# Pistes pour l'optimisation de modèles de parsing syntaxique

Rayan Ziane[1, 2] Natasha Romanova[1]
(1) Centre de Recherches Inter-langues sur la Signification en COntexte (CRISCO), Caen, France
(2) Laboratoire Ligérien de Linguistique (LLL), Orléans, France
rayan.ziane@univ-orleans.fr, natalia.romanova@unicaen.fr

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

## 1 Introduction

Avec l'avènement de technologies d'apprentissage profond, toute analyse syntaxique automatique d'un corpus textuel (dans le but de constitution de *treebanks* exploitables pour la recherche en syntaxe) présuppose l'utilisation de "systèmes d'annotation" (Renwick & Kraif 2024). Un système d'annotation consiste d'un analyseur syntaxique automatique (*parseur*, un logiciel), un gros modèle de langue (*Large Language Model*, ou LLM, issu d'un apprentissage non-supervisé) et un corpus d'entraînement (une collection de phrases annotées dans un formalisme pris en charge par le *parseur*) le plus adapté au corpus cible, le tout servant à produire un modèle d'analyse syntaxique (Figure 1). Le corpus d'entraînement est utilisé pour l'entraînement supervisé ; le schéma d'annotation (en l'occurrence, parsing et tagging dans le système Universal Dependencies (UD), de Marneffe et al., 2021) et les traits linguistiques du corpus d'entraînement sont acquis à cette étape. Quatre plans d'action sont alors envisageables et se répartissent en deux groupes :

A. Annotation purement automatique

1) analyser automatiquement le corpus cible complet et évaluer la performance du modèle (ou plusieurs modèles résultants de différents systèmes) par rapport à un échantillon du corpus corrigé manuellement; l'évaluation servira à alerter les utilisateurs au taux de fiabilité des annotations ;

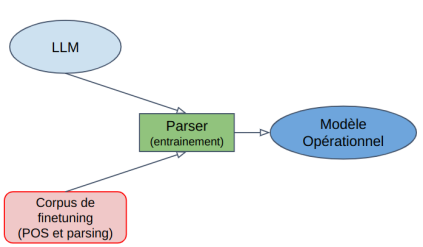


FIGURE 1: Système d'annotation syntaxique de base

B. Annotation automatique corrigée

2) analyser automatiquement le corpus entier et corriger manuellement (et/ou à l'aide de règles);

3) utiliser les prédictions de multiples systèmes d'annotation pour prioriser les corrections en utilisant la convergence des systèmes de parsing comme indicateur de fiabilité (Renwick & Kraif 2024);

4) adopter une approche agile à l'annotation (=par *bootstrapping*) qui permette de réentraîner (=*finetuner*, ou peaufiner) le système d'annotation via une série d'entraînements et campagnes de correction en ajoutant progressivement des phrases corrigées du corpus cible au corpus d'entraînement (que nous appellerons les "sous-corpus de finetuning") afin d'adapter le système de parsing au corpus analysé (Miletić, 2018; Peng et al., 2022) (Figure 2).

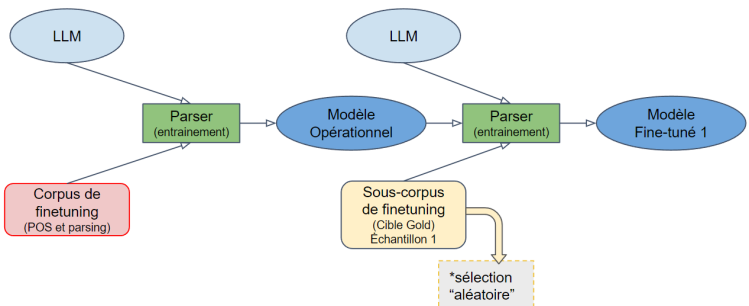


FIGURE 2 : Adaptation du système d'annotation de base par finetuning

## 2 Problématique

Il est désormais admis que la performance des systèmes d'annotation dépend fortement des caractéristiques du corpus utilisé pour l'entraînement (Grobol et al., 2021; Guibon et al., 2015; Seghier et al., 2023). La distance en ce qui concerne la variété linguistique (dont les variétés diachronique et diatopique) ainsi que celle de genre ou de registre a un impact sur la performance. La sélection du corpus d'entraînement requiert donc, autant que possible, la réduction de cette distance, dans la mesure où les corpus annotés et vérifiés sont disponibles pour l'entraînement. L'attention portée à la qualité du corpus d'entraînement pour l'apprentissage supervisé, néanmoins, peut être étendue aux autres étapes du processus d'annotation agile afin d'optimiser l'injection des caractéristiques du corpus cible dans le système de parsing. Nous avançons deux hypothèses : 1) que la qualité et échantillonnage des "sous-corpus de finetuning" (post-apprentissage du schéma d'annotation) peut augmenter les performances du système d'annotation lors d'entraînement progressifs; 2) que l'introduction d'une étape supplémentaire avant l'acquisition du schéma d'annotation peut également apporter un gain de performance. Ces deux propositions sont présentées dans Figure 3.

L'objectif de cette étude est double: 1) commencer à formuler des recommandations d'échantillonnage des "sous-corpus de finetuning" pour améliorer les performances du système d'annotation et limiter les interventions humaines (en utilisant l'exemple de la taille de la phrase, pertinent pour l'annotation des corpus écrits, notamment en diachronie) b) trouver des pistes pour la réutilisation et mise à profit des corpus existants partiellement annotés (notamment en POS).

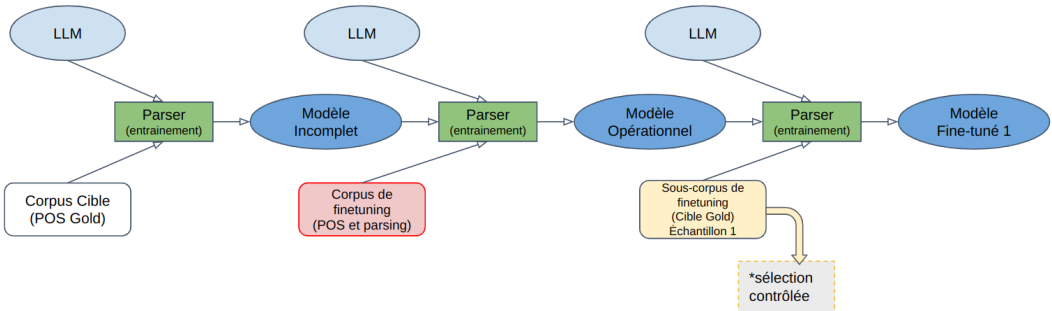


FIGURE 3 : Proposition de révision du processus de *bootstrapping*

# 3 Méthodologie et résultats

Les deux expériences menées dans le cadre de cette étude utilisent des modèles basés sur les embeddings de BERT multilingue (Delvin et al., 2019), le corpus Profiterole (PRFT) (Prévot et al., 2024) et l'analyseur syntaxique BertForDeprel (Guiller, 2020). Le corpus cible est la transcription du texte juridique du 16ème siècle, le registre Crime I de l'île de Guernesey, produite et annotée dans le cadre du projet ANR-DFG MICLE (2021-2024), contenant 1271 phrases et 40996 tokens. Le corpus d'entraînement et le corpus cible présentent une variation diachronique et diatopique du français, mais également en genre et (en partie) "modalité d'écriture" (prose/vers). La longueur moyenne de la phrase dans le corpus d'entraînement est de 11.79 tokens contre 32.25 dans le corpus cible.

1. Longueur des phrases et finetuning

Dans la première expérience, nous avons d'abord divisé le texte de Guernesey (GNS) en deux moitiés, chacune équilibrée par nombre et longueur de la phrase : les corpus "train" et "test". Le premier a été divisé par groupes selon la longueur des phrases (court, moyen, long et mix) afin de tester si l'adaptation sur des phrases de différentes longueurs affecte les performances des modèles, lors de trois itérations de finetuning - avec 100, 150 et 200 phrases. La Figure 4 montre l'évolution de la performance des modèles, finetunings du modèle de base (uniquement PRFT) avec les douze échantillons répartis par longueur et nombre de phrases, testés sur le corpus "test" divisé en dix groupes de longueur croissante. On observe, indépendamment du groupe, que la performance générale est supérieure au modèle de base. Néanmoins, la répartition par longueur dans les "sous-corpus de

finetuning" laisse apparaître des tendances où l'utilisation des phrases longues semble plus pertinente pour l'adaptation du modèle, suivi de près par le groupe "mix". À l'inverse, les modèles entraînés sur les phrases courtes sont moins efficaces, notamment sur les phrases les plus longues ce qui pourrait s'expliquer par un effet de sur-spécialisation à ce type de phrase, la complexité syntaxique des phrases longues étant absente du "sous-corpus de finetuning".

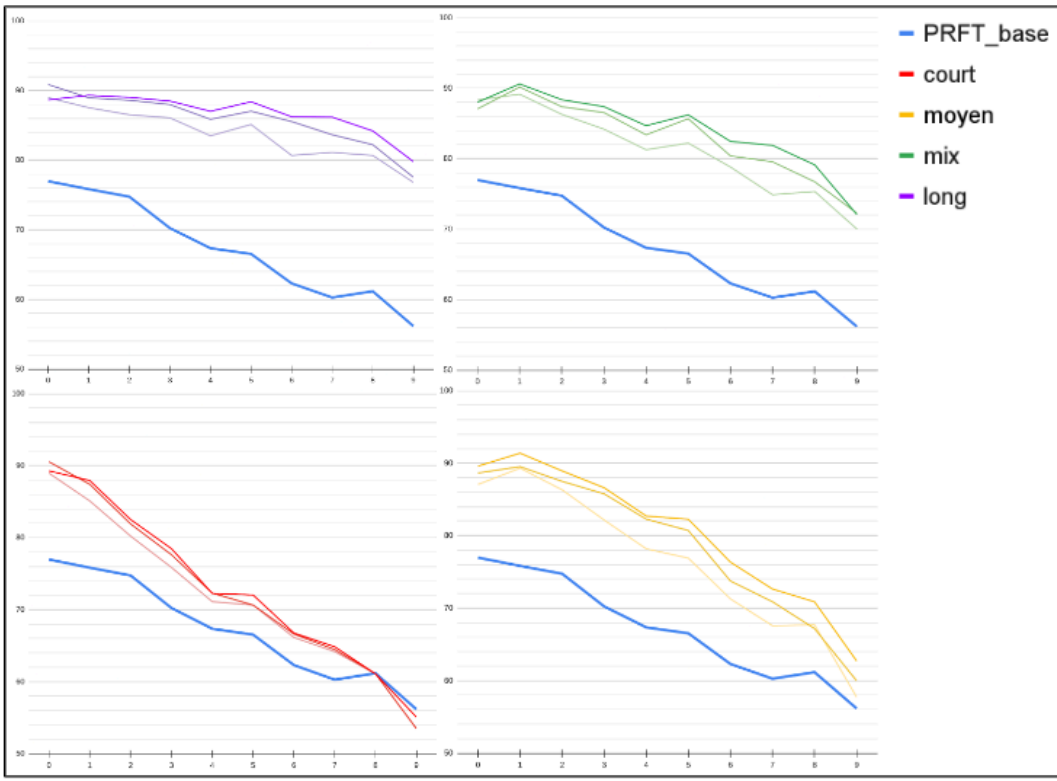


FIGURE 4 : Performance des échantillons répartis par longueur (LAS)

2. Pré-finetuning avec des annotations POS

Simultanément, nous avons exploré si le fait de fournir au modèle des bases d'informations morphosyntaxiques et/ou les traits linguistiques du corpus "cible" pouvait augmenter les performances lors du traitement de la diversité linguistique. Premièrement, quatre modèles ont été préalablement entraînés : (1) sur un corpus[1] non annoté contenant des textes de Guernesey (MICLE_noPOS), (2) sur le même corpus annoté en POS 'Gold' contenant des

[1] Corpus MICLE, ensemble de 14 textes juridiques entre le 13ème et 17ème siècles (422 117 tokens). Le corpus est consultable sur le portail TXM du Laboratoire CRISCO. URL: https://txm-crisco.huma-num.fr/txm/ (consulté 4 octobre 2024).

textes de Guernesey (MICLE_gPOS), (3) uniquement sur le texte de Guernesey annoté en POS 'Gold' (GNS_gPOS), et (4) uniquement sur le texte de Guernesey non annoté (GNS_noPOS). Les quatre ont, ensuite, été finetunés avec le corpus PRFT. Les performances des modèles résultants sont comparées à celle du "modèle de base", entraîné directement sur le corpus PRFT (Table 1). L'évaluation des modèles sur l'ensemble du texte de Guernesey confirme l'impact positif du pré-finetuning avec le corpus cible annoté en POS, en amont de l'entraînement avec un corpus contenant des fonctions et têtes syntaxiques. Le pré-finetuning avec seulement le texte (les traits linguistiques) par contre n'améliore pas les performances. Par ailleurs, les résultats très proches entre le pré-finetuning sur le texte seul et son ajout dans un corpus plus large montrent que la qualité des données utilisées semble être plus importante que la quantité utilisée.

| | **PRFT** | **MICLE_noPOS+PRFT** | **MICLE_gPOS+PRFT** | **GNS_noPOS+PRFT** | **GNS_gPOS+PRFT** |
|---|---|---|---|---|---|
| LAS | 69.72 | 68.81 | **73.61** | 69.00 | 72.84 |
| UAS | 76.02 | 75.37 | **78.65** | 75.44 | 78.17 |
| DEPREL | 83.75 | 83.48 | **86.88** | 83.41 | 86.07 |
| UPOS | 90.33 | 90.43 | **93.93** | 90.47 | 93.59 |

TABLE 1 : Performance des modèles pré-finetunés par rapport au modèle de base

3. Combinaison des deux méthodes

Dans une approche combinée, les deux méthodes décrites ci-dessus sont appliquées successivement: d'abord, un pré-finetuning avec annotations morphosyntaxiques (POS) pour doter le modèle d'une base linguistique adaptée (modèle GNS_gPOS), puis un fine-tuning sur des phrases de longueur variable après un entraînement avec le corpus PRFT. Les résultats (Table 2) montrent une amélioration des scores LAS avec cette combinaison après des finetunings sur 100 phrases. Ce gain confirme l'intérêt de renforcer le modèle avec des informations morphosyntaxiques, quand les données existantes le permettent, avant d'adapter le modèle sur des sous-corpus variés, offrant une meilleure spécialisation aux données du corpus cible.

| | **Sans Pré-finetuning** | **Avec Pré-finetuning** |
|---|---|---|
| PRFT | 69,72 | **72,84** |
| PRFT+100_court | 72,44 | **75,09** |
| PRFT+100_moyen | 78 | **79,82** |
| PRFT+100_mix | 81,86 | **83,96** |
| PRFT+100_long | 84,06 | **85,77** |

TABLE 2 : Performance des modèles finetunés par longueur de phrase avec et sans pré-finetuning

# 4 Conclusion

Dans le processus d'annotation syntaxique par *bootstrapping*, nous identifions deux étapes où une intervention plus contrôlée qui prend en compte la qualité plutôt que la quantité des données d'entraînement peut augmenter le taux de succès de l'annotation automatique. Pour maximiser la performance des modèles d'analyse syntaxique dans des contextes de diversité linguistique, nous recommandons d'intégrer les deux stratégies présentées dans cette étude :

1. Pré-finetuning avec des annotations morphosyntaxiques comme les POS, afin de doter le modèle de bases solides en le biaisant. Ceci est d'autant plus pertinent dans le contexte actuel où il existe de nombreux corpus annotés en POS qui nécessitent l'ajout de fonctions et têtes syntaxiques.
2. Finetuning sur un corpus de phrases variées, incluant des phrases longues, pour permettre au modèle de s'adapter à la complexité syntaxique.

Par ailleurs, nos résultats montrent qu'une approche hybride, mêlant ces deux dimensions, peut non seulement améliorer les performances sur des corpus diachroniques et diatopiques, mais également ouvrir une palette de possibilités selon les corpus disponibles afin d'y ajouter une couche d'annotation syntaxique en dépendances à moindre coût par la réutilisation de l'existant.

# Remerciements

Nous remercions les Archives du Tribunal de Guernesey (Guernsey Greffe) et l'ancien archiviste de l'île Daryl Ogier pour nous avoir accordé l'accès au manuscrit de Crime I. Nous remercions aussi la Direction du Système d'Information de l'Université de Caen pour avoir soutenu ce travail en fournissant l'accès aux ressources informatiques pour l'entraînement des modèles. Cette recherche a été menée dans le cadre du projet ANR Franco-allemand MICLE (2021-2024) et du projet AUTOMATED (2023-2024) financé par la région Normandie, sous la direction de Professeur Pierre Larrivée (CRISCO, Université de Caen).

# Références

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

DEVLIN, J., CHANG, M.-W., LEE, K., & TOUTANOVA, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

GROBOL, L., PRÉVOST, S., & CRABBÉ, B. (2021). Is Old French tougher to parse? *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, 27‑34. https://aclanthology.org/2021.tlt-1.3

CRIME I, 1563-1567. Guernsey Greffe (manuscrit).

GUIBON, G., TELLIER, I., PRÉVOST, S., CONSTANT, M., & GERDES, K. (2015). Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ? *TALN* *22*, http://www.atala.org/taln_archives/TALN/TALN-2015/taln-2015-long-017.pdf. https://hal.science/hal-01251006

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*.

MILETIC, A. (2018). *Un treebank pour le serbe : constitution et exploitations* [Phdthesis, Université Toulouse le Mirail - Toulouse II]. https://theses.hal.science/tel-02639473

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

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

RENWICK, A., & KRAIF, O. (2024). Annotation de textes d'états de langue anciens : pour le redéploiement de l'existant. *Corpus*, *25*. https://doi.org/10.4000/corpus.8286

SEGHIER, M., MILLOUR, A., & ANTOINE, J.-Y. (2023). Descripteurs Linguistiques et Caractérisation Objective des Catégories Textuelles. In K. Fort, C. Gardent, & Y. Parmentier (Éds.) (Éds.), *5èmes journées du Groupement de Recherche CNRS " Linguistique Informatique, Formelle et de Terrain "* (p. 106‑112). GdR LIFT, CNRS. https://hal.science/hal-04303374