

Extension des ressources polonaises pour NooJ

Krzysztof Bogacki
Institut de Philologie Romane
Université de Varsovie – Pologne
kbogacki@gmail.com

Plan de la présentation

- 1. Les textes de démonstration**
- 2. Trois spécificités du polonais**
 - existence des cas → conséquences pour les prépositions et les numéraux
 - l'aspect et le temps → conséquences pour les formes en présence
 - la catégorie du genre : trois masculins, féminin, neutre
- 3. Les codes utilisés**
- 4. Les graphes**
 - la syntaxe des prépositions
 - la syntaxe du GN avec les numéraux de 1 à 99.
- 5. Extension des dictionnaires**
- 6. Grammaire morphologique et déclinaison des substantifs de type RUCH**
- 7. Une grammaire syntaxique du passif**

Le module polonais est le dernier venu dans la communauté NooJ

Opérationnel et accessible sur le web à partir du 30 mai 2007, il comporte

- deux textes de démonstration avec les dictionnaires complets (fichiers .nod, .dic et .flx) qui permettent de reconnaître toutes les formes que ces textes contiennent
- deux séries de graphes qui illustrent deux phénomènes linguistiques que l'on retrouve dans les textes de démonstration

Premier texte de démonstration: Bruno Schulz – *Ulica Krokodyli*

- Court récit datant de 1933, tiré d'un recueil de textes de Bruno Schulz, écrivain et peintre, né dans une petite ville en Galicie, l'actuelle Ukraine dans une famille juive polonaise assimilée
- Il comporte 18601 caractères (68 différents), dont 15570 lettres (57 différentes), 2567 espaces, 2561 word forms
- Les fichiers correspondants sont: Krokodyl.not, Dico_pl.dic, Fleks.flx

Deuxième texte de démonstration:

Bolesław Prus - *Faraon*

- Le texte proposé correspond au premier chapitre d'un roman de Bolesław Prus publié en 1897. Son auteur est un écrivain très connu, candidat au Prix Nobel pour deux de ses romans : *Lalka* ('Poupée') et *Faraon* ('Pharaon'). Il a été sélectionné à cause d'une accumulation de noms propres et de numéraux peu représentés dans le texte de Bruno Schulz
- Le texte contient 10002 caractères (dont 65 différents), 8286 lettres (dont 57 différentes), 1449 espaces, 1671 tokens dont 1406 word forms
- Les fichiers correspondants sont: Dico_Faraon.nod (qui ajoute les mots ne figurant pas dans Dico_pl.nod et qui sont nécessaires à l'analyse de *Faraon*), Dico_NPr.nod (qui regroupe les noms propres), Pro_Num.nod (qui contient les pronoms et les numéraux). La flexion est décrite dans Dico_Faraon.flx, Dico_NPr.flx, Pro_Num.flx

Trois spécificités du polonais.

(1) Les cas

- Les cas – 7 pour chacun des deux nombres: Nominatif (Mi), Génitif (Do), Datif (Ce), Accusatif (Bi), Instrumental (In), Locatif (Lo), Vocatif (Wo)
- Ils ont diverses conséquences au niveau syntaxique notamment en ce qui concerne la syntaxe des prépositions et celle des syntagmes nominaux contenant les numéraux

Trois spécificités du polonais.

(2) L'aspect

- Il existe en polonais deux aspects : perfectif et imperfectif
- L'existence des aspects a une influence, entre autres, sur la sélection des catégories temporelles et leur valeurs
 - Les formes morphologiques: présent sémantique avec les imperfectifs (*robię* 'je fais') vs futur avec les perfectifs (*zrobię* 'je ferai')
 - La sélection des formes morphologiques compatibles avec les verbes perfectifs mais non avec les imperfectifs et inversement

Trois spécificités du polonais.

(3) Le genre

La catégorie du genre comporte

– trois masculins:

- masculin-animé-non-humain (*pies* ‘chien’)
- masculin-non-animé-concret (*stół* ‘table’)
- masculin-animé humain (*chłopiec* ‘garçon’)

– féminin

– neutre

Les codes utilisés: les parties du discours

- A - adjectif
- ADV - adverbe
- CONJC - conjonction de coordination
- CONJS - conjonction de subordination
- X - invariable
- N - nom
- NUM - numéral
- PREP - préposition
- PRON – pronom
- PTCL - particule
- V – Verbe
- On ajoutera, dans le dictionnaire complet, un code pour les interjections, absentes dans les deux textes de démonstration (INTERJ)

Les codes utilisés: la morphologie nominale

- f - féminin **(5 genres)**
- mo – masculin-humain
- mr – masculin-non-animé-concret
- mz – masculin-animé-non-humain
- n – neutre

- Bi – accusatif **(7 cas)**
- Ce - datif
- Do - génitif
- In - instrumental
- Lo - locatif
- Mi - nominatif
- Wo – vocatif

- p – pluriel **(2 nombres)**
- s – singulier

Les codes utilisés: la morphologie verbale

- F - infinitif
 - G – impératif
 - H – présent de l’indicatif
 - J – participe présent actif adjectival
 - K – participe présent actif adverbial
 - P – passé de l’indicatif
 - Q - participe passé adjectival
 - S - forme impersonnelle du verbe
 - T – participe passé adverbial
 - U – futur morphologique du verbe *być* ‘être’
 - Z – conditionnel présent
-
- 1 – première personne
 - 2 – deuxième personne
 - 3 – troisième personne

Les codes utilisés: la sémantique

- Abstr, Alim, AnimLoc, Atm, CollHum, CollImmeub, Conc, ConcColl, Cpmc, Geom, Hum, Immeub, Liq, Mach, Mat, Org, Pdc, Pr, Qual, Quant, Sent, Text, Tmp, Vehicl
- DK, NDK (l'un des deux est obligatoire avec les verbes)
- Cette classe de codes risque de s'allonger au fur et à mesure que le dictionnaire sera enrichi de nouvelles entrées

Les graphes

Les ressources du polonais contiennent:

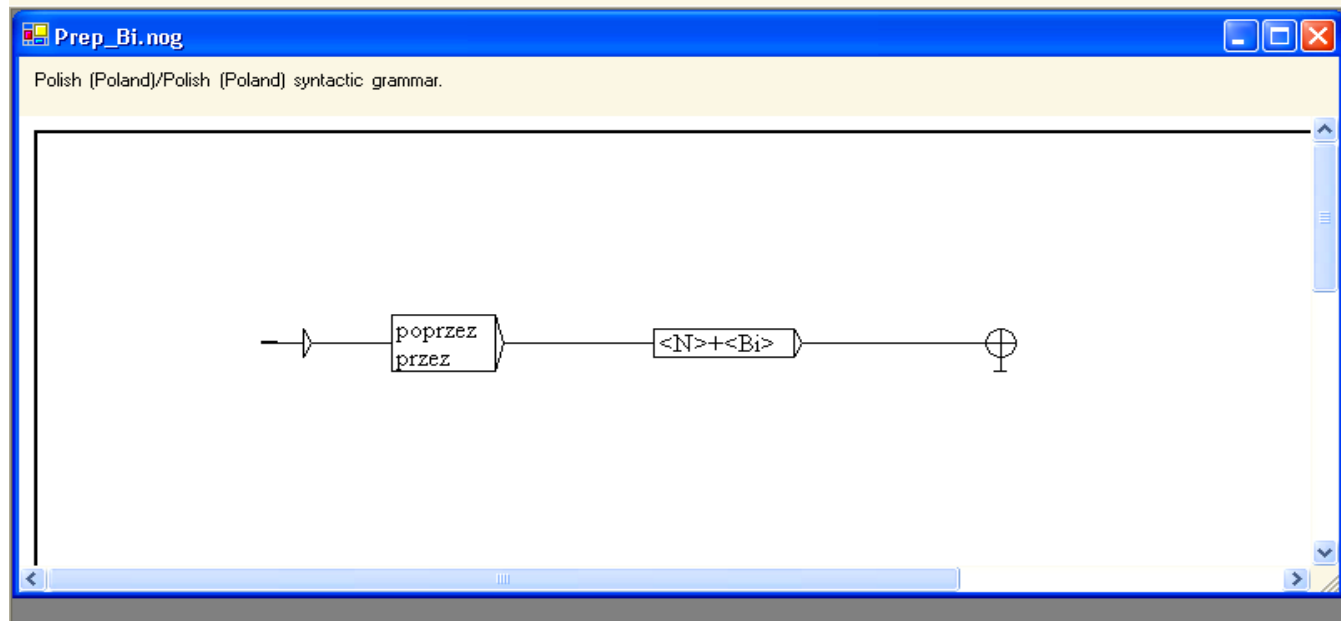
- Une série de graphes qui décrivent certains aspects du fonctionnement syntaxique des prépositions
- Une grammaire syntaxique des numéraux de 1 à 99.

Les graphes des prépositions

- Chaque préposition impose un cas particulier à son régime substantival
- Parfois, une seule et même préposition se combine avec deux, trois ou même quatre cas
- Le plus souvent le choix se fait en fonction du sémantisme du verbe – centre de la phrase
- Les prépositions simples font partie des locutions prépositionnelles (environ 1200) ce qui devrait donner lieu à un traitement approprié dans le cadre offert par NooJ.

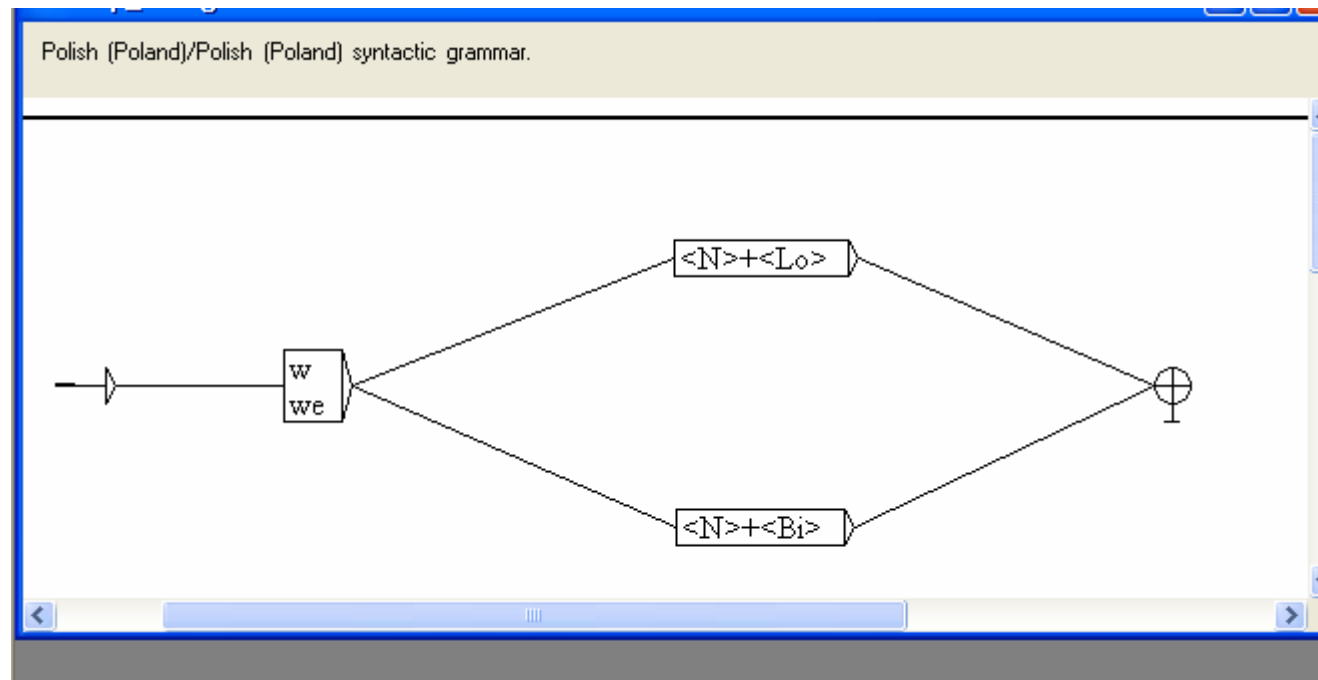
Préposition à régime unique

- Le régime de *(po)przez* se met à l'accusatif



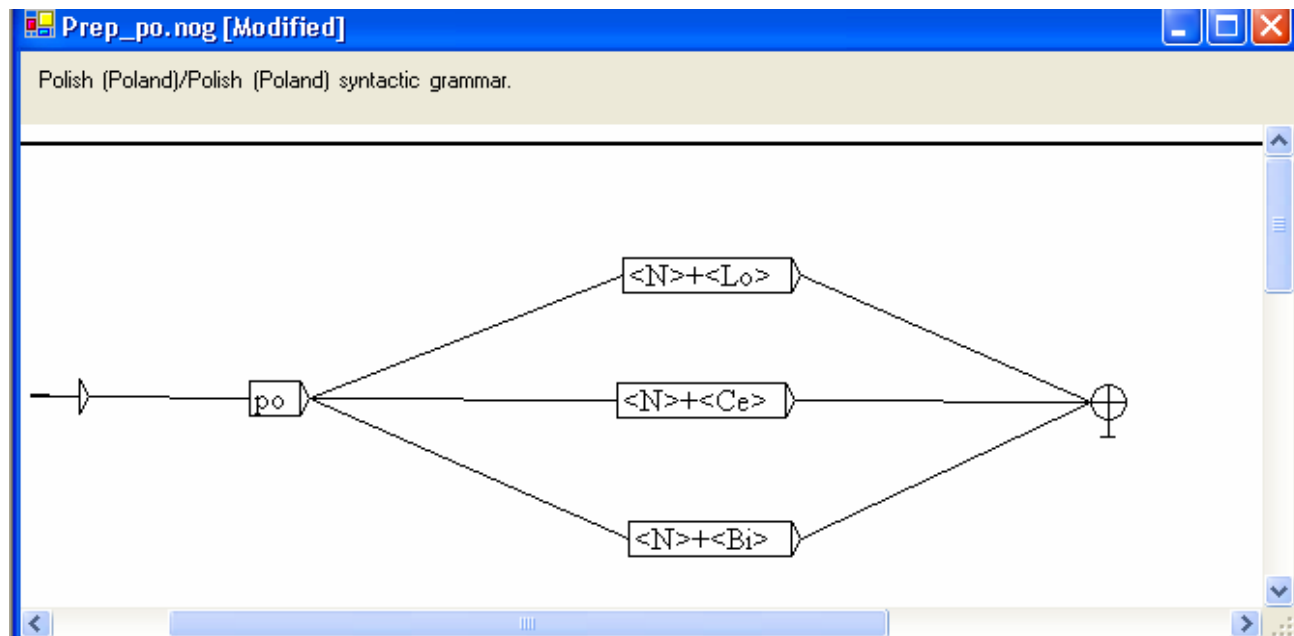
Préposition à deux régimes

- La préposition $w(e)$ impose soit l'accusatif soit le locatif



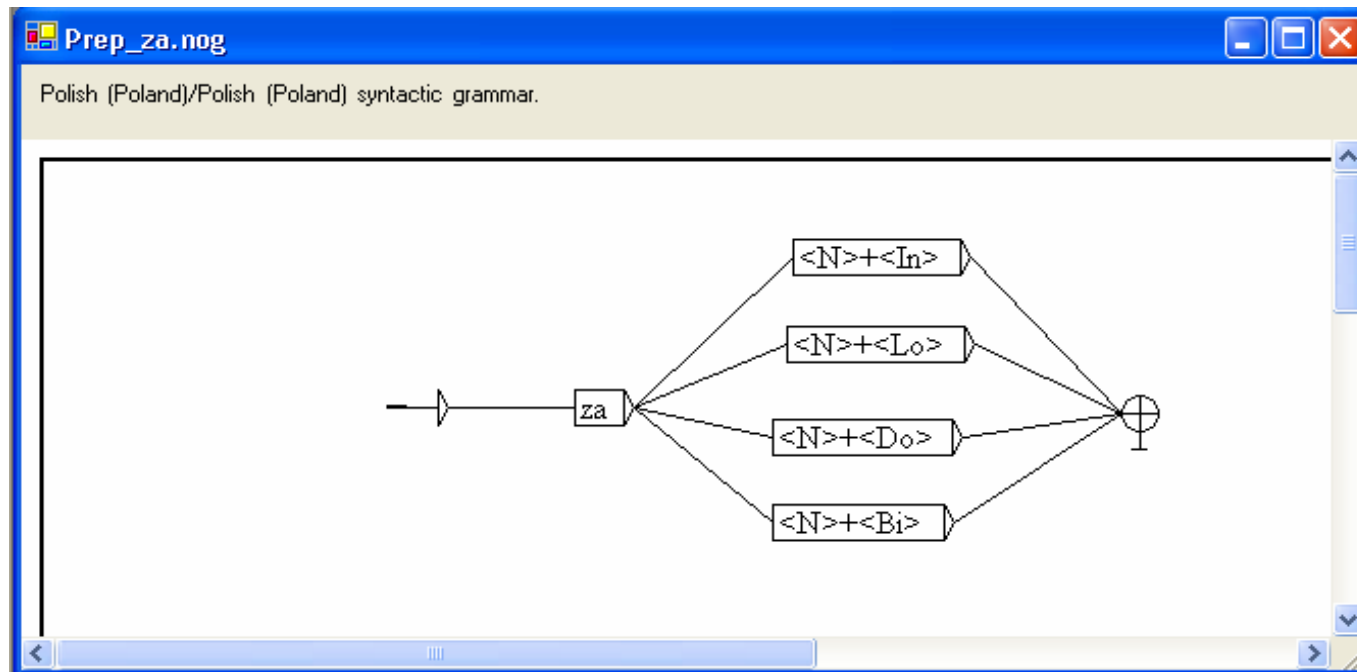
Préposition à trois régimes

- La préposition *po* impose soit l'accusatif soit le locatif soit le datif



Préposition à quatre régimes

- La préposition *za* impose soit l'accusatif soit le locatif soit le datif soit l'instrumental

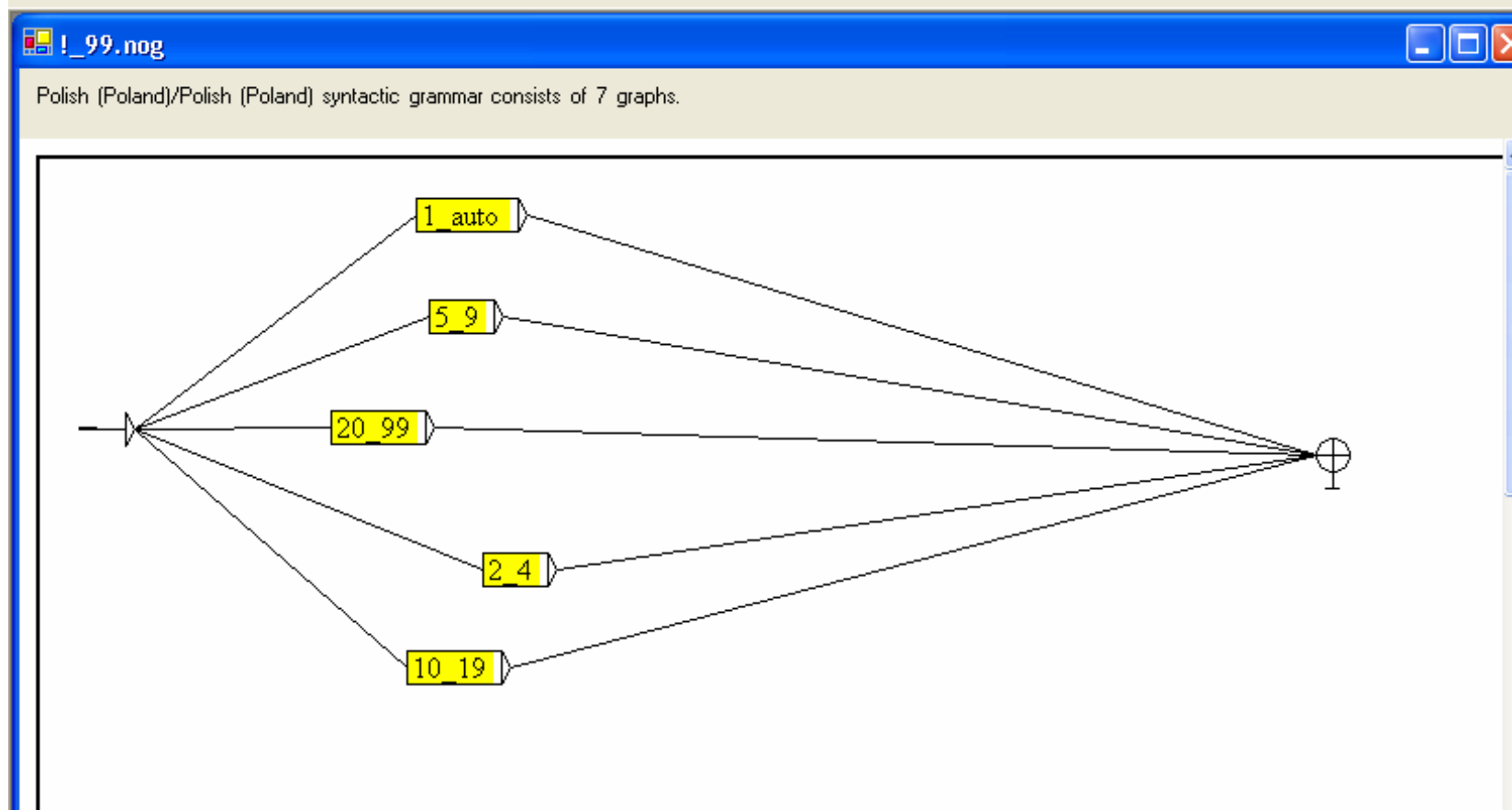


Les numéraux

Les numéraux (1 – 99) demandent 7 graphes qui décrivent:

- La syntaxe de <jeden> (1) en position autonome et en fin d'un numéral complexe
- Celle de <dwa>, <trzy>, <cztery> (2 – 4)
- Celle de <pięć>...<dziewięć> (5 – 9)
- Celle de <dziesięć>...<dziewiętnaście> (10 – 19)
- Celle de 20 à 99 (avec des graphes imbriqués)
- La structure globale des numéraux

Les numéraux: structure globale

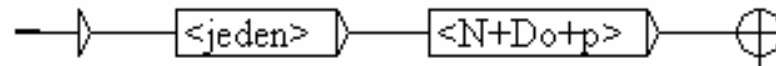


Les numéraux: <jeden>

- <jeden> autonome

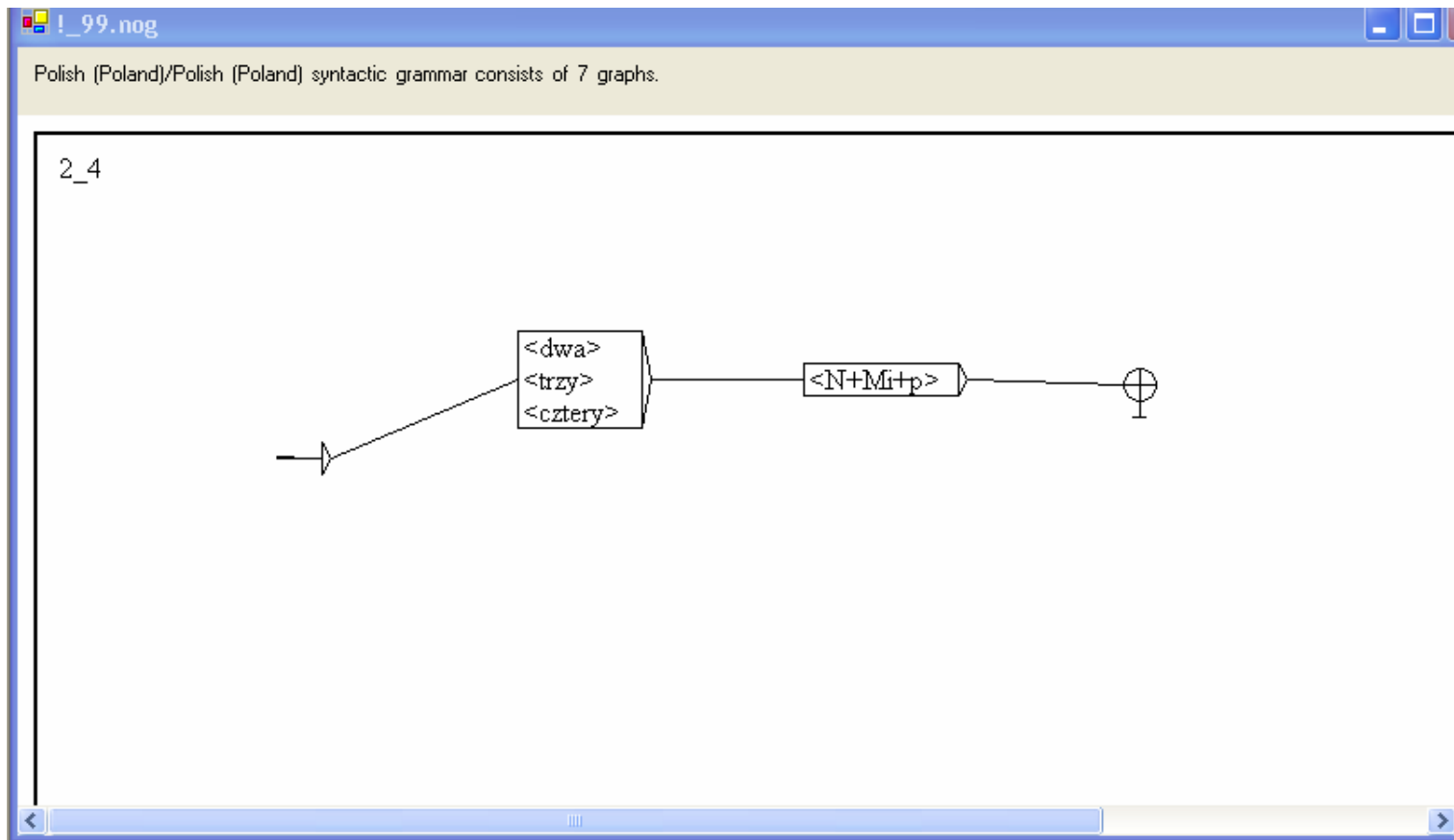


- en fin d'un numéral complexe



Les numéraux: 2 ... 4

Les numéraux <dwa>, <trzy>, <cztery>

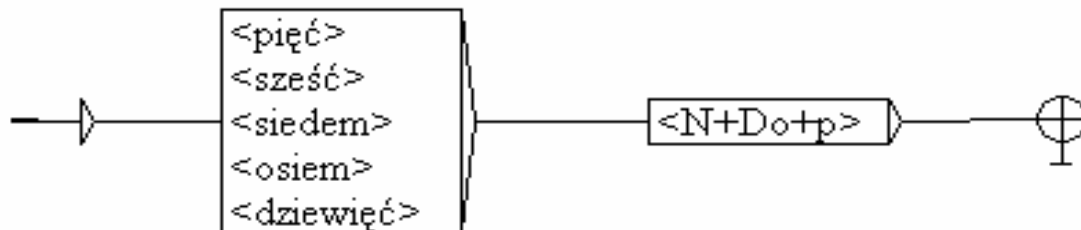


Les numéraux: 5 ... 9

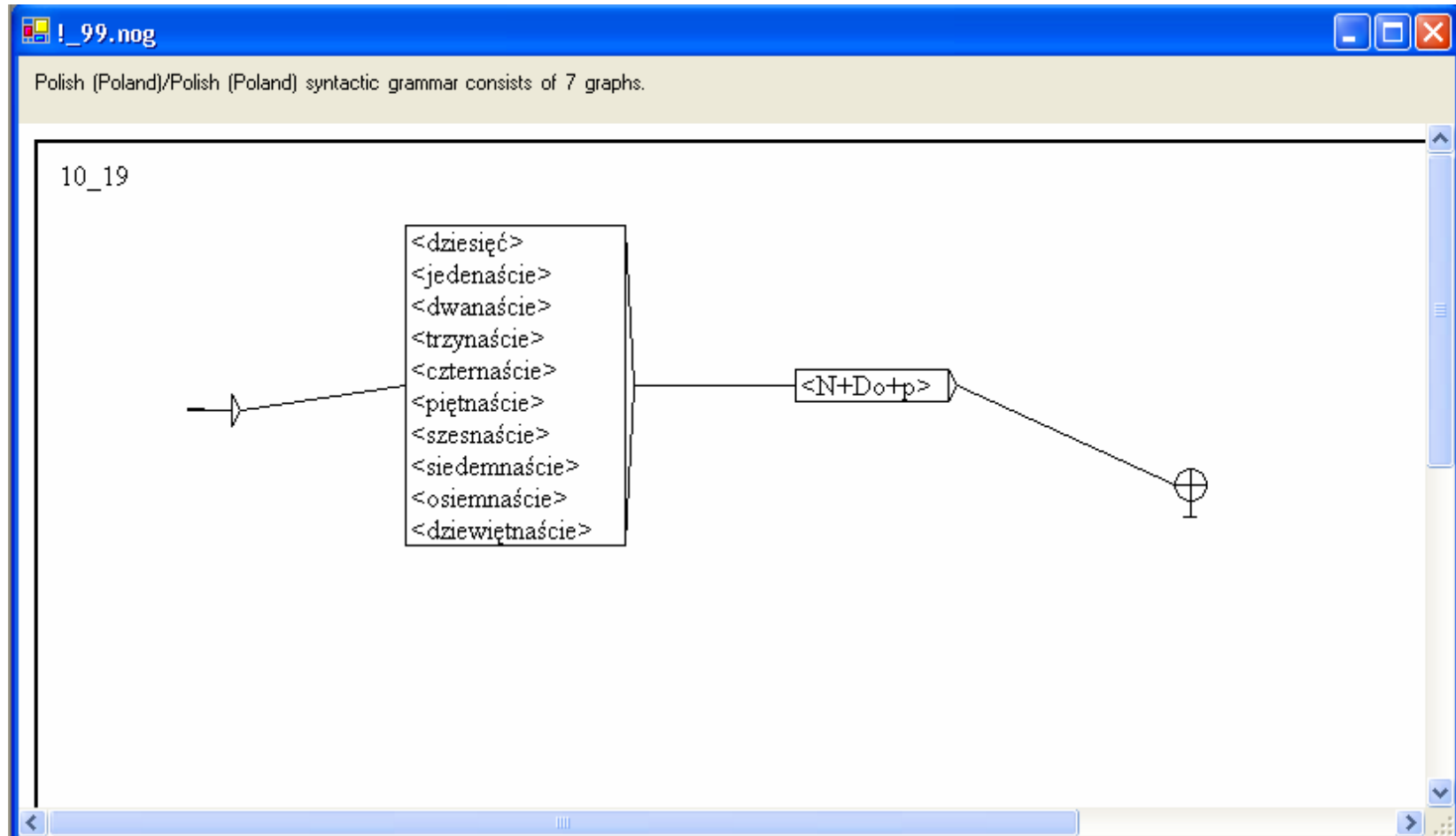
Les numéraux: <pięć>, <sześć>, <siedem>, <osiem>, <dziewięć>

Polish (Poland)/Polish (Poland) syntactic grammar consists of 7 graphs.

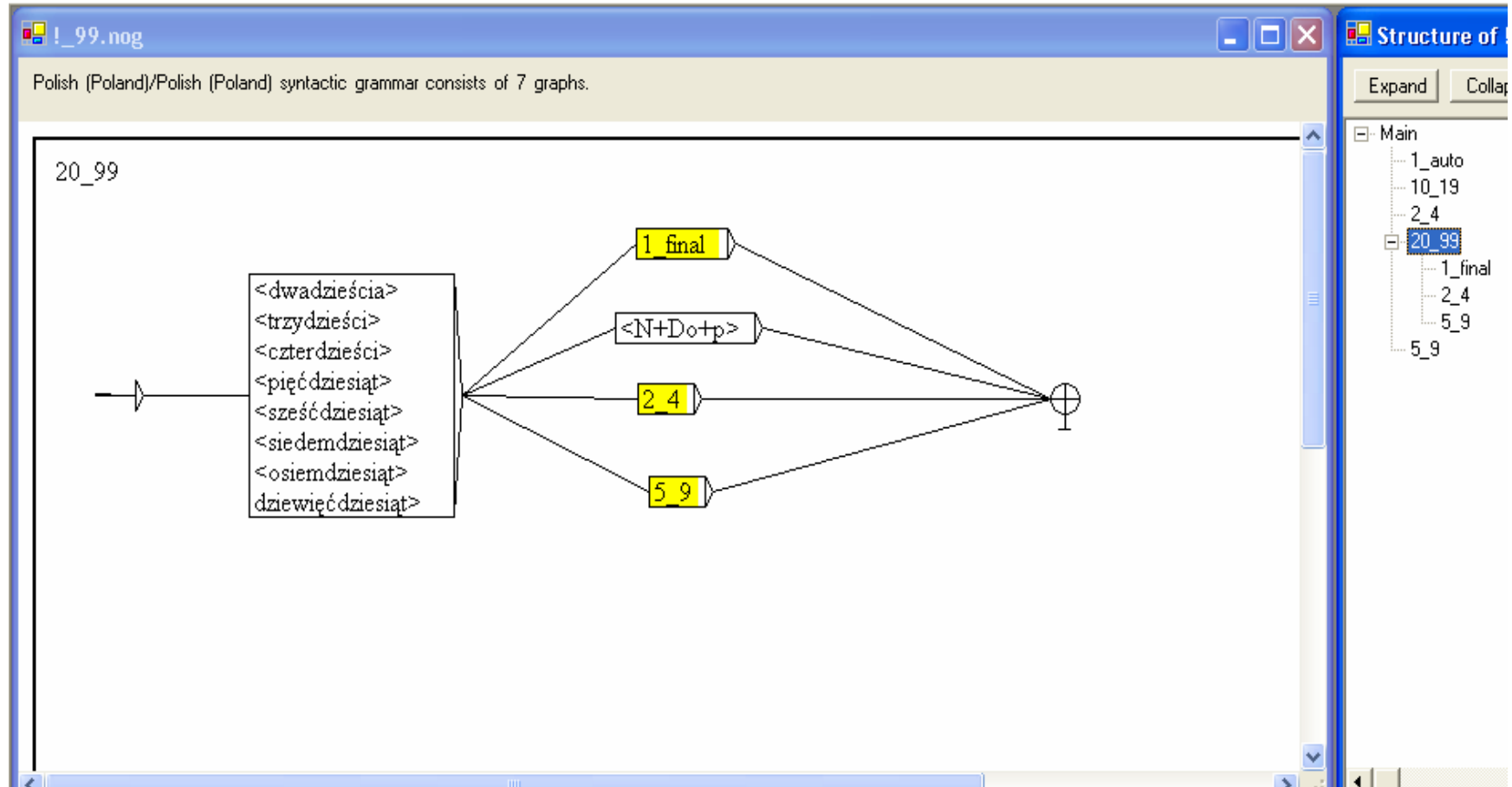
5_9



Les numéraux: 10 ... 19



Les numéraux: 20...99



A faire – juin 2007

- Elargissement du dictionnaire
- Constitution d'une bibliothèque de grammaires locales couvrant les phénomènes les plus répandus

Fait dans le courant de l'année scolaire 2007/2008

- * **Une série de dictionnaires:**

- **noms propres (Dico_NPr),**
- **appellatifs simples (Dico_simple)**
- **composés (Dico_cmp)**

- * **Illustration des phénomènes dérivationnels**

- * **Construction de grammaires morphologiques
pour décliner les substantifs de type *RUCH***

- * **Construction de grammaire syntaxiques pour
gérer le passif**

Les dictionnaires (1)

Ils exploitent un fond existant à l'Université de Varsovie sous forme d'un dictionnaire morphologique (**POLLEX**) comptant presque 140.000 formes canoniques.

Les ressources converties au format NooJ contiennent 75140 formes simples dont

- - Adverbes – 1126
- - Adjectifs – 18301
- - Substantifs – 55081
- - Prépositions - 104
- - Verbes – 280
- - Conjonctions – 65
- - Interjections – 129
- - Pronoms – 26
- - X – 11, numéraux, particules.....

Dictionnaires (2)

Les formes X :

- mots étrangers trouvés dans les textes de démonstration :

- *King of England*
- *definitivum*
- *papier mâché*
- *cliché*

mais aussi des mots polonais difficiles à classer:

- *bogdajby*

Dictionnaires (3)

- Dico_Npr : 1676 items
- Nom propres déclinables : 30
- Noms propres indéclinables: 1646

Types de flexion

- Substantifs – 75 (+ de 400 dans Pollex)
- Adjectifs -16 (29)
- Verbes – 24 (142)

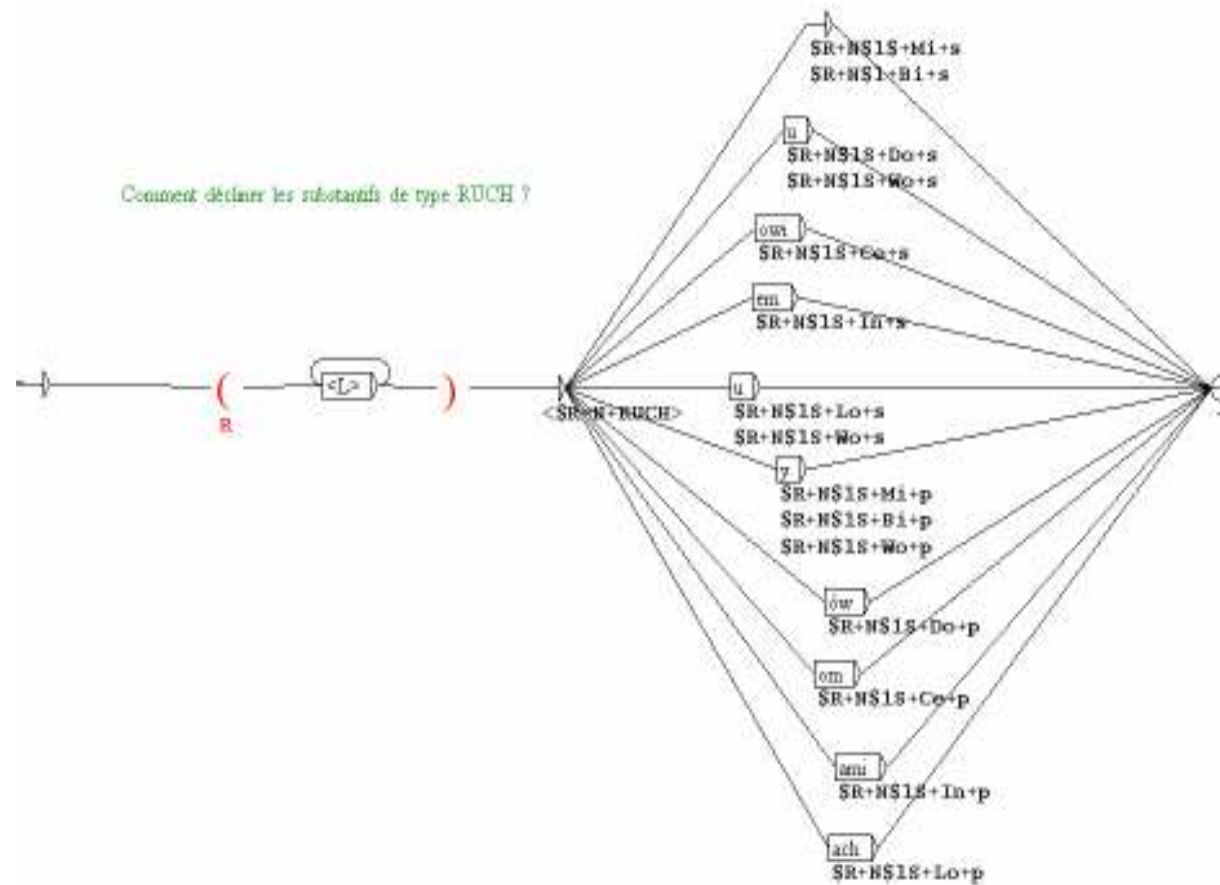
Dans les noms propres: 9

- Structures de composés : 45

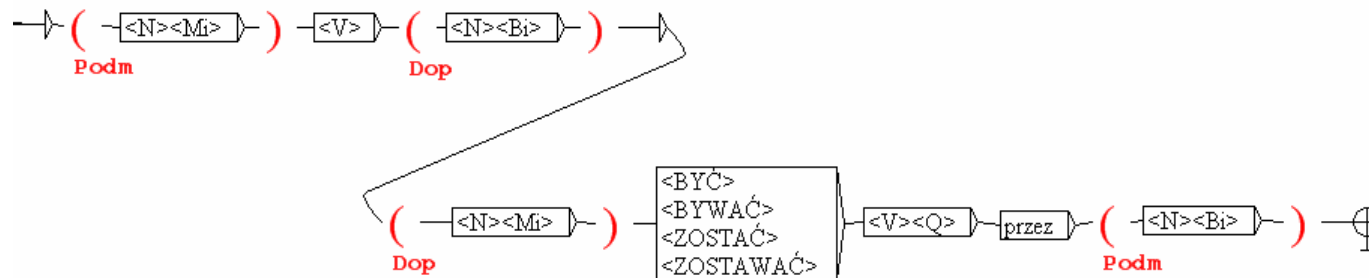
Nombre de formes par modèle flexionnel

- Substantif – max. 12, moyenne: moins de 9
- Verbe – 169 – entre 90 et 100
- Adjectif – 72 – moins de 30

Grammaire morphologique et déclinaison des substantifs de type RUCH



Les grammaires syntaxiques et les transformations: le passif



A faire – entre juin 2008 et juin 2009

- Extension du dictionnaire des mots simples communs
 - verbes
 - substantifs

(total a atteindre : environ 90.000 mots)

- Extension du dictionnaire des composés (→3000)
- Extension du dictionnaire des noms propres (→5000)
- Grammaires syntaxiques transformationnelles (→une dizaine au total)

MERCI DE VOTRE ATTENTION !

Nombre de lexèmes dans tous les dictionnaires nombre de schémas flexionnels

L'archive **1_99.zip** contient 6 graphes:

auto.nog, 2_4.nog, 5_9.nog, 10_19.nog, 20_99.nog

et le graphe-mère **1_99.nog** qui permettent de reconnaître les numéraux de 1 à 99 introduisant leurs régimes nominaux employés au cas approprié :

1 employé tout seul → N au sg. (le cas dépend de la position syntaxique dans la phrase)

1 terminant un numéral complexe (21, 31, 41.....) → N + Do + p

2, 3 et 4 tout seuls ou en fin de numéral complexe (2, 22, 3, 23, 4, 24 etc.) → N + Mi + p

5, 6, 7, 8, 9 tout seuls ou en fin de numéral complexe (5, 25, 6, 26, etc.) → N + Do + p

10.....19 → N + Do + p

20, 30, 40,90 → N + Do + p

L'archive **Prepositions.zip** contient 9 graphes indépendants:

Prep_Bi.nog reconnaît 2 prépositions qui exigent l'accusatif

Prep_Do.nog reconnaît 21 prépositions qui exigent le génitif

Prép_In_Bi.nog reconnaît 11 prépositions qui exigent l'instrumental ou l'accusatif

Prep_ku décrit la syntaxe de la préposition **ku** qui exige le datif

Prep_mimo décrit la syntaxe de la préposition **mimo** qui exige le génitif ou l'accusatif

Prep_niedaleko décrit la syntaxe de la préposition **niedaleko** qui exige le génitif, l'accusatif ou le locatif

Prep_po décrit la syntaxe de la préposition **po** qui exige le locatif, le datif ou l'accusatif

Prep_w décrit la syntaxe de la préposition **w(e)** qui exige le locatif ou l'accusatif

Prep_za décrit la syntaxe de la préposition **za** qui exige le génitif, l'accusatif, l'instrumental ou le locatif