CoSwID, a Code Switching Identification Method Suitable for Under-Resourced Languages

Laurent KEVERS – University of Corsica (France)

SIGUL 2022 Workshop @ LREC 2022 Marseille

Saturday, June 25, 2022

Let me introduce ...

The "Banque de Données Langue Corse" (BDLC) project

Corsican Language Database: https://bdlc.univ-corse.fr/

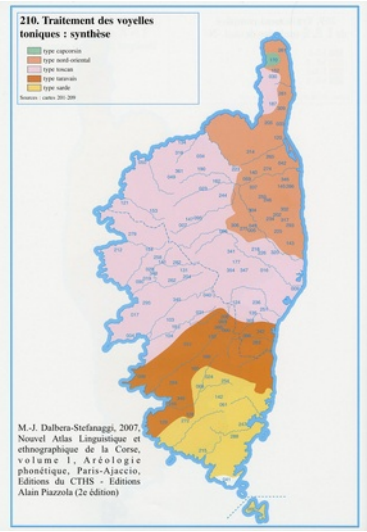
Collects linguistic data on know-how and cultural traditions all over Corsica

Illustrate and study the linguistic variation

Oral interviews (field surveys) with native speakers

- Lexical data
- Ethnotexts: mainly in Corsican but code switching with French occurs

From 2019, starting NLP for Corsican



The Challenge of building and using corpora for under-resourced langages (UL)

Language Identification (LgID) is an important tool to build uniform and high quality corpora with regard to the language of the documents

→ LID, a classical NLP component with high and satisfactory results

BUT: ... sometimes a little less – to much less – good for UL !

→ It is possible to train a document-level LID module with UL support given some raw initial corpora [Kevers 2022]

BUT: in the real life, many documents are not expressed in one single language !

The case of multilingual documents

BELGISCH STAATSBLAD | MONITEUR BELGE

Publicatie overeenkomstig artikelen 472 tot 478 van de programmawet van 24 december 2002, gewijzigd door de artikelen 4 tot en met 8 van de wet houdende diverse bepalingen van 20 juli 2005 en artikelen 117 en 118 van de wet van 5 mei 2019.

Dit *Belgisch Staatsblad* kan geconsulteerd worden op :
www.staatsblad.be

Bestuur van het Belgisch Staatsblad, Antwerpsesteenweg 53, 1000 Brussel - Directeur : Wilfried Verrezen

Gratis tel. nummer : 0800-98 809

192e JAARGANG

N. 154

Publication conforme aux articles 472 à 478 de la loi-programme du 24 décembre 2002, modifiés par les articles 4 à 8 de la loi portant des dispositions diverses du 20 juillet 2005 et les articles 117 et 118 de la loi du 5 mai 2019.

Le *Moniteur belge* peut être consulté à l'adresse :
www.moniteur.be

Direction du Moniteur belge, chaussée d'Anvers 53, 1000 Bruxelles - Directeur : Wilfried Verrezen

Numéro tél. gratuit : 0800-98 809

192e ANNEE

MAANDAG 13 JUNI 2022
TWEEDE EDITIE

LUNDI 13 JUIN 2022
DEUXIEME EDITION

INHOUD

SOMMAIRE

Wetten, decreten, ordonnanties en verordeningen

Federale Overheidsdienst Werkgelegenheid, Arbeid en Sociaal Overleg

24 APRIL 2022. — Koninklijk besluit waarbij algemeen verbindend wordt verklaard de collectieve arbeidsovereenkomst van 14 oktober 2021, gesloten in het Paritair Subcomité voor de beschutte werkplaatsen van het Waalse Gewest en van de Duitstalige Gemeenschap, betreffende het stelsel van werkloosheid met bedrijfstoeslag ten laste van het fonds voor bestaanszekerheid (FBZ ETAW) voor sommige oudere werknemers met een aandoening of werknemers met ernstige lichamelijke problemen in de Duitstalige Gemeenschap bij ontslag, bl. 50607.

Lois, décrets, ordonnances et règlements

Service public fédéral Emploi, Travail et Concertation sociale

24 AVRIL 2022. — Arrêté royal rendant obligatoire la convention collective de travail du 14 octobre 2021, conclue au sein de la Sous-commission paritaire pour les entreprises de travail adapté de la Région wallonne et de la Communauté germanophone, relative au régime de chômage avec complément d'entreprise à charge du fonds de sécurité d'existence (FSE ETAW) pour certains travailleurs âgés moins valides ou ayant des problèmes physiques graves en Communauté germanophone, en cas de licenciement, p. 50607.

Federale Overheidsdienst Werkgelegenheid, Arbeid en Sociaal Overleg

24 APRIL 2022. — Koninklijk besluit waarbij algemeen verbindend wordt verklaard de collectieve arbeidsovereenkomst van 14 oktober 2021, gesloten in het Paritair Subcomité voor de beschutte werkplaatsen van het Waalse Gewest en van de Duitstalige Gemeenschap, betreffende het stelsel van werkloosheid met bedrijfstoeslag van het fonds voor bestaanszekerheid FBZ ETAW op 60 jaar met een loopbaan van ten minste 40 jaar in de Duitstalige Gemeenschap, bl. 50611.

Service public fédéral Emploi, Travail et Concertation sociale

24 AVRIL 2022. — Arrêté royal rendant obligatoire la convention collective de travail du 14 octobre 2021, conclue au sein de la Sous-commission paritaire pour les entreprises de travail adapté de la Région wallonne et de la Communauté germanophone, relative au régime de chômage avec complément d'entreprise à charge du fonds de sécurité d'existence FSE ETAW pour les travailleurs de 60 ans ayant une carrière professionnelle d'au moins 40 ans en Communauté germanophone, p. 50611.

Officiële berichten

Federale Overheidsdienst Justitie

Rechterlijke Orde. — Vacante betrekkingen, bl. 50615.

Avis officiels

Service public fédéral Justice

Ordre judiciaire. — Places vacantes, p. 50615.

NLD

Wetten, decreten, ordonnanties en verordeningen

Federale Overheidsdienst Werkgelegenheid, Arbeid en Sociaal Overleg

24 APRIL 2022. — Koninklijk besluit waarbij algemeen verbindend wordt verklaard de collectieve arbeidsovereenkomst van 14 oktober 2021, gesloten in het Paritair Subcomité voor de beschutte werkplaatsen van het Waalse Gewest en van de Duitstalige Gemeenschap, betreffende het stelsel van werkloosheid met bedrijfstoeslag ten laste van het fonds voor bestaanszekerheid (FBZ ETAW) voor sommige oudere werknemers met een aandoening of werknemers met ernstige lichamelijke problemen in de Duitstalige Gemeenschap bij ontslag, bl. 50607.

FRA

Lois, décrets, ordonnances et règlements

Service public fédéral Emploi, Travail et Concertation sociale

24 AVRIL 2022. — Arrêté royal rendant obligatoire la convention collective de travail du 14 octobre 2021, conclue au sein de la Sous-commission paritaire pour les entreprises de travail adapté de la Région wallonne et de la Communauté germanophone, relative au régime de chômage avec complément d'entreprise à charge du fonds de sécurité d'existence (FSE ETAW) pour certains travailleurs âgés moins valides ou ayant des problèmes physiques graves en Communauté germanophone, en cas de licenciement, p. 50607.

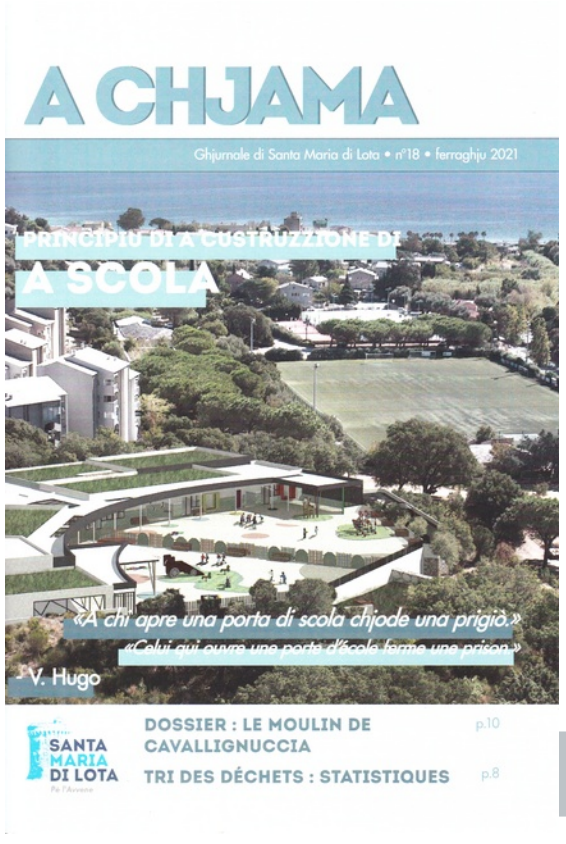
Example 1: Belgian Official Journal

- Intentional
- Well-organised and delimited

http://www.ejustice.just.fgov.be

The case of multilingual documents

Example 2: Santa Maria Di Lota municipality journal (Corsica, France)


A CHJAMA

Ghjurnale di Santa Maria di Lota • n°18 • ferraghju 2021

PRINCIPIU DI A CUSTRUZZIONE DI A SCOLA

«A chi apre una porta di scola chjode una prigiò.»

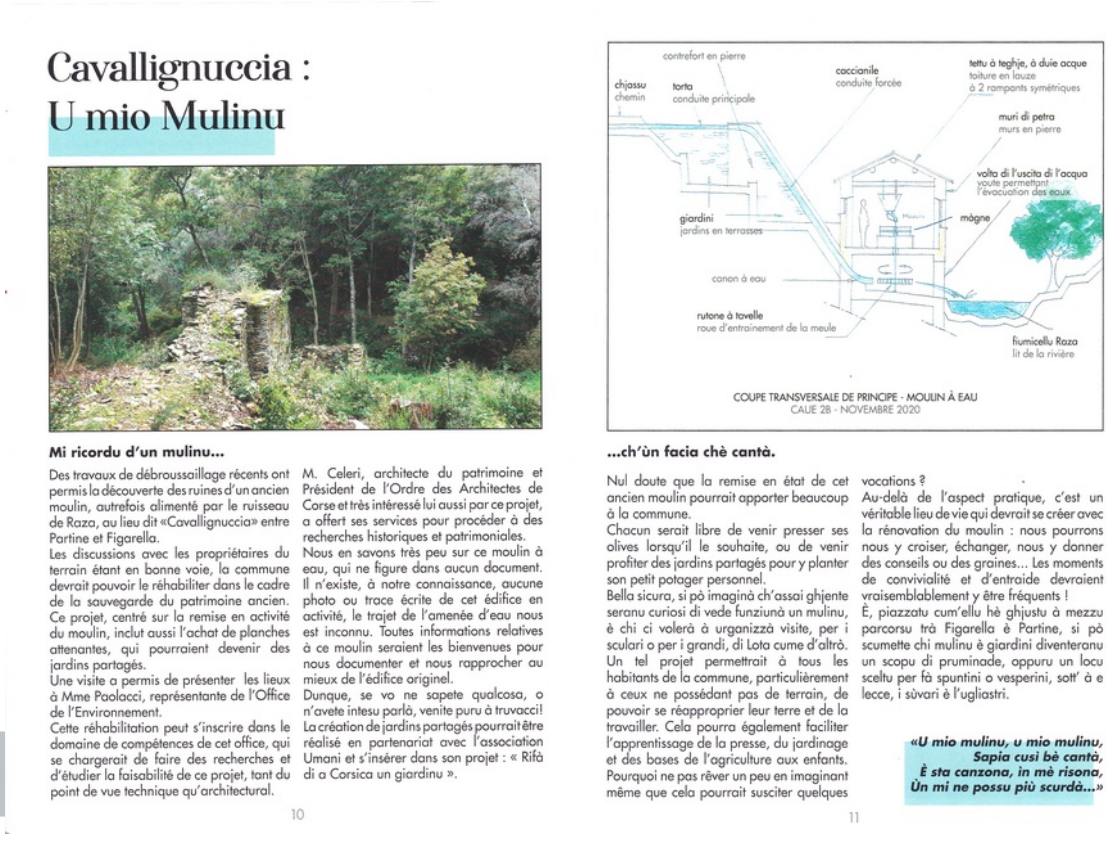
«Celui qui ouvre une porte d'école ferme une prison»

- V. Hugo

SANTA MARIA DI LOTA

DOSSIER : LE MOULIN DE CAVALLIGNUCCIA p.10

TRI DES DÉCHETS : STATISTIQUES p.8


Cavallignuccia :
U mio Mulinu

Mi ricordu d'un mulinu...

Des travaux de débroussaillage récents ont permis la découverte des ruines d'un ancien moulin, autrefois alimenté par le ruisseau de Raza, au lieu dit «Cavallignuccia» entre Partine et Figarella.
Les discussions avec les propriétaires du terrain étant en bonne voie, la commune devrait pouvoir le réhabiliter dans le cadre de la sauvegarde du patrimoine ancien.
Ce projet, centré sur la remise en activité du moulin, inclut aussi l'achat de planches attenantes, qui pourraient devenir des jardins partagés.
Une visite a permis de présenter les lieux à Mme Paolacci, représentante de l'Office de l'Environnement.
Cette réhabilitation peut s'inscrire dans le domaine de compétences de cet office, qui se chargerait de faire des recherches et d'étudier la faisabilité de ce projet, tant du point de vue technique qu'architectural.
M. Celeri, architecte du patrimoine et Président de l'Ordre des Architectes de Corse et très intéressé lui aussi par ce projet, a offert ses services pour procéder à des recherches historiques et patrimoniales.
Nous en savons très peu sur ce moulin à eau, qui ne figure dans aucun document. Il n'existe, à notre connaissance, aucune photo ou trace écrite de cet édifice en activité, le trajet de l'amenée d'eau nous est inconnu. Toutes informations relatives à ce moulin seraient les bienvenues pour nous documenter et nous rapprocher au mieux de l'édifice originel.
Dunque, se vo ne sapete qualcosa, o n'avete intesu parlà, venite puru à truvacci!
La création de jardins partagés pourrait être réalisé en partenariat avec l'association Umani et s'insérer dans son projet : « Rifà di a Corsica un giardinu ».

10

COUPE TRANSVERSALE DE PRINCIPE - MOULIN À EAU
CAUE 2B - NOVEMBRE 2020

...ch'ùn facia chè cantà.

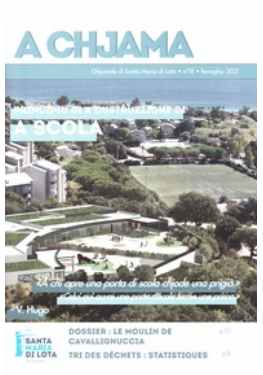
Nul doute que la remise en état de cet ancien moulin pourrait apporter beaucoup à la commune.
Chacun serait libre de venir presser ses olives lorsqu'il le souhaite, ou de venir profiter des jardins partagés pour y planter son petit potager personnel.
Bella sicura, si pò imaginà ch'assai ghjente seranu curiosi di vede funziunà un mulinu, è chi ci volerà à urganizzà visite, per i sculari o per i grandi, di Lota cume d'altrò.
Un tel projet permettrait à tous les habitants de la commune, particulièrement à ceux ne possédant pas de terrain, de pouvoir se réapproprier leur terre et de la travailler. Cela pourra également faciliter l'apprentissage de la presse, du jardinage et des bases de l'agriculture aux enfants.
Pourquoi ne pas rêver un peu en imaginant même que cela pourrait susciter quelques vocations ?
Au-delà de l'aspect pratique, c'est un véritable lieu de vie qui devrait se créer avec la rénovation du moulin : nous pourrons nous y croiser, échanger, nous y donner des conseils ou des graines... Les moments de convivialité et d'entraide devraient vraisemblablement y être fréquents !
È, piazzatu cum'ellu hè ghjustu à mezzu parcorsu trà Figarella è Partine, si pò scumette chi mulinu è giardini diventeranu un scopu di pruminade, appuru un locu sceltu per fà spuntini o vesperini, sott' à e lecce, i sùvari è l'ugliastri.

«U mio mulinu, u mio mulinu,
Sapia cusì bè cantà,
È sta canzona, in mè risona,
Un mi ne possu più scurdà...»

11

The case of multilingual documents

Example 2: SMDL municipality journal



Corsican

French

...ch'ùn facia chè cantà.

Nul doute que la remise en état de cet ancien moulin pourrait apporter beaucoup à la commune.
Chacun serait libre de venir presser ses olives lorsqu'il le souhaite, ou de venir profiter des jardins partagés pour y planter son petit potager personnel.

Bella sicura, si pò imaginà ch'assai ghjente seranu curiosi di vede funziunà un mulinu, è chi ci volerà à urganizzà vìsite, per i sculari o per i grandi, di Lota cume d'altrò.

Un tel projet permettrait à tous les habitants de la commune, particulièrement à ceux ne possédant pas de terrain, de pouvoir se réapproprier leur terre et de la travailler. Cela pourra également faciliter l'apprentissage de la presse, du jardinage et des bases de l'agriculture aux enfants. Pourquoi ne pas rêver un peu en imaginant même que cela pourrait susciter quelques vocations ?
Au-delà de l'aspect pratique, c'est un véritable lieu de vie qui devrait se créer avec la rénovation du moulin : nous pourrons nous y croiser, échanger, nous y donner des conseils ou des graines... Les moments de convivialité et d'entraide devraient vraisemblablement y être fréquents !

È, piazzatu cum'ellu hè ghjustu à mezzu parcorsu trà Figarella è Partine, si pò scumette chi mulinu è giardini diventeranu un scopu di pruminade, oppuru un locu sceltu per fà spuntini o vesperini, sott' à e lecce, i sùvari è l'ugliastri.

«U mio mulinu, u mio mulinu,
Sapia cusì bè cantà,
È sta canzona, in mè risona,
Ùn mi ne possu più scurdà...»

- **Intentional**
- **Moderately organised and delimited**
- **Large segments (paragraphs)**
- **Evenly balanced**

The case of multilingual documents

Example 3: Ethnotext from the BDLC

BDLC # 368 / Lento / Cultures – Le jardin

— Allora, quand'è t'ai u to ortu, chì faci ? In prima cume l'orti, cume i pigliavate, vogliu dì, à sulana o...?

— L'ortu ? Ah, l'ortu u megliu, pè a robba u megliu hè à sulana. Eh, a robba hè più bona. U megliu, a robba hè più sfangata.

— È allora, quand'è tù ai da preparà u to terrenu, cume fecii ?

— Allora u terrenu, di novembre cusì, u volti. Di novembre li dai una vultulata à... cume si dice, dimu à volta tonda, chì nu a terra. Allora... tutti i microbes è i cosi si morenu.

— È, certe terre sò più fàciule à travaglià ? Nò ? Cume si dice ?

— Ah iè. Oui. Certe sò più lene. È certe sò più dure, sò più carche. Cume u chjamanu, ci hè u calcaire chì, sò più, sò più dure.

—È... bon allora quandu sò più, quandu sò fàciule à travaglià, dici : sò lene ? [...]

[English]

- So when you have your garden, what do you do? First of all how the gardens, how do you choose them, I mean, in the sun or...?

- The garden? Ah, the ideal for the garden, for the products, the ideal is exposure to the sun. Eh, the products are better. It's better, the products are not in contact with the mud.

- And then when you have to prepare the ground, how do you do it?

- In November, you turn the soil. In November, you turn over the soil a little, you... how do you say, we say "à volta tonda", in the soil. Then... all the microbes and so on die.

- And some land is easier to work? No ? How do you say?

- Oh yes. Yes. Some are softer. And some are harder, they're more compact. How- do they call it? There's limestone and they're harder.

- And... well then when they are more, when they are easy to work, you say: they are soft?

The case of multilingual documents

BDLC # 368 / Lento / Cultures – Le jardin

— Allora, quand'è t'ai u to ortu, chì faci ? In prima cume l'orti, cume i pigliavate, vogliu dì, à sulana o...?

— L'ortu ? Ah, l'ortu u megliu, pè a robba u megliu hè à sulana. Eh, a robba hè più bona. U megliu, a robba hè più sfangata.

— È allora, quand'è tù ai da preparà u to terrenu, cume fecii ?

— Allora u terrenu, di novembre cusì, u volti. Di novembre li dai una vultulata à... cume si dice, dimu à volta tonda, chì nu a terra. Allora... tutti i microbes è i cosi si morenu.

— È, certe terre sò più fàciule à travaglià ? Nò ? Cume si dice ?

— Ah iè. Oui. Certe sò più lene. È certe sò più dure, sò più carche. Cume u chjamanu, ci hè u calcaire chì, sò più, sò più dure.

—È... bon allora quandu sò più, quandu sò fàciule à travaglià, dici : sò lene ? [...]

Example 3: Ethnotext from the BDLC

- Unintentional
- Completely mixed
- Small segments
- Not very frequent (isolated)

→ Code switching

The case of multilingual documents

Language Identification (LgID) at the word level is an important tool to build uniform and high quality corpora with regard to the language of the documents

→ Only a few systems are available…

→ Generally, no annotated corpus to carry out specific and massive machine learning

BUT: we argue that it is possible to use a LID module that performs well on short texts to carry out this work

→ only requires raw monolingual corpora

→ optionally uses monolingual dictionaries

→ LID = ldig [Nakatani 2012]

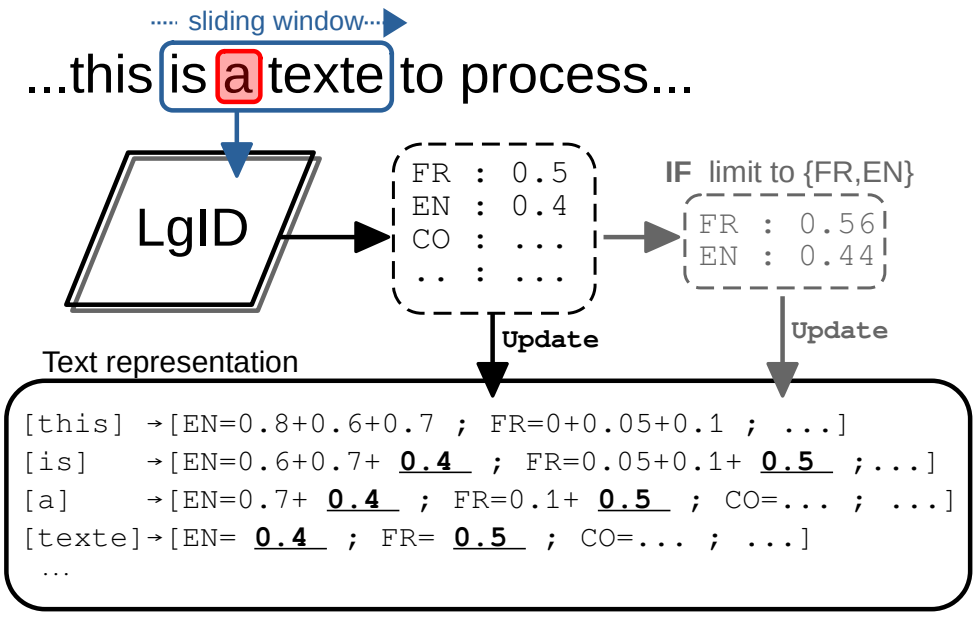
CoSwID, a tool to identify monolingual segments into multilingual documents

General parsing:

- Split text into tokens
- Analyse it through a sliding window
- Use LgID on each snippet

  → returns probability distribution
- Optionally, limit the results to a subset of langages
- Consolidate results

  → for a 3-tokens window, each token will get 3 scores to be added up



CoSwID, a tool to identify monolingual segments into multilingual documents

Final Annotation:

- The language with the highest probability is assigned to the token
- If the difference between first languages is small → further verification
  - Thresholding ("*indecision gap*")
  - Verification of remaining languages (three possible methods) :
    - Dictionary
    - LgID on the token
    - Dictionary + LgID

If *indecision gap* = 0.10:

Text representation

```
[this] →[EN=0.70 ; FR=0.05 ; CO=... ; ... ]
[is]   →[EN=0.57 ; FR=0.22 ; CO=... ; ... ]
[a]    →[EN=0.47 ; FR=0.40 ; CO=... ; ... ]
[texte]→[EN=0.39 ; FR=0.51 ; CO=... ; ... ]
 ...
```

No direct decision for `[a]`

→ keep `EN` and `FR`

→ use dictionnary and/or LgID to choose

→ if not sucessful, keep the original result

CoSwID, a tool to identify monolingual segments into multilingual documents

Tests and evaluation – Data

- Selection of nine languages
- LgID training data
  - Three corpora for Corsican (*Wikipedia*, the *Bible*, *A Piazzetta* blog)
  - *Tatoeba* (eight languages)
- Dictionaries
  - Lists of inflected forms

Language		*Base*	*Filter*	*Filter2*
eng	en	67 948 293	64 653 131	58 824 526
ita	it	32 022 121	22 815 185	20 813 785
deu	de	29 987 665	29 106 064	28 126 760
fra	fr	22 482 372	19 062 318	17 833 313
por	pt	17 399 633	13 688 730	13 054 128
spa	es	15 437 547	12 294 945	11 069 048
cos	co	11 868 620	10 483 557	10 402 975
nld	nl	5 968 644	5 455 944	5 034 300
ron	ro	1 045 723	862 135	782 957

Language			*Items*
English	eng	en	398 417
Italian	ita	it	95 038
German	deu	de	8 277
French	fra	fr	794 286
Portuguese	por	pt	890 193
Spanish	spa	es	477 976
Corsican	cos	co	43 051
Dutch	nld	nl	401 575
Romanian	ron	ro	19 946

CoSwID, a tool to identify monolingual segments into multilingual documents

Tests and evaluation – Data

- Evaluation corpora
  - UDHR-parag: synthetic ; full paragraphs in one language
  - UDHR-sent: synthetic ; full sentences in one language
  - UDHR-word: synthetic
    - Select a sentence in a main langage (chosen randomly)
    - Every 3 to 7 tokens, replace 1 to 4 token in an other langage (chosen randomly)

    → Will give a "worst-case scenario" performance measure
  - BDLC-ethno: authentic ; code switching like alternations

CoSwID, a tool to identify monolingual segments into multilingual documents

Tests and evaluation – Experiments

CoSwID parameters	TESTED VALUES
Training set for LgID	*Base, Filter, Filter2*
Sliding window size	*1, 3, 5, 7* tokens
Indecision gap	*0, 0.5, 0.1, 0.2*
Verification methods	*dico, lgid, full*
Limitation to a subset of langages (for BDLC-ethno corpus only)	*{Corsican,French}*

→ All the combinations of these parameters were tested

CoSwID, a tool to identify monolingual segments into multilingual documents

Tests and evaluation – Experiments

Third party systems	USE CASE	LANGUAGES	USABLE BUILT-IN DATA	CUSTOMISATION
SegLang [Yamaguchi and Tanaka-Ishii 2012]	Multilingual (1-N) documents	222 *Our 9 lg. covered*	- UDHR - Wikipedia	Data from Filter2
LangId [King and Abney 2013]	Bilingual documents	30 *Corsican missing*	n/a	Data from Filter2
Codeswitchador [Lignos and Mitch 2013]	Bilingual documents	2 *Spanish/English*	n/a	Data from Filter2

CoSwID, a tool to identify monolingual segments into multilingual documents

Results

UDHR-parag / UDHR-sent

CoSwID : 98.15% / 98.00% – $\text{SegLang}_{\text{Wikipedia}}$: 92.11% / 90.53% – $\text{SegLang}_{\text{Custom}}$: 99.54% / 99.61%

UDHR-word

CoSwID : 87.29% – $\text{SegLang}_{\text{Wikipedia}}$: 70.61% – $\text{SegLang}_{\text{Custom}}$: 88.07%

BDLC-ethno

CoSwID : 96.31% – $\text{SegLang}_{\text{Wikipedia}}$: 94.60% – $\text{SegLang}_{\text{Custom}}$: 97.54%

$\text{BDLC-ethno}_{\{\text{cos,fra}\}}$

CoSwID : 97.97% – $\text{SegLang}_{\text{Custom}}$: 97.70% – LangId : 97.38 % – Codeswitchador : 96.70 %

CoSwID, a tool to identify monolingual segments into multilingual documents

Results

- A short window is suitable for dynamic language alternations and/or for short segments.
- A longer window will maximise accuracy for texts containing mainly long monolingual segments.
- Verification mechanism has an added value (but quite marginal when only two languages)
  - Generally dictionary based
  - Optimal gap size is not obvious (0.2 ?)
- Filtering is generally benefical (but sometimes marginally) → data quality matters

CoSwID, a tool to identify monolingual segments into multilingual documents

Conclusions

→ Achieved results – Acc_o [87.29% , 97.97%] – in line with third-party systems (using our custom data)

→ LgID module can be used to detect language alternations such as code switching

→ This approach allows to include without difficulty under-resourced languages with a minimum of resources

→ The release of the code and data will contribute to improve the offer of open and reusable systems for language and code switching identification

Aknowledgments & References

This work was carried out thanks to CPER funding:

"Un outil linguistique au service de la Corse et des Corses: la Banque de Données Langue Corse (BDLC)".

King, B. and Abney, S. (2013). Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia, June. Association for Computational Linguistics.

Lignos, C. and Mitch, M. (2013). Toward web-scale analysis of codeswitching. In *87th Annual Meeting of the Linguistic Society of America*.

Nakatani, S. (2012). Short Text Language Detection with Infinity-Gram, May 2012. https://www.slideshare.net/shuyo/short-text-language-detection-with-infinitygram-1294944.

Yamaguchi, H. and Tanaka-Ishii, K. (2012). Text Segmentation by Language Using Minimum Description Length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 969–978, Jeju Island, Korea, July 2012. Association for Computational Linguistics.