Formalizing the Morphology of Rromani Adjectives

Masako Watabe, University of Franche-Comté, CRIT, UFR SLHS,
30-32 rue Mégevand, F-25000 Besançon, France, masako.watabe@univ-fcomte.fr
Max Silberztein, University of Franche-Comté, CRIT, UFR SLHS,
30-32 rue Mégevand, F-25000 Besançon, France, max.silberztein@univ-fcomte.fr

Abstract

This paper presents a set of linguistic resources that formalizes the morphological behavior of simple Rromani adjectives. We describe the formalization of the adjectives' morphology and the implementation with the NooJ linguistic platform of an electronic dictionary associated with a formal morpho-syntactic grammar. We can then apply this set of resources to a corpus to evaluate the resources and automatically annotate adjectival forms in Rromani texts. The final set of resources can then be used to identify each Rromani dialectal variant and can be used as a pedagogical tool to teach Rromani as a second language.

1 Introduction

1.1 Rromani language

Rromani is the language of the Rromani people; it is an Indo-Aryan language. The number of Rromani speakers is estimated at 5.5 million (Gurbetovski, M. et al. 2010). UNESCO's "Atlas of the World's Languages in Danger" classifies Rromani as a "definitely endangered[1]" language (UNESCO. 2010). There are four Rromani dialects, formed by two isoglosses combining with each other (Courthiade, M. 2016):

- The first isoglossal criterion concerns the opposition between "o" and "e," *e.g.*, *phirdom vs. phirdem* [I walked], *o Rroma vs. e Rroma* [the Rroms].

- The second isoglossal criterion concerns the phonetic mutation of two consonants: the alveolar affricates "ʧʰ" and "ʤ" transform into alveolar-palatal fricatives [ɕ] and [ʑ], *e.g.*, "ʧʰavo" *vs*. "ɕavo" [Rromani boy, son], "ʤukel" *vs*. "ʑukel" [dog].

These four dialects are not areal: Rromani speakers living in nearby regions do not necessarily speak the same dialects, and the same dialect is used in distant countries.

The Rromani alphabet was standardized at the International Rromani Union Congress in 1990, see Figure 1.

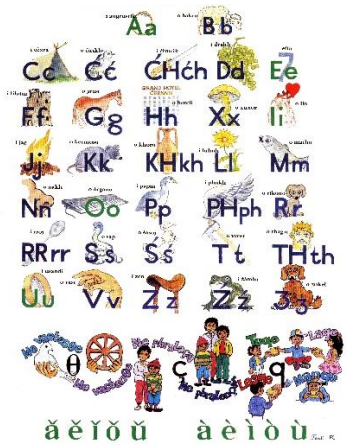

Figure 1: The Rromani standardized alphabet

If all Rromani speakers transcribe, for example, the word *ćhib* [language] using their local alphabets, there can be up to 60 different spellings. The written word *ćhib* is an underlying form including four possible pronunciations: [ʧʰb], [ʧʰp], [ɕib], and [ɕip], see Figure 2. The standardized alphabet enables speakers of different dialects to understand each other in writing, giving them comfort in pronunciation.

No other standardization exists: neither lexical, nor grammatical, nor phonetic.

[1] The UNESCO list has six categories of danger: Stable yet threatened, vulnerable, definitely endangered, severely endangered, critically endangered, extinct.

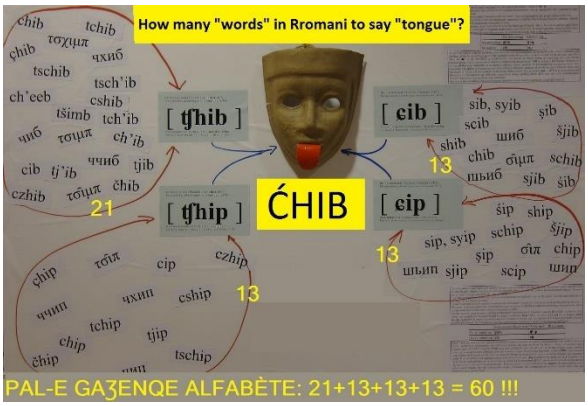

Figure 2: 60 different spellings of the word *ćhib* [language]

1.2 Natural Language Processing (NLP) applications in Rromani

Most of the NLP applications, such as Sketch Engine and DeepL, do not support Rromani. Very few NLP applications support Rromani, but always unsatisfactorily. For example, Rromani has been integrated into Google Translate in 2024. Translation quality is misleading at all levels: lexical, grammatical, orthographic, and dialectal. For example, translating "*thule bakră*" which means either "fat ewes" in the direct plural or "fat ewe" in the oblique singular, Google produces the single translation "fat goats[2]", see Figure 3.

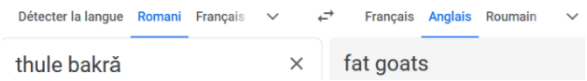

Figure 3: *thule bakră* translated by Google Translate (accessed September 12, 2024).

Conversely, Google Translate produces the translation "*thulo bakro*" for "fat sheep" correctly, whereas its translation for *fat ewe*: "*thuli bakro*" is incorrect at both lexical and grammatical levels, because the noun "ewe" should be translated as *bakri*, and there is a disagreement between the feminine adjective form *thuli* and the masculine noun *bakro*, see Figure 4.

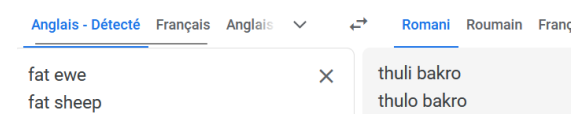

Figure 4: "fat ewe" and "fat sheep" translated by Google Translate (accessed October 2, 2024).

In addition, Google Translate incorrectly translates "my daughter, my daughters, the daughter, the daughters" as *miri ćhaj, mire čhaja, e shej, ćhaja* in Rromani. There are several problems at the grammatical, orthographic, and dialectal levels; the definite article in the plural (*e* or *le* depending on the dialect) is omitted, three different graphemes "ćh," "čh," and "sh" are confused, and two dialectal variants *ćhaj* and *ćhej*[3] are confused. Users, especially learners, would be lost.

Other NLP resources for Rromani are Russian Romani Corpus and ROMLEX. However, they do not adopt the standardized alphabet and do not clearly show the correspondences of dialectal variants, therefore, non-scientists and learners of Rromani cannot easily use them.

Facebook tries to process Rromani, but fails. When one posts a text in Rromani, Facebook will incorrectly recognize the source language (for example as French [4]), and then translate it incorrectly and partially (Watabe, M. 2024). For example, Facebook "translated" in French the Rromani text "Baxtalo 8 Aprilo savore Romenge, oven saste!" meaning [Happy April 8, for all the Rroms, may you be healthy!] but its translation "Baxtalo 8 avril Romenge au flaveur, saste de four!" is not a French text, see Figure 5.

Figure 5: A Rromani text incorrectly translated by Facebook (accessed April 8 2023).

2 Our project

We aim to describe the Rromani language by developing linguistic resources in the form of formalized dictionaries and grammar. Our initial dictionary, based on two small corpora; a two-page story written by a Rromani teacher (Duka, J.

[2] Rromani distinguishes lexically "ewe," "sheep," "she-goat," and "he-goat." Their equivalent words in Rromani are *bakri*, *bakro*, *buzni*, and *buzno*.

[3] This variant is transcribed as *shej* by Google Translate.

[4] French is the default language of the Facebook account of one of the authors, *i.e.*, Watabe, M.

MS.) and a one-page poem (Đurić, R. 2006), contained only 747 lexical entries associated with a well-developed morphological grammar that includes 179 inflectional paradigms and 11 derivational paradigms [5] for nouns, verbs, adjectives, and grammatical words. A feature of these resources is that they take into account Rromani four dialects, as well as a few vernaculars.

An editorial dictionary (Courthiade, M. et al. 2009) including the four dialects of Rromani explains Rromani morphology in the grammar section. It is our principal lexical and grammatical resource.

In this paper, we are addressing the problem of describing adjectives and their inflection, which causes massive ambiguities.

3 Rromani adjectives

The inflectional morphology of Rromani adjectives is governed by two genders (masculine and feminine), two numbers (singular and plural), and two cases (direct and oblique[6]). Adjectival forms are according to noun genders, numbers, and cases. The basic form of adjectives is the masculine singular direct. Combining these three properties produces eight possibilities; in practice, however, most adjectives have no more than three forms (Courthiade, M. et al. 2009. Sarău, G. 2009). Consequently, there are many inflectional homonyms.

Most Rromani words are oxytonic; *i.e.*, the tonic stress is on the last syllable. One uses a grave accent to mark the stress when it is not on the last syllable. For example, *bakri* [ewe] and *thulo* [fat, thick, dense] are oxytonic, whereas *profesòri* [professor] and *sociàlo* [social] are non-oxytonic. The opposition oxytonic *vs.* non-oxytonic plays a role in inflectional morphology.

3.1 Oxytonic adjectives

Oxytonic adjectives are classified into four classes: large adjectives, narrow adjectives, plural adjectives, and invariable adjectives.

Large adjectives[7]: Large adjectives are vocalic and have three suffixes: "-o" in the masculine singular direct, "-i" in the feminine singular direct, and "-e" in the plural direct for both genders, as well as oblique for both genders and numbers, see an example of the adjective *thulo* [fat, thick, dense] in Table 1:

Form	Gender	Number	Case
thulo	masculine	singular	direct
thuli	feminine	singular	direct
thule	masculine	plural	direct
thule	feminine	plural	direct
thule	masculine	singular	oblique
thule	feminine	singular	oblique
thule	masculine	plural	oblique
thule	feminine	plural	oblique

Table 1: Inflected forms and properties of the adjective *thulo* [fat, thick, dense]

The form *thule* is therefore 6-time ambiguous. This high level of ambiguity is general in Rromani; as a matter of fact, we do not know any Rromani adjective that would inflect to eight different forms, each for each combination of properties.

Narrow adjectives [8] : Narrow adjectives are consonant, and the direct forms of both genders are identical in each number: "-∅" in the direct singular and "-a" in the direct plural. The suffix of the oblique is "-e" for both genders and numbers, as in "large" adjectives, see an example of the adjective *godăver* [intelligent] in Table 2:

Form	Gender	Number	Case
godăver	masculine	singular	direct
godăver	feminine	singular	direct
godăvera	masculine	plural	direct
godăvera	feminine	plural	direct
godăvere	masculine	singular	oblique
godăvere	feminine	singular	oblique
godăvere	masculine	plural	oblique
godăvere	feminine	plural	oblique

Table 2: Inflected forms and properties of the adjective *godăver* [intelligent]

[5] Only the diminutive and abstract nouns with the suffix "-pen" are described in the current Rromani module.

[6] In Rromani, the direct case of human and most animal nouns is used as a subject, while the oblique case is used as an object complement. The direct case of inanimate object nouns is used as a subject and an object complement.

[7] The Rromani adjective *buxlo* [large] is the origin of this designation.

[8] The Rromani adjective *tang* [narrow] is the origin of this designation.

Plural adjectives: Plural adjectives are used specifically with plural nouns. In the direct case, they have the suffix "-∅" for both genders, and in the oblique "-e" for both genders, see an example of the adjective *but* [many, numerous] in Table 3:

Form	Gender	Number	Case
but	both	plural	direct
bute	both	plural	oblique

Table 3: Inflected forms and properties of the adjective *but* [many, numerous]

Invariable adjectives: So-called "international" adjectives have a tendency to be invariable. International oxytonic adjectives are completely invariable, see an example of the adjective *bordo* [Bordeaux-colored] in Table 4:

Form	Gender	Number	Case
bordo	both	both	both

Table 4: Invariable form and properties of the adjective *bordo* [Bordeaux-colored]

3.2 Non-oxytonic adjectives

Borrowed adjectives and suffixed adjectives: Inflectional paradigms of borrowed non-oxytonic adjectives and suffixed non-oxytonic adjectives are identical. Their suffix is "-o" in the direct singular for both genders, "-a" in the direct plural for both genders, and "-one" in the oblique for both genders and numbers. Oblique forms are oxytonic, therefore the stress is not marked, see an example of the adjective *veśìtko* [of the woods, wild] in Table 5:

Form	Gender	Number	Case
veśìtko	masculine	singular	direct
veśìtko	feminine	singular	direct
veśìtka	masculine	plural	direct
veśìtka	feminine	plural	direct
veśitkone	masculine	singular	oblique
veśitkone	feminine	singular	oblique
veśitkone	masculine	plural	oblique
veśitkone	feminine	plural	oblique

Table 5: Inflected forms and properties of the adjective *veśìtko* [of the woods, wild]

International adjectives: Compared to international oxytonic adjectives (*e.g., bordo* [Bordeaux-colored]), international non-oxytonic adjectives are not completely invariable. International non-oxytonic adjectives have two suffixes: "-o" in the direct for both genders and numbers and "-one" in the oblique for both genders and numbers. Oblique forms are oxytonic, therefore the stress is not marked, see an example of the adjective *sociàlo* [social] in Table 6:

Form	Gender	Number	Case
sociàlo	masculine	singular	direct
sociàlo	feminine	singular	direct
sociàlo	masculine	plural	direct
sociàlo	feminine	plural	direct
socialone	masculine	singular	oblique
socialone	feminine	singular	oblique
socialone	masculine	plural	oblique
socialone	feminine	plural	oblique

Table 6: Inflected forms and properties of the adjective *sociàlo* [social]

3.3 Conclusion

We have defined six classes of Rromani adjectives, according to their morphological properties:

- Oxytonic vocalic adjectives ending in "-o": *e.g., thulo* [fat, thick, dense], *buxlo* [large],
- Oxytonic consonant adjectives: *e.g., godăver* [intelligent], *tang* [narrow],
- Oxytonic consonant adjectives used only in the plural: *e.g., but* [many, numerous],
- Oxytonic vocalic international adjectives totally invariable: *e.g., bordo* [Bordeaux-colored], *pane* [breaded],
- Non-oxytonic borrowed adjectives and non-oxytonic suffixed adjectives: *e.g., zèleno* [green], *veśìtko* [of the woods, wild],
- Non-oxytonic vocalic international adjectives ending in "-o": *e.g., sociàlo* [social], *interesànto* [interesting].

4 Formalization of the Rromani vocabulary

4.1 The NooJ linguistic platform

NooJ is a linguistic development environment linguists use to describe natural languages, by constructing linguistic resources in the form of electronic dictionaries and formal grammars from the Chomsky-Schützenberger hierarchy: regular, context-free, context-sensitive, and unrestricted

grammars. NooJ can formalize twelve levels of linguistic phenomena, from the typographical to the semantic level (Silberztein, M. 2016).

To formalize the Rromani adjectives vocabulary, one needs to construct the following linguistic resources:

- a dictionary containing the affixes, simple words, compound words, and discontinuous expressions that make up the Rromani vocabulary of adjectives
- a grammar containing the description of adjectives inflectional paradigms

One could describe Rromani's four dialectal variants in the dictionary and morphological grammar. The following sections present these levels of description.

4.2 Electronic dictionary

For example, the adjective *thulo* is represented by the following lexical entry in NooJ formalized (aka electronic) dictionary:

thulo,ADJ+EN="fat, thick, dense"+FLX=BUXLO +DRV=ĆAĆIPEN:SASTIPEN

In this extract, each lexical entry is composed of a lemma, its category "ADJ" (adjective), its English translation "+EN," and the name of its inflectional paradigm "+FLX".

The lexical entry *thulo* is associated with derivational paradigm ĆAĆIPEN and its derivative's inflectional paradigm SASTIPEN. ĆAĆIPEN describes the derivation of abstract nouns with the suffix "-pen," which applies to words of various categories, such as adjectives, nouns, and verbs.

The four main dialects are encoded using the following double codes:

- O-bi dialect: rro+rrbi
- O-mu dialect: rro+rrmu
- E-bi dialect: rre+rrbi
- E-mu dialect: rre+rrmu

In addition, we have added a third code to label specific language variants. For example, the northern speech used in Russia and Poland is defined by the extra label: "+rrn." This is the case of the entry *veśìtko* [of the woods, wild].

veśitko,ADJ+rro+rrbi+rrn+EN="of the woods, wild"+FLX=VEŚÌTKO+SYN= "veśutno"

The entry above shows that the adjective *veśìtko* belongs to the dialects O-bi (+rro+rrbi), and is used specifically in Russia and Poland (+rrn), its English translation is "of the woods, wild" (+EN="of the woods, wild"), it is inflected according to the paradigm named VEŚÌTKO (+FLX=VEŚÌTKO [9]), and it has the synonym "veśutno" (+SYN=veśutno) used in most Rromani dialects except the vernacular in Russia and Poland.

4.3 NooJ morphological grammar

In NooJ, inflectional paradigms are represented by regular or context-free grammars built over suffix/property factors: suffixes are added to the lexical entry to construct forms, which are associated with the corresponding properties (Silberztein, M. 2003-). For example, the following is the grammar rule that describes the inflectional paradigm RROM:

RROM = <E>/sg+dr | a/pl+dr | es/sg+ob | en/pl+ob;

This rule states that if one adds the empty string (<E>) to the lexical entry, one produces a singular (+sg) direct (+dr) form; if one adds an "a" to the lexical entry, one produces a plural (+pl) direct (+dr) form; if one adds "es" to the lexical entry, one produces a singular (+sg) oblique (+ob) form; if one adds "en" to the lexical entry, one produces a plural (+pl) oblique (+ob) form.

Suffixes may contain stack operators. For instance, operator <B> (for "Backspace") is used to delete the current letter. In the following paradigm:

BUXLO = <E>/m+sg+dr | <B>i/f+sg+dr | <B>e/m+pl+dr | <B>e/f+pl+dr | <B>e/m+sg+ob | <B>e/f+sg+ob | <B>e/m+pl+ob | <B>e/f+pl+ob | <B>:CMP/comparative ;

The second term states that if one deletes the last letter of a lexical entry and then adds an "i" (suffix

[9] The lexical entry (*i.e., veśìtko*) is in lower case, whereas the paradigm name (*i.e., VEŚÌTKO*) is in upper case. It prevents confusion between two distinct values represented by identical writing.

<B>i), one produces the feminine, singular, direct form (f+sg+dr) of the lexical entry.

For example, when the paradigm BUXLO is applied to the lexical entry *thulo* [fat, thick, dense], there will be no change (<E>) to the direct masculine singular, one final letter will be deleted (<B>) and an "i" will be added to the direct feminine singular, and one final letter will be deleted and an "e" will be added to produce the direct plural forms in both genders, and to the oblique forms in both genders and numbers. This paradigm then represents the three forms of *thulo*:

- *thulo*: masculine singular direct
- *thuli*: feminine singular direct
- *thule*: masculine plural direct, feminine plural direct, masculine singular oblique, feminine singular oblique, masculine plural oblique, or feminine plural oblique

It means that the wordform *thule* is associated with six potential linguistic analyses.
The last term of the BUXLO paradigm is used to produce the comparative forms of the lexical entries. :CMP refers to the name of the following rule (similarly to auxiliary symbols in generative context-free grammars):

CMP = eder/m+sg+dr | eder/f+sg+dr | edera/m+pl+dr | edera/f+pl+dr | edere/m+sg+ob | edere/f+sg+ob | edere/m+pl+ob | edere/f+pl+ob ;

The comparative suffix "-eder" is added in place of the final letter "o" in the *thulo* lexical entry (see "<B>:CMP" in the paradigm BUXLO above). The comparative is declined like the narrow adjectives: without suffix in the direct singular of both genders, "-a" in the direct plural of both genders, and "-e" in the oblique of both genders and numbers. This rule produces three forms: *thuleder*, *thuledera*, and *thuledere* for eight linguistic analyses.

Beside a dozen generic operators such as <B> that are available for any language, NooJ offers linguists the possibility of creating specific operators for each language. For instance, the Spanish operator <Á> is used to add an acute accent to the current vowel; the Hebrew operator <F> is used to de-finalize the last consonant of a word; the Tamazight operator <T> replaces letter "ḍ" with "ṭ", etc. For the Rromani language, we have implemented two specific operators:

- <A> deletes a grave accent, regardless the position and returns to the initial position,
- <À> adds a grave accent to the current letter.

For example, operator <A> is used in the following paradigms:

VEŚÌTKO = <E>/m+sg+dr | <E>/f+sg+dr | <B>a/m+pl+dr | <B>a/f+pl+dr | <A>ne/m+sg+ob | <A>ne/f+sg+ob | <A>ne/m+pl+ob | <A>ne/f+pl+ob ;

In the VEŚÌTKO paradigm, the fifth term states that if one removes the grave accent of the lexical entry, and then adds the suffix "ne" (<A>ne), one produces the masculine singular oblique form (+m+sg+ob) of the lexical entry. The operator <A> is typically used in paradigms associated with non-oxytonic words. For example, if the paradigm VEŚÌTKO is applied to the lexical entry *zèleno* [green], its oblique form will be *zelenone*, *i.e.*, without the grave accent.

By applying all inflectional NooJ paradigms to the dictionary, NooJ produces all the inflected forms for each lexical entry automatically. When applying these resources to a text, NooJ annotates all recognized word forms. For example, the wordform *veśitkone* will be annotated as an adjective (ADJ), its basic form is *veśìtko*, its inflectional value is the oblique (ob), its dialect value is a Northern vernacular belonging to the O-bi dialect (rro+rrbi+rrn), its synonym in other dialects is *veśutno*. That helps users pedagogically recognize the relationship between the basic form, an inflected form, and a dialectal variant. However, there are four sets of annotations because oblique forms are identical for both genders and numbers, see Figure 6. We need syntactic grammar to resolve ambiguities.

It is often better for pedagogical applications, to use NooJ graphical grammars to describe some phenomena. For instance, the paradigm BUXLO shown above can equally be described with the following graph, see Figure 7.

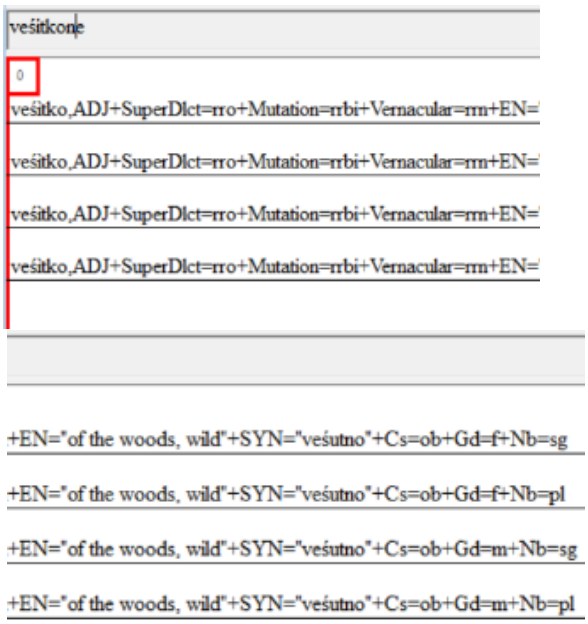

Figure 6: Inflected form *veśitkone* annotated automatically

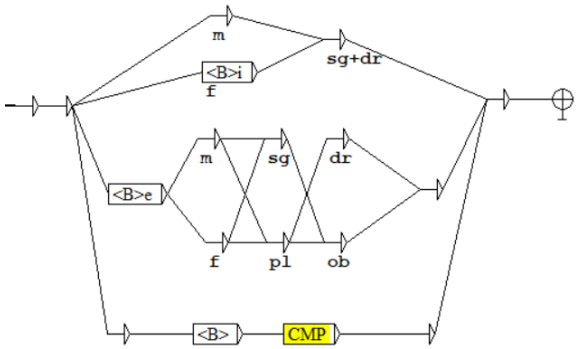

Figure 7: Inflectional grammar for adjective paradigm BUXLO.

The yellow node CMP refers to the embedded graph that represents the paradigm of the same name, see Figure 8.

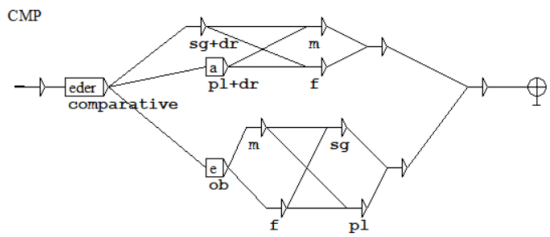

Figure 8: Inflectional grammar of the comparative suffix CMP.

The "DRV" property allows us to describe derivations. For example, the lexical entry *thulo* mentioned above:

thulo,ADJ+EN="fat, thick, dense"+FLX=BUXLO
+DRV=ĆAĆIPEN:SASTIPEN

states that the word *thulo* can be derived according to paradigm ĆAĆIPEN, and the resulting derived form can be inflected according to inflectional paradigm SASTIPEN. Following is the definition of the derivational paradigm ĆAĆIPEN:

ĆAĆIPEN = <B><A>ipen/N+ina+m+abstract ;

From the lexical entry *thulo*, one deletes the final letter (<B>), removes the accent (<A>[10]), and then adds suffix "ipen." The resulting derivative *thulipen* [fatness, thickness, density], is a masculine inanimate noun (N+ina+m), that belongs to the semantic class "abstract."

The abstract noun *thulipen* is a generic masculine singular direct form, and has seven dialectal variants: *thulipe* used in the two dialects O-bi and E-bi belonging to a dialectal subgroup "without mutation (rrbi)," *thulipo* used in the O-mu dialect (rro+rrmu), *thulimos* and *thulimo* used in the E-mu dialect (rre+rrmu), *thuliben*, *thulibe*, and *thulibo* used in the Carpathian vernacular belonging to the O-bi dialect (rro+rrbi+rrc), all of which mean "fatness, thickness, density."

All these variants share the same inflectional morphology; *thulimàta* is the masculine plural direct form, *thulimas-* and *thulipnas-*[11] are the masculine singular oblique forms, and *thulimaten-* is the masculine plural oblique form. The following grammar rule describes the inflectional forms of the derivative as well as all its dialectal variants:

SASTIPEN = <E>/sg+dr | <B3>màta/pl+dr |
<B3>(mas|pnas)<E>/sg+ob |
<B3>maten<E>/pl+ob |
<B>/sg+dr+rrbi | <B2>o/sg+dr+rro+rrmu |
<B3>(mos|mo)<E>/sg+dr+rre+rrmu|
<B3>(ben|be|bo)<E>/sg+dr+rro+rrbi+rrc ;

NooJ automatically produces the derived forms for each lexical entry, and annotates all recognized derived forms in texts. For example,

[10] The <A> operator concerns only non-oxytonic words. Thus, it produces no change for the lexical entry *rromano*.

[11] This is an archaic form.

thulipe will be annotated as an abstract masculine inanimate noun (N+m+ina+abstract), its inflectional values are the singular direct (sg+dr), its dialect value is the "without mutation (rrbi)," and its initial category is the adjective meaning "fat, thick, dense." That helps users pedagogically recognize the relationship between the base word (*e.g.*, the adjective *thulo* [fat, thick, dense]) and its derivative (*e.g.*, the noun *thulipe* [fatness, thickness, density]), see Figure 9.

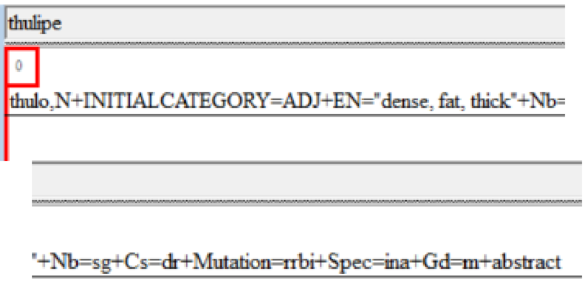

Figure 9: Derived form *thulipe* annotated automatically

4.4 Automatic Natural Language Processing

NooJ can use the same linguistic resources both to parse and generate texts. For example, one can apply a dictionary and its corresponding inflectional grammar to automatically produce all the forms associated with each lexical entry of the dictionary, see an extract in Figure 10.

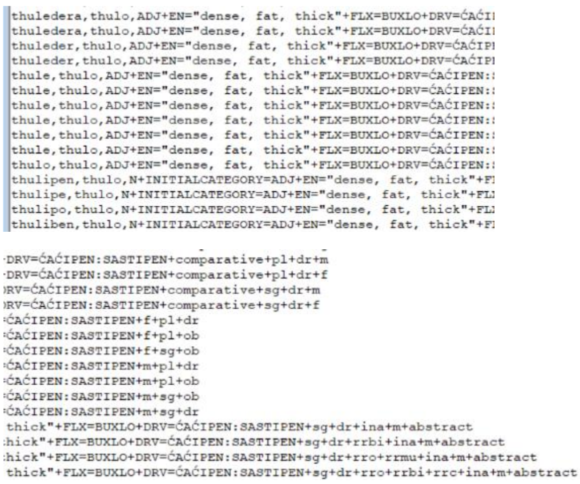

Figure 10: Inflected and derived forms of the adjective *thulo* [fat, thick, dense] generated automatically.

NooJ uses the same resources to parse texts, lemmatize and annotate their wordforms, to apply queries in the form of regular, context-free, context-sensitive, or unrestricted grammars, perform statistical analyses, compute semantic analyses in Predicative or XML format, translations (if accessing multilingual dictionaries), etc. For example, Figure 11 displays the query "<but>", which stands for "all inflected and derived forms of lexical entry *but* [many, numerous]" and the correspondence concordance computed by NooJ.

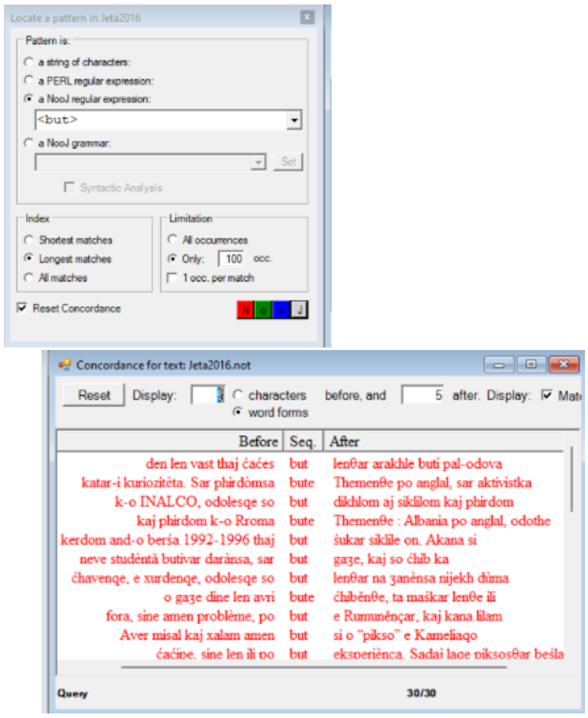

Figure 11: The query "<but>" and its resulting concordance.

The wordform *but* corresponds to three semantic values: the adjective "many, numerous," *e.g.*, *phirdòmsa bute Themenθe* [I traveled to many countries], the adverb "a lot, much," *e.g.*, *but dikhlom* [I have seen a lot], and another adverb "very," *e.g.*, *but śukar siklile on* [They learned very well].

Syntactic grammar is underdeveloped in the current NooJ module for Rromani, this is why wordforms remain ambiguous in general.

However, the query "<but,ADV>" will not retrieve the adjectival inflected forms *bute* because adverbs are invariable, see Figure 12.

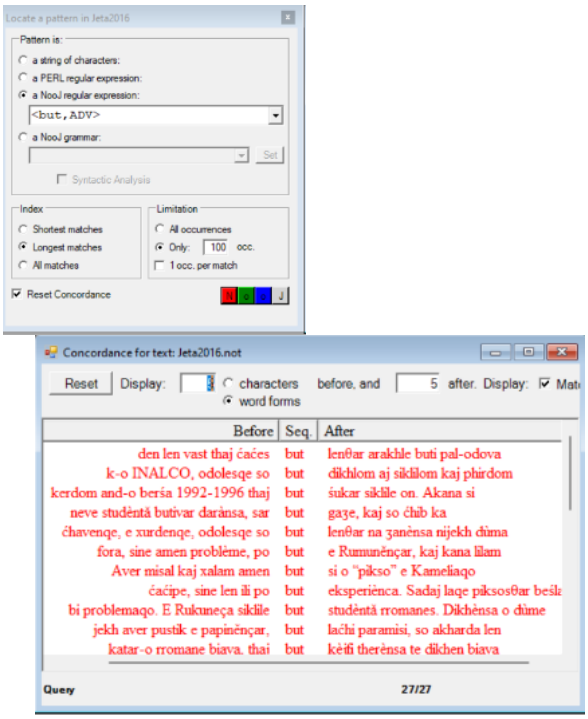

Figure 12: The query "<but,ADV>" and its resulting concordance.

If NooJ recognizes ambiguous forms, NooJ will show the annotation of all of them and not choose one randomly, as often happens in empirical applications. For example, as mentioned above, the adjectival inflected form *thule* is ambiguous because of six inflectional homonyms and the nominal inflected form *bakră* is ambiguous because of two inflectional homonyms, see Figure 13.

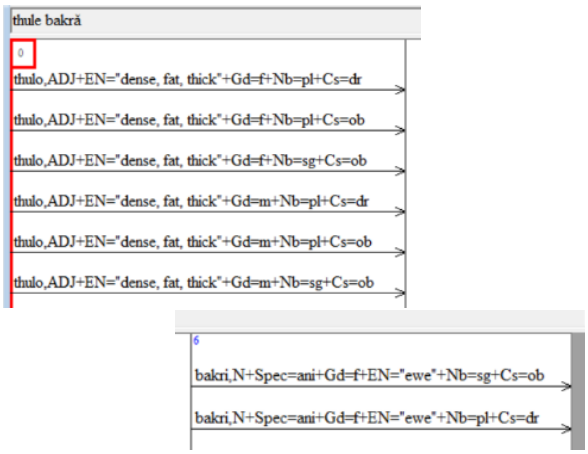

Figure 13: Inflected forms *thule* and *bakră* annotated automatically with ambiguity

4.5 Evaluation

There are 88 adjectives out of 747 lexical entries in the current NooJ dictionary for Rromani. Applying these lexical entries and their corresponding inflectional grammars generated 1,278 inflected and derivational forms.

In our corpus, all wordforms that might correspond to potential adjectives have been recognized and annotated correctly, *i.e.*, we have reached a 100% recall, which is expected as we are specifically constructing our linguistic resources from the corpus. However, without any syntactic grammar, wordforms that might function as adjectives or as adverbs (*e.g.,* the wordform *but*) and be associated with different properties (*e.g.*, the wordforms *thule* and *bakră*) remain ambiguous until we can apply a syntactic grammar.

5 Conclusion, perspective

The current Rromani module recognizes all 170 adjectival forms from a small corpus that contains 708 wordforms. We are currently importing around 4,500 lexical entries from an editorial dictionary (Courthiade, M. et al. 2009) into a formalized NooJ format.

Removing ambiguities is our current challenge. We are constructing syntactic local grammars to disambiguate frequent adjectives.

The resulting linguistic resources will be downloadable from the NooJ website. The NooJ dictionary for Rromani will use the standard Rromani alphabet and include dialectal variants at the lexical and morphological levels. It will be available as a new digital and linguistic tool for all speakers of Rromani: native speakers and learners, regardless of their dialects.

We believe this polylectal dictionary is valuable from a dialectological point of view. Furthermore, as the declaration of the first Congress of the International Rromani Union in 1971 stated that "no dialect is better than another," the dictionary will describe all dialects.

Unlike empirical methods, the NooJ platform produces analyses based on handcrafted linguistic resources, and thus offers linguists to describe and understand its properties. We believe that carefully and precisely handcrafting linguistic resources for Rromani is a worth scientific project, and will have many applications in Natural Language Processing, second-language teaching and corpus linguistics.

References

Gheorghe Sarău. 2009. *Struktùre rromane ćhibăqe*. Editura Universității din București, Bucharest.

Jeta Duka. Deś berś vaś-i rromani ćhib and-o INALCO. MS.

Marcel Courthiade. Structure dialectale de la langue rromani. Études tsiganes, 22-2005, pages 14-26. Le Centre de documentation, Paris.

Marcel Courthiade. 2016. The nominal flexion in Rromani. In Marcel Courthiade and Delia Grigore (eds.) *Professor Gheorghe Sarău: a life devoted to the Rromani language*. pages 157-211. Editura Universității din București, Bucharest.

Marcel Courthiade et al. 2009. *Morri angluni rromane ćhibăqi evroputni lavustik*. Romano Kher, Budapest.

Masako Watabe. 2024. A polylectal linguistic resource for Rromani. In Max Silberztein. (ed.) *Linguistic Resources for Natural Language Processing: On the Necessity of Using Linguistic Methods to Develop NLP Software*. pages 147-172. Springer, Cham.

Max Silberztein. 2003-. NooJ manual. https://nooj.univ-fcomte.fr

Max Silberztein. 2016. *Formalizing Natural Languages: the NooJ approach*. Wiley Ed.: Hoboken NJ.

Medo Gurbetovski, Mozes Heinschink, and Daniel Krasa. 2010. *Guide de conversation rromani de poche*. ASSIMIL, Paris.

Rajko Đurić. 2006. E rromani ćhib. In Marcel Courthiade (ed.) *La littérature des Rroms, Sintés et Kalés*. pages 67-68. INALCO, Paris.

2010. *Atlas of the World's Languages in Danger*. UNESCO, Paris.

Facebook. https://www.facebook.com/

Google Translate. https://translate.google.com/

NooJ platform. https://nooj.univ-fcomte.fr/

ROMLEX. http://romani.uni-graz.at/romlex/

Russian Romani Corpus. http://web-corpora.net/RomaniCorpus/search/

La langue romani - un atout pour l'éducation et la diversité (exhibition). 2014. Council of Europe, Strasbourg.