Une langue peu dotée dans l'horizon numérique : le développement d'une ressource linguistique pour le rromani sur la plateforme NooJ

Masako WATABE
masako.watabe@univ-fcomte.fr
Université de Franche-Comté, ED592 LECLA, UA3224 C.R.I.T.
Journée d'études TEDonnées, le 12 avril 2024, MSHE, Besançon, France

1-Langue rromani

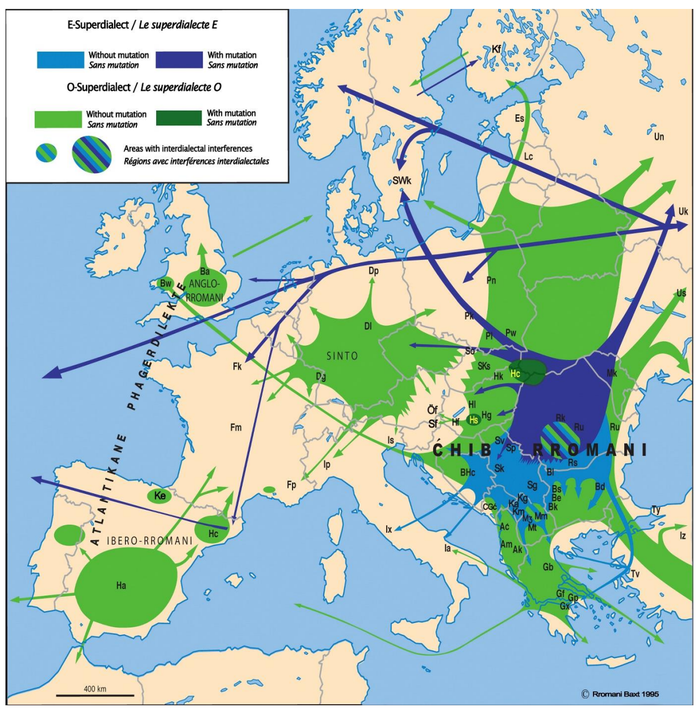
La langue des Rroms. La population est estimée 10-12 millions en Europe et 3 millions d'outre Atlantique, dont 6-8 millions de locuteurs.

Une langue indo-aryenne. Le lexique et la grammaire sont similaires à des langues de l'Inde du Nord. Le lexique est principalement constitué de racines indiennes, grecques et persanes.

Alphabet rromani. Standardisé lors du Congrès de l'Union Rromani Internationale en 1990.

Alphabet rromani standardisé

Les quatre dialectes. Deux isoglosses non aréales et croisées forment les quatre dialectes (O-bi, O-mu, E-bi et E-mu) du rromani. La distance géographique ne correspond pas à la distinction linguistique entre les dialectes du rromani.



Structure dialectale du rromani en Europe

2-Contexte

Le rromani dans l'horizon numérique. Les Rroms ne bénéficient pas pleinement de nouvelles technologies adaptées à leur langue. Comment les Rroms qui vivent dans de différents pays et parlent de différents dialectes pourraient-ils se communiquer facilement de façon numérique ?

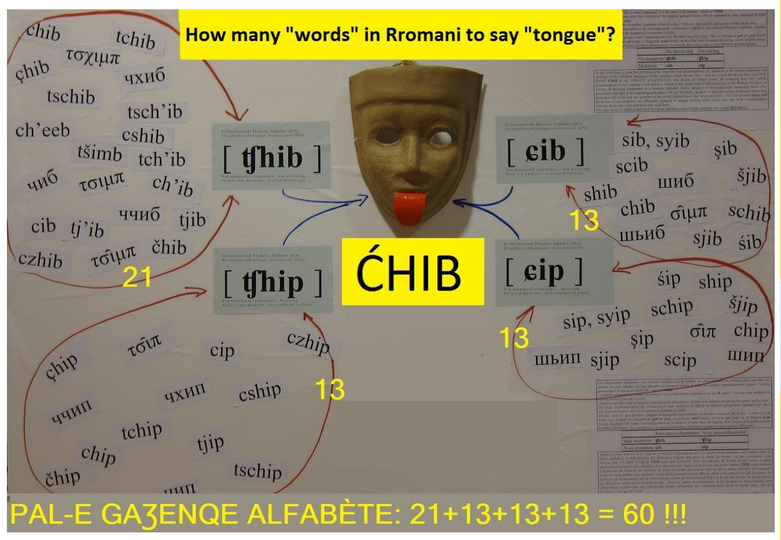
Référence
[1] Courthiade M. Structure dialectale de la langue rromani, in *études tsiganes*, No. 22, le Centre de documentation, Paris, 2005.
[2] Courthiade M. The nominal flexion in Rromani, in *Professor Gheorghe Sarău: a life devoted to the Rromani language*. Editura Universității din București, Bucarest, 2016.
[3] Courthiade, M. et al. *Morri angluni rromane čhibăqi evroputni lavustik* [Mon premier dictionnaire européen de la langue rromani]. Romano Kher, Budapest, 2009.
[4] Gurbetovski, M. et al. *Guide de conversation rromani de poche*, ASSIMIL, Paris, 2010.

Applications existantes.
Facebook ne traite pas le rromani. Pourtant, lorsqu'un usager publie quelque chose en rromani, Facebook *reconnaîtra* automatiquement une langue source (qui n'est pas le rromani), et *traduira* en langue cible ! C'est un problème typique des méthodes empiriques lorsqu'elles n'ont pas accès à des corpus suffisamment importants.

Un texte en rromani et sa *traduction* en français sur Facebook

Rusian Romani Corpus et *ROMLEX* sont des bases de données développées par des linguistes, utiles à des fins de recherche, en revanche, ils n'adoptent pas l'alphabet standardisé et ne permettent pas d'analyser et d'annoter de nouveaux textes.

Caractéristique et problématique.
Si tous les locuteurs transcrivent en utilisant les alphabets locaux, il y aura jusqu'à 60 orthographes différentes ! L'alphabet standardisé permet aux locuteurs de différents dialectes de se comprendre à l'écrit en leur donnant un confort de prononciation.



60 orthographes possibles du mot *čhib* [langue]

Nous développons une ressource linguistique numérique pour le rromani 1) unique à tous les locuteurs, quels que soient les dialectes, 2) consistante mais simple à utiliser même pour les usagers non scientifiques, et 3) accessible en ligne.

3-Méthodologie

Plateforme NooJ est un environnement linguistique pour formaliser les langues naturelles. NooJ est prêt à s'adapter à toutes les langues écrites. Le système NooJ et ses modules de diverses langues sont téléchargeables en ligne.

Caractéristiques innovantes. Un module unique et commun incluant les quatre dialectes du rromani. Chaque dialecte est défini par la combinaison de deux étiquettes dialectales dont chacune représente une isoglosse.

	O « rro »	E « rre »
Sans mutation phonétique « rrbi »	Dialecte O-bi « rro+rrbi »	Dialecte E-bi « rre+rrbi »
Avec mutation phonétique « rrmu »	Dialecte O-mu « rro+rrmu »	Dialecte E-mu « rre+rrmu »

Les quatre dialectes et leurs étiquettes dans le module NooJ

[5] Silberztein, M. *La formalisation des langues: l'approche de NooJ*. ISTE Eds., Londres, 2015.
[6] Watabe, M. A polylectal linguistic resource for Rromani, in : Silberztein, M. (édits.), *Linguistic Resources for Natural Language Processing - On the Necessity of Using Linguistic Methods to Develop NLP Software*. Springer, Cham, 2024.

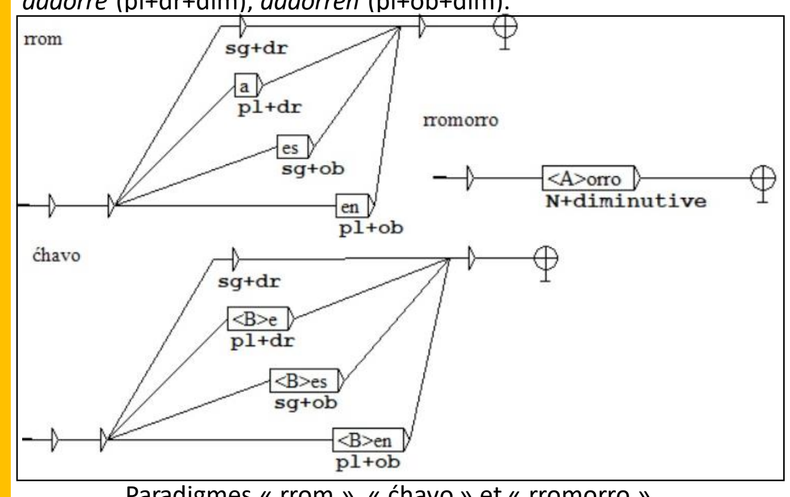
4-Application

Dictionnaire NooJ pour le rromani.

```
dad,N+hum+m+EN="father"+FLX=rrom+DRV=rromorro:ćhavo
daj,N+hum+f+rro+EN="mother"+RRE="dej"+FLX=ćhaj
    +DRV=ćhajorri:rromni
dej,N+hum+f+rre+EN="mother"+RRO="daj"+FLX=ćhej
    +DRV=ćhajorri:rromni
ça,PSTP+ins+EN="with"
```

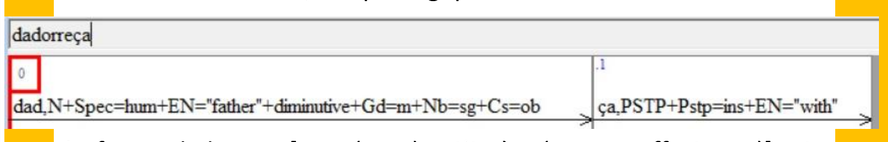
Chaque ligne commence par un mot d'entrée. Par exemple, *dad* est un nom masculin humain (N+hum+m) dont la traduction en anglais (EN) est *father*, le paradigme flexionnel (FLX) s'appelle « rrom », le paradigme dérivationnel (DRV) du diminutif s'appelle « rromorro » et le paradigme flexionnel du diminutif s'appelle « ćhavo ».
Il y a deux types de propriétés dialectales. Par exemple, *daj* est une variante en superdialecte O (rro) et son équivalente en superdialecte E (RRE) est *dej*.

Morphologie flexionnelle.
Lorsque des grammaires morphologiques (de flexion et de dérivation) sont appliquées à une entrée de dictionnaire, NooJ donnera automatiquement toutes les formes fléchies.
L'entrée *dad* [père] et ses paradigmes « rrom », « rromorro » et « ćhavo » donnera huit formes : *dad* (sg+dr), *dades* (sg+ob), *dada* (pl+dr), *daden* (pl+ob) *dadorro* (sg+dr+dim), *dadorres* (sg+ob+dim), *dadorre* (pl+dr+dim), *dadorren* (pl+ob+dim).

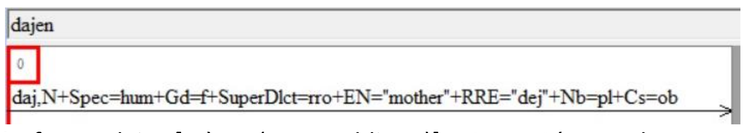


Paradigmes « rrom », « ćhavo » et « rromorro »

Annotation.
Lorsque le dictionnaire NooJ des formes fléchies pour le rromani est appliqué à un texte, NooJ donnera l'annotation avec les informations lexicales, morphologiques et dialectales.



La forme *dadorreça* [avec (mon) petit père (au sens affectueux)] est bien reconnue comme deux unités linguistiques (*dadorres* [petit père] et *ça* [avec]) alors que le « s » final de *dadorres* est contracté.



La forme *dajen* [mères (au cas oblique)] est annotée avec la propriété dialectale « rro » et la variante dialectale *dej*.

5-Conclusion et perspective

Le module NooJ pour le rromani serait un outil pragmatique sur deux aspects (dialectologique et didactique) pour tous les locuteurs du rromani ; quels que soient les dialectes, les locuteurs natifs et les apprenants.
Nous envisageons la diffusion de nos travaux 1) sur le site web NooJ, 2) au sein de communautés rrom à l'occasion de la journée mondiale des Rroms (le 8 avril) et la journée mondiale de la langue rromani (le 5 novembre), 3) aux étudiants et aux enseignants du rromani à l'occasion de séminaires à l'INALCO et de l'école d'été du rromani en Roumanie, et 4) par l'intégration du module NooJ dans un autre système, tel que « R.E.D.-RROM » (un site web autodidactique de la langue et la culture rromani).

[7] Exposition, *La langue romani - un atout pour l'éducation et la diversité*. Conseil de l'Europe, Strasbourg, 2014.
[8] Facebook. https://www.facebook.com/
[9] Plateforme NooJ. http://www.nooj4nlp.org.
[10] R.E.D.-RROM (Restoring the European Dimension of Rromani Language and Culture). http://www.red-rrom.com/home.page
[11] ROMLEX. http://romani.uni-graz.at/romlex/.
[12] Russian Romani Corpus. http://web-corpora.net/RomaniCorpus/search/?interface_language=en.