Adaptation of models for parsing of Old Gascon

Natasha Romanova[1], Rayan Ziane[2], Barbara Francioni[1,3]
(1) Centre de Recherches Inter-langues sur la Signification en COntexte (CRISCO), Caen, France
(2) Laboratoire Ligérien de Linguistique (LLL), Orléans, France
(3) Università degli studi Aldo Moro, Bari, Italy
natalia.romanova@unicaen.fr,rayan.ziane@univ-orleans.fr,
barbara.francioni@outlook.it

MOTS-CLÉS : parsing automatique ; adaptation de modèles ; variation linguistique
KEYWORDS: automatic parsing ; models finetuning ; linguistic variation

1 Introduction

Syntactic parsing of a corpus in a poorly resourced language for which no model or annotated corpora exist is a difficult and time-consuming task, requiring expert intervention. This is even more relevant for corpora of ancient languages. In this paper, we present a corpus of Old Gascon (CorAG), that has been manually annotated in the Universal Dependencies (UD) framework. We further provide results of the evaluation of different models based on existing Romance corpora for the annotation of our data in a bootstrapping scenario.

2 Presentation of CorAG

2.1 Old Gascon

Old Gascon is a romance variety spoken and written in the Middle Ages in Gascony, in the south-west of contemporary France. It is the precursor of today's Gascon dialect, considered a variety of contemporary Occitan. Old Gascon is distinguished by a number of phonetic, morphological and syntactic features, including the emergence of enunciative particle "que" to mark declarative main clauses (Ledgeway, 2020; Morin 2005, Pusch, 2001; Rohlfs, 1970; Romieu & Bianchi, 2005). Medieval Gascon mainly survives in texts of the legal genre (Glessgen, 2021).

2.2 Digital corpora of Occitan, medieval and modern

Occitan is a minoritized language. The UNESCO lists it among endangered languages (Moseley, 2010). In recent years, several important scholarly efforts have been made to create digital resources for modern and medieval varieties of Occitan.[1] For Modern Occitan, BaTelÒc text base[2] can be queried by word form, sequence of words and parts of a word (Bras & Vergez-Couret, 2016). A first PoS-tagged and lemmatised multi-dialect corpus of Modern Occitan was published as a dataset as part of RESTAURE corpus (Bras et al, 2018; Bernhard et al, 2018). CorpusArièja is PoS-tagged and lemmatized (Pujade et al, 2023). A first *treebank* for modern Occitan, Tolosa Treebank (TTB), was released in May 2025 in the Universal Dependencies collection (v. 2.16).[3] It is lemmatized and annotated in PoS and syntactic functions in the Universal Dependencies framework (de Marneffe et al, 2021). TTB contains material for four dialects of Occitan, including Gascon (Miletic et al, 2020). For Old Occitan, *Concordance de l'occitan médiéval* (COM2), accessible on CD-ROMS, with search possible by prefix and suffix, word form and sequence of words, remains a reference corpus (Ricketts et al, 2005). *Corpus dell'Antico Occitano (CAO)* is a corpus of diplomatic transcriptions and interpretative editions of troubadour chansonniers that can be searched by word form (Asperti et al, 2020). *Documents linguistiques galloromans* (DocLing) database contains some Old Occitan and Gascon texts searchable by word form and phonetic structure (Glessgen et al, 2021). The online *Corpus linguistique de l'ancien occitan gascon* gives access to editions of the charters but not the PoS annotation (Field, 2016).[4] The recently published COMETA corpus dataset offers new transcriptions of Old Occitan narrative texts and has PoS annotation in UD, partially manually checked, but no functions (Schöffel et al, 2025; Wiedner, 2025).

2.3 The corpus

Corpus d'Ancien Gascon (CorAG) is, to our knowledge, the only corpus of any variety of medieval Occitan annotated both in PoS and syntactic functions.[5] As of October 2025, it contains six medieval texts: *Coutumes et Privilèges de l'Entre-Deux-Mers* (1214-1342), *Coutume de Banières* (1251) and *Coutume de Banières* (1260), *Charte des Boucheries d'Orthez* (1270), *Charte d'Herrère* (1278) and *Les Fors Anciens de Béarn* (1460). Two of

[1] For full list of Occitan resources for linguistics and Natural Language Processing, see Bras&Vergez-Couret, 2024 and Bras et al, 2024.

[2] http://redac.univ-tlse2.fr/bateloc.

[3] https://github.com/UniversalDependencies/UD_Occitan-TTB

[4] At the time of writing the corpus was no longer accessible at the University of Maryland website (https://mllidev.umbc.edu/gascon/French/description/index.html).

[5] At the time of writing, we became aware that a new UD parsed corpus of troubadour poetry is in preparation by Mariagrazia Staffieri. The corpus will soon be made available via https://github.com/UniversalDependencies/UD_Old_Occitan-OOT.

the texts were published in the v. 2.16 release of the UD collection.[6] Automatic annotation with a model based on the Profiterole corpus (Prévost et al, 2024) was manually corrected and the model was retrained iteratively in a bootstrapping scenario (Peng et al, 2022).

3 Parsing experiments

The goal of the experiments presented below was to establish whether using an existing UD-annotated corpus of a Romance language for prefinetuning parsing models would improve parsing performance on our data. We used two corpora of medieval Romance languages and one corpus of Modern Occitan with a small proportion of modern Gascon data.

3.1 Experimental set-up

For training of models, we used BertForDeprel parser (Guiller, 2020), that is an implementation of Dozat & Manning, (2018) architecture, and FacebookAI/xlm-roberta-large[7] (Conneau et al, 2020), providing contextualized multilingual embeddings. Fine-tunings were performed with a batch size of 16, maximum sequence length 512, and 8 data-loading workers, using AdamW optimization (initial learning rate of 5e-5). Training was run for up to 100 epochs without early stopping. The same hyperparameter configuration and random seed were applied to all experiments to ensure comparability between all models using pre-finetuning corpora and CorAG for finetuning. CorAG corpus (version May 2025) was divided into two halves, one used for finetuning and one for validating of models (50% train, 25% dev and 25% train). For the experiments described below, the groups of sentences used were selected to be of different sizes to avoid biases linked to parsing performance on longer sentences (Ziane & Romanova, 2024).[8]

Stage	Training Data Romance Corpus	CorAG "From Scratch"	Strategy
E0 (baseline)	PRFT or OI or TTB		Zero-shot on CorAG
E1	E0 + 100 sent CorAG	100 sent CorAG	Finetuning 1
E2	E0 + 200 sent CorAG	200 sent CorAG	Finetuning 2
E3	E0 + 300 sent CorAG	200 sent CorAG	Finetuning 3
E4	E0 + 400 sent CorAG	400 sent CorAG	Finetuning 4
E5	E0 + 500 sent CorAG	500 sent CorAG	Finetuning 5
E6	E0 + 607 sent CorAG	607 sent CorAG	Finetuning 6

TABLE 1 : Types of models trained for the experiments.

[6] *Coutumes et Privilèges de l'Entre-Deux-Mers* and *Les Fors Anciens de Béarn*. The corpus can be downloaded from https://github.com/UniversalDependencies/UD_Old_Occitan-CorAG. The rest of the corpus will be available in future UD releases.

[7] https://huggingface.co/FacebookAI/xlm-roberta-large.

[8] This version of the CorAG corpus can be downloaded at https://zenodo.org/records/17285100.

We trained four sets of models. Firstly, "from scratch" models, using CorAG sentences (100, 200, 300, 400, 500 and 607 sentences). Then models pre-finetuned on each of the three UD corpora of Romance languages: Tolosa Treebank (TTB, modern Occitan), Profiterole (PRFT, Medieval French) and Old Italian (OI) (Corbetta et al, 2023). Each of these models was finetuned with 100, 200, 300, 500 and 607 sentences of CorAG (Table 1). The Labelled Attachment Score of the resulting models was then tested on the remaining half of CorAG reserved for testing (see Section 3 Results).

3.2 Finetuning with Romance corpora

We ran two series of experiments. Firstly, we trained models with all of the corpus data available in the UD collection for the three Romance corpora. Secondly, we repeated the experiment scaling down OI and PRFT data to the number of tokens in the smallest corpus, TTB (Table 2). For OI, PRFT and TTB the UD dev and test files were concatenated to constitute a train corpus for our experiments.

	Number Of Tokens In The Corpus	**Number Of Tokens Scaled Down To Ttb**
CorAG	46,333 (train 20,599/ dev 12,521/ test 13,213)	N/A
TTB	26,117 (train 24,551/dev 1,566)	N/A
OI	124,988 (train 112,493/dev 12,495)	26,125 (train 24,559/dev 1,566)
PRFT	237,913 (train 214,185/dev 23,728)	26 131 (train 24,565/dev 1,566)

TABLE 2 : Corpora size in tokens.

3.3 Results

Figures 1 and 2 illustrate the progression of LAS scores on the test set during bootstrapping and the convergence of baseline models, measured through the standard deviation across pre-finetuning corpora.

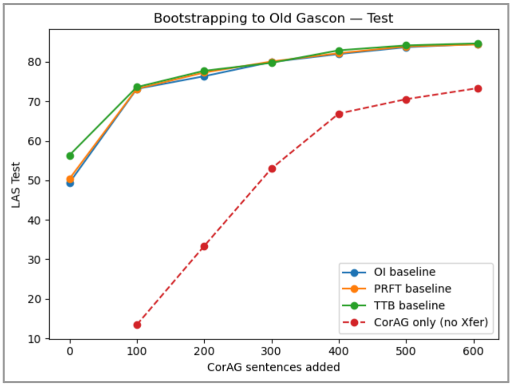

Figure 1 : Bootstrapping progression on Old Gascon CorAG test set

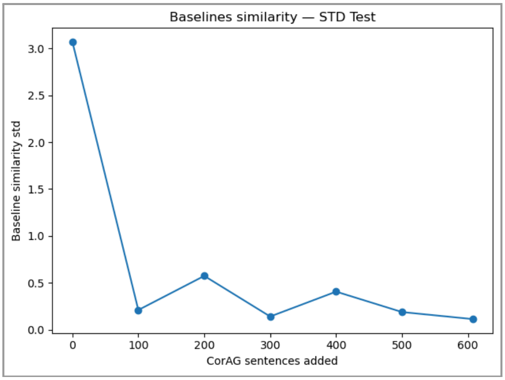

Figure 2 : Baseline convergence (standard deviation) on the test set

In Annex 1 Table 3, we present results of the evaluation of models trained "from scratch" with CorAG data; in Annex 1 Table 4, results for TTB; in Annex 1, Tables 5 and 6, results for OI and PRFT entire corpus and corpus scaled down to the number of tokens in TTB. For each table, results on both dev and test set are provided.

4 Discussion

We observe that all romance corpora used for pre-finetuning models give remarkably similar results irrespective of the size in tokens (PRFT has ten times more tokens than TTB). Around 50%-55% LAS "zero shot", with TTB giving the best performance despite its smaller size. Performances rise to around 75% LAS with further finetuning on 100 CorAG sentences and to 85% after finetuning with 600 sentences. Only 300 CorAG sentences are needed to achieve the same performance as training a zero-shot model on a large Romance corpus, yet with further addition of lots of 100 sentences that performance does not rise as sharply as when a pre-training corpus is used. We therefore recommend using a pretraining corpus of a similar language for parsing of a new language (here the modern state of the language, despite considerable difference in vocabulary, gives slightly better performance in zero shot, 56% LAS for TTB against 49% for IO and 50% for PRFT with the same number of tokens).

5 Future work

In future, we plan to 1) explore using the number of sentences and sentence complexity and not just the number of tokens for the constitution of training corpora; 2) identify the minimal number of tokens for each of the Romance corpora to achieve the 50% LAS benchmark in zero shot; 3) experiment with mixing training corpora; 4) test the developed protocol on a non-Romance corpus; 5) conduct error-analysis of different models.

6 Acknowledgements

This work was funded by the Institut Universitaire de France and Normandy Region as part of projects led by Professor Pierre Larrivée at the University of Caen. We would like to thank Professor Martin Glessgen (University of Zurich) and his team for letting us use their transcriptions of *Coutume de Banières* (1251 and 1260), *Charte des Boucheries d'Orthez* (1270) and *Charte d'Herrère* (1278). The Tolosa Treebank team's help and advice have been invaluable in setting up the annotation process. The Direction du Système d'Information at the University of Caen supported this work by giving us access to the computational resources to train models. Khensa Daoudi provided assistance with uploading CorAG data in the UD collection.

References

Asperti, S., Carapezza, F., Careri, M., Di Girolamo, C., Di Luca, P., Lachin, G., Meliga, W. & Squillacioti P. (2020). *Corpus dell'Antico Occitano (CAO) 2*. Online resource. caodiweb.ovi.cnr.it/(S(mfjltjocunk0htxhqfysv3mr))/CatForm01.aspx

Bernhard, D., Ligozat, A.L., Martin, F., Bras, M., Magistry, P., Vergez-Couret, M., Steiblé, L., Erhart, P., Hathout, N., Huck, D., Rey, C., Reynés, P., Rosset, S., Sibille, J. & Lavergne, T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. *11th edition of the Language Resources and Evaluation Conference (LREC), May 2018*. Miyazaki, Japan. 3917-3924. https://aclanthology.org/L18-1619.pdf

Bras, M. & Vergez-Couret, M. (2024). Traitement automatique de l'occitan. In Esher, L. & Sibille, J. (Eds.) *Manuel de linguistique occitane*. Berlin, De Gruyter. 543-561

Bras, M., Esher, L., Sibille, J. & Vergez-Couret, M. (2018). *Annotated Corpus for Occitan (RESTAURE)*. Dataset. https://doi.org/10.5281/zenodo.1182949

Bras, M., Vergez-Couret, M. & Sibille, J. (2024). Corpus et bases de données. In Esher, L. & Sibille, J. (Eds.) *Manuel de linguistique occitane*. Berlin, De Gruyter, pp. 523-542

Bras, M. & Vergez-Couret, M. (2016). BaTelÒc: A text base for the Occitan language. In Ferreira V. & Bouda P. (Eds.) *Language Documentation and Conservation in Europe*. *Special Publication No. 9 of the Journal Language Documentation & Conservation*. Honolulu, University of Hawai'i Press. 133-149. https://hal.science/hal-00987241

Conneau,A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th annual meeting of the association for computational linguistics. association for computational linguistics.* 8440–8451. https://aclanthology.org/2020.acl-main.747.pdf

Corbetta, C., Passarotti, M., Ceccini F. M. & Moretti, G. (2023). Highway to Hell. Towards a Universal Dependencies Treebank for Dante Alighieri's Comedy. Boschetti, F., Lebani, G., agnini, B. & Novielli, N. (Eds.), *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023).* Associazione italiana di linguistica computazionale (AILC). https://aclanthology.org/2023.clicit-1.20.pdf

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2). 255-308. https://doi.org/10.1162/coli_a_00402

Dozat T. & Manning, C.D. (2017). Deep Biaffine Attention for Neural Dependency Parsing. *ICRL (International Conference on Learning Representations)*. https://arxiv.org/abs/1611.01734

Field, T.T. (ed). (2016). *Corpus linguistique de l'ancien occitan gascon*. Baltimore. Online resource. https://mllidev.umbc.edu/gascon/French/description/index.html

Glessgen, M. (2021) Pour une histoire textuelle du gascon médiéval. *Revue de linguistique romane*, 85. 325-384

Glessgen, M. with Duval, F., Videsott, P. & Carles, H. (eds). (2024). *DocLing: Documents linguistiques galloromans*. Electronic edition. https://gallrom.linguistik.uzh.ch/#/docling

Guiller, K. (2020). *Analyse syntaxique automatique du pidgin-créole du Nigeria à l'aide d'un transformer (BERT): Méthodes et Résultats.* Mémoire de Master, Sorbonne Nouvelle.

Ledgeway, A. (2020). Variation in the Gallo-Romance left periphery: V2, complementizers, and the Gascon enunciative system. Wolfe, S. and Maiden M. (eds) *Variation and Change in Gallo-Romance Grammar*, 71-99. Oxford, Oxford University Press

Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C. & Sibille J. (2020). A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*. International Committee on Computational Linguistics (ICCL). Barcelona, Spain (Online). 140–149. https://aclanthology.org/2020.vardial-1.13

Morin, A. (2005) Syntaxe de la particule *que* en gascon, *Linguistica Occitana*, 4. 60-67

Moseley, C. (2010). Atlas des langues en danger dans le monde. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000189451

Peng, Z., Gerdes, K., & Guiller, K. (2022). Pull your treebank up by its own bootstraps. In L. Becerra, B. Favre, C. Gardent, & Y. Parmentier (Eds.) *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*. 139-153. https://hal.science/hal-03846834

Poujade, C., Bras, M., Urieli, A. (2024). CorpusArièja: Building an Annotated Corpus with Variation in Occitan. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, ELRA and ICCL. Torino, Italia. 66–71. https://aclanthology.org/2024.sigul-1.9/

Prévost, S., Grobol, L., Dehouck, M., Lavrentiev, A., & Heiden, S. (2024). Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval. *Corpus*, *25*. https://doi.org/10.4000/corpus.8538

Pusch, C.D. (2000). The attitudinal meaning of preverbal markers in Gascon: Insights from the analysis of literary and spoken language data. In Andersen G. & Fretheim T. (Eds). *Pragmatic markers and propositional attitude*. Amsterdam/Philadelphia, Benjamins. 189-206

Ricketts, P.T., Reed, A., Akehurst, F.R.P., Hathaway, J. & Van der Horst, C.H.M. (2005). *Concordance de l'Occitan médiéval : COM 2 : les troubadours, les textes narratifs en vers = The concordance of medieval Occitan : the troubadours, narrative verse.* Turnhout, Belgique: Brepols Publishers. CD-ROM

Romieu M. & Bianchi A. (2005). *Gramatica de l'occitan gascon contemoranèu*, Bordeaux, Presses Universitaires de Bordeaux

Rohlfs, G. (1977). *Le Gascon. Études de philologie pyrénéenne*. Tübingen, Max Niemeyer Verlag, 1977

Schöffel, M., Wiedner, M., Arias, E.G., Ruppert P., Heumann C. & Assenmacher M. (2025) Modern Models, Medieval Texts: A POS Tagging Study of Old Occitan. *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics. 334–349. https://aclanthology.org/2025.nlp4dh-1.30.pdf

Wiedner, M. (ed.) (2025). COMETA : Corpus de l'occitan médiéval comparatif et annoté: Provence et Languedoc. Dataset. https://zenodo.org/records/15300719

Ziane, R. & Romanova, N. (2025). Pistes pour l'optimisation des modèles de parsing syntaxique. *Proceedings of LIFT 2-2024.* Orléans. https://lift2-2024.sciencesconf.org/590561/document

Annex 1: Results tables

Models	LAS on CorAG dev	LAS on CorAG test
CorAG_E1	14.46607477	13.44264892
CorAG_E2	33.63199479	33.34125475
CorAG_E3	53.49026635	52.97053232
CorAG_E4	67.65496457	66.88846641
CorAG_E5	71.62987701	70.53231939
CorAG_E6	74.26895821	73.3365019

TABLE 3 : “From scratch” models for CorAG

Models	LAS on CorAG dev	LAS on CorAG test
TTB_E0	56.30039912	56.35297845
TTB_E1	74.16306915	73.58998733
TTB_E2	79.68559094	77.71704689
TTB_E3	81.92555184	79.72908745
TTB_E4	84.067769	82.89765526
TTB_E5	85.30585648	84.14131812
TTB_E6	86.16111428	84.65621039

TABLE 4 : Models trained on TTB corpus

	Models trained on OI corpus		Models trained on OI corpus scaled down to TTB	
Models	LAS on CorAG dev	LAS on CorAG test	LAS on CorAG dev	LAS on CorAG test
OI_E0	51.55982732	51.56051965	49.28728517	49.39797212
OI_E1	75.80842225	75.49904943	73.51144416	73.12262357
OI_E2	80.80149874	79.2617237	78.3416144	76.3387199
OI_E3	82.67492058	81.16286439	80.60601124	79.94296578
OI_E4	85.15924086	83.72940431	83.53017838	81.94708492
OI_E5	85.63166897	84.2126109	84.71124868	83.68187579
OI_E6	86.38918303	85.30576679	85.69683147	84.50570342

TABLE 5 : Models trained on OI corpus

	Models trained on PRFT corpus		Models trained on the PRFT corpus scaled down to TTB	
Models	LAS on CorAG dev	LAS on CorAG test	LAS on CorAG dev	LAS on CorAG test
PRFT_E0	55.45328663	55.61628644	50.43577421	50.39607098
PRFT_E1	76.51706443	76.13276299	74.11419728	73.17015209
PRFT_E2	80.34536124	79.64987326	78.35790503	77.26552598
PRFT_E3	83.22065651	81.68567807	81.70562841	80.06970849
PRFT_E4	85.26512992	83.76108999	83.90486275	82.16888466
PRFT_E5	86.18555022	84.11755387	84.94746274	83.87991128
PRFT_E6	86.6987049	85.30576679	85.92490022	84.37896071

TABLE 6 : Models trained on PRFT corpus