Formalizing the Morphology of Rromani Adjectives

Masako WATABE
masako.watabe@univ-fcomte.fr
Université de Franche-Comté, ED592 LECLA, UA3224 C.R.I.T.

Max SIBERZTEIN
max.silberztein@univ-fcomte.fr
Université de Franche-Comté, ED592 LECLA, UA3224 C.R.I.T.

ComputEL-8 Workshop, 4-5 March 2025, Hawaii, USA (Hybrid).

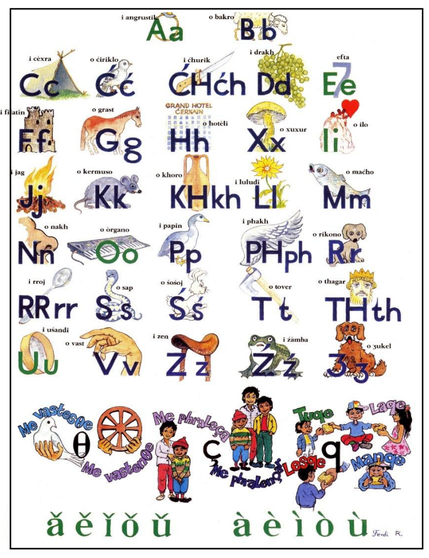
1-Our project

We aim to describe the Rromani language by developing linguistic resources in the form of formalized dictionaries and grammar; 1) common to all speakers, regardless of dialect, 2) consistent but easy to use for all users, and 3) accessible online.

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

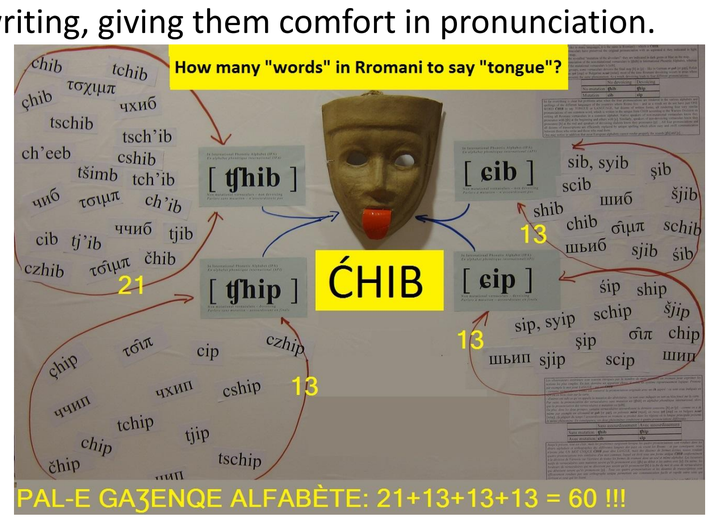
2-Rromani Language

- It is the language of the Rromani people.
- It is a "definitely endangered" language according to UNESCO's "Atlas of the World's Languages in Danger."
- There are about 5.5 million of Rromani speakers among 13-15 million population (Gurbetovski, M. et al. 2010).
- The four Rromani dialects are not areal.
- The Rromani alphabet was standardized at the International Rromani Union Congress in 1990.



Rromani standardized Alphabet

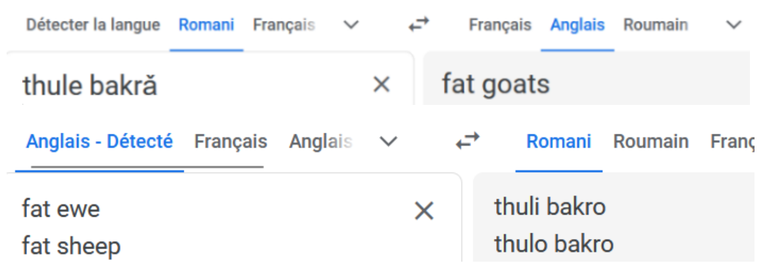
If all Rromani speakers transcribe, for example, the word *ćhib* [language] using their local alphabets, there can be up to 60 different spellings. The written word *ćhib* is an underlying form including four possible pronunciations: [ʧʰb], [ʧʰp], [ɕib], and [ɕip]. The standardized alphabet enables speakers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



60 possible spellings of the word *ćhib* [language, tongue]

3-NLP applications in Rromani

Very few NLP applications support Rromani, but always unsatisfactorily. Rromani has been integrated into Google Translate in 2024. Its translation quality is misleading at all levels: lexical, grammatical, orthographic, and dialectal.



Google Translate making errors (accessed October 2, 2024)

"Russian Romani Corpus" and "ROMLEX" developed by linguists do not adopt the standardized alphabet and do not clearly show the correspondences of dialectal variants. Non-scientists and learners of Rromani cannot easily use them.

4-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). Adjectival forms are according to noun genders, numbers, and cases. Combining these three properties produces eight possibilities. However, most adjectives have no more than three forms (Courthiade, M. et al. 2009. Sarău, G. 2009). Consequently, there are many inflectional homonyms. For example, an inflected form *thule* of the adjective *thulo* [fat, thick, dense] is 6-time ambiguous.

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

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

5-The NooJ platform

NooJ < https://nooj.univ-fcomte.fr/ > is a linguistic development environment linguists use to describe natural languages, by constructing linguistic resources in the form of electronic dictionaries and formal grammars.

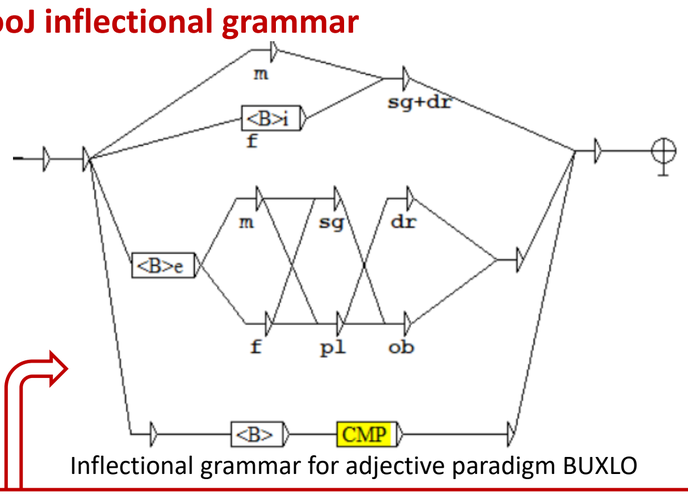
- a dictionary containing the affixes, simple words, compound words, and discontinuous expressions.
- a grammar containing the description of inflectional and derivational paradigms.

NooJ Electronic Dictionary

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

Each lexical entry is composed of a lemma, its category "ADJ" (adjective), its English translation "+EN," its inflectional paradigm name "+FLX", its derivational paradigm name "+DRV," and its derivative's inflectional paradigm name following a colon. The derivational paradigm ĆAĆIPEN describes the derivation of abstract nouns with the suffix "-pen."

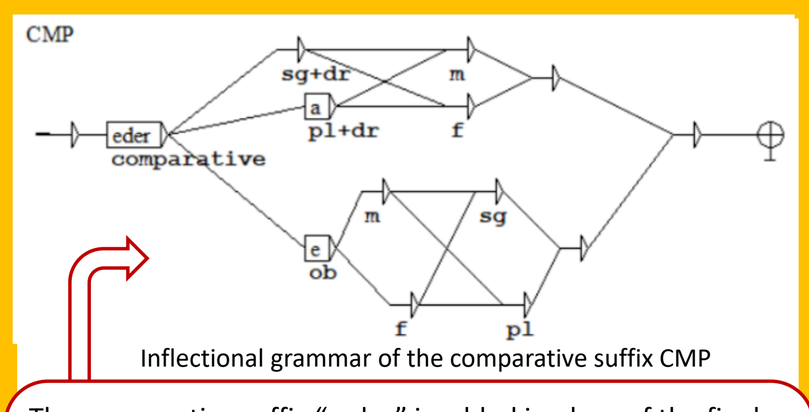
NooJ inflectional grammar



Inflectional grammar for adjective paradigm BUXLO

The BUXLO paradigm states that if one does not add anything to a lexical entry, one produces a masculine, singular, direct (m+sg+dr) form; if one deletes (<B> for "Backspace") the last letter of a lexical entry and adds the suffix "i," one produces the feminine, singular, direct form (f+sg+dr); and if one deletes the last letter of a lexical entry and adds the suffix "e," one produces the plural, direct form (pl+dr) in both genders and the oblique form (ob) in both genders and numbers.

The last line of the BUXLO paradigm produces the comparative forms of lexical entries. The embedded paradigm name (CMP) is highlighted in yellow.



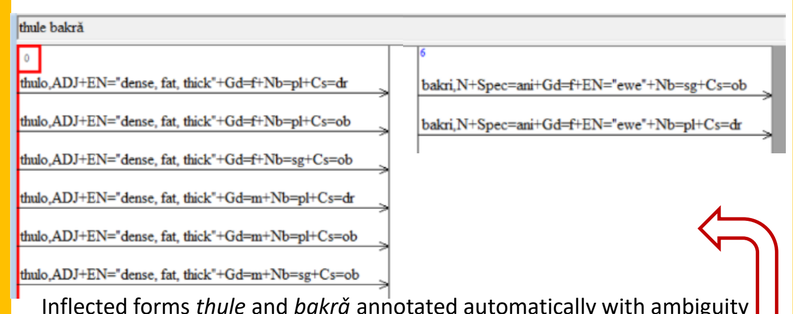
Inflectional grammar of the comparative suffix CMP

The comparative suffix "-eder" is added in place of the final letter (<B> in the BUXLO paradigm) of a lexical entry. The suffix "-eder" is declined as an adjective: 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.

When the BUXLO paradigm is applied to the lexical entry *thulo* [fat, thick, dense], NooJ produces automatically all inflected forms: *thulo, thuli*, and *thule* associated with eight different properties. Then, when the CMP paradigm is applied to the lexical entry *thulo*, NooJ produces automatically all inflected comparative forms: *thuleder, thuledera*, and *thuledere* associated with eight different properties.

Automatic Natural Language Processing

All inflected forms are stocked in NooJ as linguistic 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.



Inflected forms *thule* and *bakră* annotated automatically with ambiguity

If NooJ recognizes ambiguous forms, NooJ will show the annotation of all of them and not choose one only based on linguistic certainty, as opposed to statistical likehood. For example, 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.

6-Conclusion, perspective

Our initial dictionary based on two small corpora is associated with a well-developed morphological grammar including 179 inflectional paradigms and 11 derivational paradigms for nouns, verbs, adjectives, and grammatical words.

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 standardized 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.

References

Gheorghe Sarău. 2009. *Struktùre rromane ćhibâqe*. Editura Universităţii din Bucureşti, Bucharest.

Jeta Duka. Deś berš 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 Universităţii din Bucureşti, Bucharest.

Marcel Courthiade et al. 2009. *Morri angluni rromane ćhibâqi 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.