Building a treebank for Occitan: what use for Romance UD corpora?

Aleksandra Miletic[1] Myriam Bras[1] Louise Esher[1] Jean Sibille[1]
Marianne Vergez-Couret[2]

[1]CLLE-ERSS UMR 5263, CNRS & University of Toulouse Jean Jaurès, France

[2]FoReLLIS (EA 3816), University of Poitiers, France

Universal Dependencies Workshop, 30 August 2019

UNIVERSITÉ TOULOUSE Jean Jaurès

Clle UMR 5263 COGNITION, LANGUES, LANGAGE, ERGONOMIE

Interreg POCTEFA LINGUATEC UNIÓN EUROPEA UNION EUROPÉENNE

UNION EUROPÉENNE Fonds Européen de Développement Régional

Outline

1. Introduction
2. Resources and tools
3. Delexicalized parsing: experiments and results
4. Manual annotation analysis
5. Conclusions and future work

Introduction

Goal

Initiate the building of the first dependency treebank for Occitan

- relatively low-resourced Romance language: no syntactically annotated data
- → need to simplify and accelerate manual annotation
- **Constraint:** Less time-consuming than full manual annotation

Methodology

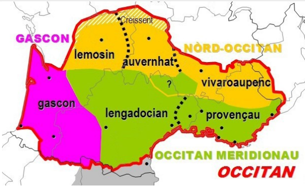
Direct delexicalized cross-lingual parsing using Romance UD treebanks

- Train a parser on these treebanks and use the models to parse Occitan
- Use best models to provide human annotators with an initial annotation

Focus

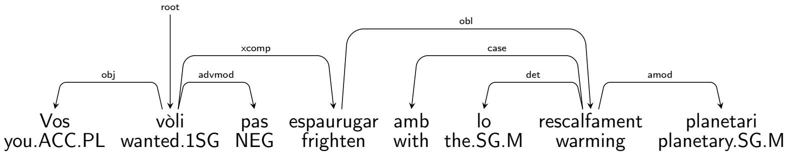
Effects of cross-lingual annotation on the work of human annotators in terms of annotation speed and ease

Occitan



- Romance language
- South of France, some areas of Italy and Spain
- Pro-drop, free word order
- Relatively under-resourced:
  - morphological lexicon (850K entries): Vergez-Couret (2016)
  - POS-tagged corpus (15K tokens): Bernhard et al. (2018)
- Rich diatopic variation, no standard dialect

(1)



Vos	vòli	pas	espaurugar	amb	lo	rescalfament	planetari
you.ACC.PL	wanted.1SG	NEG	frighten	with	the.SG.M	warming	planetary.SG.M

‘I didn’t want to scare you with global warming.’

Direct delexicalized cross-lingual parsing

Parsing a low-resourced language with insuffcent treebank data:

- Training a delexicalized model on a related language
  - training based typically on POS tags and morphosyntactic traits
  - tokens and lemmas (i.e., lexical information) are ignored
- Using the delexicalized model to parse the target language

Essential condition: harmonized annotations between the source and the target corpus
(cf. McDonald et al., 2011, 2013) → utility of the UD corpora
Already used in similar experiments: Lynn et al. (2014) ; Tiedemann (2015) ; Duong et al. (2015)

Outline

1 Introduction

2 Resources and tools

3 Delexicalized parsing: experiments and results

4 Manual annotation analysis

5 Conclusions and future work

Resources and tools

Training corpora

- Universal Dependency Treebanks v2.3
- Catalan, French, Galician, Italian, Old French, Portuguese, Romanian and Spanish
- 14/23 available corpora: selected for content compatibility (no spoken language, no tweets) and annotation quality (manual annotation or conversion from manual annotation)
- No morphosyntactic traits, only one-level syntactic labels used

Test sample

- 1152 tokens of newspaper texts (Languedocian and Gascon dialects)
- Gold-standard UD POS tags converted from an existing Occitan corpus based on the GRACE tagset (Miletic et al., 2019)
- Manual gold-standard syntactic annotation (one-level labels)

Parser

- Talismane NLP suite (Urieli, 2013) (SVM algorithm used here)

Outline

1. Introduction
2. Resources and tools
3. Delexicalized parsing: experiments and results
4. Manual annotation analysis
5. Conclusions and future work

Parsing experiments setup

Three-step evaluation:

1. Establishing the baseline: training models on each corpus and testing them on their designated test sample
2. Intrinsic evaluation: testing all models from **Step 1** on the manually annotated Occitan sample
3. Extrinsic evaluation: parsing a new Occitan sample using the best performing models from **Step 2**
    - Manual annotation speed and ease evaluation
    - Recurrent error analysis based on annotator feedback

Step 1: Baseline evaluation

Corpus	Train size	Test size	LAS	UAS
ca_ancora	418K	58K	77.82	82.20
es_ancora	446K	52.8K	76.75	81.29
es_gsd	12.2K	13.5K	74.88	78.81
fr_partut	**25K**	**2.7K**	**82.41**	**84.60**
fr_gsd	364K	10.3K	78.51	81.81
fr_sequoia	52K	10.3K	78.29	80.71
fr_ftb	470K	79.6K	68.93	73.08
gl_treegal	16.7K	10.9K	73.91	78.79
it_isdt	**294K**	**11.1K**	**81.03**	**84.19**
it_partut	**52.4K**	**3.9K**	**82.66**	**85.22**
ofr_srcmf	136K	17.3K	69.41	79.09
pt_bosque	222K	10.9K	77.41	81.27
pt_gsd	**273K**	**33.6K**	**80.2**	**83.2**
ro_rrt	185K	16.3K	71.87	78.92
ro_nonstandard	155K	20.9K	65.59	75.45
es_ancora+gsd	458.2K	66.3K	73.14	78.24
fr_partut+gsd+sequoia	441K	23.3K	73.69	77.57
fr_partut+gsd+sequoia+ftb	911K	102.9K	74.87	78.55
it_isdt+partut	**346.4K**	**15K**	**81.78**	**84.66**
pt_bosque+gsd	495K	44.5K	76.09	81.47
ro_nonstand+rrt	340K	37.2K	67.21	76.06

LAS: 65.59 (ro_nonstandard) – 82.41 (fr_partut)
UAS: 73.08 (fr_ftb) – 85.22 (it_partut)
Merging corpora didn't improve best individual result per language. Merging = annotation incoherence?
All models tested in **Step 2**

Step 2: Evaluation on the Occitan sample

Train corpus	LAS	UAS
it_isdt	71.6	76.0
it_isdt+partut	71.3	75.9
fr_partut+gsd+sequoia	70.8	75.7
fr_gsd	70.4	75.9
pt_bosque	70.0	75.3
it_partut	69.7	74.1
fr_partut+gsd+sequoia+ftb	69.6	74.4
fr_partut	69.4	74.6
es_ancora+gsd	69.1	74.9
es_ancora	69.0	75.3
gl_treegal	68.7	73.4
ca_ancora	68.6	75.2
fr_sequoia	68.6	73.3
es_gsd	67.8	73.4
fr_ftb	67.4	72.5
ro_rrt	67.1	72.2
ro_nonstand+rrt	66.6	72.0
pt_bosque+gsd	66.4	74.3
pt_gsd	63.1	73.3
ro_nonstand	60.2	72.7
ofr_scmrf	59.2	66.0

Test: manually annotated Occitan sample (1000 tokens)
LAS: 59.2 (ofr_scmrf) – 71.6 (it_isdt)
UAS: 66.0 (ofr_scmrf) – 76.0 (it_isdt)

Top 5 models:

- 3 based on French and Portuguese (not close to Occitan)
- All based on large corpora (smallest: 222K tokens)
- Smallest loss compared to baseline: fr_partut+gsd+sequoia. Merging = robustness?

Step 3: Parsing new texts in Occitan

Which model is the most useful as a pre-annotation tool for human annotators?

Setup: parse test sample → filter dependencies → submit to human annotators → measure annotation speed

Models: best model for each language among top 5 from Step 2: it_isdt, fr_partut+gsd+sequoia, pt_bosque

Test sample: 3 x 300 tokens of literary text with gold-standard POS

Dependency filter: parser's decision probability score >0.7

Results:

Sample	Model	Size	Coverage at	LAS	UAS	Man.
		(tokens)	prob. >0.7	(filtered deps)		time
viaule1	it_isdt	352	84.7 %	81.2	88.7	30'
viaule2	fr_partut+gsd+sequoia	325	86.5 %	74.8	85.2	32'
viaule3	pt_bosque	337	88.3 %	84.5	89.4	21'

- Comparable results for the three models
- Mean annotation speed increase: 340 tok/h → 730 tok/h
- Positive ergonomic effect reported by the annotator: preannotation (although partial) makes the task less daunting compared to dealing with a blank text

Outline

1 Introduction

2 Resources and tools

3 Delexicalized parsing: experiments and results

4 Manual annotation analysis

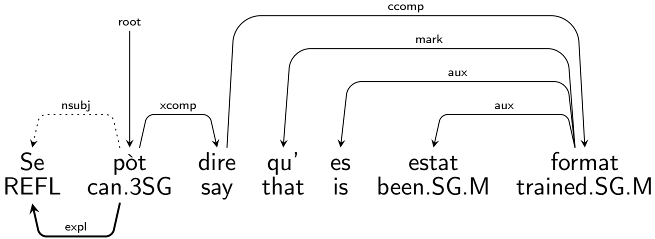
5 Conclusions and future work

Step 3: Recurrent error analysis

Reflexive clitics:

- POS=PRON, no morphosyntactic traits in the Occitan sample → indistinguishable from other pronouns
- Most often annotated as nsubj, obj or iobj rather than expl

(2)



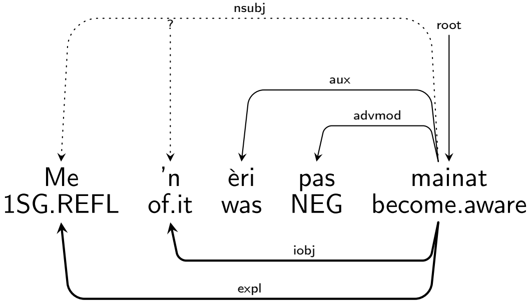
'You could say that he has been trained.'

Step 3: Recurrent error analysis

Pronoun clusters:

- Sentence-initial `PRON` often annotated as `nsubj`
- Other `PRON`s in the cluster without annotation (filtered out)
- Can be explained for the model based on French (obligatory subject), but not for the other two: Italalian and Portuguese allow for subject dropping

(3)



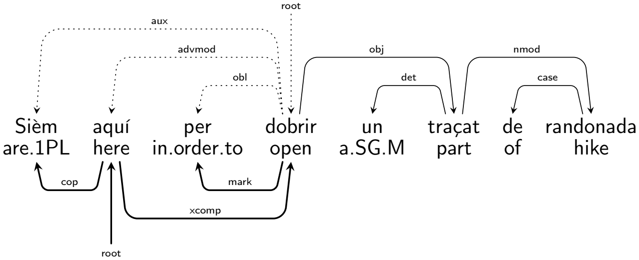
'I hadn't noticed it.'

Step 3: Recurrent error analysis

Auxiliaries vs copulas:

- Copula *èsser* 'to be' annotated as aux in proximity of a main verb
- Creates error propagation (copula dependents, root identification) requiring time-consuming corrections

(4)



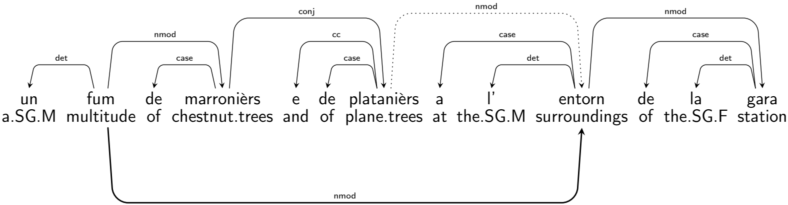
'We are here to open a part of a hike.'

Step 3: Recurrent error analysis

Long-distance dependencies:

- All models produced relatively few long-distance dependencies with relatively low accuracy
- Well-known issue in parsing

(5)



'a multitude of chestnut trees and plane trees around the station'

Outline

1. Introduction
2. Resources and tools
3. Delexicalized parsing: experiments and results
4. Manual annotation analysis
5. Conclusions and future work

Conclusions and future work

Recap

- 14 UD corpora in 8 Romance langauges used to train 21 models
- Models tested on a manually annotated Occitan sample
- 3 of the 5 best performing models used to preannotate new texts
- Manual annotation speed increase from 340 tok/h to 730 tok/h

New directions

- Improving `PRON` and `AUX` processing: adding `PronType` and `VerbForm`
- Given output consistency, test combining the corpora of the 3 models

General conclusions

- Clear positive impact of delexicalized cross-lingual parsing on the manual annotation of Occitan: speed increase, but also positive ergonomic effect reported by the annotator
- Reasonably quick and straightforward process

Acknowledgments

This work is part of the Linguatec Project financed by the European Regional Development Fund.

Interreg
POCTEFA
LINGUATEC

UNIÓN EUROPEA
UNION EUROPÉENNE

UNION EUROPÉENNE
Fonds Européen de Développement Régional

References

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, et al. Corpora with part-of-speech annotations for three regional languages of france: Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*, 2018.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 845–850, 2015.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, 2014.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics, 2011.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97, 2013.

Aleksandra Miletic, Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, and Marianne Vergez-Couret. Transformation d'annotations en parties du discours et lemmes vers le format universal dependencies: étude de cas pour l'alsacien et l'occitan. In *Actes du Traitement Automatique de Langage (TALN2019)*, 2019.

Jörg Tiedemann. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349, 2015.

Assaf Urieli. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université Toulouse le Mirail-Toulouse II, 2013.

Marianne Vergez-Couret. Description du lexique Loflòc. Technical report, 2016.

Contact information

Aleksandra Miletic

- aleksandra.miletic@univ-tlse2.fr
- aleksandramiletic1207@gmail.com
- www.linkedin.com/in/aleksandra-miletic-1207