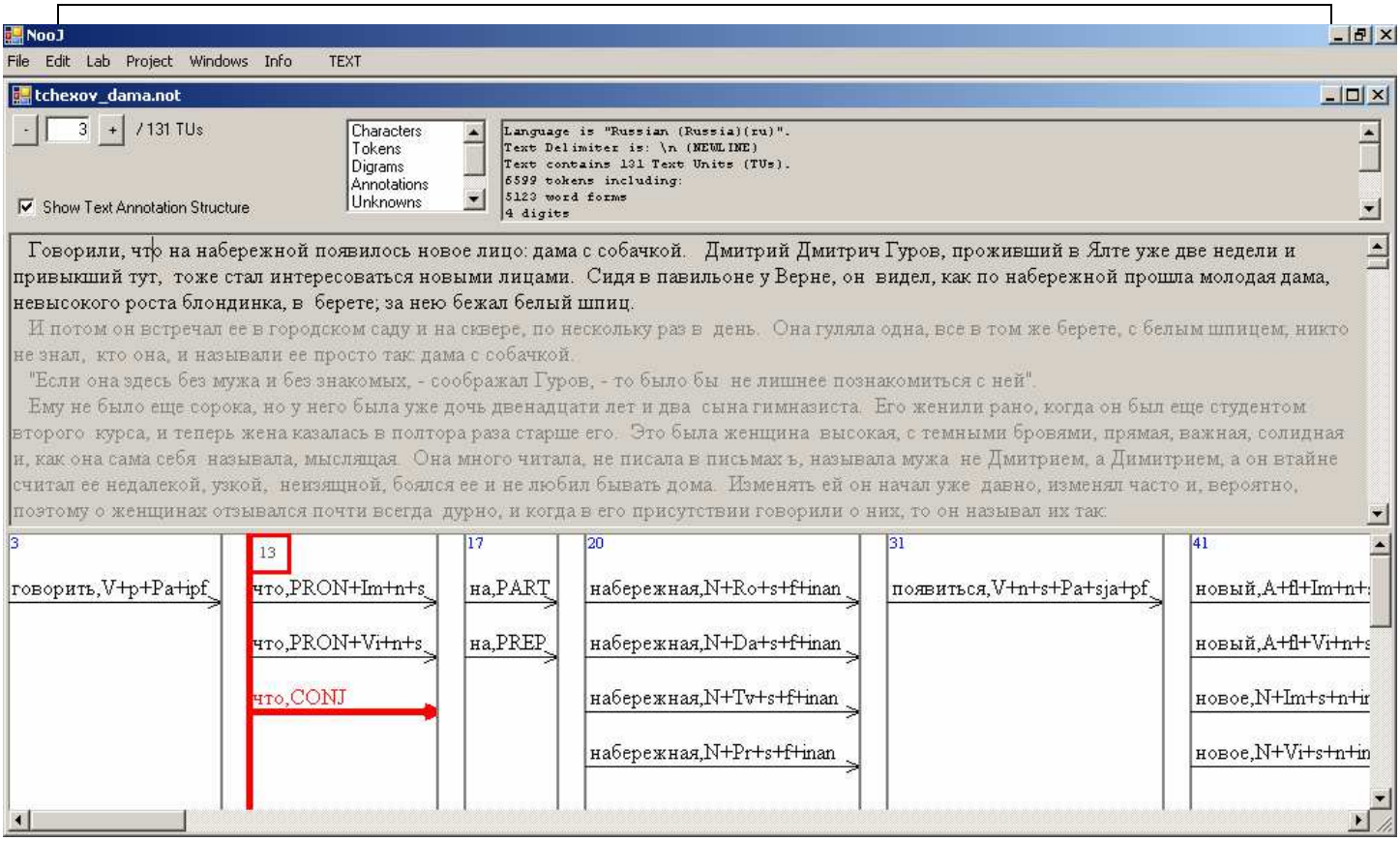
Travail sur le texte :

3) repérage des mots ambigus --> amélioration de la reconnaissance / suppression du dictionnaire



Travail sur le texte :

4) désambiguïsation



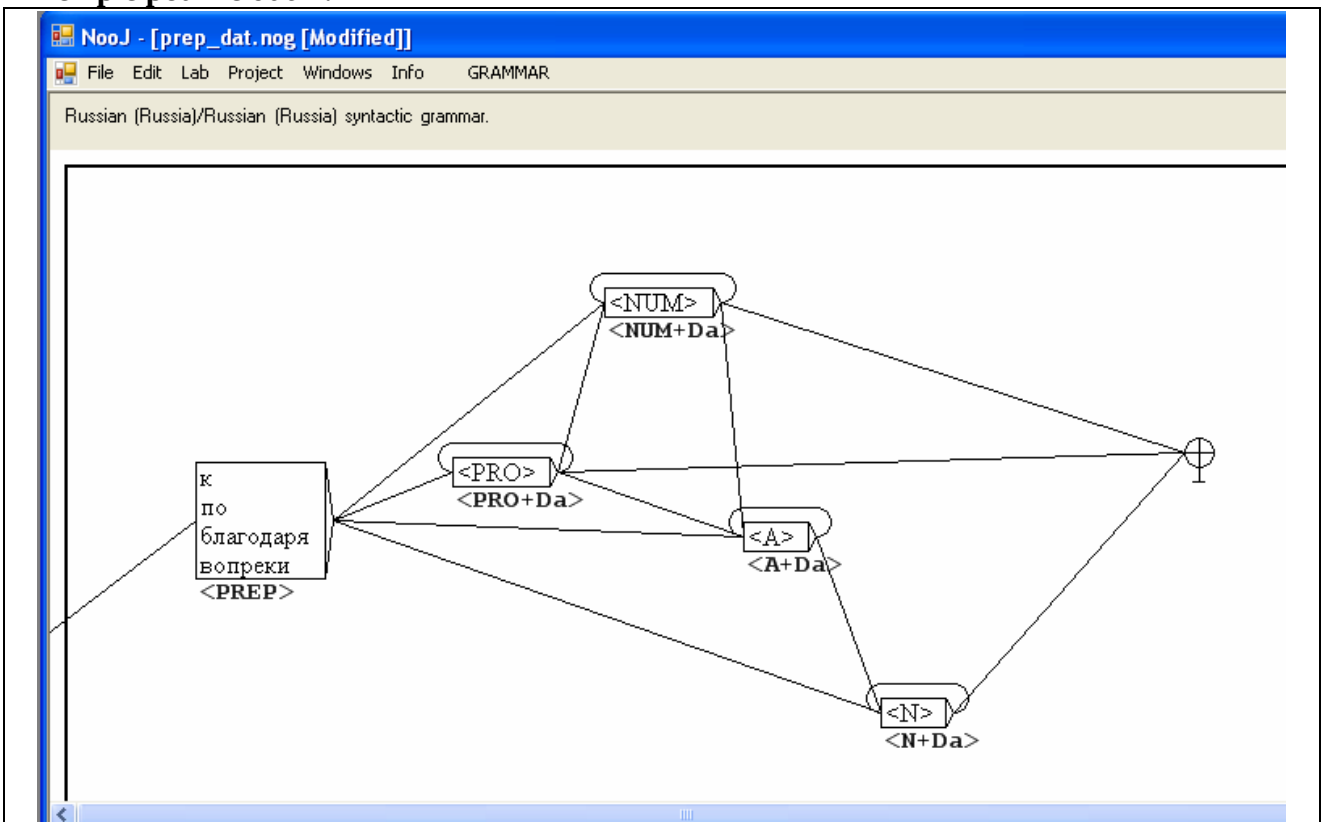
LES GRAMMAIRES NOOJ

- **désambiguisation automatique (contraintes et limites)**

prépositions
construction verbales
pronoms conjonction

Ecriture de grammaires syntaxiques de désambiguisation (fichier nog : nooj **g**rammar**)**

Exemple pour le datif :



TRAVAIL ULTERIEUR

- **simplification et enrichissement du dictionnaire (critères sémantiques à ajouter + éventuellement traductions)**
- **vérification des types associés et des paradigmes**
- **écriture de grammaires de base**

Constituer un corpus de textes avec un minimum de mots ambigus ou non reconnus.

Nooj : outil complémentaire aux corpus existants (ruscorpora, cfrl, narusco)

→ choix du texte / des textes
→ prépration du texte
→ application des sources lexicales
→ vérification correctin enrichissement des dictionnaires et des paradigmes

→ constitution de lexique
→ recherche d'occurrences
→ établissement des concordances
→ mise en forme des résultats