

Ressources lexicales et grammaticales pour le russe

**Conception, réalisation et
utilisation
des ressources lexicales et
grammaticales
pour le russe**



**avec le logiciel Nooj
de Max Silberztein**

Séminaire
24/10/12

Vincent BÉNET
CREE
CERRUS - RAO
Recherche assistée par ordinateur

1

Historique des ressources existantes en russe

LOGICIELS DE TRAITEMENT DE CORPUS

pour la langue russe
1980-2000

UNILEX (sous DOS)

MAK (sous Windows)

avec un corpus d'environ 4000 pages
de textes russes lemmatisés

2

Historique des ressources existantes en russe

SOURCES TEXTUELLES POUR CORPUS RUSSES

La bibliothèque de Moshkov

<http://lib.ru>

Le fonds informatisé de la langue russe

<http://cfri.ru>

Le corpus national de la langue russe

<http://www.ruscorpora.ru>

Le corpus national de la langue littéraire* russe

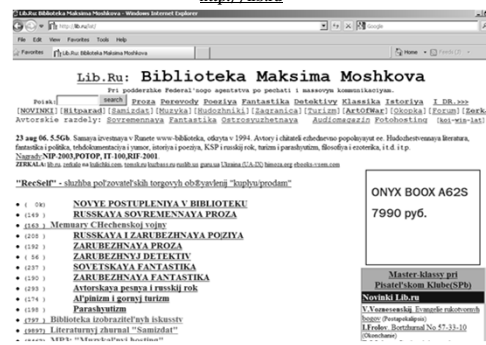
<http://www.narusco.ru>

3

Corpus de textes russes

La Bibliothèque de Moshkov

<http://lib.ru>

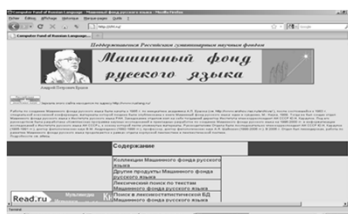


4

Corpus de textes russes

Le fonds informatisé de la langue russe

<http://cfri.ru>



- 1600 œuvres littéraires (romans, poésies, théâtre du XIX et XX, environ 100 auteurs)
- Corpus de textes de journaux (années 1991-2000)
- Dictionnaires
- avec logiciel intégré de recherche d'occurrences

5

Ressources pour Nooj pour la langue russe

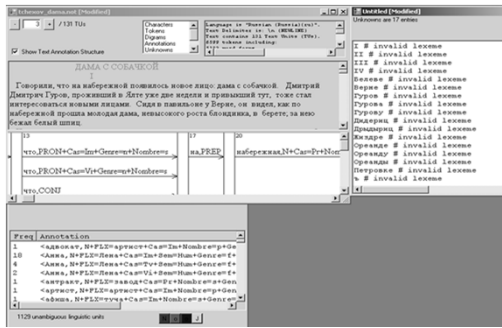
- **NOOJ** <http://www.nooj4nlp.net/> M. Silbersztein



6

Lemmatiseur du russe pour Nooj

travail avec des textes : « la dame au petit chien » de Tchekhov



7

Ecrire des ressources pour le russe pour Nooj

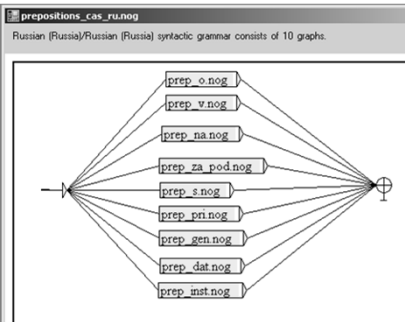
concevoir des grammaires de désambiguisation

- régime des prépositions
- régime des verbes courants
- accords adjectifs / substantifs
- rection des numéraux
- expression des dates / heures etc.
- reconnaissance des toponymes, des prénoms, des diminutifs etc.

8

Ecrire des ressources pour le russe pour Nooj

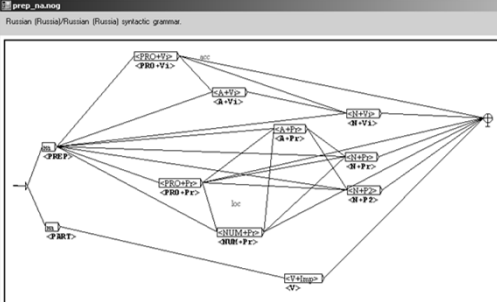
concevoir des grammaires de désambiguisation



9

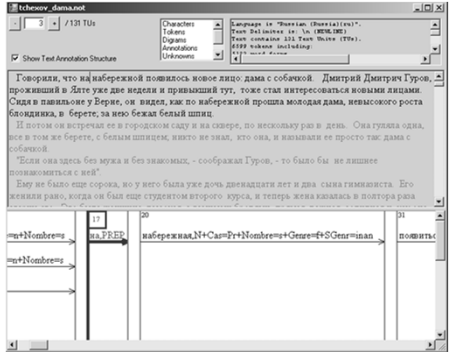
Ecrire des ressources pour le russe pour Nooj

concevoir des grammaires de désambiguisation



10

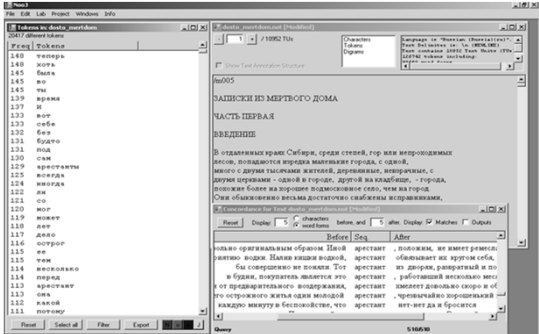
Constituer un corpus de textes vérifiés et désambiguïsés



11

Constituer un corpus de textes vérifiés

■ NOOJ



12