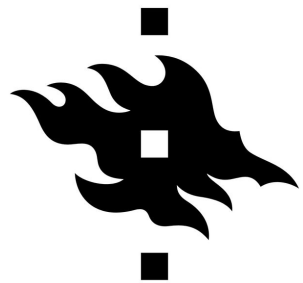
Occitan in Wikipedia discussions: Corpus extraction, language identification and user practices

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.

Talk pages: discussions about content and editing policies on Wikipedia

Talk pages: discussions about content and editing policies on Wikipedia

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

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

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

Transformed into a corpus using a Python script:

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.

Four versions of the corpus:

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

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

Is Occitan confused with other Romance 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	Pfalzisch	3

Top-10 erroneous labels for fasttext2 and HeLI

Romance languages are also frequent in the ltered 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_un ltered	155	Occitan, Catalan, French, Spanish, Portuguese, Galician, Italian, Korean, Lombard, English, Asturian

LID issues or Wikipedia practices?

Occitan Wikipedia rules regarding spelling norms:
https://oc.wikipedia.org/wiki/Wikip%C3%A8dia:Carta_ling%C3%BCistica

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

Occitan Wikipedia rules regarding spelling norms:
https://oc.wikipedia.org/wiki/Wikip%C3%A8dia:Carta_ling%C3%BCistica

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

4 The classical norm glosses over some dialectal di erences:

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

Occitan Wikipedia rules regarding dialects:

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

Dialect distribution in the 100-message test sample

Out of 68 messages containing Occitan:

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

Dialect distribution in the 100-message test sample

Out of 68 messages containing Occitan:

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

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

New, freely available corpus in Occitan (600K words)

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

Language identi cation 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

In the future:

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

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

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steibel, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynes, 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. BaTeLOc : 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 Hawai'i Press, 2016.

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benazet Dazeas. Looc : Lexic obert echit occitan. In Jean-François Courouau, editor, Fidelités et dissidences (Actes du xii e congrès de l'Association Internationale d'Etudes Occitanes), pages 141–156, Albi, 2020. Centre d'Etude de la Littérature Occitane.

Ander Corral, Igor Leturia, Aure Seguier, Michael Barret, Benaset Dazeas, Philippe Boula de Mareuil, 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 ¾Language Resources and Evaluation Conference Marseille 11 16 May 2020¿ , pages 53 60. European Language Resources Association (ELRA), 2020.

Marta R. Costa-jussa, James Cross, Onur Celebi, Maha Elbayad, Kenneth Hea eld, Kevin He ernan, 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 Ho man, 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 Guzman, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Sa yyah Saleem, Holger Schwenk, and Je 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 .

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 Linden. HeLI-OTS, o -the-shelf language identi er for text. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3912 3922, 2022.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for e cient text classi cation. 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 Ljubesic and Filip Klubicka. f bs,hr,srgWaC - 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 o -the-shelf language identi cation tool. In Proceedings of the ACL 2012 system demonstrations, pages 25 30, 2012.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamenca Poujade, and Jean Sibille. A four-dialect treebank for Occitan: Building process and parsing experiments. InProceedings 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 Suarez, Benoit Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Banski, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lungen, and Caroline Iliadi, editors, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardi , 22nd July 2019, pages 9 16, Mannheim, 2019. Leibniz-Institut fur 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 Guzman. 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 presentation et de synthese. In Dominique Caubet, Salem Chaker, and Jean Sibille, editors,Les langues de France et leur codi cation. 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 .

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

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