ОЧУВАЊЕ, ЗАШТИТА И ПЕРСПЕКТИВЕ РОМСКОГ ЈЕЗИКА У СРБИЈИ

SERBIAN ACADEMY OF SCIENCES AND ARTS

SCIENTIFIC MEETINGS
Book CLXXV

DEPARTMENT OF SOCIAL SCIENCES
Book 40

COMMITTEE FOR THE STUDY OF THE LIFE
AND CUSTOMS OF THE ROMA

PRESERVATION, SAFEGUARDING AND PROSPECTS OF THE ROMANI LANGUAGE IN SERBIA

PROCEEDINGS OF THE SCIENTIFIC CONFERENCE
HELD ON OCTOBER 20–21, 2016

Accepted for publication at the 6th Session of the Department of Social Sciences, held on June 1, 2018, after being reviewed by Academicians *Tibor Varady*, Professors *Ranko Bugarski, Dragoljub B. Đorđević, Ivana Vučina Simović, Ana Jovanović, Maja Miličević Petrović, Dragan Todorović, Sanja Zlatanović*, PhD, *Stana Ristić*, PhD, *Mirjana Mirić*, PhD

E d i t o r s

Academician TIBOR VARADY
BILJANA SIKIMIĆ, PhD

B E L G R A D E 2 0 1 8

СРПСКА АКАДЕМИЈА НАУКА И УМЕТНОСТИ

НАУЧНИ СКУПОВИ
Књига CLXXV

ОДЕЉЕЊЕ ДРУШТВЕНИХ НАУКА
Књига 40

ОДБОР ЗА ПРОУЧАВАЊЕ ЖИВОТА И ОБИЧАЈА РОМА

ОЧУВАЊЕ, ЗАШТИТА И ПЕРСПЕКТИВЕ РОМСКОГ ЈЕЗИКА У СРБИЈИ

ЗБОРНИК РАДОВА СА НАУЧНОГ СКУПА
ОДРЖАНОГ 20–21. ОКТОБРА 2016.

Примљено на VI скупу Одељења друштвених наука 1. јуна 2018. године
на основу рецензија академика *Тибора Варадија*, професора *Ранка Бугарског, Драгољуба Б. Ђорђевића, Иване Вучина Симовић, Ане Јовановић, Маје Миличевић Петровић, Драгана Тодоровића*, др *Сање Златановић*, др *Стане Ристић*, др *Мирјане Мирић*

У р е д н и ц и

академик ТИБОР ВАРАДИ
др БИЉАНА СИКИМИЋ

Б Е О Г Р А Д 2 0 1 8

Издаје
Српска академија наука и уметности
Београд, Кнеза Михаила 35

Коректура
Невена Ђурђевић

Технички уредник
Никола Стевановић

Тираж
300 примерака

Штампа
Colorgrafx, Београд

САДРЖАЈ

ПРЕДГОВОР

Српска академија наука и уметности и њен Одбор за проучавање живота и обичаја Рома Одељења друштвених наука на скуповима и трибинама посвећеним Ромима редовно се баве и питањима ромског језика и језика Рома. Тако је у тематским зборницима радова, почевши од 1992. године до данас, увек било места за лингвистичке студије или студије које су се из угла других хуманистичких дисциплина на неки начин дотицале и лингвистичких тема. Зборник радова *Очување, заштита и перспективе ромског језика у Србији* – у целини је посвећен ромском језику и први је такав у издањима Одбора.

У првом објављеном зборнику Академијиног Одбора за проучавање живота и обичаја Рома *Развитак Рома у Југославији. Проблеми и тенденције* (М. Мацура, ур.) (Београд: САНУ, 1992) објављене су лингвистичке студије Марсела Куртијадеа, Јанардан Синга Патаније и Шаипа Јусуфа. У следећем по реду зборнику *Друштвене промене и положај Рома* (М. Мацура, А. Митровић, ур.) (Београд: САНУ, 1993), није било лингвистичких радова.

Зборник *Цигани/Роми у прошлости и данас* (М. Мацура, ур.) (Београд: САНУ, 2000), објавио је радове са трећег по реду ромолошког скупа, одржаног 1996. Овај зборник садржи посебно поглавље посвећено теми „Језик и образовање“, у оквиру које су и две лингвистичке студије: Трифуна Димића и Ибрахима Османија. Уводну реч за зборник написао је лингвиста Павле Ивић (стр. 5–7). Два кључна питања која ова студија покреће и сада су, после више од двадесет година, актуелна у савременој ромологији. Први став Павла Ивића тиче се стандардизације ромског језика и имплицитно указује на званичан став лингвиста у Србији крајем 20. века:

> Учење стандардног ромског, који неизбежно мора садржати мноштво импровизованих речи из цивилизацијске и апстрактне сфере, стављаће

> сваког појединца пред озбиљан задатак, знатно тежи од учења стандардног језика средине, који се не мора посебно савладавати, него се, пошто испуњава човеково окружење, из њега неосетно упија у човека. (...) Ромски језик је од свих језика најмање подесан за амбициозну, друштвено релевантну стандардизацију. Нема изгледа да би неки ромски стандардни језик, било општи или регионални, могао постати заједнички комуникациони медиј Рома у Европи, или у некој европској земљи. (...) Бојим се да би наметање неког ромског стандардног језика у школама могло само отежати ионако не баш једноставан процес школовања ромске деце.

Лингвистичка ромологија, бележи даље Павле Ивић, веома је захвално поље истраживања: она од истраживача захтева вансеријска знања која превазилазе оквире опште и посебне лингвистике, њој се као науци предвиђа лепа будућност. Друго питање тицало се односа дијалеката и стандарда:

> Управо оне околности које отежавају његову стандардизацију чине га фасцинантним предметом лингвистичких истраживања. Када наука буде располагала подробним описима свих ромских дијалеката, то ће омогућити да се утврде њихове сличности и разлике, да се реконструишу путеви и донекле хронологија њиховог гранања. (...) једном речју, за лингвистичка проучавања језичка ситуација европских Рома остаће елдорадо.

После скоро десетогодишње паузе, следећи зборник у овом низу *Друштвене науке о Ромима у Србији* (Београд: САНУ, 2007) објављује и радове нове генерације лингвиста – Биљане Сикимић, Светлане Ћирковић и Мирјане Мандић. У још једном зборнику са скупа *Промене идентитета, културе и језика Рома у условима планске социјално-економске интеграције* (Београд: САНУ, 2012) већ у наслову се појављује „језик“ као једна од основних тема, па садржи и четири лингвистичка рада: Игора Лакића, Биљане Сикимић, Петра Радосављевића и Анамарије Сореску Маринковић, ови радови се тичу како ромског језика, тако и румунског којим говоре Бањаши. У зборнику *Прилози стратегији унапређења положаја Рома* (Т. Варади, Д. Б. Ђорђевић, Г. Башић, ур.) (Београд: САНУ и Заштитник грађана Републике Србије, 2014), рад о румунофоним Бањашима објављује Анамарија Сореску Маринковић. У зборнику *Роми Србије у XXI веку* (Т. Варади, ур.) (Београд: САНУ, 2018), могу се наћи и лингвистички радови Марсела Куртијадеа (о стандардизацији ромског језика) и Биљане Сикимић (о спорном идентитету Ковача у Санџаку).

Тек радови настали у 21. веку отварају питање изједначавања ромског језика и језика којим говоре Роми, што даље, са своје стране, отвара још једно широко дискутовано питање – ко су све Роми у Србији и које све језике говоре као свој први и други језик (будући

да постоји широко прихваћен став, како у академској заједници, тако и међу ромским активистима, да су сви Роми у Србији – двојезични). Ситуација је у пракси ипак сложенија: двојезични или вишејезични су сви Роми осим оних који су већ напустили ромски језик, било да су у питању целе заједнице или само поједине породице.

Сви ови радови, мада садрже вредне научне увиде, у целини узевши ипак не дају репрезентативну слику о стању ромског језика у Србији, нити су довољни да се може говорити о постојању ромолошке лингвистике као развијене дисциплине у Србији.

Одбор за проучавање живота и обичаја Рома САНУ покушао је да, прво организацијом међународног научног скупа „Очување, заштита и перспективе ромског језика у Србији“ у Београду 21–22. 10. 2016, и сада издавањем тематског зборника радова под истим именом, допринесе афирмацији ромологије у Србији и да представи реално стање у заштити ромског језика. На скупу је било изложено 18 радова, од којих се 15 објављује у овом зборнику.

Идеја организатора скупа била је да окупи истраживаче који би изложили своје ставове на основу актуелног пресека стања у заштити ромског језика и унапред понуђеног сета питања. Скуп је остао отворен и за теме ван задатог оквира које су као актуелне предложили сами истраживачи. Међу темама је и листа угрожених језика Унеска која класификује ромски језик у Србији као „дефинитивно угрожен“ (*definitely endangered*). У целом свету, према проценама Унеска, има око три и по милиона говорника ромског језика. Међутим, у Србији не постоје тачни научни подаци о броју говорника ромског језика, нити о стању његове социолингвистичке виталности. Не постоје ни тачни научни подаци о ромским дијалектима и локалним говорима (о броју говорника и географским ареалима). Не постоје ни тачни подаци о Ромима мигрантима и стању њиховог језика, као ни увиди у социолингвистичку ситуацију репатрираних Рома.

Организатори су желели да радови на скупу предложе основне мере заштите. Међу такве мере спадало би, на пример, снимање локалних говора и њихово мапирање, а у следећој фази постављање базе података о ромским говорима. База података би укључивала усмене и писане изворе како за све дијалекте, тако и за ромски стандард. Поставило се и питање конкретних начина обједињавања свих постојећих, већ обављених истраживања, у једну централну базу података и умрежавање истраживача. Организатори су имали у виду чињеницу да у Србији не постоји Завод за културу Рома, који би могао да буде носилац таквог посла, као и да је број компетентних истраживача веома мали. Очекивало се да скуп иницира оспособљавање тима истраживача који би могли да обаве сложене теренске задатке (дијалектолошке и социолингвистичке природе), а који би истовремено радили на другим обли-

цима очувања ромског језика, пре свега у настави. Жеља је била да се на основу радова помогне увођењу ромског језика и културе на академском нивоу у оквиру факултетских наставних планова и програма.

Још један од циљева скупа било је стицање увида у лингвистичка истраживања ромских говора у Србији и мапирање свих варијаната ромских говора уз евентуалну процену броја говорника сваког од њих. Други постављени циљ била је подршка академској настави ромског језика: успостављање консензуса око језичког стандарда, преглед постојећих искустава у настави ромског језика и препоруке за савремену методику наставе ромског језика као матерњег и нематерњег и, посебно, српског као нематерњег језика репатрираних Рома.

Учесницима скупа биле су предложене следеће оквирне теме: Ромски говори у Србији: дијалектолошки и социолингвистички увиди, Граматички опис ромског језика, Документовање ромских говора у Србији, Лингвистичка виталност ромског језика у Србији, Језичка политика у Србији о ромском језику, Инструменти заштите ромских говора у Србији, Ромски стандардни језик у Србији, Настава ромског језика на свим образовним нивоима, Ромски као нематерњи језик и Ромски језички пејзаж (*linguistic landscape*). А како је увид у стање заштите ромског језика у земљама у региону неопходан да би се боље осветлила ситуација у Србији, на скуп су били позвани и истраживачи из Босне и Херцеговине, Мађарске, Словеније, Хрватске и Црне Горе, уз истраживаче који су омогућили увиде у европске и светске оквире заштите ромског језика.

Зборник радова је уредднички обликован према научним изазовима које су понудили сами радови. У оквиру језичке политике посебно се издвајају радови са још увек нерешеном темом стандардизације ромског језика, уз посебан преглед актуелне језичке политике са акцентом на положај ромског језика у Србији, аутора Горана Башића. Следе радови који се баве очувањем и заштитом ромског језика у земљама у региону (Игора Лакића и Вере Клопчич, донекле и прилог Хедине Тахировић Сијерчић).

Настава ромског језика била је планирана као једна од основних тема: ипак, стањем наставе ромског језика у Србији бави се само коауторски рад Јелене Филиповић и Јулијане Вучо, а студија Масако Ватабе аутоматском језичком обрадом ромске граматике и њеном применом у формалној настави.

У зборнику следи тематски блок радова који се баве анализом садашњег стања у ромолошкој лингвистици у Србији: применом корпусне лингвистике бави се студија Свенке Савић, актуелним теренским истраживањима ромског језика студија Светлане Ћирковић, а стањем у ромолошкој лексикографији студија Љубице Ђурић. Нека конкретна питања лексикографске обрађености ромских модалних

глагола проблематизује прилог Рајка Ђурића. Домену историје ромологије у Србији (са анализираним узорцима говора београдских Рома са краја 19. века) припада студија Биљане Сикимић, а прилог Анамарије Сореску Маринковић из угла антрополошке лингвистике прати феномене употребе ромског језика у литургијској пракси малих верских заједница и тиме анализира звучни ромски језички пејзаж из савремене, неспорно ефемерне перспективе.

Радови који се у коначној форми објављују у зборнику припадају следећим лингвистичким поддисциплинама: језичка политика и заштита језика, методика наставе ромског језика, историја и преглед ромолошке лингвистике у Србији као и лингвистичка антропологија. Тематски блок језичке политике почиње уводним радом Горана Башића о службеној употреби ромског језика у Србији. Неке радове прожима одјек актуелне расправе о различитим задацима ромологије као науке и као примењене дисциплине, односно разматрања о приоритету лингвистичког описа у односу на активности у раду на стандардизацији. У зборник су укључена различита, понекад међусобно супротстављена мишљења о стандардизацији и начинима стандардизације ромског језика као и посебне улоге ромске (а не само ромолошке) академске заједнице у тим процесима.

Осим ових, у савременој ромологији и иначе актуелних тема, око којих ни данас у светској академској заједници нема јединственог става, кроз радове се може сагледати и слика знатно узнапредовалог стања у ромологији на некадашњем југословенском простору који, између осталог, одликује пролиферација аматерских лексикографских покушаја, употреба и (зло)употреба лексикографије. О овом феномену посебно реферишу студије Вере Клопчич, Љубице Ђурић, Рајка Ђурића, Свенке Савић, али и других аутора који се ове теме узгредно дотичу, свако на свој начин.

Већина радова у зборнику обухвата по неколико лингвистичких и сродних ромолошких тема, а аутори се често међусобно допуњавају и тематски надовезују: питањима стандардизације ромског језика, свака из свог угла, баве се студије Свенке Савић, Баје Саитовића, Љатифа Демира и Марсела Куртијадеа. Значајне предлоге за рад на корпусу дају Свенка Савић и Светлана Ћирковић. Валоризацијом преводилаштва са ромског и на ромски баве се посебно Свенка Савић и Хедина Тахировић Сијерчић.

Зборник као целина показује и то да је у Србији и у региону стасала нова генерација лингвиста ромолога, способна да се бави ромолошком лингвистиком као универзалном научном дисциплином, која, и када се бави анализом локалног, мора уважавати достигнућа светске науке.

Академик Тибор Варади и др Биљана Сикимић

THE NOOJ APPROACH OF AUTOMATIC LANGUAGE PROCESSING AS A TOOL FOR SYSTEMATIZATION OF RROMANI GRAMMAR IN BOTH DESCRIPTION AND FORMAL TEACHING

Masako Watabe*

A b s t r a c t . – NooJ (http://www.nooj-association.org/) is a linguistic environment that includes tools to create and maintain large-coverage lexical resources, as well as morphological and syntactic grammars, and parses corpora in real time. NooJ dictionaries and grammars are applied to texts in order to locate lexical, morphological and syntactic patterns and tag simple and compound words.

This article is aimed at showing how a NooJ module for the Rromani language works through examples of the NooJ dictionary, morphology, syntax and annotation, and how this module could contribute for systematization of Rromani grammar. With NooJ, even complex grammars may be formalized clearly and simply. Then linguistic resources formalized with NooJ may be employed for other projects such as formal teaching especially in the case of classes with dialectal diversity.

The Rromani module has a characteristic to cover all four dialects of Rromani (superdialects O and E, without or with phonetic mutation), namely polylectal, in order to formalize the entire Rromani language without favouring or disfavouring any dialect. This could contribute to dialect studies, and also to preserving the richness and diversity of the Rromani language and culture.

Another characteristic of the Rromani module is to constitute a diasystem, a modelization of the entire of dialectal subsystems into a single grammar. Different types of correspondence of diasynonyms (dialectal equivalents) explain the diasystemic unity of the Rromani. In fact, the dialectal diversity of Rromani is not an obstacle for the mutual understanding between good speakers of different dialects. The similarity between dialects justifies the indivisible unity of the Rromani language.

* Paris-Sorbonne University, Paris, France, masako.watabe@paris-sorbonne.fr

Key words: Rromani language, NooJ, NLP (Natural Language Processing), dialectology, diasystem, linguistics

1.INTRODUCTION: NOOJ

1.1 *NooJ, a linguistic environment*

NooJ (http://www.nooj-association.org/) is a linguistic environment. NooJ includes tools for creating and maintaining lexical resources with a wide coverage, as well as morphological and syntactic grammars to analyse corpora of various characters in real-time. NooJ grammars are applied to texts in order to locate lexical, morphological and syntactic units, and to recognize simple and compound words. NooJ is also used by social scientists to carry out literary, journalistic, didactic or sociological analysis of corpora, and by industry as a tool for extracting information from technical texts.

NooJ tools adapt to a wide range of languages; not only European languages, but also Asian languages without spaces between words in writing or Semitic languages whose writing direction is from right to left. To make NooJ function for a given language, it is necessary to create a module that corresponds to the particular system of that language. If there are specific needs, new functions can be created. NooJ develops continuously.

1.2 *NooJ modules*

NooJ modules are already developed for several languages belonging — to different language families:

- Austroasiatic (Vietnamese),
- Indo-European (Belarusian, Bulgarian, Croatian, English, French, German, Greek, Italian, Polish, Portuguese, Russian, Serbian, Slovenian, Spanish etc.),
- Semitic (Arabic, Hebrew),
- Ural-Altaic (Hungarian, Turkish),

to different typological categories:

- Isolating (Vietnamese),
- Agglutinative (Hungarian, Japanese, Turkish),
- Inflectional (Belarusian, Bulgarian, Croatian, English, French, German, Greek, Italian, Polish, Portuguese, Russian, Serbian, Slovenian, Spanish etc.),

and also for minority languages:

- Kabyle (Berber), Runa Simi (Quechuan), Sorani (Central Kurdish), Western Armenian.

But, all these modules are monolectal, i.e. including one language or one dialect in the system. Only the module for the Rromani language is polylectal, thus including the four main dialects of the Rromani, respectively called O-bi, O-mu, E-bi and E-mu; superdialects O and E, without or with mutation.

2. NOOJ MODULE FOR THE RROMANI LANGUAGE

The NooJ module for the Rromani language was born with a small dictionary and morphologies (inflectional and derivational) to describe the main paradigms of verbs, nouns and adjectives, including exceptions and diasynonyms (namely dialectal equivalents). The first NooJ dictionary for the Rromani language contained few entries, yet each entry had been intentionally chosen to represent an inflectional paradigm. To enrich the NooJ dictionary, an existing paper dictionary (Courthiade Marcel et alii, *Morri angluni rromane ćhibăqi evroputni lavustik*, 2009) will be imported. Currently it is a compact dictionary but rich in resources for morphology and dialectology. NooJ applies morphologies to dictionaries, then creates and conserves the dictionary of inflected forms. Through this dictionary, NooJ can recognize and annotate even inflected forms in texts to analyse these texts.

2.1. *Single and common module for the Rromani language*

The Rromani module has a characteristic to cover all four dialects of Rromani, namely polylectal. A single and common module for Rromani is aimed at formalizing the entire Rromani language without favouring or disfavouring any dialect. As we find in the declaration of the first World R̂omani Congress at London in 1971; "No one dialect is superior to any other dialect.[1]"

1 The full quotation goes: "(i) *Language.* It was recognized that the R̂omani language played an important rôle [*sic*] both as one of the distinctive features of the R̂omani people in each country in which they lived and as a link between different groups. The efforts of the English and Spanish Gypsies to restore their language to active use were approved.

It was recognized that all spoken R̂omani dialects are of equal merit and that no one dialect is superior to any other dialect. Nevertheless there was a need for an international standardized dialect which could be used in periodicals and in congresses. It was hoped that at the next Congress R̂omani could be used much more and less translation required.

It was agreed to start a journal, *Rômani Ćhib*, to discuss language problems and print poems, stories and articles in the language. This journal would contain a vocabulary of the

The polylectal module could contribute to dialect studies, and also to preserving the richness and diversity of the Rromani language and culture. We should notice the existence of locally forgotten vocabulary that causes the loss of specific concepts, and also the loss of cultures as practice of these concepts — as a former president of the International Rromani Union (IRU) Stanisław Stankiewicz mentioned, "Dialect is not an issue, oblivion is the issue[2]" (2003). But serious cultural projects and didactic efforts could help in reversing this process. A multilingual and polylectal dictionary for the Rromani language, *Morri angluni rromane ćhibăqi evroputni lavustik*[3] and a transnational web portal designed for school teaching of Rromani language and culture, R.E.D.-RROM[4] (Restoring the European Dimension of Rromani Language and Culture), share the same philosophy and objectives as the NooJ module for Rromani.

2.2. *Diasystem of the Rromani language*

Another characteristic of the Rromani module is to constitute a diasystem (a modelization of the entire of dialectal subsystems into a single grammar). Today, there are very few polylectal materials (computerized or otherwise) for the Rromani, yet the dialectal diversity of this language is not an obstacle for the mutual understanding between good speakers of different dialects. It would be therefore beneficial to gather the four dialects into a single Rromani module and establish the diasystemic unity of the Rromani. Furthermore, the similarity between dialects justifies the indivisible unity of the Rromani language.

2.2.1. *Opposition 1:* o *and* e

The division into four dialects of the Rromani language is defined by two types of isoglosses which are non-areal and crossed.

---

less common words. It was agreed that standardization of the alphabet was an urgent problem. In informal discussion later between members of the commission the following alphabet was proposed: a b ć ćh d e f g h ĥ i j k kh l m n o p ph r r̂ s ś t th u v z ź (This corresponds to the script used in JGLS with exception of ĥ and r̂)." (*Journal of the Gypsy Lore Society*, 1971).

[Note that these two letters were eventually replaced respectively by *x* and double *rr*, while ʒ standing for both [ʤ] and [ʒ] was added at the Warsaw Congress in 1990].

[2] Stanisław Stankiewicz mentioned that during the seminar held at the Council of Europe in September 2003 on "The cultural identities of Roma, Gypsies, Travellers and related groups in Europe".

[3] Courthiade Marcel et alii, 2009.

[4] www.red-rrom.com

The criteria of the first isogloss is the opposition *o/e* forming the two superdialects:

- the superdialect O,
- the superdialect E.

This opposition is systematically marked on endings of the verbs (1SG.PST.IND), the copula (1SG.PRS.IND) and the definite article (PL.DRT).

Correspondence	Superdialect O	Superdialect E	
Opposition *o/e*	**phirdom** **phirdŏm**, **phirdŭm**	**phirdem**	(*I*) *walked*
	sinom, **som** **sinŏm**, **hom**, **hium**	**sem**	(*I*) *am*
	o ćhave	**e ćhave**	*the Rromani boys*, *the sons*

Fig. 1. Opposition *o/e* between the two superdialects: O and E

In fact, this isogloss is associated with other several isoglosses, and thus between the two superdialects there are not only the correspondence *o/e*, but also other types of correspondences (morphological, phonological and lexical ones) which are systematically applied to the concerned words to make dialectal equivalents (diasynonyms).

Correspondence	Superdialect O	Superdialect E	
Morphological	**daràndilăs**	**daràjas**	(*he*/*she*) *was afraid*
Phonological	**pani**	**paj**	*water*
	phen	**phej**	*sister*
	ćhaj	**ćhej**	*Rromani girl*, *daughter*
	daj	**dej**	*mother*
Lexical	**puzgal**	**istral**	*to slip*
	ćulal	**pitălj**	*to drip*

Fig. 2. Various types of correspondences between the two superdialects

2.2.2. *Opposition 2: without or with phonetic mutation*

The criteria of the second isogloss is a phonetic mutation which forms the two dialectal subgroups:

- without phonetic mutation,
- with phonetic mutation.

The two consonants *ćh* and ʒ are originally affricates [ʧʰ] and [ʤ], but these are transformed into alveolo-palatal fricatives [ɕ] and [ʑ] according to the phonetic mutation. This mutation concerns only the phonetic realization and there is no change in the writing.

Correspondence	Without mutation	With mutation	
Phonetic mutation	**ćhavo** [ʧʰavo]	**ćhavo** [ɕavo]	*Rromani boy, son*
	ʒukel [ʤukel]	**ʒukel** [ʑukel]	*dog*

Fig. 3. Opposition "affricates vs. alveolo-palatal fricatives" between the two dialectal subgroups: without or with phonetic mutation

This isogloss is also associated with other types of correspondence (phonological, morphological ones).

Correspondence	Without mutation	With mutation	
Phonological	**tiro**, **to**, **klo**, **ko**	**tĭro**, **tŏ**	*your*
	sinom, **som**, **sem**	**hom**, **som**, **sem**	(*I*) *am*
Morphological	**po**, **piro**	**pesqo**	*one's own*

Fig. 4. Various types of correspondences between the two dialectal subgroups: without or with phonetic mutation

2.2.3. *Four dialects of the Rromani language*

The entire of the Rromani language is divided into the two superdialects (O and E), then each of the superdialects is divided into the two subgroups (without or with phonetic mutation). The four dialects of the Rromani language have been formed as a result of the two crossed isoglosses.

	Without mutation	With mutation
Superdialect O	O-bi[5] (superdialect O without mutation)	O-mu[6] (superdialect O with mutation)
Superdialect E	E-bi (superdialect E without mutation)	E-mu (superdialect E with mutation)

Fig. 5. The four dialects of the Rromani language

Beyond this double dichotomy, certain phenomena may be common to the two units separated by one of the isoglosses; for example, in the north region of the Danube all verbs have a common ending *-ă* for 3SG.PST.IND; **avilă** (*he*/*she*) *came*, while in the south region of the Danube certain intransitive verbs have endings *-o*/*-i* for 3SG.PST.IND each of which agrees with its subject in gender; **avilo** (*he*) *came*, **avili** (*she*) *came*. This isogloss lying on the Danube crosses the two main isoglosses. The diasystem of the Rromani language is composed of a polynomic structure.

[5] In Rromani, **bi** means *without*.

[6] The indication "mu" comes from the word mutation.

3. NOOJ DICTIONARY

As the Rromani module is part of the NooJ system, certain specific rules to NooJ must be respected. In the NooJ dictionary each line starts with an entry word. Then, the entry word is followed by a comma "," and a lexical category in capital letters and, if it is necessary, morphological properties in small letters. An example in Fig. 6 shows that the entry word (i.e. lemma) **ruv** is an animal masculine noun (ruv,N+ani+m).

Then, "EN" precedes the translation in English *wolf*, "FLX" precedes the inflectional paradigm **rrom**, and finally "DRV" precedes the derivational paradigm **rromorro** and its inflectional paradigm **ćhavo**. In Fig. 6, only **ruv** *wolf* is the lemma, on the other hand, **rrom** *husband*, **rromorro** *hubby*, **ćhavo** *Rromani boy*, *son* are paradigm names. In the Rromani module, each paradigm is called with a representative Rromani word. For example, the name of a nominal paradigm **rrom** has been chosen to represent all of human or animal masculine consonant and oxytonic nouns (e.g. **rrom** *husband*, **dad** *father*, **ruv** *wolf*, **grast** *horse*) which have a common paradigm. According to the indication "FLX=**rrom**" in the dictionary, NooJ users understand that the inflectional paradigm **rrom** is applied to the lemma **ruv** *wolf*.

```
ruv,N+ani+m+EN="wolf"+FLX=rrom+DRV=rromorro:ćhavo
```

Fig. 6. Entry **ruv** in the NooJ dictionary for Rromani

On the first line in Fig. 7, the word **rrom** appears twice, but only the first one is a lemma, while the second is a paradigm name. The entry **rrom** is a human masculine noun (rrom,N+hum+m), the translation in English is *husband*[7], the inflectional paradigm is **rrom**, the derivational paradigm is **rromorro** and the inflectional paradigm of the derivative is **ćhavo**.

```
rrom,N+hum+m+EN="husband"+FLX=rrom+DRV=rromorro:ćhavo
ćhavo,N+hum+m+EN="Rromani boy, son"+FLX=ćhavo+DRV=ćhavorro
```

Fig. 7. Entries **rrom** and **ćhavo** in NooJ dictionary for Rromani

On the second line with the entry **ćhavo** *Rromani boy*, *son* in Fig. 7, the inflectional paradigm of the derivative **ćhavorro** *Rromani little boy*, *little son* is not indicated. This is because the lemma **ćhavo** and its derivative (i.e. diminutive) **ćhavorro** have the same inflectional paradigm. If the inflectional paradigm of a derivative is not indicated, its lemma's paradigm will be automatically applied. The entry **ćhavo** is a human

[7] As the noun **Rrom** meaning *Rromani man* always has the capitalized initial, it is another entry word in the NooJ dictionary for the Rromani language.

masculine noun (**ćhavo**,N+hum+m), the translation in English is *Rromani boy*, *son*, the inflectional paradigm is **ćhavo**, the derivational paradigm is **ćhavorro** and the inflectional paradigm of the derivative is **ćhavo** too.

4. NOOJ INFLECTIONAL MORPHOLOGY

4.1. *Basic inflectional morphology*

Also in morphology, specific rules to NooJ must be respected. Each paradigm begins with its name (e.g. **rrom** in Fig. 8), the content of the paradigm begins with an equal sign "=", and ends with a semicolon ";". A vertical bar "|" separates different forms, a slash "/" separates an ending and its morphological properties, parentheses "( )" group endings that have common morphological properties, and "<E>" means "Empty String" (i.e. without change).

```
rrom = <E>/sg+dr | a/pl+dr |
       es/sg+ob | en/pl+ob |
       (!a|e!a|e!ana)<E>/sg+voc | a!len/pl+voc ;
```

Fig. 8. Basic inflectional paradigm **rrom**

If the paradigm **rrom** *husband* is applied to the lemma **ruv** *wolf*, what will happen? The lemma **ruv** will be treated as the initial form. First, nothing will be added to the initial form **ruv** to make the singular direct form (sg+dr) **ruv**, the ending **-a** will be added to the initial form **ruv** to make the plural direct form (pl+dr) **ruva**, the ending **-es** will be added to the initial form **ruv** to make the singular oblique form (sg+ob) **ruves**, the ending **-en** will be added to the initial form **ruv** to make the plural oblique form (pl+ob) **ruven**, each of the three variants of endings **-!a -e!a -e!ana** will be added to the initial form **ruv** to make the singular vocative forms (sg+voc) **ruv!a**, **ruve!a**, **ruve!ana**, and the ending **-a!len** will be added to the initial form **ruv** to make the plural vocative form (pl+voc) **ruva!len**. Then the lemma **ruv** will give eight forms.

ruv + <E>	=> **ruv**	*wolf*	(sg+dr)
ruv + **-a**	=> **ruva**	*wolves*	(pl+dr)
ruv + **-es**	=> **ruves**	*wolf*	(sg+ob)
ruv + **-en**	=> **ruven**	*wolves*	(pl+ob)
ruv + **-!a**, **-e!a**, **-e!ana**	=> **ruv!a**, **ruve!a**, **ruve!ana**	*wolf!*	(sg+voc)
ruv + **-a!len**	**=> ruva!len**	*wolves!*	(pl+voc)

In fact, the paradigm **rrom** has a more complex structure including embedded paradigms of postpositions.

4.2. *Embedded paradigms*

In the present module for Rromani, even the postpositions are programmed as nominal endings,[8] and the paradigms of the postpositions are embedded in each nominal paradigm.

```
rrom = <E>/sg+dr | a/pl+dr |
       es(<E>|:pstpS|:possS|:possL)<E>/sg+ob |
       en(<E>|:pstpX|:possS|:possL)<E>/pl+ob |
       (!a|e!a|e!ana)<E>/sg+voc | a!len/pl+voc ;
```

Fig. 9. Complete inflectional paradigm **rrom**

The difference between the two paradigms **rrom** in Fig. 8 and Fig. 9 is marked in the oblique forms. For example, the ending of the singular oblique form **-es** is followed by parentheses (<E>|:pstpS|:possS|:possL), that means each of the elements between parentheses "<E>" ":pstpS" ":possS" ":possL" will be applied to the ending **-es**. However, there is not a form ***ruves:pstpS** because "pstpS" followed by a colon ":" is not an ending, but an embedded paradigm. There are four embedded paradigms in the nominal paradigm **rrom**:

- "pstpS" is the paradigm of the invariable postpositions[9] which is agglutinated to the endings with *-s* at the end (e.g. **-es**),
- "pstpX" is the paradigm of the invariable postpositions which is agglutinated to the endings without *-s* at the end (e.g. **-en**),
- "possS"" is the paradigm of the possessive[10] in short forms (e.g. **-qo** *of*),
- "possL" is the paradigm of the possessive in long forms (e.g. **-qoro** *of*).

[8] In fact, the postpositions in Rromani are not nominal endings, but these were treated as nominal endings for a technical reason. In the latest version of the Rromani module, the postpositions are programmed as agglutinations.

[9] The main invariable postpositions in Rromani are: **-qe** *for*, **-ça** *with*, **-θar** *from*, **-θe** *at*.

[10] The possessive in Rromani consists of the variable postposition (e.g. **-qo**/**-qi**/**-qe** *of*), except the possessive of the personal pronouns of the first and second persons (e.g. **amaro** *our*, **tumaro** *your*). In both, the possessive agrees with its possessed in number, gender and case.

If the paradigm **rrom** is applied to the lemma **ruv** *wolf*, for example the singular oblique form **ruves** will be with nothing (<E>), with one of the invariable postpositions (:pstpS), with the postposition of the possessive in short form (:possS), or with the postposition of the possessive in long form (:possL). Then, NooJ will make automatically 519 inflected forms[11] from the lemma **ruv**.

5. NOOJ INFLECTIONAL MORPHOLOGY IN GRAPH

5.1. *Inflectional paradigm* rrom *in graph*

To program grammars in the NooJ system, we can choose between two ways: rule editor (as we saw in the previous chapter) or graphical editor. Despite the visual difference, the two paradigms **rrom** (Fig. 9 by rule editor and Fig. 10 by graphical editor) have the same morphological value and, if these are applied to the lemma **ruv** *wolf*, each of them will give exactly the same inflected forms.

In the graphical morphology, each linear sequence corresponds to the ending of a form. The "Empty String" sign "<E>" is transformed into a triangle and morphological properties are indicated under nodes. Embedded paradigms (pstpS, pstpX, possS, and possL) are marked with the yellow colour in the background instead of a colon ":". For example, if we apply the graphical paradigm **rrom** to the lemma **ruv** *wolf* and follow the first line, the entry form **ruv** will not change anything in the singular direct (sg+dr) and the form **ruv** *wolf* will be given; then if we follow the second line, the ending **-a** in the plural direct (pl+dr) will be added to the entry form **ruv** *wolf* and the form **ruva** *wolves* will be given; these are exactly the same process as in the paradigm programmed by rule editor.

The contents of the embedded paradigms is not visible in the supra-paradigm **rrom**, because these are programmed separately and are stored on different layers of the main grammar ("MAIN" in Fig. 10) including the entire of the Rromani morphology. If we need to see the structure of the whole grammar, it is possible to display that, as we see on the left in Fig. 10. The paradigm **rrom** and its embedded paradigms are linked in the NooJ morphology for the Rromani, and most of these embedded paradigms are reused in other nominal paradigms.[12]

[11] Half of these forms concern the diminutive.

[12] Four paradigms (pstpS, pstpX, possS, possL) are embedded in other paradigms of masculine nouns (e.g. **čhavo** *Rromani boy*, *son*) too, however only three paradigms (pstpX, possS, possL) are embedded in paradigms of feminine nouns because the oblique forms of the

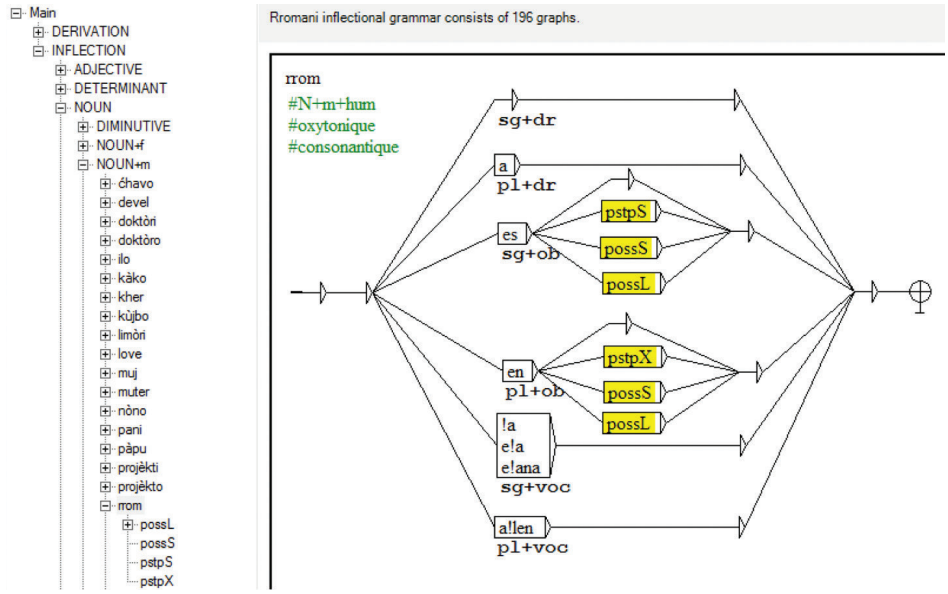

Fig. 10. Paradigm **rrom** in graph and structure of NooJ morphology

In the NooJ morphology by rule editor, all the paradigms are displayed in a single window without superimposed structure, and we have to go up or down on a two-dimensional window to consult different paradigms. As the NooJ morphology by graph editor can display the paradigms in a superimposed structure, reconstitute in a simpler but more effective way a complex morphological system, and give more visibility to locate a paradigm in the entire morphology, it seems appropriate to use the graphical morphology for the Rromani module.

5.2. *Paradigms of invariable postpositions*

The principle of the embedded paradigms is the same as the supra-paradigms. For example, the paradigm "pstpX" (Fig. 11 to the right) includes four invariable postpositions and a part of the circumposition, and the whole content of this paradigm follows the ending of the plural oblique **-en**. If the paradigms **rrom** and "pstpX" are applied to the lemma **ruv**, NooJ will give, in addition to the eight forms without postposition which we have seen above, also six forms with invariable postpositions below:

---

feminine nouns never end with **-s** and the paradigm of invariable postpositions is identical to the singular and the plural.

ruvença	*with wolves*	(instrumental),
ruvençar	*with wolves*	(instrumental in the dialect O-bi),
ruvenθar	*from wolves*	(ablative),
ruvenθe	*at wolves' place*	(locative),
ruvenqe	*for wolves*	(dative),
***ruvenqo**		(part of the abbessive).

The ending **-qo** in the paradigm "pstpX" is not the postposition of the possessive **-qo** *of*, but it is part of the circumposition of the abbessive **bi -qo** *without*. Therefore the form ***ruvenqo** is incomplete. The capital letter "X" in NooJ marks an element which cannot exist alone as ***ruvenqo** meaning nothing without **bi**. There are homographic[13] forms; a part of the abbessive ***ruvenqo**, and the possessive **ruvenqo** *of wolves*. To remove this ambiguity, we need a syntactic grammar presented in the next chapter (Fig. 18).

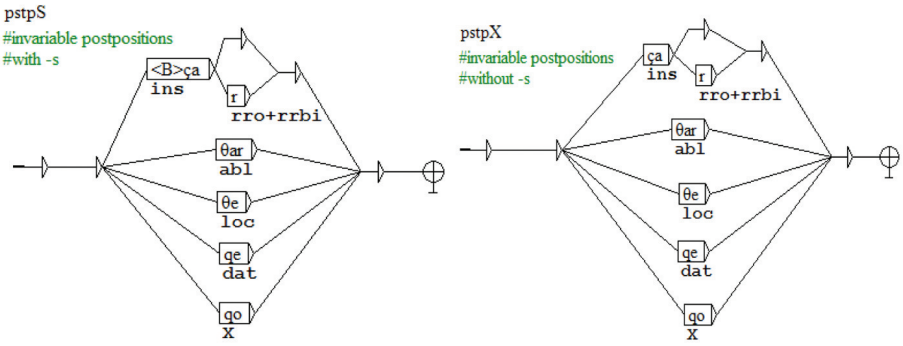

Fig. 11. Embedded paradigms of invariable postpositions

The only difference between the paradigms "pstpS" and "pstpX" is marked in the sequence of the instrumental; "<B>ça" in "pstpS" and "ça" in "pstpX". The command <B> meaning "Back Space" indicates deleting of a letter starting from the end of the preceding form. It is because, when the postposition **-ça** *with* is added to a form with **-s** (e.g. **ruves** *wolf* in the oblique), this **-s** will be contracted. For example, if the command "<B>ça" is applied to the form **ruves**, first **ruves** will be ***ruve**[14], then **-ça** will be added to ***ruve**. Therefore the singular instrumental form of the lemma **ruv** *wolf* will not be ***ruvesça**, but **ruveça** *with wolf.*

[13] NooJ is a system exclusively written. NooJ does not recognize any ambiguity between homonyms that are written in different ways. Besides it is very rare in Rromani (e.g. **dasa** *Christians* vs. **daça** *with mother*).

[14] The form **ruve** *wolves* exists as the plural direct form, but it does not exist as the singular oblique form.

If the paradigms **rrom** and "pstpS" are applied to the lemma **ruv**, NooJ will give eight forms without postposition and six forms with invariable postpositions; as the result with the paradigms **rrom** and "pstpX".

5.3. *Short and long forms of the possessive*

The only variable postposition in Rromani is the possessive, it is agglutinated to oblique forms of nouns and personal pronouns.[15] There are three basic forms: **-qo**/**-qi**/**-qe** each of which agrees with the genders, the numbers and the inflectional cases[16] of possessed.

gurumnăqo śing	*cow's horn*	(possessed is m+sg+dr)
gurumnăqi ćhib	*cow's tongue*	(possessed is f+sg+dr)
gurumnăqe danda	*cow's teeth*	(possessed is m+pl+dr)
gurumnăqe dandença	*with cow's teeth*	(possessed is m+pl+ob)

The basic position of the possessive is before its possessed object (samples 1 & 3), however the possessive can be placed after its possessed object, then the possessive will be a long form and have an emphatic value[17] (samples 2 & 4).

• Short form of the possessive before its possessed:

1) **gurumn-ă-q-o** **śing**
cow-F.SG.OBL-POSS-M.SG.DRT horn[M.SG.DRT]
cow's horn (neutral)

• Long form of the possessive after its possessed :

2) **śing** **gurumn-ă-q-oro**
horn[M.SG.DRT] cow-F.SG.OBL-POSS-M.SG.DRT
horn of cow (emphatic)

---

[15] It concerns only the third persons: **les** *him*, **la** *her*, **len** *them* (in the oblique).

[16] In Rromani there are two types of cases: inflectional ones which are completely morphological (direct and oblique), and functional ones which are expressed by combinations of one of the inflectional cases and a preposition or a postposition. For example, **ruv** *wolf* (direct) and **ruves** *wolf* (oblique) are inflectional cases, on the other hand, **katar o ruv** *from the wolf* composed with a preposition **katar** *from* and a noun **ruv** *wolf* in the direct and **e ruvesθar** *from the wolf* composed with a noun **ruves** *wolf* in the oblique and a postposition **-θar** *from* are functional cases (i.e. ablative).

[17] The position of the possessive and the choice between short or long form depend on dialects too.

• Long form of the possessive before its possessed with a postposition:

3)	**kangl-ă**	**gurumn-ă-q-e**	**śing-es-θar**
	comb-F.PL.DRT	cow-F.SG.OBL-POSS-M.SG.OBL	horn-M.SG.OBL-ABL
	combs in cow's horn (neutral)		

• Long form of the possessive with the same postposition as its possessed's one :

4)	**kangl-ă**	**śing-es-θar**	**gurumn-ă-q-eres-θar**
	comb-F.PL.DRT	horn-M.SG.OBL-ABL	cow-F.SG.OBL-POSS-M.SG.OBL-ABL
	combs in horn of cow (emphatic)		

Moreover, if the possessed has an invariable postposition and is followed by a long form of the possessive, this possessive will have a nominal[18] ending and the same postposition of its possessed (sample 4). This invariable postposition added to the possessive is for a purely grammatical reason and without any functional value. For example, the form **gurumnăqeresθar** *of cow* includes an invariable postposition **-θar** *from*, but its functional value is the possessive and not the ablative; this is because its possessed **śing** *horn* includes the postposition of the ablative **-θar** *from*.

5.4. *Paradigm of the possessive in short forms*

In the NooJ morphology for Rromani, the postposition of the possessive is programmed in two paradigms:

- possS: paradigm of the possessive in short forms,
- possL: paradigm of the possessive in long forms.

There are only three short forms (**-qo**/**-qi**/**-qe**) at the inflectional level, yet the paradigm "possS" (Fig. 12) will give eight "forms" according to the morphological properties of the possessed. It is aimed at better annotating texts in the NooJ system.

-qo : Dm+Dsg+Ddr,	**-qi** : Df+Dsg+Ddr,
-qe : Dm+Dpl+Ddr,	**-qe** : Df+Dpl+Ddr,
-qe : Dm+Dsg+Dob,	**-qe** : Df+Dsg+Dob,
-qe : Dm+Dpl+Dob,	**-qe** : Df+Dpl+Dob.

[18] The inflection of the possessive in Rromani is basically adjectival.

The capital letter "D" precedes the morphological properties of determined (i.e. possessed). For example, **gurumnăqo** *cow's* (sample 1) is the possessive in short form of the noun **gurumni** *cow*, the properties of the possessor are f+sg+ob, on the other hand, the properties of the postposition of the possessive **-qo** *of* are Dm+Dsg+Ddr agreeing to the possessed **śing** *horn*.

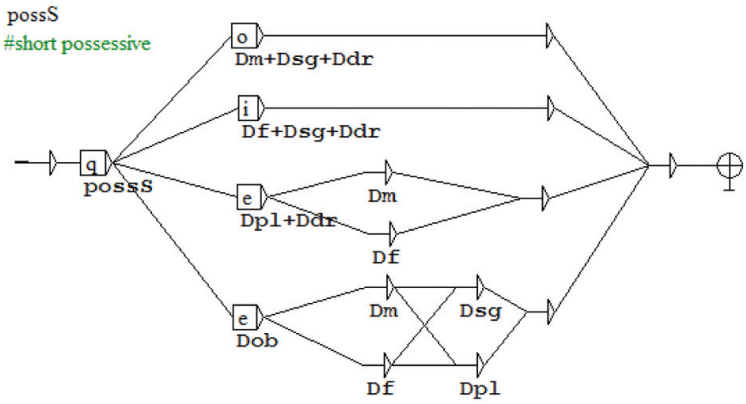

Fig. 12. Paradigm of the possessive in short forms

The long forms of the possessive are used in the dialect O-bi[19]. The final vowels (-o/-i/-e) are common to short forms, on the other hand, there are variants of radicals according to the vernaculars:[20]

-qero/-qeri/-qere	in the Carpathians (rrc[21]),
-qëro/-qëri/-qëre or **-qro/-qri/-qre**	in Russia and North of Poland (rrn),
-qoro/-qiri/-qere	in the Balkans (rrs).

[19] The dialect O-bi meaning the superdialect O without mutation is expressed by the combination of two properties "rro+rrbi" in the Rromani module.

[20] In the Rromani dialectology, the four dialects whose isoglosses are not areal are distinguished from the vernaculars as the results of the contact with local languages.

[21] The last letters of these codes (i.e. "c" "n" and "s") correspond to geographic characteristics of these vernaculars; "c" as Carpathians, "n" as nord, and "s" as south.

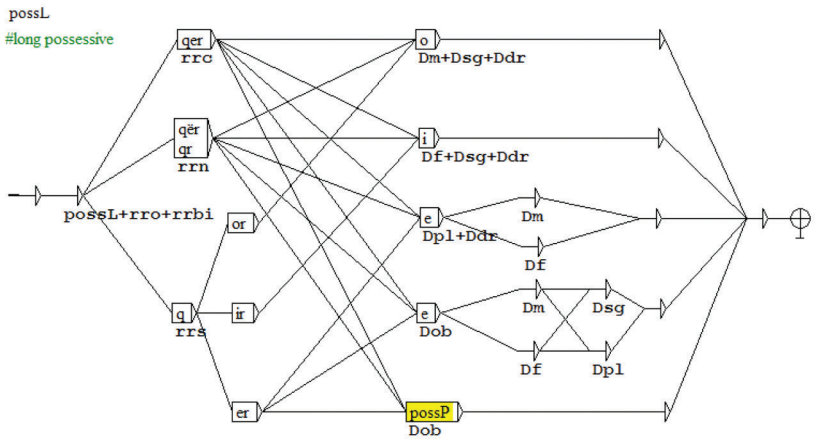

Fig. 13. Paradigm of the possessive in long forms

Each of these variants has eight "forms" according to the properties of possessed; 4×8=32, and these 32 forms are applied to the two numbers (singular, plural) of the possessor; 32×2=64. The paradigm "possL" (Fig. 13) gives 64 long forms of the possessive.

In fact, the paradigm "possL" includes another paradigm "possP" (Fig. 14) which is the paradigm of the endings of the postposed possessive. As we have seen above (sample 4), if an invariable postposition is added to a long form of the possessive, the ending of the possessive will be not adjectival, but nominal.

• In vernaculars "rrc" and "rrs"

gurumnăqeres-	*of cow*	(possessed is m+sg+ob and postposed),
gurumnăqeren-	*of cows*	(possessed is m+pl+ob and postposed),
gurumnăqeră-	*of cow*	(possessed is f+sg+ob and postposed),
gurumnăqerĕn-	*of cows*	(possessed is f+pl+ob and postposed).

• In vernacular "rrn"

gurumnăqëres-	or **gurumnăqres-**	*of cow*	(possessed is m+sg+ob and postposed),
gurumnăqëren-	or **gurumnăqren-**	*of cows*	(possessed is m+pl+ob and postposed),
gurumnăqëră-	or **gurumnăqră-**	*of cow*	(possessed is f+sg+ob and postposed),
gurumnăqërĕn-	or **gurumnăqrĕn-**	*of cows*	(possessed is f+pl+ob and postposed).

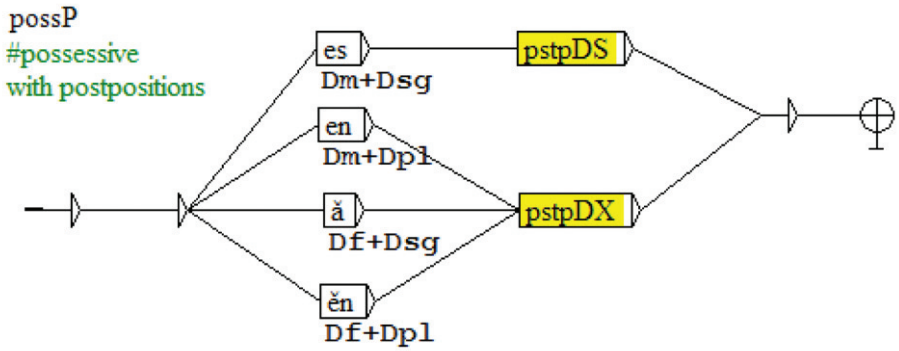

Fig. 14. Paradigm of the endings of postposed possessive

Each of these forms will be followed by four invariable postpositions. We have seen above that if the postposition of the instrumental **-ça** *with* follows the masculine singular oblique forms with **-s**, this **-s** will be contracted. There are therefore two paradigms of the invariable postpositions for the possessive in long form (Fig. 15), but contrary to Fig. 11, there is not "**-qo** X" being part of the circumposition **bi -qo** *without*, as **-qo** being part of the circumposition cannot follow the possessive in long form (e.g. ***bi gurumnăqeresqo**, **bi gurumnăqo** *without cow*).

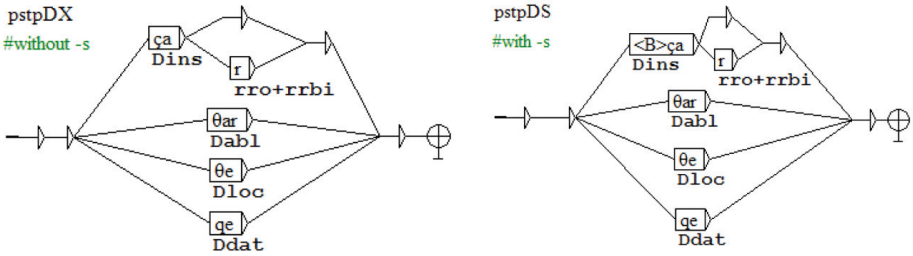

Fig. 15. Embedded paradigms of the invariable postpositions for the possessive in long form

Each of the 16 long forms of the possessive with the nominal endings can be followed by the four invariable postpositions whose instrumental has a variant **-çar** *with*; 16×5=80, then these 80 forms of the postposed possessive are applied to the two numbers (singular, plural) of the possessor; 80×2=160.

When the paradigm **rrom** and these seven embedded paradigms are applied to the lemma **ruv** *wolf*, NooJ will systematically give:

- 8 inflectional forms (direct, oblique, vocative) without postposition,
- 12 forms with invariable postpositions (including **-qo** being part of the circumposition),

- 16 short forms of the possessive,
- 64 long forms of the possessive,
- 160 long forms of the postposed possessive,
=> 260 forms[22] in total.

6. SYNTAX AND ANNOTATION OF TEXTS

E RROMANI ĆHIB Rajko Đurić
E rromani ćhib si ćhib e jagaqi thaj e balvalaqi
Voj si e khamesqo disŏpen, e ruvesqo thomupen.
E veśa an laθe baron, e parne-balenqe berśa vàʒe aśundon.
An rromani ćhib dikhol sar e ile kaśtenθar peren
vi sar e raja an ćhonutesqi angali śïlesθar roven (…)

The Rromani language Rajko Đurić
The Rromani language is a language of fire and wind
it is the rising sun and the howl of the wolf.
In it forests grow, and we still hear the years gone by.
In the Rromani language, we see the hearts falling from trees
we see gentlemen sobbing with cold in the arms of the moon (…)

6.1. *Two types of annotation: morphological and syntactic*

In fact, in the NooJ system there are two types of dictionary:
- lexical dictionary whose entry words are lemmas,
- inflectional dictionary whose entry words are inflected forms.

NooJ morphology is applied to the lexical dictionary in order to give inflected forms of all entry words of the lexical dictionary. On the other hand, the inflectional dictionary is applied to a text in order to annotate the text at morphological level. Then, to annotate a text at syntactic level, NooJ syntax will be applied to the text. In this chapter, by extracting from the poem **E RROMANI ĆHIB** *The Rromani language* written by Rajko Đurić, two types of NooJ annotation will be shown:

- morphological annotation on **(e) ruvesqo** *of* (*the*) *wolf*,
- syntactic annotation on **an laθe** *in her*[23].

[22] If we also count the forms of the diminutive, there will be 519 forms; one form is short of the double. It is because the diminutive **ruvorro** *little wolf* has only two variants (instead of three variants of the non-diminutive **ruv** *wolf*) of the singular vocative; ***ruvorr!a**, **ruvorre!a**, **ruvorre!ana** *little wolf!*

[23] This phrase means literally *in her*, yet the equivalent in English is *in it*, because *her* corresponds to **e rromani ćhib** *the Rromani language*, and **ćhib** is a feminine noun in Rromani.

6.1.1. *Morphological annotation*

In Fig. 16, each of three annotations is shown above a long arrow "→"; the first and second annotations are morphological and written in black, while the third one is syntactic and written in green. As the form **ruvesqo** can be, without context, the possessive **ruvesqo** *of wolf* or a part of the abbessive **bi ruvesqo** *without wolf*, NooJ keeps two possible annotations at morphological level.

Most of these two morphological annotations is common;

ruv,N+Spec=ani+Gd=m+EN="wolf"+Nb=sg+Cs=ob.

Each morphological annotation begins with the corresponding lemma **ruv**, it is a noun (N), the specificity (Spec) is animal (ani), the English translation (EN) is *wolf*, and the morphological properties of the inflected form **ruves** are masculine in gender (Gd=m), singular in number (Nb=sg), oblique in case (Cs=ob).

The difference between these two annotations begins with the properties of postpositions (Pstp):

- "X" means that **ruvesqo** being part of the abbessive **bi ruvesqo** *without wolf* cannot exist alone,
- "possS" means that **ruvesqo** is the possessive in short form, whose possessed (i.e. determined[24]) is masculine, singular, and direct (DGd=m, DNb=sg, DCs=dr).

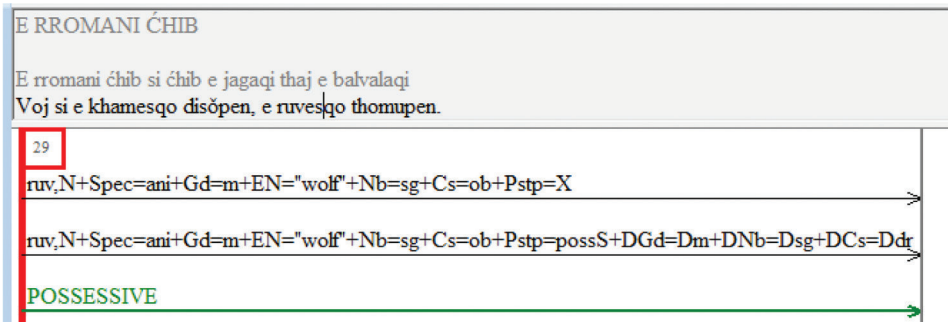

Fig. 16. Two types of annotation on **ruvesqo**

The NooJ system does not exclude morphological ambiguities between "X" and the possessive, but at the same time indicates the syntactic annotation "POSSESSIVE". And thus the NooJ users will understand that the form **ruvesqo** in the context **e ruvesqo thomupen** *the howl of the wolf* is not a part of the abbessive **bi ruvesqo** *without wolf*, but it is indeed the possessive.

[24] The capital letter "D" precedes the properties of the possessed (i.e. determined).

6.1.2. *Syntactic annotation*

Concerning the annotation on the abbessive, we will see, as an example, the title of another poem of Rajko Đurić **BI KHERESQO BI LIMORESQO** *without house*, *without grave* including two abbessive phrases: **bi kheresqo** *without house* and **bi limoresqo** *without grave*. In this case, each of **kheresqo** and **limoresqo** can be a homograph of the possessive, but in fact, these are parts of the abbessive, because **bi** (another part of the circumposition of the abbessive) precedes them.

In Fig. 17, NooJ keeps an ambiguity on **kheresqo** at the morphological level for the same reason which we have seen in Fig. 16. But, there is not any ambiguity at the syntactic level and the module recognizes the phrase **bi kheresqo**, the abbessive with a circumposition.

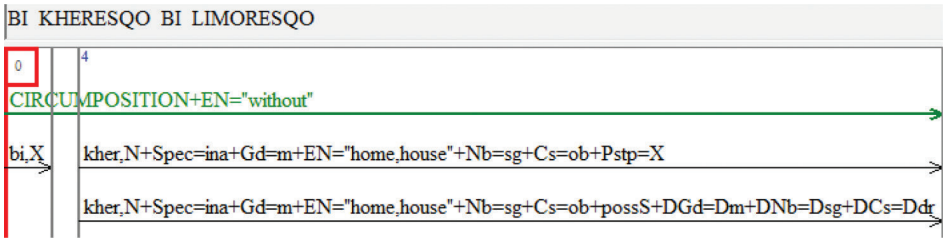

Fig. 17. Annotation of abbessive **bi kheresqo**

6.2. *NooJ syntax 1: abbessive*

How can the NooJ system distinguish the possessive and the abbessive?

First, we have to know that a linear sequence in NooJ morphology corresponds to a form and the space between two nodes (e.g. “es” and “pstpS” in Fig.10) is not recognized, while a linear sequence in NooJ syntax corresponds to a phrase and each node corresponds to a word (e.g. “<DET>[25]” in Fig.18 corresponding to one of any determiners).

The first sequence in Fig. 18 (i.e. “<N+possS+Dm+Dsg+Ddr>”) represents the possessive in short form of any nouns whose possessed is masculine singular direct (e.g. **ruvesqo** *of wolf*). Therefore, the syntactic annotation is “<POSSESSIVE>[26]”. On the other hand, the second sequence

[25] In NooJ, the angle brackets “< >” in a node give the value of inclusiveness, for example, “<DET>” represents any determiners, “<ruv,N>” represents any forms of the noun ruv, while “ruv” without angle brackets in a node means only the form ruv.

[26] In NooJ, the angle brackets “< >” outside the nodes indicate the beginning and the end of a syntactic sequence. In Fig. 18, these are below the arrows placed at the extreme left and right at the bottom.

(i.e. "<bi,X>" and "<N+X>") corresponds to a phrase including two words: **bi** whose lexical category is "X", and any nouns with **-qo** whose morphological property is "X" (e.g. **ruvesqo**). Therefore, the syntactic annotation is "<CIRCUMPOSITION>" meaning the abbessive *without* (e.g. **bi ruvesqo** *without wolf*).

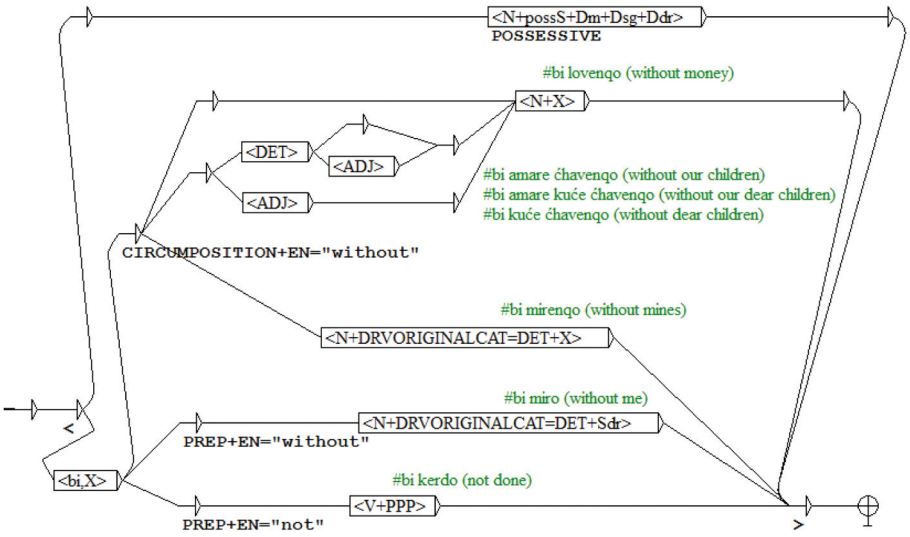

Fig. 18. NooJ syntax of the abbessive

The example **bi kheresqo** is a simple phrase, but in fact, other words can be inserted between **bi** and **kheresqo** such as a possessive and/or an adjective. Even if **bi** is not placed just before **kheresqo**, an abbessive phrase can be formed:

bi kheresqo	*without house*	bi N-**qo**,
bi amare kheresqo	*without our house*	bi DET N-**qo**,
bi kuće kheresqo	*without precious house*	bi ADJ N-**qo**,
bi amare kuće kheresqo	*without our precious house*	bi DET ADJ N-**qo**.

Then, **bi** can function as a preposition with a verb in the past passive participle (e.g. **kerdo** *done* => **bi kerdo** *not done*), or with a pronominal possessive (e.g. **miro** *my*, *mine* => **bi miro** *without me*), in this case what is absent is not the possessed *mine* but its possessor *me*. However, when the circumposition **bi -qo** is added to a pronominal possessive, what is absent is the possessed (e.g. **mire** *my*, *mine* in the plural => **bi mirenqo** *without mine* in the plural).

Through this syntactic grammar, NooJ can annotate correctly without confusing the possessive and the abbessive.

6.3. *NooJ syntax 2: locative without or with preposition*

In Rromani, there are three types of the locative which are often synonyms:

- with a preposition	**k-o**[27] **kher**	*at the home*,
- with a postposition	**e kheresθe**	*at the home*,
- with a suffix	**khere**	*at home*.

Basically, the prepositions precede a noun in the direct case in Rromani, as we see **k-o kher** above. But, when a spatial preposition (e.g. **pala**[28] *behind*) precedes a personal pronoun (e.g. **me** *I*), this personal pronoun will be locative (e.g. **manθe** *at my place*), then the equivalent phrase *behind me* in Rromani is **pala manθe**, and not ***pala me** with the pronoun in the direct, nor ***pala man** with the pronoun in the oblique. Off course, the personal pronouns in the locative case (e.g. **manθe** *at my place*) can function alone without any other preposition too. Therefore, as soon as a personal pronoun in the locative is detected in a text, NooJ will check if this pronoun is preceded by a spatial preposition or not.

The phrase **an laθe** *in her* from the poem of Rajko Đurić is correctly annotated as "LOCATIVE" (Fig. 19).

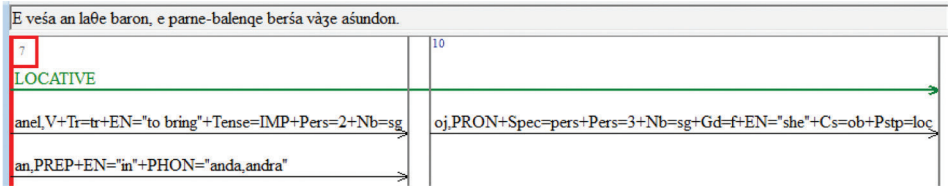

Fig. 19. Annotation of the locative phrase **an laθe**

NooJ syntax of the locative (Fig. 20) can recognize not only the locative phrase with a pronoun, but also several patterns of the locative phrases with a noun:

---

[27] The basic form of this preposition is **ka** *at*, but when a preposition ending with a vowel (e.g. **ka** *at*, **pala** *behind*) is followed by a definite article (e.g. **o** *the*) or a personal pronoun (e.g. **amen** *we*) beginning with a vowel, the final vowel of the preposition will be contracted and a hyphen will be inserted between the preposition and the definite article (e.g. **k-o kher** *at the house*, **pal-o kher** *behind the house*) or personal pronoun (e.g. **pal-amenθe** *behind us*).

[28] This preposition has a temporal value too, in this case it means *after*.

pal-o kher	*behind the house*	PREP DET+def N,
pal-o nevo kher	*behind the new house*	PREP DET+def ADJ N,
pala tumaro kher	*behind your house*	PREP DET+poss N,
pala tumaro nevo kher	*behind your new house*	PREP DET+poss ADJ N,
pal-e kakosqo kher	*behind uncle's house*	PREP DET+def DET+poss N,
pala tumare kakosqo nevo kher	*behind your uncle's new house*	PREP DET+poss DET+poss ADJ N.

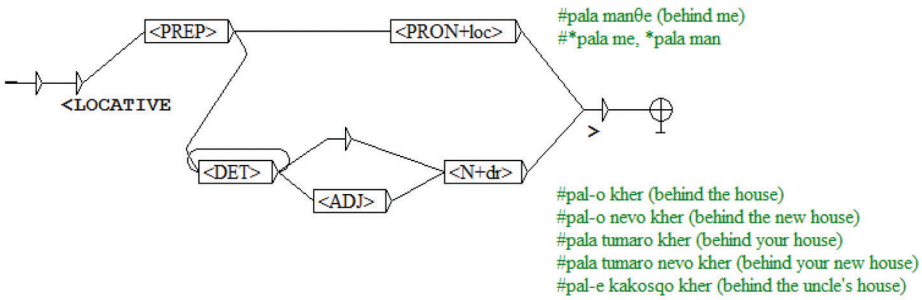

Fig. 20. NooJ syntax of the locative phrase

The round line above the node means the possibility of repeating the concerned word (i.e. "<DET>" in Fig. 20) without limit of number. To make a locative phrase with a noun, we need at least a preposition "<PREP>", a determiner "<DET>" and a noun in the direct "<N+dr>", but the number of the determiner may be more than one and an adjective can be inserted between the final determiner and the noun.

7. CONCLUSION – A STEP FOR AN AUTOMATIC TRANSLATOR?

Let's see the annotation of the entire phrase **E veśa an laθe baron** *In it forests grow* from the poem in Fig. 21. Wouldn't it be a first step for an automatic translator?[29]

5)	**e**	**veś-a**	**an**	**l-a-θe**	**bar-o-n**
	the.M.PL.DRT.EBI[29]	forest-M.PL.DRT	in	3-F.SG.OBL-LOC	big-V.MP-3PL.PRS
	=>the forests		=>in her		=>grow

[29] There are three variants of the definite article in the plural direct: **o** *the* in superdialect O (including two dialects O-bi and O-mu), **e** *the* in dialect E-bi, **le** *the* in dialect E-mu.

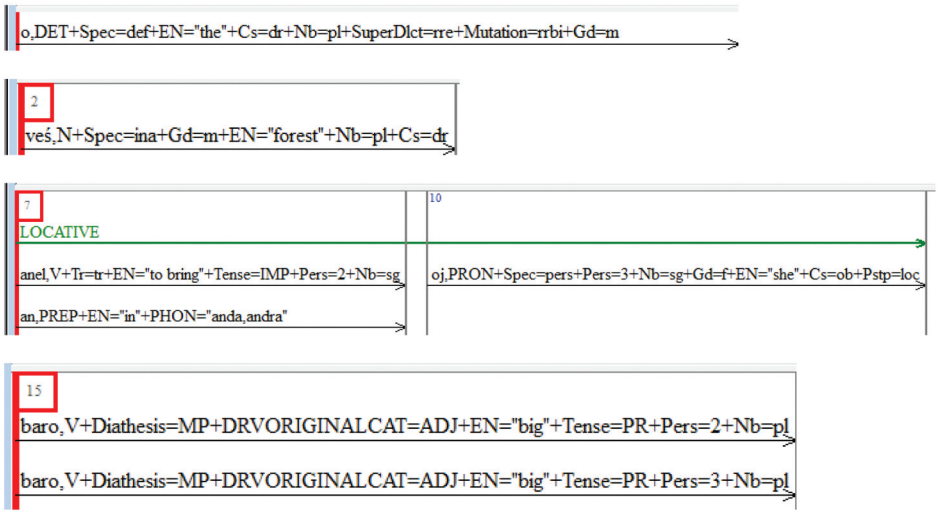

Fig. 21. Annotation of the phrase **E veśa an laθe baron**

In fact, **an** may be the imperative (2SG) of the verb **anel** *to bring* (Fig. 19), and the phrase **an laθe** include an ambiguity:

an laθe *in her*,
an laθe! *bring* (*something*) *at her place!*

This ambiguity is not resolved for the moment. Is it possible? It would be a question for a future project.

With the NooJ system, even complex grammars may be formalized clearly and simply. Then linguistic resources formalized with NooJ may be employed for other projects such as the development of several IT tools such as on-line dictionaries and spell checkers. The NooJ module could contribute to the Rromani language surviving and prospering in a hyper-connected world as well as "bigger" languages which are already equipped with various modern materials.

The NooJ module for the Rromani language, whose innovative characteristic is polylectal in a diasystemic approach, has the prospect to be employed for formal teaching, especially in the case of classes with dialectal diversity such as in Serbia, and in preserving the dialectal diversity of Rromani, while justifying the indivisible unity of the Rromani language.

REFERENCES

Courthiade, Marcel. *La littérature des Rroms, Sintés et Kalés*, Paris: INALCO, 2006.

Courthiade, Marcel et alii. *Morri angluni rromane ćhibăqi evroputni lavustik* (My first European dictionary of the Rromani language), Budapest: Cigány Ház, 2009.

Kenrick, Donald. "The World Ĵomani Congress—April 1971", *Journal of the Gypsy Lore Society* 1971/50 v., 1971, 102–103.

NooJ, available on http://www.nooj-association.org/.

R.E.D.-RROM (Restoring the European Dimension of Rromani Language and Culture), available on www.red-rrom.com.

Silberztein, Max. *La formalisation des langues: l'approche de NooJ*, London: ISTE Eds., 2015.

Zoli, Carlo. "'Smallcodes', a Unified computational linguistics toolbox for minority languages", in: *Les Technologies de l'Information et de la Communication* (*TICs*) *au service de l'amazighe*, ⴰⵙⵉⵏⴰⴳ-Asinag N° 9, Rabat, 2013.

ABBREVIATIONS

Lexical categories:

ADJ	adjective	PREP	preposition
DET	determiner	PRON	pronoun
N	noun	V	verb
X	linguistic unit being never employed alone		

Lexical subcategories:

def	definite article	pers	personal (pronoun)

Morphosyntactic properties:

abl	ablative	pl	plural
ani	animal	poss	possessive
dr	direct	possL	possessive in long form
dat	dative	possS	possessive in short form
f	feminine	PPP	past passive participle
hum	human	PR	present
IMP	imperative	PS	past
ina	inanimate	Pstp	postposition
ins	instrumental	sg	singular
itr	intransitive	tr	transitive
loc	locative	voc	vocative
m	masculine	1	1st person
MP	medio-passive	2	2nd person
ob	oblique	3	3rd person

Grammatical categories:

Cs	case	Nb	number
Gd	gender	Pers	person

Dialectal properties:

rrbi	without mutation	rrn	vernacular in Russia and north of Poland
rrc	vernacular in the Carpathians		
rre	superdialect E	rro	superdialect O
rrmu	with mutation	rrs	vernacular in the Balkans

NooJ commands:

DRV	derivational paradigm	FLX	inflectional paradigm
DRVORIGINALCAT	original category of derivate	<B>	back space
EN	translation in English	<E>	empty string

Масако Ватабе

NOOJ ПРИСТУП АУТОМАТСКОЈ ЈЕЗИЧКОЈ ОБРАДИ КАО АЛАТКА ЗА СИСТЕМАТИЗАЦИЈУ РОМСКЕ ГРАМАТИКЕ У ОПИСУ И ФОРМАЛНОЈ НАСТАВИ

Р е з и м е

NooJ (http://www.nooj-association.org/) представља лингвистичко окружење које укључује алате за креирање и одржавање лексичких извора широког опсега, као и морфолошких и синтаксичких граматика и служи за парсирање корпуса у реалном времену. NooJ речници и граматике примењени су на текстове да би се лоцирали лексички, мофолошки и синтаксички модели и обележиле просте и сложене речи.

Овај рад има за циљ да покаже како NooJ модул за ромски језик ради кроз примере NooJ речника, морфологије, синтаксе и анотације, и како овај модул може допринети ситематизацији ромске граматике. Применом NooJ-а, чак и комплексне граматике могу бити формализоване јасно и једноставно. Затим се лингвистички ресурси, формализовани путем NooJ-а, могу користити за друге пројекте као што је то формална настава, посебно у случају дијалектолошки разноликих разреда.

Ромски модел покрива сва четири дијалекта ромског језика (супердијалекте О и Е, са и без фонетске мутације), односно, он је полилекталан да би формализовао целокупни ромски језик без фаворизовања или занемаривања неког од дијалеката. То може допринети дијалектолошким истраживањима, као и очувању богатства и диверзитета ромског језика и културе.

Друга карактеристика ромског модула јесте да конституише дијасистем, моделовањем целине од дијалекатских субсистема у једну граматику. Различити типови кореспонденције дијасинонима (дијалекатских еквивалената) објашњавају дијасистемско јединство ромског језика. У ствари, дијалекатски диверзитет ромског језика није препрека за узајамно разумевање међу добрим говорницима различитих дијалеката. Сличност између дијалеката оправдава нераздвојиво јединство ромског језика.

Кључне речи: ромски језик, NooJ, NLP (обрада природних језика), дијалектологија, дијасистем, лингвистика