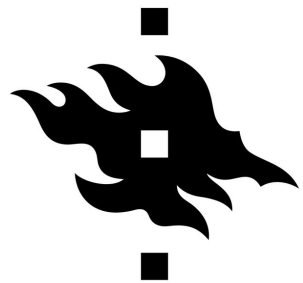
Occitan in Wikipedia discussions: Corpus extraction, language identification and user practices

Aleksandra Miletić
(joint work with Yves Scherrer)

Department of Digital Humanities, University of Helsinki



UNIVERSITY OF HELSINKI

Introduction

Occitan:

- Regional/minority language spoken in the South of France and in parts of Italy and Spain
- Rich system of dialects, not standardized, not recognized as an official language in France
- Recent and still limited digital presence

In this talk, I will:

- present a new freely available corpus for Occitan (OcWikiDisc)
- report on results of language identification experiments (LID) on the corpus
- discuss LID difficulties with respect to closely related languages
- offer insights about user practices when writing in Occitan on this medium

Introduction

Occitan:

- Regional/minority language spoken in the South of France and in parts of Italy and Spain
- Rich system of dialects, not standardized, not recognized as an official language in France
- Recent and still limited digital presence

In this talk, I will:

- present a new freely available corpus for Occitan (OcWikiDisc)
- report on results of language identification experiments (LID) on the corpus
- discuss LID difficulties with respect to closely related languages
- offer insights about user practices when writing in Occitan on this medium

CorCoDial Project

"Corpus-based computational dialectology" – Academy of Finland project No. 342859

Outline

Linguistic properties

Occitan:

- Romance group
- null subject language
- tense, person and number inflection marks on finite verbs
- number and gender inflection on all components of the noun phrase

(1) T' aviái laissat un messatge
you.DAT have.1SG.IMPF leave.PST.PTCP a.SG.M message
'I had just left you a message'

Dialects

Six main dialect groups (Bec, 1995):

No dialect considered as a standard variety

Spelling norms

No standardized spelling norm, but two are dominant (Sibille, 2002):

- *classical*, based on medieval troubadours' spelling
- *mistralian*, closer to French spelling convetions
- others: *grafia febusiana*, *Nòrma de l'Escòla dau Pò*

Nòrma classica (transcripcion)	**Nòrma mistralenca (originau)**
Mirèlha, Cant I (F. Mistral)	*Mirèio, Cant I (F. Mistral)*
Cante una chata de Provença.	Cante uno chato de Prouvènço.
Dins leis amors de sa jovença,	Dins lis amour de sa jouvènço,
A travèrs de la Crau, vèrs la mar, dins lei blats,	A travès de la Crau, vers la mar, dins li blad,
Umble [Umil] escolan dau grand *Omèra* [Omèr],	Umble escoulan dóu grand Oumèro,
Ieu la vòle seguir. Coma èra	Iéu la vole segui. Coume èro
Ren qu'una chata de la tèrra,	Rèn qu'uno chato de la terro,
En fòra de la Crau se n'es gaire parlat.	En foro de la Crau se n'es gaire parla.

(source: `https://oc.wikipedia.org/wiki/Proven%C3%A7au`)

Occitan in the NLP era

Still low-resourced when it comes to natural language processing tools and resources:

- Searchable text database (3.4M words) (Bras and Vergez-Couret, 2016)
- Morphological lexicon Lofloc (850K entries) (Vergez-Couret, 2016; Bras et al., 2020)
- PoS-tagged corpus (12K words) (Bernhard et al., 2018)
- Syntactic treebank (25K words) (Miletic et al., 2020)
- Text-to-speech model (Corral et al., 2020)
- **No large, downloadable corpus**

User Generated Content in Occitan

Popular solution: relying on user-generated content found on the web (Ljubešić and Klubička, 2014)

Difficulties:

- Comparatively recent and limited online presence
- No dedicated top-level domain
- General crawling methods yield small amounts of data (31K words in the OSCAR corpus) (Ortiz Suárez et al., 2019)

Adopted solution:

- Targeted extraction: Occitan Wikipedia
- Focus on the talk pages: direct user-to-user interactions
- Preserve linguistic integrity of the data
- Provide as much metadata as possible

→ Make the corpus suited for NLP, but also for corpus linguistics

Outline

Using Occitan Wikipedia

Previous work:

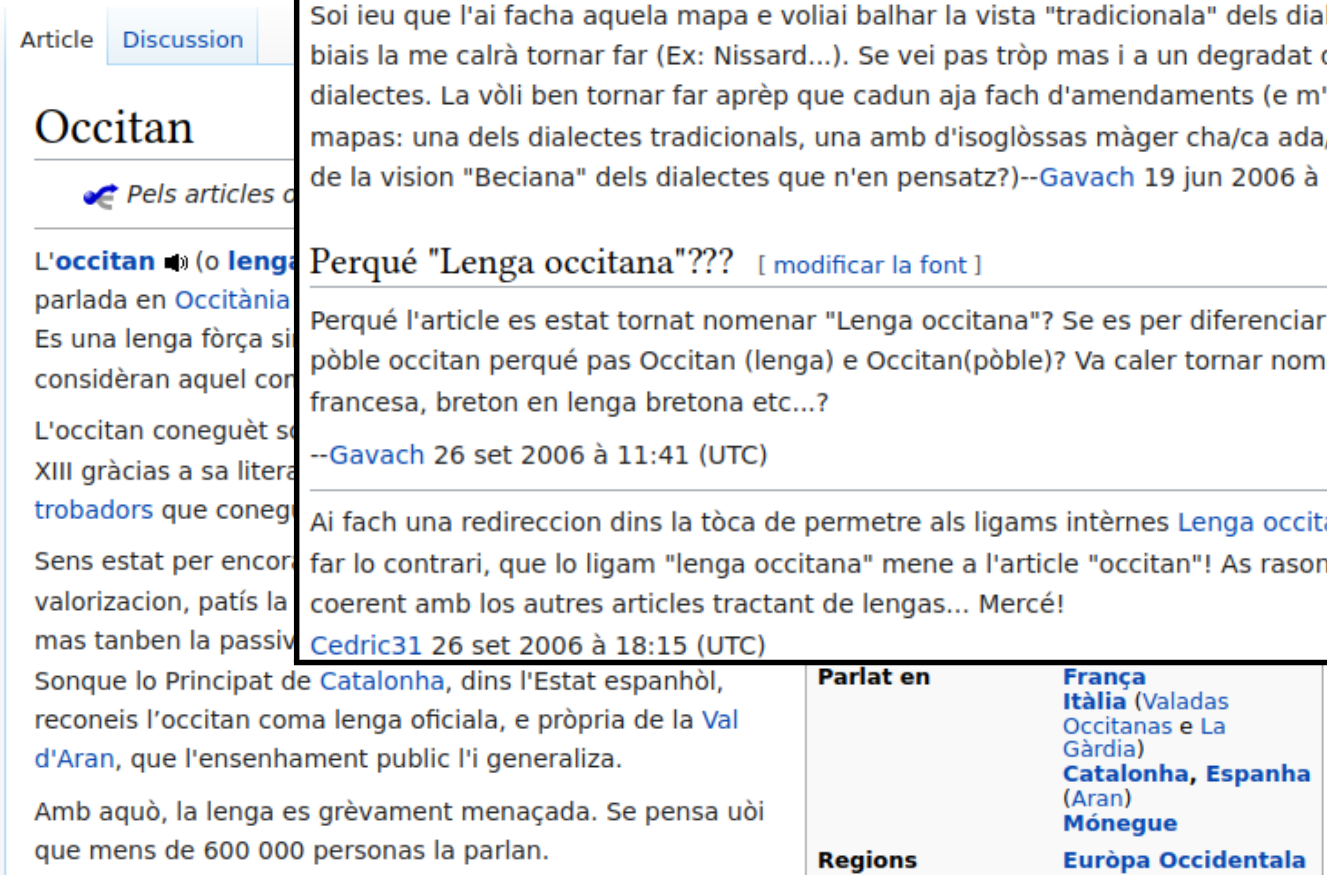
- Used as training material for mBERT (Devlin et al., 2019), as well as for several language identification tools (Lui and Baldwin, 2012; Joulin et al., 2017; Jauhiainen et al., 2022; Costa-jussà et al., 2022) ⚠ **Not distributed.**
- Present in WikiMatrix, a collection of parallel corpora extracted from Wikipedia (Schwenk et al., 2021) ⚠ **Isolated sentences, little metadata.**

Focusing on Talk Pages

Talk pages: discussions about content and editing policies on Wikipedia

Focusing on Talk

Talk pages: discussi



Carta dei dialèctes [modificar la font]

Nòti que la carta mete en contacte dirècte lo gavaudanés e lo vivaroaupenc, çò que significa que lo l'occitan bas-vivarés i es considerat coma de gavaudanés... Remandi a l'article sus lo bas-vivarés : deuriá permetre de precisar lei causas. Vivarés 18 jun 2006 à 21:30 (UTC)

Soi ieu que l'ai facha aquela mapa e voliai balhar la vista "tradicionala" dels dialectes occitans. De tot biais la me calrà tornar far (Ex: Nissard...). Se vei pas tròp mas i a un degradat de colors entre los dialectes. La vòli ben tornar far aprèp que cadun aja fach d'amendaments (e m'agradaria melhor tres mapas: una dels dialectes tradicionals, una amb d'isoglòssas màger cha/ca ada/aa h/f t/0 etc..., e una de la vision "Beciana" dels dialectes que n'en pensatz?)--Gavach 19 jun 2006 à 17:03 (UTC)

Perqué "Lenga occitana"??? [modificar la font]

Perqué l'article es estat tornat nomenar "Lenga occitana"? Se es per diferenciar lenga occitana de pòble occitan perqué pas Occitan (lenga) e Occitan(pòble)? Va caler tornar nomenar francés en lenga francesa, breton en lenga bretona etc...?

--Gavach 26 set 2006 à 11:41 (UTC)

Ai fach una redireccion dins la tòca de permetre als ligams intèrnes Lenga occitana d'èsser valids... Vai far lo contrari, que lo ligam "lenga occitana" mene a l'article "occitan"! As rason! Es mai logic, e mai coerent amb los autres articles tractant de lengas... Mercé!

Cedric31 26 set 2006 à 18:15 (UTC)

article source: `https://oc.wikipedia.org/wiki/Occitan`

discussion source: `https://oc.wikipedia.org/wiki/Discutir:Lenga_occitana`

Extracting a Corpus

Talk pages: distributed as part of data dumps by Wikimedia (XML + wikitext)

```
==Sientista/Scientific==
Adiu! Ai creat la categoria [[:Categoria:Scientific alemand]] e escafat la categoria [[:Cat
egoria:Scientista alemand]] perque en occitan, un &quot;scientific&quot; es una persona que
 practica una sciéncia e un &quot;scientista&quot; es una persona que fa de scientologia (q
u'aparten a la Glèisa de scientologia)! Coralament!&lt;br&gt;
[[Utilisator:Cedric31|Cedric31]] 19 set 2006 à 11:06 (UTC)

== Es fach! ==

Es fach e n'aprofiechèri per agachar 2 o 3 tres articles sus Brasil. Òsca! per ton trabalh
sus la wiki occitana. Se vòs vaqui un article corregit (preposicions &quot;lusitanistas&quo
t;) que pòdes utilizar per ne far un copiat/pegat (cut and past).

'''Santos''' es una ciutat del [[Brasil]], sul litoral de l'[[estats del Brasil|estat]] de
[[São Paulo (estat)|São Paulo]]. Sa populacion èra estimada a 418 316 abitants en [[2005]].
 Sa superficia totala es de 280,3 km². Es lo [[pòrt]] màger d'[[America del Sud]] e tanben
una ciutat [[torisme|toristica]] e [[comèrci|comerciala]].

--[[Utilisator:Gavach|Gavach]] 26 set 2006 à 14:31 (UTC)

== Adiu!==
&quot;second round of presidential election&quot; = &quot;Segond torn de las eleccions pres
idencialas&quot;&lt;br&gt;
[[Utilisator:Cedric31|Cedric31]] 1 oct 2006 à 07:04 (UTC)

==&quot;Acontecerá&quot;==
&quot;Acontecerá&quot; = &quot;Aurà luòc&quot; o &quot;Tendrà luòc&quot;&lt;br&gt;
[[Utilisator:Cedric31|Cedric31]] 2 oct 2006 à 18:03 (UTC)
----
```

Exracting a Corpus

Transformed into a corpus using a Python script:

message	discussion_type	discussion_title	thread_title	username	signature	timestamp
Adiu! Ai creat la categoria Scientific alemand e escafat la categoria Scientista alemand perque en occitan, un "scientific" es una persona que practica una sciéncia e un "scientista" es una persona que fa de scientologia (qu'aparten a la Glèisa de scientologia)! Coralament!	Discussion Utilizaire	Joao Xavier	Sientista/Scientific	Cedric31	Cedric31	2006-09-19 11:06:00+00:00
Es fach e n'aprofiechèri per agachar 2 o 3 tres articles sus Brasil. Òsca! per ton trabalh sus la wiki occitana. Se vòs vaqui un article corregit (preposicions "lusitanistas") que pòdes utilizar per ne far un copiat/pegat (cut and past).\rSantos es una ciutat del Brasil, sul litoral de l'estat de São Paulo. Sa populacion èra estimada a 418 316 abitants en 2005. Sa superficia totala es de 280,3 km². Es lo pòrt màger d'America del Sud e tanben una ciutat toristica e comerciala.\r	Discussion Utilizaire	Joao Xavier	Es fach!	Gavach	Gavach	2006-09-26 14:31:00+00:00
"second round of presidential election" = "Segond torn de las eleccions presidencialas"	Discussion Utilizaire	Joao Xavier	Adiu!	Cedric31	Cedric31	2006-10-01 07:04:00+00:00
"Acontecerá" = "Aurà luòc" o "Tendrà luòc"	Discussion Utilizaire	Joao Xavier	"Acontecerá"	Cedric31	Cedric31	2006-10-02 18:03:00+00:00

Language Filtering

Multilingual content:

I administrate the Catalan version (this is the reason that I can understand Occitan) of the project. And I have the code that let delete the pages of all the wikipedia old software (Catalan and Occitan are using the same software).

Salut. Félicitations pour le changement de l'interface. Bonne chance

Vaya, vaya...., ¿en qué lengua prefieres que nos comuniquemos? ;)

Adieu/καλήμέρα Xaris, With pleasure, just provide me the text. Ειλικρινείς χαιρετισμούς

Guten Tag, La wikipèdia occitana n'est pas un lieu de querelle linguistique. Le fait d'utiliser le ter
en langue romane n'implique pas le rejet du terme en allemand. Cordialement, -

Need for language-based filtering → language identification experiments

Language Filtering

Task definition: **identify posts containing Occitan**

Language identification tools used:

- langid (Lui and Baldwin, 2012) and py3langid (by A. Barbaresi)
- fasttext1 (Joulin et al., 2017) and fasttext2 (Costa-jussà et al., 2022)
- HeLI (Jauhiainen et al., 2022)

Individual tools		**Strategies**	
Tool	**Accuracy (%)**	**Strategy**	**Accuracy (%)**
fasttext1	62.30	heli_top2	93.22
langid	66.64	fasttext2_top2	95.00
py3langid	70.00	fasttext2_heli	95.20
heli	90.70		
fasttext2	93.22		

Baseline evaluation: Occitan only (n=1520)

Language Filtering

	Occitan		
	Precision	Recall	F1-score
fasttext2_top2	84.75	73.53	78.74
heli_top2	93.33	61.76	74.34
fasttext2_top5	79.49	**91.18**	84.93
heli_top5	88.06	86.76	**87.41**
fasttext2_heli_top1	**100.00**	57.35	72.90
fasttext2_heli_top2	85.00	75.00	79.69

Realistic evaluation: Multilingual OcWikiDisc sample (n=100)

Detailed account

Aleksandra Miletić and Yves Scherrer. 2022. OcWikiDisc: a Corpus of Wikipedia Talk Pages in Occitan. *In Proceedings of VarDial - Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70-79.

Corpus statistics

Four versions of the corpus:

	Messages	Tokens	Tokens/Message	Users	Messages/User
ocwikidisc_precision	8149	618,153	75.86	206	33.69
ocwikidisc_balanced	9032	756,922	83.80	323	23.19
ocwikidisc_recall	9394	804,959	85.69	347	22.39
ocwikidisc_unfiltered	11025	1,186,239	107.60	522	17.07

Downloadable here: `https://doi.org/10.5281/zenodo.7079580`

Outline

Closely Related Languages

Long-standing issue: low-resourced languages confused with better-resourced closely related varieties

Is Occitan confused with other Romance languages?

Closely Related Languages

Long-standing issue: low-resourced languages confused with better-resourced closely related varieties

Is Occitan confused with other Romance languages?

Baseline evaluation confusion matrix (Occitan only, n=1520):

fasttext2		heli	
Catalan	23	**Catalan**	38
French	11	**Spanish**	11
Vietnamese	10	*Interlingua*	7
Portuguese	8	**Lombard**	6
Spanish	6	**French**	6
English	5	**Extremaduran**	5
Asturian	5	**Piemontese**	4
Galician	5	**Portuguese**	4
Standard Malay	4	*Haitian*	3
Italian	4	Pfälzisch	3

Top-10 erroneous labels for fasttext2 and HeLI

Other Languages in the Filtered Corpus

Romance languages are also frequent in the filtered corpus:

	No.lang.	Top 11
ocwikidisc_precision	54	Occitan, **Catalan**, **French**, English, German, **Spanish**, **Portuguese**, **Lombard**, **Romanian**, **Piemontese**, **Galician**
ocwikidisc_balanced	124	Occitan, **Catalan**, **Extremaduran**, **Lombard**, **Spanish**, *Interlingua*, **French**, **Galician**, **Piemontese**, **Portuguese**, Lingala
ocwikidisc_recall	114	Occitan, **Catalan**, **French**, **Spanish**, **Galician**, **Portuguese**, **Lombard**, **Italian**, **Asturian**, Korean, **Romanian**
ocwikidisc_unfiltered	155	Occitan, **Catalan**, **French**, **Spanish**, **Portuguese**, **Galician**, **Italian**, Korean, **Lombard**, English, **Asturian**

LID issues or Wikipedia practices?

Collaboration between Wikipedias

Bonjorn, Cedric! Podètz revisar la gramatica del tèxte Colómbia? E tanben Giuseppe Garibaldi (adicionei algumas informações). Bon coralament,

Adiu, Jiròni! Bon Nadal e un Excellent 2011, a tu e a la tiá familha. (Espero que el próximo año yo pueda contribuir más con oc:wiki). Coralament,

Adiu, Pasha! How are you? I'm also working on oc:wiktionary, and I ask if you can help me translating into Occitan at least some of these extremely common words: fork - spoon - knife - dish - doll - baby - toy - wheel - chair - refrigerator - bed - bedroom (at least for the beginning, in order to improve it). Bon coralament,

Outline

Use of Spelling Norms

Occitan Wikipedia rules regarding spelling norms:

`https://oc.wikipedia.org/wiki/Wikip%C3%A8dia:Carta_ling%C3%BCistica`

Nòrma classica

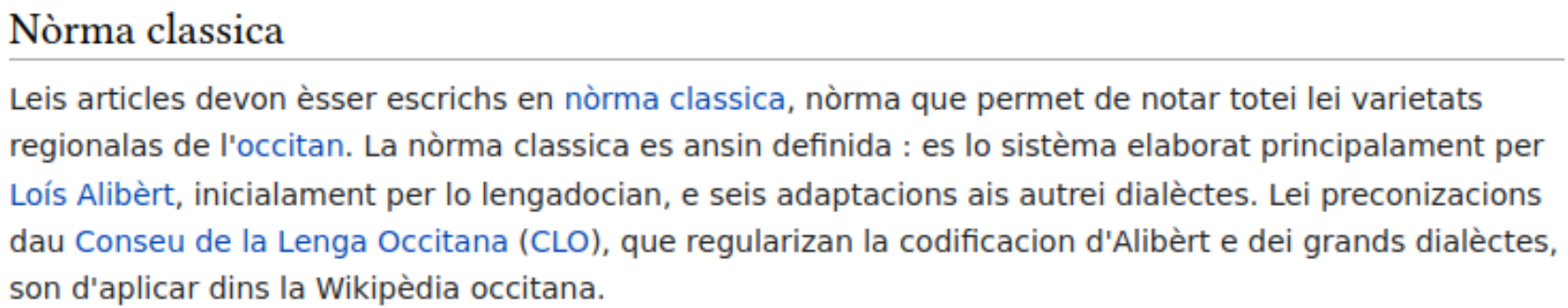
Leis articles devon èsser escrichs en nòrma classica, nòrma que permet de notar totei lei varietats regionalas de l'occitan. La nòrma classica es ansin definida : es lo sistèma elaborat principalament per Loís Alibèrt, inicialament per lo lengadocian, e seis adaptacions ais autrei dialèctes. Lei preconizacions dau Conseu de la Lenga Occitana (CLO), que regularizan la codificacion d'Alibèrt e dei grands dialèctes, son d'aplicar dins la Wikipèdia occitana.

Respected in the analysed sample: other spelling norms only as quotes from published works

Use of Spelling Norms

Occitan Wikipedia rules regarding spelling norms:

`https://oc.wikipedia.org/wiki/Wikip%C3%A8dia:Carta_ling%C3%BCistica`



Nòrma classica

Leis articles devon èsser escrichs en nòrma classica, nòrma que permet de notar totei lei varietats regionalas de l'occitan. La nòrma classica es ansin definida : es lo sistèma elaborat principalament per Loís Alibèrt, inicialament per lo lengadocian, e seis adaptacions ais autrei dialèctes. Lei preconizacions dau Conseu de la Lenga Occitana (CLO), que regularizan la codificacion d'Alibèrt e dei grands dialèctes, son d'aplicar dins la Wikipèdia occitana.

Respected in the analysed sample: other spelling norms only as quotes from published works

⚠ The classical norm glosses over some dialectal differences:

- some final consonants (e.g. plural marker *-s*) are always written in this norm, but not pronounced in all varieties
- some graphemes can have different pronunciations across varieties

Use of Dialects

Occitan Wikipedia rules regarding dialects:

Varietats regionalas de l'occitan

Tot article de la Wikipèdia occitana deu èsser redigit dins una dei grandei varietats regionalas de l'occitan, ansin definidas : auvernhat, gascon (inclús l'aranés), lemosin, lengadocian, provençau (inclús lo niçard), vivaroaupenc.

Toteis aquelei varietats regionalas son egalas, e la Wikipèdia occitana valoriza la diversitat dialectala.

Also seems to be respected in the corpus: multi-user discussions in which everyone uses their own dialect

Use of Dialects

Dialect distribution in the 100-message test sample
Out of 68 messages containing Occitan:

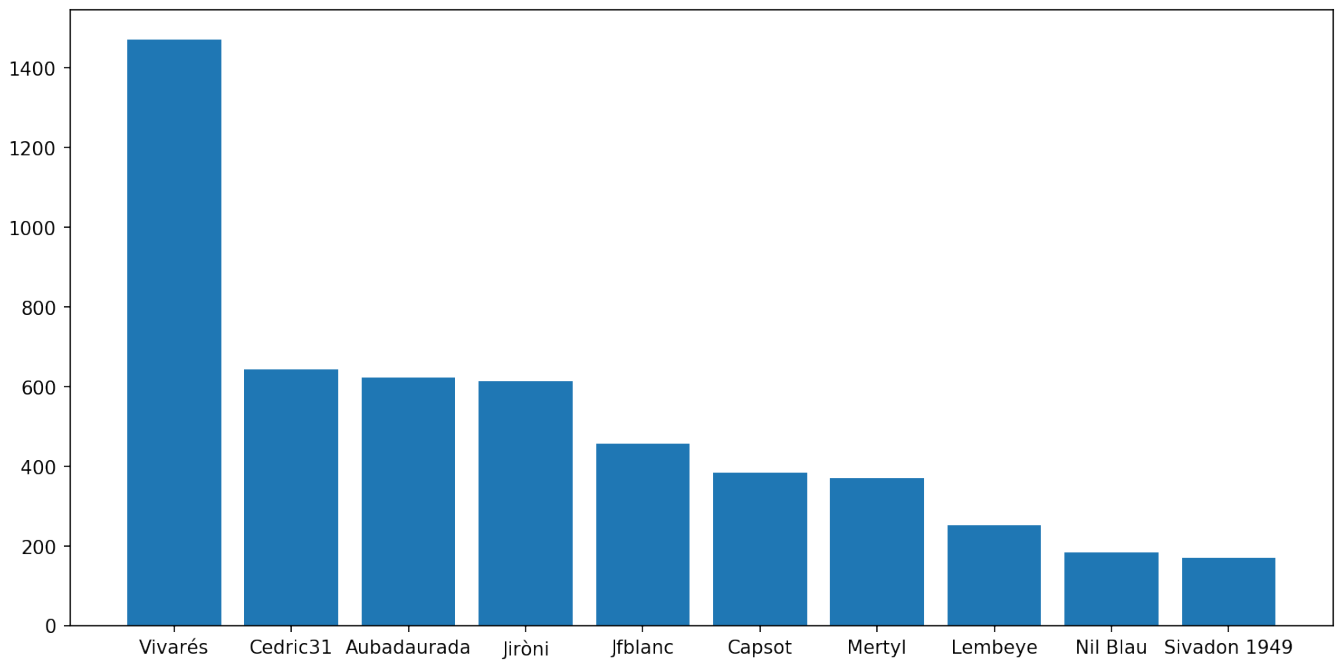
- Lengadocian: 36
- Gascon: 6
- Provençau: 5
- impossible to specify: 21

Use of Dialects

Dialect distribution in the 100-message test sample
Out of 68 messages containing Occitan:

- Lengadocian: 36
- Gascon: 6
- Provençau: 5
- impossible to specify: 21 $\rightarrow$ **classical spelling norm?**

The prevalence of Lengadocian may stem from the dialect usage of the 10 most active users (over 50% of content):



Outline

Summary

- New, freely available corpus in Occitan (600K words)
  - Downloadable here: `https://doi.org/10.5281/zenodo.7079580`
- Language identification experiments show fasttext and HeLI as top performers in identifying Occitan content
- Frequent presence of other Romance languages among predictions raises questions about LID of closely related languages
- Observations about the use of Occitan in the discussions:
  - Multilingual messages: evidence of new users and new types of interactions
  - Spelling norms: selecting the classical norm drives towards standardisation and facilitates searches, but it also hides some of the distinctive properties of individual dialects
  - Dialect use: each user seems to preserve their own dialect. Lengadocian is dominant, possibly due to uneven distribution of content across users

Future Work

In the future:

- Looking into LID of related languages
- Dialect identification for top users
- Dialect identification with NLP tools
- Using the corpus in computational dialectology

References I

Pierre Bec. *La langue occitane*. PUF, 6th edition, 1995.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3917–3924, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1619`.

Myriam Bras and Marianne Vergez-Couret. BaTelÒc : a text base for the Occitan language. In Vera Ferreira and Peter Bouda, editor, *Language Documentation and Conservation in Europe*, pages 133–149. Honolulu: University of Hawaï Press , 2016.

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benazet Dazéas. Loflòc : Lexic obèrt flechit occitan. In Jean-François Courouau, editor, *Fidélités et dissidences (Actes du* XIIe *congrès de l'Association Internationale d'Études Occitanes)*, pages 141–156, Albi, 2020. Centre d'Etude de la Littérature Occitane.

References II

Ander Corral, Igor Leturia, Aure Séguier, Michael Barret, Benaset Dazéas, Philippe Boula de Mareüil, and Nicolas Quint. Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of Gascon Occitan. In *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop «Language Resources and Evaluation Conference–Marseille–11–16 May 2020»*, pages 53–60. European Language Resources Association (ELRA), 2020.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. arXiv preprint: https://arxiv.org/abs/2207.04672, 2022. URL `https://arxiv.org/abs/2207.04672`.

References III

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3912–3922, 2022.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017.

Nikola Ljubešić and Filip Klubička. {bs,hr,sr}WaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0405. URL https://aclanthology.org/W14-0405.

Marco Lui and Timothy Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30, 2012.

References IV

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. A four-dialect treebank for Occitan: Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). URL `https://aclanthology.org/2020.vardial-1.13`.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL `http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215`.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, 2021.

Jean Sibille. Ecrire l'occitan : essai de présentation et de synthèse. In Dominique Caubet, Salem Chaker, and Jean Sibille, editors, *Les langues de France et leur codification. Ecrits divers – Ecrits ouverts*, Paris, France, May 2002. Inalco / Association Universitaire des Langues de France, L'Harmattan. URL `https://hal.archives-ouvertes.fr/hal-01296986`.

References V

Marianne Vergez-Couret. Description du lexique Loflòc. Research report, CLLE-ERSS, Apr 2016. URL https://hal.archives-ouvertes.fr/hal-01338774.

Mercés!

Adiu, Lembeye! 13 de julhet es l'anniversari de Cedric31. Daissatz un messatge (I left a message and a birthday cake!). Coralament,

Adiu, Vivarés! Podètz revisar la gramatica de l'article? (I've obtained all the informations from CIA - The World Factbook. First, I've translated it to Portuguese. Then I've used Apertium to translate it to Spanish. Afterwards I've corrected the translation and then, used the Spanish version to obtain the Catalan version and also the Occitan version, this one using the "unstable" version of Apertium). Bona annada e fòrt coralament,

Adishatz Nicolas ! Sorry I cannot address you in Occitan. I was sad to see no Occitan language Wikimedian is attending the Wikimedia Conference Berlin 2017. We wanted to gather minority language representatives in a meeting, see here, to address common grounds and challenges to possibly make a diagnosis to provide the grounds for constructive collaboration among minority language WPs. However, I will be happy to listen to any concern you may want to raise to present it in the Berlin meeting. A totara :)

Contact:
aleksandra.miletic@helsinki.fi

Individual results

Individual results for fasttext2 and heli on the multilingual sample:

	Occitan		
	Precision	Recall	F1-score
fasttext2_top1	100.00	55.88	71.70
heli_top1	100.00	52.94	69.23