République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Ferhat Abbas de Sétif 1

Faculté des Sciences

Thèse de Doctorat

En Sciences

Option : **Informatique**

Etude et exploitation des caractéristiques linguistiques de la langue arabe pour une nouvelle approche de traitement automatique du langage naturel

Présentée par

MAHDJOUBI Roussafi

Devant le jury :

Pr. Faiza TITOUNA	Université de Batna 2	Président
Pr. Mohamed Tayeb LASKRI	Université Kofi ANNAN, Guinée	Directeur de thèse
Pr. Abdelaziz LAKHFIF	Université Ferhat ABBAS, Sétif 1	Examinateur
Pr. Khaled REZEG	Université Mohamed KHIDER, Biskra	Examinateur
Pr. Chahrazed Mediani	Université Ferhat ABBAS, Sétif 1	Examinateur
Dr. Adel KERMI	Ecole Supérieure d'Informatique, Alger	Examinateur

Année Universitaire 2025-2026

Résumé

Cette thèse propose une approche hybride pour le traitement automatique de la langue arabe, fondée sur l'intégration explicite de ses caractéristiques morpho-syntaxiques au sein des modèles neuronaux. Elle s'inscrit dans le paradigme du NLP hybride, où les connaissances linguistiques structurées viennent enrichir et guider les représentations distribuées apprises par les architectures neuronales profondes.

Nous introduisons LiPAr (Linguistic Prompting for Arabic), un cadre méthodologique novateur qui permet l'injection contrôlée d'informations morpho-syntaxiques dans les modèles de langage. Ce cadre est évalué sur deux tâches fondamentales du TALN arabe : l'étiquetage morpho-syntaxique (POS) et la reconnaissance d'entités nommées (NER).

Les expérimentations ont été conduites sur trois modèles pré-entraînés (mBERT, AraBERT et MARBERT). Nous comparons systématiquement une configuration de base à une configuration enrichie intégrant 32 traits morpho-syntaxiques (genre, nombre, cas, définitude, etc.), extraits à l'aide de l'analyseur linguistique Stanza. Les résultats, obtenus sur les corpus UD_Arabic-PADT pour le POS et ANERcorp pour le NER, montrent que cette intégration produit des améliorations systématiques et statistiquement significatives pour la tâche de POS, avec des gains allant de +1,5 à +6,6 points F1 selon le modèle. Les effets observés en NER se révèlent plus variables et étroitement dépendants des caractéristiques de pré-entraînement de chaque modèle.

Les analyses quantitatives et qualitatives révèlent des comportements différenciés selon les architectures : mBERT bénéficie particulièrement de l'apport des traits en POS, ce qui s'explique vraisemblablement par une couverture plus limitée de l'arabe standard moderne dans ses données d'entraînement ; AraBERT enregistre les gains les plus marqués en NER, traduisant une forte synergie entre son pré-entraînement et les informations linguistiques injectées ; MARBERT présente des performances intermédiaires, influencées en partie par la variabilité dialectale de ses données d'apprentissage.

Dans l'ensemble, ces travaux démontrent que l'intégration explicite de connaissances linguistiques renforce la robustesse, l'interprétabilité et les performances des modèles Transformer appliqués à l'arabe standard moderne. Ils ouvrent également des perspectives prometteuses pour l'extension du cadre LiPAr à d'autres variétés de l'arabe ainsi qu'à d'autres tâches du TALN, notamment l'apprentissage en few-shot et l'analyse de dépendances.

Mots clés : Traitement automatique du langage naturel, arabe standard moderne, NLP hybride, morphologie arabe, modèles Transformer, étiquetage morpho-syntaxique (POS), reconnaissance d'entités nommées (NER), Enrichissement linguistique (Linguistic Prompting).

Abstract

This dissertation introduces a hybrid approach to Arabic natural language processing that explicitly integrates morpho-syntactic features into neural models. Anchored in the hybrid NLP paradigm, it seeks to complement the distributed representations learned by deep neural architectures with structured linguistic knowledge.

We present LiPAr (Linguistic Prompting for Arabic), an original methodological framework enabling the controlled injection of morpho-syntactic information into language models. The framework is evaluated on two core Arabic NLP tasks: part-of-speech tagging (POS) and named entity recognition (NER).

Experiments were conducted with three pre-trained models (mBERT, AraBERT and MARBERT). We compare a baseline configuration against an enriched setup incorporating 32 morpho-syntactic features (gender, number, case, definiteness, etc.) extracted using the Stanza linguistic analyzer. Results on the UD_Arabic-PADT corpus (POS) and the ANERcorp dataset (NER) demonstrate consistent and statistically significant improvements in POS tagging, ranging from +1.5 to +6.6 F1 points depending on the model. The effects on NER prove more variable and closely tied to each model's pre-training characteristics.

Quantitative and qualitative analyses reveal distinct architectural behaviours: mBERT benefits most notably from morpho-syntactic enrichment in POS tagging, likely due to its more limited exposure to Modern Standard Arabic during pre-training; AraBERT achieves the strongest gains in NER, indicating a robust synergy between its training data and the injected linguistic cues; MARBERT exhibits intermediate performance, partly shaped by the dialectal diversity of its training corpus.

Overall, these findings demonstrate that the explicit integration of linguistic knowledge enhances the robustness, interpretability and performance of Transformer-based models applied to Modern Standard Arabic. They also open promising avenues for extending the LiPAr framework to other Arabic varieties and to additional NLP tasks, particularly few-shot learning and dependency parsing.

Keywords: Natural Language Processing, Modern Standard Arabic, Hybrid NLP, Arabic Morphology, Transformer Models, Part-of-Speech Tagging, Named Entity Recognition, Linguistic Prompting.

ملخّص

تقدم هذه الأطروحة منهجًا هجينًا للمعالجة الآلية للغة العربية، يقوم على الدمج الصريح لخصائصها الصرفية والنحوية داخل النماذج العصبية. يندرج هذا العمل في إطار النمذجة اللغوية الهجينة، حيث تُثري المعرفة اللغوية البنيوية التمثيلات التوزيعية التي تتعلمها البنى العصبية العميقة.

نقدم إطارًا منهجيًا جديدًا يُسمى ليبار(LiPAr – Linguistic Prompting for Arabic) ، يهدف إلى إثراء نماذج اللغة عن طريق حقن محكوم للمعلومات الصرفية والنحوية. وقد تم تقييم هذا الإطار على مهمتين أساسيتين في معالجة اللغة العربية: الوسم الصرفي (POS Tagging) والتعرف على الكيانات المسماة (NER).

أُجريت التجارب باستخدام ثلاثة نماذج مدربة مسبقًا (mBERT و AraBERT وMARBERT)، حيث قورنت تهيئة أساسية بتهيئة مُثراة تضم 32 سمة صرفية-نحوية (مثل الجنس والعدد والإعراب والتعريف)، مستخلصة بواسطة المحلل اللغوي Stanza. أظهرت النتائج على مدونتي UD_Arabic-PADT لمهمة الوسم الصرفي و ANERcorp لمهمة التعرف على الكيانات تحسنات منهجية وذات دلالة إحصائية في مهمة الوسم الصرفي (تراوحت الزيادة في درجة F1 بين 1.5 و6.6 نقطة حسب النموذج)، في حين كانت التأثيرات في مهمة NER أكثر تباينًا وارتباطًا بخصائص التدريب المسبق لكل نموذج.

كشفت التحليلات الكمية والنوعية عن سلوكيات متمايزة حسب البنى: استفاد نموذج mBERT بشكل خاص من السمات في مهمة الوسم الصرفي، وهو ما يُعزى إلى تمثيله المحدود نسبيًا للعربية الفصحى الحديثة؛ حقق AraBERT أكبر المكاسب فيNER ، مما يشير إلى انسجام قوي بين تدريبه والمعلومات المضافة؛ أما MARBERT فقد أظهر أداءً متوسطًا، متأثرًا جزئيًا بتنوع اللهجات في بيانات تدريبه.

في المجمل، تُثبت هذه النتائج أن الدمج الصريح للمعرفة اللغوية يُعزز متانة النماذج المحولة (Transformer) وقابليتها للتفسير وأدائها عند تطبيقها على العربية الفصحى الحديثة. كما تفتح آفاقًا واعدة لتوسيع إطار LiPAr ليشمل أصنافًا أخرى من العربية ومهام أخرى في المعالجة الآلية، لا سيما التعلم بقليل من الأمثلة والتحليل الإعرابي.

الكلمات المفتاحية : المعالجة الطبيعية للغة، اللغة العربية الفصحى الحديثة، المعالجة الهجينة، الصرف العربي، النماذج المحولة (Transformer) ، الوسم الصرفي (POS) ، التعرف على الكيانات المسماة (NER) ، التوجيه اللغوي (Linguistic Prompting).

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à mon directeur de thèse, **Mr le Professeur Mohamed Tayeb LASKRI**, pour son accompagnement, sa confiance et ses précieux conseils tout au long de ce travail. Sa rigueur scientifique et sa disponibilité ont été déterminantes dans l'aboutissement de cette recherche.

Je remercie sincèrement les membres du jury qui ont accepté d'évaluer ce travail et d'enrichir ma réflexion par leurs remarques et leurs questionnements.

Un merci tout particulier à Mr le **Professeur Lekhfif**, membre de ce jury et enseignant-chercheur au département d'informatique de l'Université de Sétif, pour toute l'aide qu'il m'a apportée depuis mon inscription jusqu'à la soutenance. Ses conseils avisés, sa disponibilité et son regard bienveillant sur mon parcours ont été d'une valeur inestimable.

J'adresse mes remerciements à l'ensemble des chercheurs et enseignants du laboratoire LIAM et du département d'informatique de l'université de M'Sila, pour leurs échanges stimulants et leur soutien au quotidien.

Un merci tout particulier à ma famille, pour leur soutien indéfectible et leur patience. Sans leur encouragement, ce travail n'aurait pas été possible.

Enfin, je remercie toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce travail, qu'il m'est difficile de citer individuellement mais dont la présence a été précieuse.

Sommaire

Liste des figures ..... **xi**
Liste des tableaux ..... **xii**
Liste des acronymes ..... **xiii**
Chapitre 1 - Introduction générale **1**
1.1 Contexte et enjeux 1
1.2 Spécificité et complexité linguistiques de la langue arabe 1
1.2.1 Productivité dérivationnelle et densité flexionnelle 2
1.2.2 Ambiguïté induite par l'absence de diacritiques 2
1.2.3 Défis computationnels et diglossie 3
1.3 TALN arabe : état actuel et défis 3
1.4 Limites des approches actuelles 4
1.5 Vers une approche hybride 6
1.6 Problématique scientifique et hypothèse de recherche 7
1.7 Objectifs de la recherche 8
1.8 Originalité et apports scientifiques 9
1.8.1 Contribution théorique : un paradigme d'hybridation linguistique 9
1.8.2 Contribution méthodologique : un pipeline d'enrichissement reproductible 10
1.8.3 Contribution empirique : Validation expérimentale rigoureuse 10
1.8.4 Contribution pratique : Ressources et outils pour la communauté 10
1.9 Organisation du manuscrit 11
Chapitre 2 - Fondements linguistiques de l'arabe standard **12**
2.1 Introduction 12
2.2 Aperçu historique et position linguistique de l'arabe 12
2.2.1 Position généalogique 12
2.2.2 Évolution et standardisation 12
2.3 Système morphologique de l'arabe 13
2.3.1 La structure racine–schème 13
2.3.2 Flexion nominale et verbale 14
2.3.3 Les affixes et les morphèmes liés 14

2.4 Caractéristiques syntaxiques de l'arabe 15
2.4.1 Typologie des phrases et ordre des constituants 15
2.4.2 Ellipses et flexibilité syntaxique 15
2.5 Ambiguïté graphique et diacritisation 15
2.6 Caractéristiques sémantiques et dérivationnelles 16
2.7 Diglossie et variation linguistique 16
2.8 Synthèse des caractéristiques fondamentales 17
2.9 Conséquences linguistiques pour la modélisation 17
2.10 Conclusion 18
Chapitre 3 - Etat de l'art des approches du TALN arabe 19
3.1 Introduction 19
3.2 Aperçu historique de l'évolution du TALN arabe 19
3.3 Les approches symboliques : la linguistique computationnelle classique 20
3.3.1 Principes et fondements 20
3.3.2 Outils emblématiques 20
3.3.3 Avantages et limites 20
3.4 Les approches statistiques et distributionnelles 21
3.4.1 Du symbolique à l'apprentissage 21
3.4.2 Contributions et limites 21
3.5 Les approches neuronales : réseaux profonds et modèles pré-entraînés 22
3.5.1 Les représentations distribuées 22
3.5.2 Les architectures profondes 22
3.5.3 L'ère des Transformers 22
3.5.4 Extensions récentes : modèles génératifs et multitâches 23
3.5.5 Vers des analyseurs neuronaux : l'émergence de pipelines intégrés 23
3.6 Limites persistantes des modèles neuronaux 24
3.6.1 Limites architecturales, représentationnelles et empiriques 24
3.6.2 Positionnement comparatif des paradigmes du TALN arabe 24
3.7 Vers les approches hybrides : intégration linguistique et apprentissage profond 25
3.7.1 Principe général 25
3.7.2 Méthodes d'intégration linguistique 25
3.8 Discussion critique et positionnement de la thèse 26
3.9 Conclusion du chapitre 27
Chapitre 4 - Conception de l'approche hybride proposée 28
4.1 Introduction 28

4.2 Principes de conception et fondements scientifiques 28
4.2.1 Motivation scientifique 28
4.2.2 Fondement conceptuel : des traits aux « prompts linguistiques » 29
4.3 Représentation linguistique intégrée 30
4.3.1 Extraction des traits morphologiques 30
4.3.2 Problèmes d'alignement token–sous-token et normalisation 31
4.4 Architecture de l'approche hybride 31
4.4.1 Schéma général unifié du pipeline LiPAr 31
4.4.2 Description modulaire du cadre LiPAr 33
4.4.3 Propriétés architecturales et compatibilité 35
4.5 Formulation mathématique du cadre LiPAr 35
4.5.1 Représentations lexicales contextuelles d'entrée 35
4.5.2 Encodage et projection des traits morpho-syntaxiques 36
4.5.3 Fusion hybride des représentations 36
4.5.4 Fonction de prédiction 37
4.5.5 Fonction de perte et masquage 37
4.6 Stratégies de fusion envisagées 38
4.6.1 Fusion précoce contrôlée (*Early Linguistic Injection*) : stratégie de référence 38
4.6.2 Fusion tardive (*Late Fusion*) : Variante exploratoire 39
4.6.3 Fusion résiduelle multi-niveaux (*Hybrid Multi-stage Injection*) 39
4.6.4 Gestion de la granularité subword et stratégies d'alignement 40
4.6.5 Décodage et fonctions de perte 40
4.6.6 Optimisation, régularisation et arrêt anticipé 40
4.7 Hypothèses de recherche et protocole d'évaluation du cadre LiPAr 41
4.8 Discussion linguistique et illustrations qualitatives 42
4.9 Limites identifiées et perspectives d'amélioration 42
4.10 Synthèse du cadre proposé et perspectives expérimentales 43
Chapitre 5 - Méthodologie expérimentale 44
5.1 Introduction générale 44
5.2 Architecture générale du pipeline expérimental LiPAr 44
5.2.1 Prétraitement et normalisation 44
5.2.2 Encodage et vectorisation 45
5.2.3 Extraction et représentation des traits linguistiques structurés 45
5.3 Architecture d'entraînement et stratégie de fusion LiPAr 46
5.3.1 Fusion précoce des représentations (stratégie principale) 46

5.3.2 Fonction de perte et optimisation 47
5.3.3 Hyperparamètres d'entraînement 47
5.4 Etude comparative des stratégies de fusion 49
5.4.1 Fusion tardive (exploration complémentaire) 49
5.4.2 Résultats comparatifs des stratégies de fusion 50
5.4.3 Analyse coût / bénéfice et discussion linguistique 50
5.5 Cadre d'évaluation et métriques 51
5.6 Synthèse du chapitre 52
Chapitre 6 - Analyse des Résultats et Validation 53
6.1 Introduction 53
6.2 Synthèse quantitative des performances 53
6.2.1 POS tagging : effets nets et systématiques 54
6.2.2 Résultats en NER : forte dépendance au pré-entraînement arabe 55
6.2.3 Comparaison transversale POS vs NER 56
6.3 Analyse par classes du POS : apports ciblés des traits morphologiques 57
6.3.1 Le modèle AraBERT : gains ciblés sur les classes morphologiques 58
6.3.2 mBERT : effet compensatoire des traits 59
6.3.3 Le modèle MARBERT : réponse hétérogène aux traits 59
6.4 Analyse par classes du NER : effets contrastés selon la nature des entités 60
6.4.1 Le modèle mBERT : gains ultra-ciblés et fragiles 61
6.4.2 Le modèle AraBERT : synergie maximale avec les traits morphologiques 62
6.4.3 Le Modèle MARBERT : entre bénéfice MSA et bruit dialectal 63
6.5 Analyse qualitative des erreurs 64
6.5.1 Dans quelles situations les traits morphologiques corrigent-ils des erreurs ? 64
6.5.2 Dans quelles situations les traits introduisent-ils du bruit ? 64
6.5.3 Synthèse qualitative 64
6.6 Comparaison triangulaire : mBERT vs AraBERT vs MARBERT 65
6.6.1 Pourquoi mBERT bénéficie-t-il le plus des traits en POS ? 65
6.6.2 Pourquoi AraBERT maximise-t-il les gains en NER ? 65
6.6.3 Pourquoi MARBERT est-il intermédiaire ? 65
6.6.4 Synthèse comparative triangulaire 66
6.7 Analyse linguistique approfondie : le rôle structurel des traits arabes 66
6.7.1 Traits nominaux (CASE, DEFINITE, GENDER/NUMBER) 66
6.7.2 Traits verbaux : segmentation morphologique et clitiques 68
6.7.3 Traits dérivationnels (racine, schème) 68

6.7.4 Traits spécialisés (SPECIAL_FEATURES) 68
6.8 Limites de l'approche 68
6.8.1 Couverture restreinte au MSA 68
6.8.2 Alignement tokenisation BERT vs traits morphologiques 69
6.8.3 Surcoût d'inférence 69
6.8.4 Sur-apprentissage sur traits rares 69
6.8.5 Synthèse 69
6.8.6 Positionnement par rapport à l'état de l'art 69
6.9 Perspectives de recherche 70
6.9.1 Extension aux dialectes de l'arabe 70
6.9.2 Few-shot learning et prompting linguistique 70
6.9.3 Application au dependency parsing 70
6.9.4 Déploiement industriel : API LiPAr open-source 71
6.10 Conclusion du chapitre 71
Chapitre 7 - Discussion et Perspectives 73
7.1 Interprétation linguistique des résultats 73
7.1.1 Le rôle central de la morphologie dans les performances de LiPAr 73
7.1.2 Différences entre POS et NER : une lecture linguistique 75
7.1.3 Synthèse interprétative 76
7.2 Mise en perspective avec l'état de l'art 77
7.2.1 Morphologie et TAL arabe : apports et limites des approches classiques 77
7.2.2 Apprentissage implicite de la morphologie dans les modèles neuronaux 77
7.2.3 Intégration explicite de la morphologie : approches morph-aware 78
7.2.4 Positionnement de LiPAr par rapport à l'état de l'art 79
7.3 Apports méthodologiques de l'approche LiPAr 80
7.4 Limites et perspectives de l'approche LiPAr 81
7.4.1 Dépendance à la qualité de l'extraction morphologique 81
7.4.2 Sensibilité au domaine et aux variétés linguistiques 81
7.4.3 Couverture limitée des phénomènes morphologiques 82
7.4.4 Perspectives méthodologiques 82
7.4.5 Perspectives linguistiques et applicatives 83
7.5 Conclusion du chapitre 83
Chapitre 8 - Conclusion générale 85
8.1 Rappel du contexte et de la problématique 85
8.2 Synthèse des contributions scientifiques 86

8.3 Principaux résultats et enseignements 88
8.4 Limites et difficultés rencontrées 90
8.4.1 Limites méthodologiques 90
8.4.2 Difficultés techniques surmontées 90
8.5 Perspectives et travaux futurs 90
8.5.1 Perspectives immédiates 90
8.5.2 Perspectives de recherche fondamentale 91
8.5.3 Perspectives applicatives 91
8.6 Réflexion finale 92
8.6.1 Un plaidoyer pour le TALN hybride 92
8.6.2 Au-delà de l'arabe : un modèle pour les langues sous-représentées 92
8.6.3 Message de clôture 92
Bibliographie 94

Liste des figures

Figure 4.1 : Architecture modulaire de LiPAr (représentation schématique). 33
Figure 6.1 : Comparaison des scores F1 (POS) par modèle (Baseline vs avec traits) 54
Figure 6.2 : Comparaison des scores F1 (NER) par modèle (Baseline vs avec traits)..... 55
Figure 6.3 : Comparaison des gains Δ F1 par modèle et tâche (POS vs NER)..... 56
Figure 6.4 : Heatmap des gains par classes POS..... 57
Figure 6.5 : AraBERT : ΔF1 (enrichi − baseline) par classe POS. 58
Figure 6.6 : mBERT : ΔF1 (enrichi − baseline) par classe POS. 59
Figure 6.7 : MARBERT : ΔF1 (enrichi − baseline) par classe POS. 59
Figure 6.8 : Heatmap comparative des gains ΔF1 par classe NER. 60
Figure 6.9 : mBERT : ΔF1 (enrichi − baseline) par classe NER. 61
Figure 6.10 : AraBERT : ΔF1 (enrichi − baseline) par classe NER..... 62
Figure 6.11 : MARBERT : ΔF1 (enrichi − baseline) par classe NER. 63
Figure 6.12 : Profil différentiel des modèles face aux traits morphologiques..... 66
Figure 6.13 : Mécanisme de désambiguïsation par les traits morphologiques : le cas de l'homographe « كتب ». 67

Liste des tableaux

Tableau 1.1 : Productivité dérivationnelle et densité flexionnelle : comparaison interlinguistique 2
Tableau 1.2 : Tableau synthétique des défis liés à la diglossie 3
Tableau 1.3 : Facteurs de complexité morphologique influençant le TALN 4
Tableau 2.1 : Morphologie dérivationnelle de la racine arabe ك–ت–ب 13
Tableau 2.2 : Morphologie affixale : préfixes, suffixes et infixes en arabe 14
Tableau 2.3 : Morphologie verbale arabe : valeurs sémantiques des formes dérivées 16
Tableau 2.4 : Distribution fonctionnelle des principales variétés de l'arabe 17
Tableau 3.1 : Synthèse des modèles de langage arabes et multilingues utilisés dans la littérature 23
Tableau 3.2 : Synthèse des limites des approches neuronales pour le TALN arabe 25
Tableau 3.3 : Principales stratégies d'intégration linguistique dans les modèles 26
Tableau 4.1 : Environnement d'implémentation et hyperparamètres expérimentaux 35
Tableau 4.2 : Illustration de l'annotation morphologique et de la représentation LiPAr .. 42
Tableau 5.1 : Traits retenus comme prompts 45
Tableau 5.2 : Hyperparamètres d'entraînement 48
Tableau 5.3 : Impact de l'injection linguistique LiPAr (fusion précoce) par rapport aux baselines BERT. 48
Tableau 5.4 : Performances comparées des stratégies de fusion (F1-macro, ensemble de test) 50
Tableau 5.5 : Comparaison synthétique des stratégies de fusion précoce et tardive dans le cadre LiPAr 51
Tableau 5.6 : Métriques d'évaluation retenues et leur interprétation 51
Tableau 6.1 : Scores F1 (POS) par modèle (Baseline vs +32 traits) 54
Tableau 6.2 : Scores F1 (NER) par modèle (baseline vs +32 traits) 55
Tableau 6.3 : Synthèse des gains ΔF1 (POS vs NER) 56
Tableau 6.4 : Performances F1 par classe (NER, mBERT, ANERCorp) 61
Tableau 6.5 : Performances F1 par classe (NER, AraBERT, ANERCorp) 62
Tableau 6.6 : Performances F1 par classe (NER, MARBERT, ANERCorp) 63
Tableau 6.7 : Analyse contrastive des trois modèles : apports morphologiques et contraintes structurelles 66
Tableau 6.8 : Limites et pistes de solution 69
Tableau 7.1 : Analyse contrastive VSO vs. SVO (sous-corpus PADT, N = 500 phrases) 74
Tableau 7.2 : Synthèse des gains F1 apportés par LiPAr 75

Liste des acronymes

ANERcorp – Arabic Named Entity Recognition corpus
API – Application Programming Interface
BERT – Bidirectional Encoder Representations from Transformers
BiLSTM – Bidirectional Long Short-Term Memory
BPE – Byte Pair Encoding
CG – Constraint Grammar
CoNLL – Conference on Natural Language Learning
CRF – Conditional Random Fields
DEPBANK – PARC 700 Dependency Bank
F1 – F-measure / Mesure harmonique précision-rappel
GPU – Graphics Processing Unit
HPSG – Head-Driven Phrase Structure Grammar
IC – Intervalle de Confiance
LiPAr – Linguistic Prompting for Arabic
MADAMIRA – Morphological Analysis and Disambiguation for Arabic
MARBERT – Multi-dialectal Arabic BERT
mBERT – Multilingual BERT
MLP – Multi-Layer Perceptron
MSA – Modern Standard Arabic / Arabe Standard Moderne
NER – Named Entity Recognition / Reconnaissance d'Entités Nommées
PADT – Prague Arabic Dependency Treebank
POS – Part-Of-Speech / Étiquetage morpho-syntaxique
SOTA – State-Of-The-Art / État de l'art
SRL – Semantic Role Labeling
TALN – Traitement Automatique de la Langue Naturelle
TB – Treebank
UD – Universal Dependencies
XCES – XML Corpus Encoding Standard

Chapitre 1 - Introduction générale

1.1 Contexte et enjeux

Le traitement automatique du langage naturel (TALN), ou *Natural Language Processing (NLP),* figure aujourd'hui parmi les domaines les plus dynamiques de l'intelligence artificielle (Manning 2022). Cette dynamique est portée par les progrès fulgurants de l'apprentissage profond et, plus particulièrement, par l'émergence des architectures Transformer (Vaswani, et al. 2017) et leur déclinaison en modèles de langages pré-entraînés (Devlin et al. 2019; Liu et al. 2019).

Ces modèles ont révolutionné la modélisation du langage pour les langues disposant de vastes ressources numériques et d'une morphologie relativement régulière. Depuis l'avènement de BERT (*Bidirectional Encoder Representations from Transformers*, (Devlin et al. 2019) et de ses variantes multilingues telles que mBERT ou XLM-R (Conneau et al. 2018), le TALN a atteint des niveaux de performance inédits sur une large gamme de tâches, de la classification textuelle à la traduction automatique.

Pourtant, cette avancée met en lumière un déséquilibre structurel profond : La révolution récente du TALN a principalement bénéficié aux langues à forte présence numérique et à morphologie régulière (l'anglais, le chinois, le français ou l'allemand) (Joshi et al. 2020)**,** tandis que les langues morphologiquement riches et à ressources limitées comme l'arabe, le turc ou l'hébreu, restent largement sous-représentées (Darwish, Habash, Abbas 2021).

Cette asymétrie constitue aujourd'hui un enjeu majeur, à la fois scientifique et sociolinguistique. L'arabe occupe une position particulièrement critique en raison de la complexité de sa morphologie non concaténative, de son ambiguïté graphique et de la diversité de ses usages écrits.

1.2 Spécificité et complexité linguistiques de la langue arabe

L'arabe standard moderne (*Modern Standard Arabic*, MSA) est une langue sémitique dont la morphologie repose sur un système non concaténatif hautement productif : le système racine–schème. Celui-ci repose sur l'interdigitation d'une racine consonantale généralement trilitère (جذر, *ǧiḏr*), porteuse du sens lexical fondamental, et d'un schème vocalique (وزن, *wazn*) qui détermine la catégorie grammaticale, le rôle morphologique et, souvent, des nuances sémantiques (Haywood, Nahmad 2009; Ryding 2005).

Cette structure génère une richesse lexicale qui défie les modèles computationnels conçus pour des langues à morphologie concaténative.

1.2.1 Productivité dérivationnelle et densité flexionnelle

La racine trilitère كتب (*k–t–b*), associée au champ sémantique de l'écriture, illustre parfaitement cette productivité. Elle engendre des formes verbales telles que كتب (*kataba*, « il a écrit »), des participes comme كاتب (*kātib*, « écrivain ») et مكتوب (*maktūb*, « écrit »), ainsi qu'à des dérivés nominaux tels que كتاب (*kitāb*, « livre ») ou مكتب (*maktab*, « bureau »), dont le pluriel interne مكاتب (*makātib*) constitue un exemple de flexion non concaténative. Elle génère également des réalisations verbales dérivées comme اكتتب (*iktataba*, « s'inscrire », forme VIII). Les verbes intègrent simultanément des marques de personne, de genre, de nombre, d'aspect, de mode, de voix, auxquelles s'ajoutent fréquemment des clitiques pronominaux.

Les grammaires descriptives de référence, notamment (Ryding 2005), ainsi que les travaux psycholinguistiques sur la structure racine-schème (Boudelaa, Marslen-Wilson 2011), mettent en évidence la forte productivité lexicale associée aux racines trilitères dans la langue arabe. La combinaison d'une racine consonantique avec différents schèmes morphologiques permet de générer un grand nombre d'unités lexicales. Dans une perspective computationnelle, cette propriété se traduit par un réseau dense de relations dérivationnelles entre racines, lemmes et formes fléchies, que l'on retrouve aujourd'hui dans les lexiques électroniques et les ressources morphologiques utilisées en NLP.

Le Tableau 1.1 compare cette densité flexionnelle à celle d'autres langues et souligne la singularité de la langue arabe :

Tableau 1.1 : Productivité dérivationnelle et densité flexionnelle : comparaison interlinguistique

Aspect	Arabe MSA	Anglais	Français
Productivité dérivationnelle	Très élevée : système racine-schème générant de nombreux dérivés	Modérée : principalement préfixes et suffixes	Modérée : principalement préfixes et suffixes
Traits flexionnels exprimés par le verbe	Nombreux : personne, nombre, genre (2e et 3e pers.), aspect, mode, voix, clitiques	Limités : temps, aspect et accord sujet restreint	Nombreux : personne, nombre, temps et mode
Types de pluriel / variation interne	Pluriels brisés, alternances vocaliques internes	Pluriel régulier + quelques irréguliers	Pluriel régulier + quelques irréguliers

***Note** : Ce tableau vise à illustrer qualitativement la complexité morphologique de l'arabe standard moderne par rapport à certaines langues indo-européennes. Il ne s'agit pas d'une mesure statistique exacte mais d'une synthèse de la littérature descriptive et des observations sur des ressources morphologiques automatisées.*

Les verbes MSA présentent une inflection simultanée de la personne, du genre (à la 2e et 3e personne), du nombre, de l'aspect, du mode et de la voix, auxquels peuvent s'ajouter des clitiques pronominaux. Par exemple, la forme كَتَبْتُهَا (*katabtuhā*, "je l'ai écrit[e]") encode à elle seule la première personne du singulier, l'aspect perfectif, la voix active, ainsi qu'un pronom objet féminin singulier. Cette complexité morphologique dépasse de manière significative celle observée dans de nombreuses langues indo-européennes (Habash, Rambow, Kiraz 2005).

1.2.2 Ambiguïté induite par l'absence de diacritiques

L'arabe standard moderne s'écrit généralement sans voyelles brèves, ce qui introduit une ambiguïté lexicale et morphologique considérable (Habash, 2010). Un même token non diacrité comme كتب peut admettre en moyenne jusqu'à 11,6 lectures diacritisées dans certaines analyses

lexicographiques (Debili, Hadhémi, Souissi 2002), et environ 8 analyses morphologiques dans d'autres travaux sur l'ambiguïté lexicale de l'arabe (Habash, Rambow 2005), contre une ambiguïté nettement plus faible dans des langues comme l'anglais (Zalmout, Habash 2019a). Cette ambiguïté, renforcée par la flexibilité syntaxique de l'arabe, autorisant notamment les ordres VSO et SVO, constitue un défi majeur pour les modèles neuronaux, qui doivent inférer la structure morpho-syntaxique correcte à partir d'indices contextuels parfois limités (Habash, 2010).

1.2.3 Défis computationnels et diglossie

A ces difficultés s'ajoute une diglossie marquée : coexistence de l'arabe standard moderne (MSA) et de multiples dialectes régionaux. Cette situation génère une forte hétérogénéité des corpus, avec des variations orthographiques, lexicales et stylistiques qui compliquent encore l'apprentissage automatique (Tableau 1.2).

Ces phénomènes accroissent la complexité du traitement automatique de l'arabe : les modèles doivent gérer non seulement l'ambiguïté intrinsèque au MSA, mais aussi l'hétérogénéité des corpus, qui peut provoquer des variations de performances selon le registre, la région ou le domaine d'application. Le concept de diglossie, introduit par (Ferguson 1959) souligne l'importance de cette dualité linguistique pour l'arabe dont l'impact demeure aujourd'hui central dans les systèmes modernes de TAL.

Nous proposerons une analyse plus approfondie des implications linguistiques et computationnelles de la diglossie arabe au chapitre 2, illustrée par des exemples de variations dialectales et de divergences orthographiques fréquentes dans les corpus contemporains.

Tableau 1.2 : Tableau synthétique des défis liés à la diglossie

Facteur	Exemple concret	Impact sur le TAL/NLP
Diglossie MSA vs dialectes	Texte MSA vs tweet dialectal	Modèles doivent reconnaître et normaliser différents registres
Variations orthographiques	Même mot écrit différemment selon auteur ou dialecte	Augmente l'ambiguïté, complique la tokenisation et la lemmatisation
Substitutions lexicales	« voiture » : سيارة (MSA) vs طوموبيل (dialecte)	Diminution de la couverture lexicale des modèles entraînés sur MSA
Corpus hétérogènes	Articles de presse, réseaux sociaux, forums	Variabilité de style et de vocabulaire, impact sur la généralisation des modèles

1.3 TALN arabe : état actuel et défis

Ces dernières années, plusieurs modèles de langages pré-entraînés spécifiquement à l'arabe ont vu le jour, à l'image d'AraBERT (Antoun, Baly, Hajj 2020), MARBERT (Abdul-Mageed, Elmadany, Nagoudi 2021), ou encore CAMeLBERT (Inoue et al. 2021). Bien qu'ils aient considérablement amélioré les performances sur des tâches aval (classification, NER), ils restent limités sur les tâches de bas niveau (désambiguïsation morphologique, lemmatisation, diacritisation) où la richesse morphologique de l'arabe demeure difficile à capturer (Zalmout, Habash 2017; Darwish, Habash, Abbas 2021).

Les modèles de type Transformer, bien qu'extrêmement puissants, se fondent sur la co-occurrence statistique des formes de surface, sans nécessairement intégrer la structure interne

des mots. Des approches telles que *FastText* (Bojanowski et al. 2017) ont cherché à enrichir les représentations vectorielles en intégrant les sous-unités, mais ces approches restent limitées par leur dépendance aux données d'entraînement et par la complexité morphologique non concaténative de l'arabe.

Un paradoxe s'annonce suite à cette situation : les modèles paraissant les plus puissants échouent souvent à modéliser la structure interne du langage, surtout si elle est riche et non concaténative, ce qui fait de la représentation morphologique un enjeu central pour le TALA.

Afin de situer l'arabe dans une perspective typologique plus large, le Tableau 1.3 compare plusieurs propriétés morphologiques influençant la complexité du traitement automatique des langues.

Tableau 1.3 : Facteurs de complexité morphologique influençant le TALN

Aspect morphologique	Arabe (MSA)	Anglais	Français
Type morphologique dominant	Non concaténatif (racine-schème) + affixation	Principalement concaténatif (affixation)	Principalement concaténatif (affixation)
Richesse flexionnelle verbale	Très élevée : personne, nombre, genre, aspect, mode, voix	Faible : peu de marques flexionnelles	Élevée : personne, nombre, temps, mode
Accord grammatical	Accord riche (genre + nombre)	Accord minimal	Accord riche (genre + nombre)
Segmentation morphologique	Clitiques fréquents (préfixes et suffixes attachés)	Rare	Modérée
Variation interne du radical	Très fréquente (schèmes, pluriels brisés)	Rare	Rare
Ambiguïté morphologique	Très élevée	Faible	Modérée

Ces constats suggèrent que les performances élevées observées sur certaines tâches masquent une fragilité structurelle des représentations neuronales lorsqu'elles sont confrontées à des phénomènes morphologiques profonds, appelant à une révision des paradigmes de représentation du langage arabe.

1.4 Limites des approches actuelles

Malgré les progrès significatifs apportés par les modèles de langage pré-entraînés appliqués à l'arabe, plusieurs limites structurelles ressortent de la littérature et des expérimentations empiriques. Ces limites ne relèvent pas uniquement de contraintes techniques ponctuelles, mais traduisent une tension plus profonde entre les propriétés morpho-syntaxiques spécifiques de l'arabe (décrites en section 1.2) et le paradigme représentationnel dominant fondé sur l'apprentissage distributionnel.

1.4.1 Dépendance aux données de surface

Les modèles neuronaux actuels apprennent principalement des régularités linguistiques à partir de co-occurrences contextuelles observées dans les données. Leur biais inductif est essentiellement distributionnel : ils capturent des corrélations statistiques entre formes de surface sans disposer d'un mécanisme explicite d'abstraction morphologique interne.

Cette dépendance rend leurs performances sensibles à la taille, à la diversité et à la qualité du corpus d'entraînement, comme l'ont montré (Soliman, Eissa, El-Beltagy 2017) avec AraVec. En l'absence de structuration morphologique explicite, chaque variation flexionnelle ou dérivationnelle tend à être traitée comme une forme distincte, ce qui conduit à une explosion du vocabulaire, à une redondance des embeddings et à une capacité de généralisation limitée, en particulier pour les formes rares ou peu représentées.

Dans le cas de l'arabe, où une même racine peut générer un grand nombre de formes dérivées et flexionnelles, cette dépendance aux occurrences observées accentue la fragmentation représentationnelle et limite la capacité des modèles à capturer les relations systématiques entre les formes apparentées.

1.4.2 Perte d'information linguistique

Les modèles de type Transformer reposent généralement sur des segmentations en sous-unités (subwords), telles que BPE ou WordPiece, construites selon des critères statistiques. Si ces segmentations permettent de limiter la taille du vocabulaire et d'améliorer la couverture lexicale, elles ne correspondent pas nécessairement aux unités morphologiques linguistiquement pertinentes.

Dans une langue à morphologie non concaténative comme l'arabe, la segmentation statistique peut rompre l'isomorphisme entre structure morphologique et structure représentationnelle. Les racines et schèmes peuvent être fragmentés de manière arbitraire, ce qui dilue les relations dérivationnelles internes et peut entraîner une perte d'information linguistique cruciale (Zalmout, Habash 2020).

Ainsi, la structure interne des mots, pourtant centrale dans l'organisation lexicale de l'arabe, n'est pas explicitement encodée dans l'espace représentationnel appris.

1.4.3 Ambiguïté non résolue et sous-détermination contextuelle

Les ambiguïtés lexicales et morphologiques, fréquentes dans l'arabe non diacrité, ainsi que la polysémie, sont souvent mal gérées par les modèles, qui peinent à désambiguïser sans contexte externe.

Les modèles neuronaux capturent des régularités probabilistes mais échouent souvent à différencier des formes morphologiquement proches mais sémantiquement distinctes. Les travaux de (Pasha et al. 2014; Zalmout, Habash 2019b) ont montré que l'intégration d'analyseurs morphologiques peut pallier en partie ce problème.

1.4.4 Hétérogénéité des corpus et décalage distributionnel

La variété des sources (tweets, presse, littérature) et des dialectes, engendre une forte hétérogénéité des corpus arabes, limitant la généralisation des modèles (Darwish, Habash, Abbas 2021). En effet, les jeux de données utilisés pour le pré-entraînement sont biaisés vers des domaines formels et standardisés, tandis que les textes dialectaux et informels sont sous-représentés.

Ces limites convergent vers un même constat : les modèles neuronaux actuels opèrent majoritairement sur des représentations de surface, sans mécanisme explicite d'accès à la structure morpho-syntaxique interne des formes linguistiques. Ces limites révèlent une

contrainte structurelle du paradigme représentationnel actuel, centré sur les formes de surface. Elles soulignent la nécessité d'un cadre permettant d'articuler explicitement représentations neuronales et structure morphologique interne.

Dans le cas de l'arabe, langue caractérisée par une morphologie non concaténative, une forte densité flexionnelle et une ambiguïté graphique élevée, cette orientation surface-centrée apparaît partiellement inadéquate. Elle met en évidence une tension entre les propriétés linguistiques intrinsèques de la langue (section 1.2) et les mécanismes d'apprentissage dominants (section 1.3).

Dans ce contexte, la question n'est plus seulement d'améliorer la capacité ou la taille des architectures existantes, mais de repenser le cadre d'intégration des connaissances linguistiques au sein des modèles neuronaux.

1.5 Vers une approche hybride

Les travaux récents en TALN arabe ont mis en évidence la complémentarité entre les approches statistiques et linguistiques. Les analyseurs morphologiques existants tels que Stanza (Qi et al. 2020), MADAMIRA (Pasha et al. 2014), Farasa (Abdelali et al. 2016) ou CALIMA avec sa version la plus récente CALIMA-Star (Taji et al. 2018), contiennent une richesse de connaissances symboliques et de règles qui permettent une segmentation et une annotation fine des unités linguistiques.

Cependant, ces systèmes reposent sur des ressources linguistiques construites manuellement, coûteuses à maintenir et parfois limitées dans leur couverture. À l'inverse, les modèles neuronaux de type Transformer apprennent des représentations contextualisées efficaces à partir de grandes masses de données, mais sans contrôle linguistique explicite.

L'intégration de traits linguistiques explicites dans des architectures neuronales a montré des résultats prometteurs dans plusieurs travaux récents (Peters et al. 2019; Liu et al. 2020; Zhang, Yu, Zhang 2020) démontrant que l'enrichissement de représentations contextuelles par des connaissances structurées améliore significativement les performances (Zalmout, Habash 2019a; Inoue, Khalifa, Habash 2022; Eskander et al. 2022; Singh, Singh, Sarmah 2023; Keersmaekers, Mercelis 2024). Ces études suggèrent que l'ajout de représentations morphologiques, syntaxiques ou sémantiques améliore la robustesse des modèles, en particulier pour les langues morphologiquement complexes.

Les modèles pré-entraînés de type BERT, bien qu'extrêmement performants pour capturer des régularités distributionnelles et sémantiques globales en arabe standard, présentent des limites récurrentes lorsqu'il s'agit de traiter les contraintes morphologiques et sémantiques internes propres à cette langue (Edmiston 2020), particulièrement pour les langues morphologiquement riches comme l'arabe standard. Ces limites se traduisent concrètement par : une confusion fréquente entre formes dérivées issues de la même racine mais différant par leur schème ; des erreurs systématiques dans l'analyse des variations flexionnelles ; une perte partielle ou totale des informations diacritiques et une persistance d'ambiguïtés lexicales, même en contexte.

Nos travaux préliminaires sur le POS tagging en arabe standard moderne, réalisés sur un corpus PADT enrichi de 400 phrases synthétiques et 200 phrases littéraires, ont confirmé et quantifié ces faiblesses : AraBERT et MARBERT atteignent environ 83 % de F1 Macro sur des textes journalistiques, mais chutent respectivement à ~66 % et ~55 % sur les données ciblant les ambiguïtés morphologiques et stylistiques, tandis que CAMeL Tools s'effondre à moins de 49 % (Mahdjoubi, Laskri 2025). L'introduction du PL-Score dans cette étude a révélé que la majorité des erreurs restantes sont linguistiquement plausibles (ex. confusion NOUN↔ADJ), renforçant l'idée qu'un guidage morphologique explicite reste indispensable pour dépasser le plafond actuel des performances pures. Ces résultats empiriques suggèrent l'existence d'un plafond structurel des approches purement distributionnelles, qui ne disparaît ni avec l'augmentation de la taille des données ni avec celle de la capacité des modèles. Ils indiquent que certaines ambiguïtés morpho-syntaxiques de l'arabe ne relèvent pas uniquement d'un manque de données, mais d'une absence de guidage structurel explicite dans le processus d'apprentissage.

Cette nécessité d'un guidage linguistique explicite ne se limite pas au POS tagging : elle a également été démontrée pour des tâches sémantiques plus complexes comme le *Semantic Role Labeling* en arabe : (Meguehout, Bouhadada, Laskri 2017) ont montré que l'intégration de traits morphologiques (cas, état construit, clitiques) dans un système de raisonnement par cas obtient des performances nettement améliorées par rapport à des approches plus simples. Ces résultats confortent l'hypothèse selon laquelle la richesse morphologique de l'arabe doit être exploitée de manière explicite pour dépasser les limites des modèles distributionnels actuels.

Cette hybridation entre apprentissage profond et connaissances linguistiques s'inscrit également dans une tendance plus large du champ de l'intelligence artificielle, celle du *Knowledge-Infused Learning*, qui vise à enrichir les modèles neuronaux par des informations symboliques ou expertes (Gaur et al. 2022; Kursuncu, Gaur, Sheth 2019).

Le présent travail propose de systématiser cette approche pour l'arabe en formalisant un cadre méthodologique que nous nommons LiPAr (*Linguistic Prompting for Arabic*).

Dans cette thèse, le terme « prompting linguistique » est employé dans un sens méthodologique élargi, désignant un mécanisme de conditionnement explicite des représentations neuronales par des connaissances linguistiques structurées, et non un prompting instructionnel tel qu'il est entendu dans le cadre des grands modèles de langage génératifs.

Ce cadre vise à guider l'apprentissage des modèles neuronaux par l'injection contrôlée de signaux morpho-syntaxiques explicites dans les représentations vectorielles apprises par les modèles Transformer.

1.6 Problématique scientifique et hypothèse de recherche

Malgré les avancées significatives apportées par les modèles de langage pré-entraînés, la question de la représentation morphologique explicite dans les architectures neuronales demeure largement ouverte pour le MSA. Les modèles de type Transformer, bien qu'efficaces pour capter des régularités distributionnelles à grande échelle, reposent principalement sur des co-occurrences de surface et sur des segmentations statistiques (subwords), sans prise en compte explicite de la structure morpho-syntaxique interne des mots.

Cette limitation apparaît de manière particulièrement marquée pour l'arabe standard moderne, dont la morphologie non concaténative, la richesse flexionnelle et l'ambiguïté induite par l'absence de diacritiques constituent des sources récurrentes d'erreurs, notamment dans les tâches linguistiques de bas niveau telles que l'étiquetage morpho-syntaxique et la reconnaissance d'entités nommées.

Les travaux existants montrent que l'augmentation de la taille des corpus ou de la capacité des modèles ne suffit pas à résoudre ces difficultés de manière systématique, suggérant l'existence d'un plafond de performance inhérent aux approches purement distributionnelles.

Cette observation est corroborée par notre étude préliminaire menée sur un corpus PADT enrichi, qui met en évidence une dégradation significative des performances des modèles AraBERT, MARBERT et CAMeL Tools dès lors que les données ciblent explicitement des ambiguïtés morphologiques et stylistiques (Mahdjoubi, Laskri 2025).

Dans ce contexte, la problématique centrale de cette thèse peut être formulée comme suit : Dans quelle mesure, et selon quelles modalités, l'intégration explicite de connaissances morpho-syntaxiques structurées peut-elle dépasser les limites des approches purement distributionnelles et améliorer la performance, la robustesse et l'interprétabilité des modèles neuronaux appliqués au traitement automatique de l'arabe standard moderne ?

Cette problématique appelle une réponse à deux niveaux complémentaires :

• **Linguistique**, en identifiant les propriétés morpho-syntaxiques de l'arabe les plus pertinentes pour guider l'apprentissage automatique ;

• **Computationnel**, en déterminant comment intégrer ces connaissances au sein de modèles de langage pré-entraînés sans compromettre leur capacité de généralisation.

1.7 Objectifs de la recherche

L'objectif principal de cette recherche consiste à concevoir, implémenter et évaluer un cadre hybride permettant l'intégration explicite de connaissances morpho-syntaxiques dans des modèles de langage pré-entraînés pour la langue arabe.

Plus précisément, ce travail vise à :

- **Formaliser un cadre conceptuel**, nommé LiPAr (Linguistic Prompting for Arabic), décrivant les principes d'intégration des traits morpho-syntaxiques arabes au sein des représentations neuronales contextuelles ;
- **Implémenter un pipeline expérimental reproductible** permettant l'extraction, la représentation et la fusion contrôlée de traits linguistiques structurés avec les représentations issues des modèles Transformer ;
- **Évaluer empiriquement l'impact** de cet enrichissement linguistique sur deux tâches fondamentales du TALN arabe, l'étiquetage morpho-syntaxique (POS) et la reconnaissance d'entités nommées (NER) ;
- **Comparer systématiquement** les performances obtenues avec celles de modèles neuronaux de référence dépourvus d'enrichissement linguistique explicite ;

- **Analyser qualitativement et quantitativement** les effets des traits morpho-syntaxiques sur les erreurs des modèles, afin d'identifier les phénomènes linguistiques effectivement capturés ou corrigés.

Ces objectifs s'inscrivent dans une démarche visant à améliorer la robustesse, l'interprétabilité et la transférabilité des modèles neuronaux pour les langues morphologiquement riches (Roy et al. 2023).

1.8 Originalité et apports scientifiques

L'originalité de cette thèse réside dans la proposition d'un cadre méthodologique unifié, dénommé LiPAr **(***Linguistic Prompting for Arabic)*, visant à intégrer explicitement des connaissances morpho-syntaxiques dans des modèles de langage pré-entraînés pour l'arabe standard moderne. Ce cadre s'inscrit dans une démarche de rapprochement entre deux paradigmes historiquement distincts du traitement automatique des langues : d'une part, les modèles neuronaux profonds de type Transformer (Vaswani, et al. 2017), capables de capturer des régularités contextuelles complexes à partir de grandes quantités de données, et, d'autre part, les ressources linguistiques symboliques, telles que les analyseurs morphologiques arabes (Pasha et al. 2014; Eskander et al. 2022), qui offrent une description fine et interprétable de la structure interne des mots.

Contrairement aux approches existantes, souvent limitées à un nombre restreint de traits linguistiques, à une seule architecture ou à une tâche spécifique, LiPAr se distingue par une exploitation morphologique à granularité fine, une évaluation sur plusieurs modèles pré-entraînés, ainsi qu'une validation sur des tâches de nature différente, permettant d'analyser l'effet différencié de l'information linguistique selon les exigences morpho-syntaxiques ou sémantiques de la tâche considérée.

De manière plus générale, ce travail s'inscrit dans la perspective du *Knowledge-Infused Learning* (Gaur et al. 2022) et du *Neuro-symbolic AI* (Kursuncu, Gaur, Sheth 2019), tout en proposant une instanciation spécifiquement adaptée aux propriétés structurelles de l'arabe. Il vise à améliorer la robustesse et la généralisabilité des modèles neuronaux face à la complexité morphologique de cette langue, en cohérence avec les constats formulés dans plusieurs travaux récents sur les limites des approches purement distributionnelles pour les langues morphologiquement riches (Inoue, Khalifa, Habash 2022; Singh, Singh, Sarmah 2023; Keersmaekers, Mercelis 2024).

Les contributions de cette thèse s'organisent autour de quatre axes complémentaires : théorique, méthodologique, empirique et pratique, détaillés dans les sections suivantes.

1.8.1 Contribution théorique : un paradigme d'hybridation linguistique

Sur le plan théorique, cette thèse formalise le paradigme LiPAr, qui conceptualise l'injection explicite de connaissances morpho-syntaxiques comme un mécanisme de guidage inductif pour les modèles neuronaux contextuels. Ce paradigme repose sur la distinction entre deux espaces complémentaires : un espace neuronal contextuel, constitué de représentations distribuées apprises par des modèles pré-entraînés (E^{BERT}), et un espace linguistique symbolique, structuré autour de traits morpho-syntaxiques explicitement annotés (E^{morph}).

L'originalité de cette approche réside dans la formulation d'un opérateur de fusion (*F*) permettant de combiner ces deux espaces sans modifier l'architecture interne des modèles sous-jacents, tout en enrichissant leurs représentations internes par une information linguistique interprétable $F(E^{BERT}, E^{morph}) \rightarrow E^{enrichi}$. Ce positionnement distingue LiPAr des approches hybrides existantes, tant par la finesse de la granularité morphologique considérée que par l'évaluation systématique menée sur plusieurs modèles et plusieurs tâches.

Bien que développé dans le contexte de l'arabe standard moderne, ce paradigme présente un potentiel de transférabilité vers d'autres langues morphologiquement riches, qu'il s'agisse de langues à morphologie non concaténative, de langues agglutinantes ou de langues caractérisées par une forte ambiguïté flexionnelle.

1.8.2 Contribution méthodologique : un pipeline d'enrichissement reproductible

Sur le plan méthodologique, cette thèse propose un pipeline modulaire et reproductible permettant l'injection contrôlée de traits morpho-syntaxiques dans des architectures Transformer. Ce pipeline couvre l'ensemble de la chaîne de traitement, depuis l'extraction automatisée des traits linguistiques jusqu'à leur intégration dans les représentations neuronales, en passant par l'alignement entre unités subword et annotations morphologiques.

Plusieurs stratégies de fusion sont explorées et comparées de manière systématique, afin d'analyser l'impact du moment et du niveau d'intégration de l'information linguistique. Une attention particulière est portée à la reproductibilité des expériences, tant au niveau des protocoles d'évaluation que des environnements logiciels et des paramètres d'entraînement, garantissant la robustesse des résultats obtenus.

1.8.3 Contribution empirique : Validation expérimentale rigoureuse

Sur le plan empirique, cette thèse propose une évaluation expérimentale approfondie de l'approche LiPAr sur plusieurs modèles de langage pré-entraînés, couvrant à la fois des modèles multilingues et des modèles spécifiquement entraînés pour l'arabe. Les expérimentations portent sur deux tâches fondamentales du TALN arabe (l'étiquetage morpho-syntaxique et la reconnaissance d'entités nommées) permettant d'analyser l'effet de l'enrichissement linguistique selon des exigences computationnelles distinctes.

Les résultats expérimentaux présentés dans cette thèse mettent en évidence des améliorations cohérentes et statistiquement significatives des performances, ainsi qu'une réduction notable de certaines catégories d'erreurs morphologiques. Au-delà des scores globaux, des analyses qualitatives approfondies permettent d'identifier les phénomènes linguistiques pour lesquels l'apport de l'information morpho-syntaxique explicite se révèle le plus bénéfique.

1.8.4 Contribution pratique : Ressources et outils pour la communauté

Enfin, cette thèse s'accompagne de contributions pratiques destinées à faciliter la réutilisation et l'extension des travaux présentés. Celles-ci incluent des modèles de langage enrichis par LiPAr, des jeux de données annotés intégrant des traits morpho-syntaxiques à granularité fine, ainsi qu'un ensemble d'outils logiciels permettant de reproduire et d'adapter le pipeline proposé.

Ces ressources visent à encourager la recherche reproductible sur le traitement automatique de l'arabe et, plus largement, sur les langues morphologiquement riches, tout en favorisant le transfert des avancées méthodologiques vers des applications industrielles et académiques.

1.9 Organisation du manuscrit

Le manuscrit est structuré comme suit :

- **Chapitre 1 – Introduction générale** : présente le contexte, la problématique, les objectifs et la structure du travail.
- **Chapitre 2 – Fondements linguistiques de l'arabe standard** : analyse en détail les caractéristiques morpho-syntaxiques et sémantiques de la langue.
- **Chapitre 3 – État de l'art des approches du TALN arabe** : propose une revue critique des outils, ressources et modèles existants.
- **Chapitre 4 – Conception de l'approche hybride** : **Le cadre LiPAr**. Décrit la modélisation théorique et les principes de la représentation morpho-sémantique enrichie, en formalisant les stratégies de fusion.
- **Chapitre 5 – Implémentation expérimentale du cadre LiPAr** : détaille la mise en œuvre du modèle, les jeux de données et le protocole expérimental.
- **Chapitre 6 – Analyse et évaluation des résultats** : interprète les performances du modèle selon la perspective du cadre LiPAr et ses limitations.
- **Chapitre 7 – Discussion et perspectives** : met en perspective les résultats obtenus avec l'état de l'art, discute les limites de l'approche proposée et explore des pistes de recherche futures.
- **Conclusion générale** : récapitule les contributions et ouvre des pistes de recherche futures.

Par cette organisation, nous visons à établir progressivement les fondements linguistiques et computationnels nécessaires à la formalisation du cadre LiPAr, puis à en démontrer empiriquement l'efficacité sur plusieurs tâches fondamentales du traitement automatique de l'arabe.

Chapitre 2 - Fondements linguistiques de l'arabe standard

2.1 Introduction

Toute entreprise de modélisation computationnelle du langage arabe requiert une compréhension fine de la structure interne de cette langue. L'arabe standard moderne (*Modern Standard Arabic*, MSA) se distingue par une cohérence morphologique et syntaxique remarquable, héritée de la tradition grammaticale arabe classique, mais également par une complexité formelle élevée qui en fait un défi singulier pour le traitement automatique du langage naturel.

Cette complexité tient principalement à la nature sémitique de la langue, fondée sur un système non concaténatif de formation des mots (racine–schème), à la richesse de sa flexion nominale et verbale, à son système d'accord et de cas, ainsi qu'à la présence implicite, et souvent omise dans l'écrit moderne, des voyelles brèves ou diacritiques. À ces caractéristiques structurelles s'ajoute la variation diachronique et dialectale, qui confère à la langue arabe une vitalité exceptionnelle mais accentue également son hétérogénéité.

Nous établissons dans ce chapitre, les fondements linguistiques nécessaires à la compréhension des choix méthodologiques opérés dans cette thèse. On vise à montrer en quoi certaines propriétés structurelles de l'arabe standard moderne constituent non seulement un défi pour le traitement automatique, mais également une opportunité pour l'intégration explicite de connaissances linguistiques dans les modèles neuronaux.

Des allusions raisonnées seront faites aux implications de ces caractéristiques sur la représentation et la formalisation linguistique, sans toutefois aborder de méthodes techniques.

2.2 Aperçu historique et position linguistique de l'arabe

2.2.1 Position généalogique

L'arabe appartient à la branche centrale des langues sémitiques, elle-même issue de la famille afro-asiatique (ou chamito-sémitique). Il partage avec l'hébreu, l'araméen et l'amharique un ensemble de traits morphologiques caractéristiques : système trilitère, dérivation interne, alternance vocalique, accord riche et structure syntaxique à ordre flexible.

Cette appartenance à la famille sémitique implique un ensemble de propriétés morphologiques et syntaxiques qui distinguent fondamentalement l'arabe des langues indo-européennes pour lesquelles la majorité des architectures neuronales ont été initialement conçues.

2.2.2 Évolution et standardisation

La langue arabe connaît historiquement trois grandes variétés :

1. **L'arabe classique (*fuṣḥā qadīma*)**, langue du Coran et des textes littéraires anciens ;

2. **L'arabe standard moderne (*fuṣḥā mu ʿāṣira*)**, codifié à partir du XIXe siècle à la faveur du mouvement de la *nahḍa* et utilisé dans la presse, l'éducation et les institutions ;
3. **Les dialectes régionaux (*ʿāmmiyyāt*)**, langues de communication quotidienne, très variées selon les aires géographiques (maghrébine, égyptienne, levantine, etc.).

Ces variétés entretiennent des relations complexes, marquées par la diglossie (Ferguson 1959), situation où deux codes linguistiques coexistent dans des fonctions sociales distinctes : la *fuṣḥā* pour les contextes formels et les dialectes pour l'oralité quotidienne. Malgré cette diversité, une continuité morphologique et dérivationnelle relie l'ensemble des variétés arabes, notamment à travers le système racine–schème et le lexique fondamental commun (Owens 2013). Cette continuité constitue un avantage pour la modélisation linguistique, mais la diglossie complique la constitution de corpus homogènes pour l'apprentissage automatique. Dans cette thèse, nous nous concentrons exclusivement sur l'arabe standard moderne afin de garantir la cohérence linguistique et méthodologique des expérimentations.

2.3 Système morphologique de l'arabe

2.3.1 La structure racine–schème

La morphologie arabe est fondamentalement non concaténative : elle repose sur un système non concaténatif fondé sur l'interaction entre une racine consonantique (*ǧiḏr*), généralement trilitère, porteuse du sens lexical abstrait, et un schème morphologique (*wazn*) qui encode des informations grammaticales (catégorie, rôle) et sémantiques (Haywood, Nahmad 2009; Ryding 2005; Holes 2004). Ce mécanisme permet la formation de familles lexicales cohérentes, mais implique une représentation hiérarchique des unités linguistiques, difficilement capturable par des modèles reposant sur une segmentation linéaire.

Le Tableau 2.1 donne une vision sur le résultat de l'application de ce principe à l'aide d'un mot exemple (ك-ت-ب).

Tableau 2.1 : Morphologie dérivationnelle de la racine arabe ك–ت–ب

Schème	Exemple	Traduction
فَعَلَ (fa'ala)	كَتَبَ (kataba)	il a écrit
فَاعِل (fāʿil)	كَاتِب (kātib)	écrivain
مَفْعُول (mafʿûl)	مَكْتُوب (maktūb)	écrit
مَفْعَل (maf'al)	مَكْتَب (maktab)	bureau
مَفْعَلة (maf'ala)	مَكْتَبة (maktaba)	bibliothèque

Du point de vue de la description linguistique, il révèle la nature abstraite et hiérarchique de la sémantique arabe, fondée sur des racines conceptuelles. Du point de vue de la formalisation, cette structure non linéaire a des conséquences majeures : elle échappe aux modèles purement segmentaux ou concaténatifs, car elle nécessite une représentation hiérarchique de l'interaction entre consonnes et patrons vocaliques.

Certaines racines sont quadrilitères (دَحْرَجَ *dahraja*, « rouler »), d'autres présentent une structure bilitère sous-jacente avec redoublement du second radical, comme (شدّ *shadda*, « serrer »), ce qui ajoute un niveau de variabilité morphologique.

Cette typologie illustre la productivité lexicale du système arabe, capable de générer de nouvelles formes par analogie à des patrons existants.

2.3.2 Flexion nominale et verbale

L'arabe possède un système flexionnel riche et régulier, affectant tant les noms que les verbes.

a) Flexion nominale : Le nom s'accorde en genre (masculin/féminin), nombre (singulier, duel, pluriel) et cas (nominatif, accusatif, génitif). Le duel (*-āni / -ayni*) constitue une catégorie morphologique particulière à l'arabe et à quelques langues sémitiques, marquant la dualité d'un référent.

b) Flexion verbale : Le verbe arabe s'accorde selon la personne, le genre, le nombre et l'aspect. Deux formes principales sont distinguées :

- **le parfait (*māḍī*)**, exprimant l'accompli ;
- **l'imparfait (*muḍāri* ')**, exprimant l'inaccompli ou le présent/futur.

La flexion nominale et verbale encode de manière explicite des informations de genre, nombre, personne, cas et aspect, produisant des formes morphologiquement denses où plusieurs traits grammaticaux sont fusionnés dans une seule unité lexicale. Les marques de flexion verbale se combinent souvent à des particules grammaticales (préfixes, suffixes, proclitiques et enclitiques), augmentant la densité morphologique d'un mot.

L'absence fréquente de diacritiques dans les textes modernes efface cependant la distinction entre plusieurs formes possibles, ce qui crée une ambiguïté intrinsèque. Du point de vue du TALN, cette densité accroît la difficulté de généralisation des modèles neuronaux, en particulier lorsque ces traits ne sont pas représentés explicitement dans l'espace vectoriel.

2.3.3 Les affixes et les morphèmes liés

Outre la dérivation interne, l'arabe emploie une riche gamme d'affixes préfixés, suffixés ou infixés. Le Tableau 2.2 en donne quelques exemples.

Tableau 2.2 : Morphologie affixale : préfixes, suffixes et infixes en arabe

Type	Exemple	Fonction
Préfixe	سـ (*sa-*)	futur proche (سـأكتب – je vais écrire)
Suffixe	ـون (*-ūn*)	pluriel masculin (يكتبون – ils écrivent)
Infixe	ت (*t*) dans *استفعل*	demande ou recherche d'action

Ce système combine régularité morphologique et productivité lexicale. Toutefois, la superposition des affixes avec la vocalisation interne produit une grande variété formelle, rendant difficile la segmentation automatique des morphèmes.

L'analyse linguistique souligne ici la distinction entre morphologie dérivationnelle (création de nouveaux lexèmes) et morphologie flexionnelle (variations grammaticales d'un même lexème).

2.4 Caractéristiques syntaxiques de l'arabe

2.4.1 Typologie des phrases et ordre des constituants

La phrase arabe peut être verbale (commence par un verbe) ou nominale (commence par un nom), une distinction fondamentale de la grammaire arabe traditionnelle qui reflète des différences de focalisation informationnelle : la phrase verbale met l'accent sur l'action, tandis que la phrase nominale exprime l'état ou la permanence. L'ordre canonique VSO peut alterner avec des structures SVO ou autres, rendues possibles par la richesse des marques casuelles et par la fonction syntaxique explicite des flexions nominales.

Un phénomène particulièrement notable est l'accord partiel du verbe en structure VSO, contrastant avec l'accord complet en structure SVO.

- En ordre VSO, le verbe ne s'accorde qu'en genre avec le sujet (كتبت البنات, *katabati l-banātu*, « les filles ont écrit ») ;
- En ordre SVO, l'accord est complet en genre et nombre (البنات كتبن, *al-banātu katabna*, « les filles ont écrit »).

Ces régularités syntaxiques, bien formalisées dans des cadres symboliques, demeurent difficiles à capturer pour des modèles neuronaux principalement sensibles aux dépendances linéaires.

Ces phénomènes d'accord et de flexibilité ordinale, inhérents à la syntaxe arabe, ont été formalisés avec succès via le cadre HPSG dans une plateforme ouverte pour l'analyse textuelle, démontrant une couverture appréciable des structures phrastiques complexes du MSA (Loukam, Balla, Laskri 2016).

Cette dissymétrie, bien connue des grammairiens arabes, illustre la sensibilité de la syntaxe arabe à la hiérarchie informationnelle plutôt qu'à l'ordre linéaire. Pour le TALN, cette propriété renforce la dépendance aux indices morpho-syntaxiques explicites, en particulier dans les tâches d'analyse syntaxique et sémantique.

2.4.2 Ellipses et flexibilité syntaxique

L'arabe tolère de nombreuses ellipses et inversions « تقديم وتأخير » (*taqdīm wa-taʾkhīr*) motivées par le style ou l'emphase. Le sujet, le verbe ou l'objet peuvent être omis ou déplacés, notamment dans la poésie et la rhétorique.

Ces libertés de construction témoignent d'une syntaxe fondée sur les rôles morphologiques et sémantiques, plutôt que sur une position fixe dans la phrase ; un aspect crucial pour toute formalisation linguistique.

Pour le TALN, cette propriété renforce la dépendance aux indices morpho-syntaxiques explicites, en particulier dans les tâches d'analyse syntaxique et sémantique.

2.5 Ambiguïté graphique et diacritisation

L'un des traits les plus singuliers de l'arabe écrit est la **sous-spécification graphique** : la plupart des voyelles brèves (*fatḥa*, *kasra*, *ḍamma*), introduites initialement pour la lecture coranique au VII^{e} siècle, ne sont pas notées et leur usage s'est restreint à l'enseignement et aux textes religieux. Ainsi, une même séquence consonantique peut correspondre à plusieurs

lectures possibles, comme l'illustre l'exemple classique de علم où l'absence de diacritiques le rend morphologiquement ambigu :

- عَلَم (ʿalam) : nom masculin singulier, schème فَعَل → 'drapeau' ;

- عِلْم (ʿilm) : nom masculin singulier, schème فِعْل → 'science' ;

- عَلِمَ (ʿalima) : verbe parfait, 3ème personne masculin singulier → 'il a su'.

Cette ambiguïté, fréquente en MSA, peut être phonologique, morphologique ou syntaxique selon le contexte, et constitue un défi majeur pour l'analyse automatique, comme illustré pour le ratio moyen de 8,3 lectures possibles par token non diacrité (Habash, Rambow 2005). Les approches purement distributionnelles peinent à résoudre ces ambiguïtés de manière fiable, ce qui motive l'intégration de connaissances linguistiques explicites pour stabiliser les représentations.

2.6 Caractéristiques sémantiques et dérivationnelles

La sémantique lexicale de l'arabe est étroitement liée à sa morphologie. Les schèmes dérivationnels véhiculent des valeurs sémantiques régulières (causativité, réciprocité, intensité), permettant une prévisibilité partielle du sens indépendamment de la fréquence des formes observées tel que l'illustre le Tableau 2**.3** à travers quelques exemples.

Ce système fonde une organisation lexicale en réseaux sémantiques, où les mots partagent une racine commune et se différencient par leur schème.

Tableau 2.3 : Morphologie verbale arabe : valeurs sémantiques des formes dérivées

Schème	Fonction sémantique	Exemple
فَعَّلَ (*faʿʿala*)	causatif / intensif	كَسَّرَ (*kassara*) – briser en morceaux
تَفَاعَلَ (*tafāʿala*)	réciproque	تَقَاتَلَ (*taqātala*) – se battre mutuellement
اِسْتَفْعَلَ (*istafʿala*)	demande d'action	اِسْتَغْفَرَ (*istaghfara*) – demander pardon

La régularité du système rend possible une prévisibilité partielle du sens, même en l'absence de contexte. Cette régularité s'étend aux réseaux sémantiques événementiels, comme le montrent les analyses FrameNet des événements de mouvement en arabe, où les schèmes dérivationnels encodent des rôles thématiques (tels que Goal ou Theme) via des verbes non concaténatifs, contrastant avec l'anglais satellite-framed (Lakhfif, Laskri 2017).

D'un point de vue théorique, cette structuration illustre le principe de dérivation sémantique, où la morphologie sert de médiation entre la forme et le sens. Cette propriété rend la langue arabe particulièrement propice à une modélisation hiérarchique des relations lexicales.

2.7 Diglossie et variation linguistique

La situation sociolinguistique du monde arabe est marquée par une diglossie stable (Ferguson 1959; Holes 2004; Badawi 1973).

L'arabe standard moderne coexiste avec une pluralité de dialectes régionaux (*ʿāmmiyyāt*), qui varient selon plusieurs axes : phonologique, morphologique, lexical et pragmatique.

Le Tableau 2.4 résume les variétés de la langue arabe.

Tableau 2.4 : Distribution fonctionnelle des principales variétés de l'arabe

Niveau de langue	Usage principal
Arabe classique	Texte religieux, littérature ancienne
Arabe standard moderne	Presse, éducation, institutions
Arabe cultivé	Communication formelle semi-spontanée
Dialectes régionaux	Conversation quotidienne
Arabe mixte	Médias, réseaux sociaux

Malgré cette diversité, le système racine–schème reste un invariant structurel entre les variétés, garantissant la cohésion du continuum linguistique arabe.

Cependant, la variation affecte la phonétique, le lexique et certaines constructions grammaticales, ce qui rend la description unifiée de la langue délicate et, par extension, toute tentative de normalisation linguistique complexe.

2.8 Synthèse des caractéristiques fondamentales

Les propriétés linguistiques présentées peuvent être résumées selon trois axes complémentaires :

1. ***Morphologique*** : non concaténative, forte densité flexionnelle, affixation riche ;
2. ***Syntaxique*** : ordre flexible, accord partiel, ellipses fréquentes ;
3. ***Sémantique*** : dérivation régulée par des schèmes, champs lexicaux hiérarchisés, transparence morpho-sémantique.

Ces caractéristiques confèrent à l'arabe une architecture linguistique hautement structurée et régulière, mais également sujette à une grande ambiguïté formelle. Elles imposent donc, à toute entreprise de formalisation (qu'elle soit grammaticale ou computationnelle), de prendre en compte la nature interdigitée de ses unités et la variabilité inhérente à ses représentations.

2.9 Conséquences linguistiques pour la modélisation

Cette richesse morphologique entraîne plusieurs implications majeures pour le TALN :

1. ***Explosion combinatoire*** : les modèles apprenant uniquement par fréquences de surface sous-segmentent mal les formes internes.
2. ***Ambiguïtés systématiques*** : l'absence de voyelles rend la désambiguïsation contextuelle indispensable.
3. ***Variabilité dialectale*** : les dialectes arabes reconfigurent la morphologie sans règle unifiée.

De ce fait, les approches standards basées sur des représentations subword (BPE, WordPiece) ne capturent que partiellement les mécanismes internes. Elles fragmentent les mots selon la distribution statistique des segments, sans respecter la structure morphologique. La conséquence est une instabilité sémantique et un apprentissage bruité dans les réseaux neuronaux.

Cette densité morphologique justifie et motive le recours à des stratégies hybrides, capables d'intégrer explicitement des traits linguistiques structurés afin de stabiliser les représentations neuronales et de réduire l'ambiguïté morphologique (cf. Chapitre 4).

2.10 Conclusion

Nous avons exposé à travers ce chapitre, les fondements linguistiques de l'arabe standard moderne, en mettant en évidence la cohérence interne et la richesse systémique de cette langue. Nous avons montré, par la même occasion, que les propriétés morphologiques, syntaxiques et sémantiques de l'arabe standard moderne constituent à la fois un défi majeur et une opportunité pour le traitement automatique du langage. Loin d'être un simple obstacle, nous pensons que cette richesse linguistique offre un potentiel inductif précieux pour guider les modèles neuronaux au-delà des approches purement distributionnelles.

Une telle compréhension linguistique est essentielle non seulement pour la description grammaticale, mais aussi, de manière indirecte, pour toute modélisation ultérieure du langage. Elle constitue ainsi un socle conceptuel indispensable pour les choix méthodologiques opérés dans cette thèse, en particulier pour la conception du cadre LiPAr présenté dans les chapitres suivants.

Chapitre 3 - Etat de l'art des approches du TALN arabe

3.1 Introduction

Le traitement automatique de la langue naturelle appliqué à l'arabe (TALA ou TALN arabe) a connu, au cours de ces deux dernières décennies, une évolution rapide sous l'effet de l'essor des corpus numériques et des avancées en apprentissage automatique allant des approches symboliques aux modèles neuronaux profonds. Cette évolution, largement inspirée par les progrès réalisés pour les langues indo-européennes, a permis l'émergence de systèmes couvrant des tâches variées, telles que l'analyse morphologique, l'étiquetage morpho-syntaxique (POS-tagging), la reconnaissance d'entités nommées (NER) ou la traduction automatique.

Cependant, la langue arabe présente des spécificités structurelles et sociolinguistiques qui rendent son traitement particulièrement complexe : morphologie non concaténative fondée sur le couple *racine–schème*, système d'écriture sans voyelles courtes, polysémie contextuelle, et coexistence d'une diglossie profonde entre l'arabe standard moderne (MSA) et les dialectes vernaculaires. Ces caractéristiques limitent la transférabilité directe des méthodes conçues pour les langues indo-européennes et exigent une adaptation fine des approches du TALN.

L'objectif ici, est de présenter une revue critique, comparative et évolutive des principales approches du TALN arabe. Nous retraçons cette évolution depuis les modèles symboliques et linguistiques classiques jusqu'aux architectures neuronales hybrides récentes. Cette synthèse permettra de dégager les tendances actuelles, d'identifier les limites persistantes, et de justifier la pertinence d'un cadre de recherche hybride intégrant des connaissances morpho-sémantiques explicites au sein de modèles neuronaux profonds. Elle constitue le socle théorique sur lequel repose la proposition méthodologique développée dans les chapitres suivants.

3.2 Aperçu historique de l'évolution du TALN arabe

L'évolution du TALN arabe peut être schématiquement divisée en quatre grandes phases, chacune marquée par un paradigme dominant :

1. ***Phase symbolique (1980–2000)*** : prédominance des modèles à base de règles et des analyseurs morphologiques ;
2. ***Phase statistique (2000–2015)*** : généralisation de l'apprentissage supervisé et des modèles probabilistes ;
3. ***Phase neuronale (2015–2020)*** : émergence des réseaux profonds et des représentations distribuées ;
4. ***Phase des modèles pré-entraînés et hybrides (depuis 2020)*** : généralisation des architectures Transformer et intégration de connaissances linguistiques dans les modèles neuronaux.

Chacune de ces phases a contribué à une meilleure compréhension des structures de l'arabe, tout en révélant les limites d'une approche isolée.

3.3 Les approches symboliques : la linguistique computationnelle classique

3.3.1 Principes et fondements

Les premières recherches en TALN arabe s'inscrivent dans la lignée de la linguistique computationnelle. L'analyse automatique reposait sur des règles linguistiques explicites, des lexiques de racines, des listes d'affixes et des grammaires formelles permettant de décrire la morphologie dérivationnelle et flexionnelle de l'arabe.

Ces systèmes, élaborés manuellement par des linguistes et des informaticiens, visaient à encoder la richesse morphologique de la langue, en particulier la dérivation trilittérale et les alternances vocaliques, afin de générer ou d'analyser des formes lexicales. Le principe était de produire toutes les analyses possibles d'un mot donné, puis de les désambiguïser à l'aide de règles contextuelles.

3.3.2 Outils emblématiques

Parmi les systèmes de référence, on peut citer :

- **Buckwalter Morphological Analyzer (BAMA)** (Buckwalter, Tim 2004), l'un des premiers analyseurs morphologiques exhaustifs pour l'arabe, combinant un lexique de racines et d'affixes avec un ensemble de règles de combinaison ;
- **SAMA (Standard Arabic Morphological Analyzer)**, version améliorée de BAMA développée au *Linguistic Data Consortium* (LDC) ;
- **MADAMIRA** (Pasha et al. 2014), fusion des systèmes MADA et AMIRA, intégrant des modules d'analyse morpho-syntaxique et de désambiguïsation contextuelle ;
- **Farasa** (Abdelali et al. 2016), une suite d'outils rapides et modulaires pour la segmentation, la lemmatisation et le POS-tagging, adapté aux grands corpus.

Au-delà des analyseurs morphologiques, des formalismes grammaticaux comme HPSG (*Head-Driven Phrase Structure Grammar*) ont été appliqués à l'arabe standard via des plateformes ouvertes intégrées, offrant un parsing syntaxique unifié et adaptable pour une large couverture des structures phrastiques (Loukam, Balla, Laskri 2016).

Ces approches symboliques ont été étendues à l'annotation sémantique via FrameNet, comme dans le développement d'un FrameNet arabe hybride intégrant AraMorph pour traiter la morphologie inflectionnelle et les clitiques pronominaux, avec une précision d'environ 77 % dans l'induction automatique des unités lexicales (Lakhfif, Laskri 2017). Ces systèmes constituent encore aujourd'hui des références incontournables pour la préparation de données linguistiquement annotées.

3.3.3 Avantages et limites

Les approches symboliques se distinguent par leur interprétabilité, leur cohérence linguistique et leur capacité à fournir des analyses détaillées. En effet, elles permettent une analyse fine de la structure morphologique de l'arabe, en distinguant racines, schèmes, affixes et clitiques, et offrent une interprétabilité élevée des résultats. Elles demeurent essentielles, notamment comme sources de connaissances linguistiques à injecter dans des systèmes neuronaux plus récents, car particulièrement efficaces pour traiter l'ambiguïté morphologique et lexicale, notamment dans des contextes où l'information diacritique est absente.

Toutefois, elles souffrent de plusieurs limites :

- **Forte dépendance aux ressources manuelles**, coûteuses à développer et difficiles à étendre aux dialectes ;
- **Manque de robustesse** face aux variations orthographiques et lexicales ;
- **Absence de généralisation** aux domaines nouveaux ou aux contextes non prévus.

De plus, leur performance reste fortement conditionnée par la couverture lexicale et la qualité des règles définies manuellement. Ces limites ont progressivement ouvert la voie à des approches plus data-driven, perçues comme plus flexibles et plus facilement extensibles.

3.4 Les approches statistiques et distributionnelles

3.4.1 Du symbolique à l'apprentissage

Le tournant statistique du TALN arabe s'amorce au début des années 2000, avec l'émergence de corpus annotés de grande taille tels que le *Penn Arabic Treebank (ATB)* (Maamouri et al. 2004), le *Prague Arabic Dependency Treebank (PADT)* (Hajič et al. 2004) ou le *Quranic Arabic Corpus* (Dukes 2009). Ces ressources ont permis l'application de modèles probabilistes capables d'apprendre des régularités à partir de données plutôt que de règles.

Les techniques les plus utilisées ont été les *Hidden Markov Models* (HMM), les *Conditional Random Fields* (CRF) (Lafferty, McCallum, Pereira 2001) et les *Maximum Entropy Models* (Ratnaparkhi 1996), qui ont permis de capturer des régularités lexicales et contextuelles à partir de grands corpus non annotés.

Ces approches ont conduit à des améliorations notables sur plusieurs tâches telles que :

- la **segmentation morphologique** (Diab, Hacioglu, Jurafsky 2007) ;
- l'**étiquetage morpho-syntaxique** (Habash, Rambow 2005) ;
- la **reconnaissance d'entités nommées (NER)** (Benajiba, Diab, Rosso 2008).

3.4.2 Contributions et limites

Ces approches ont marqué un progrès considérable, car elles ont permis aux modèles d'apprendre directement à partir des exemples observés. Cependant, elles présentent certaines faiblesses :

- **Dépendance à la qualité et la taille des corpus annotés**, souvent limités en arabe ;
- **Incapacité à capturer la structure interne des mots** (racine et schème) ;
- **Difficulté à modéliser les dépendances à long terme** dans la phrase.

En conséquence, les ambiguïtés morpho-sémantiques profondes persistent, en particulier lorsque plusieurs analyses linguistiques sont compatibles avec une même forme de surface. Ainsi, bien que les approches statistiques aient marqué une transition importante vers des modèles plus flexibles, elles n'ont pas permis de dépasser certaines limites structurelles inhérentes à la langue arabe, ce qui a ouvert la voie à l'essor des réseaux neuronaux, capables d'apprendre des représentations distribuées riches et hiérarchiques.

3.5 Les approches neuronales : réseaux profonds et modèles pré-entraînés

3.5.1 Les représentations distribuées

L'introduction des représentations distribuées de mots (*word embeddings)* tels que Word2Vec (Mikolov et al. 2013) et GloVe (Pennington, Socher, Manning 2014) a révolutionné la représentation lexicale. Les mots ne sont plus symboles discrets mais vecteurs continus dans un espace latent, où la proximité géométrique reflète la similarité sémantique.

Des versions pour l'arabe sont rapidement apparues :

- ***AraVec*** (Soliman, Eissa, El-Beltagy 2017), entraîné sur divers corpus arabes (Wikipedia, actualités, tweets) ;
- ***FastText Arabic*** (Bojanowski et al. 2017), exploitant les *n-grammes* de caractères pour modéliser la morphologie interne.

Ces représentations ont permis d'atténuer la fragmentation lexicale induite par la morphologie complexe de l'arabe.

3.5.2 Les architectures profondes

L'étape suivante dans l'évolution des approches neuronales a été marquée par l'introduction de réseaux récurrents profonds, en particulier les architectures LSTM (*Long Short-Term Memory*) et leurs variantes bidirectionnelles. Ces modèles permettent de capturer des dépendances contextuelles de longue portée au sein des séquences textuelles.

Dans le domaine du TALN arabe, les architectures BiLSTM-CRF ont démontré des performances particulièrement élevées pour plusieurs tâches séquentielles, notamment l'étiquetage morpho-syntaxique (POS tagging), la reconnaissance d'entités nommées (NER) et la diacritisation automatique. Les travaux de (Al-Sallab et al. 2017) ont notamment montré l'efficacité de ces architectures pour modéliser les dépendances contextuelles dans des corpus arabes.

3.5.3 L'ère des Transformers

À partir de 2019, les modèles basés sur le Transformer (Vaswani, et al. 2017) dominent le TALN arabe. Leur mécanisme d'auto-attention bidirectionnelle permet de capturer des relations contextuelles complexes entre les tokens d'une séquence. Dès lors, plusieurs modèles pré-entraînés spécialisés pour l'arabe ont été proposés.

Ces modèles constituent aujourd'hui le socle des approches neuronales actuelles pour le TALN arabe moderne, utilisés dans la majorité des applications avales (QA, traduction, résumé, NER…).

Récemment, (Albahli 2025) a proposé un cadre hybride avancé pour le NER arabe, intégrant explicitement la richesse morphologique et les entités imbriquées, démontrant des gains significatifs sur des corpus comme ANERcorp. Toutefois, ces avancées reposent majoritairement sur des représentations contextuelles implicites, laissant ouverte la question de la modélisation explicite et contrôlée des phénomènes morpho-syntaxiques propres à l'arabe.

Le Tableau 3.1., illustre les principaux modèles de langage arabes et multilingues.

Tableau 3.1 : Synthèse des modèles de langage arabes et multilingues utilisés dans la littérature

Modèle	Année	Corpus d'entraînement	Particularité
mBERT	2019	Wikipedia de 104 langues	1e Modèle BERT multilingue
AraBERT	2020	≈ 23 GB (Wikipedia, news, OSIAN,...)	1e BERT optimisé pour MSA
MARBERT	2021	1 milliard de tweets (~128 GB)	Forte couverture dialectale
CAMeLBERT	2021	Jusqu'à 167 GB (MSA, dialectal, classique)	Variantes spécialisées (MSA, dialectal, mixte)
ARBERT	2021	≈ 61 GB MSA	Optimisé pour l'arabe standard
AraELECTRA	2020	≈ 77 GB de texte arabe	Pré-entraînement ELECTRA plus efficace
QARiB	2021	corpus multi-domaines arabes	Amélioration des performances multi-tâches
AraGPT2	2020	grands corpus web arabes	Génération de texte arabe
AraT5	2022	grands corpus arabes	Architecture T5 adaptée à l'arabe
Jais	2023	corpus arabe et anglais massif	LLM arabe-centré pour multitâches

3.5.4 Extensions récentes : modèles génératifs et multitâches

Depuis le début des années 2020, l'émergence de grands modèles de langage génératifs (*LLM*) a ouvert la voie à de nouvelles perspectives pour le TALN arabe. Des modèles tels que AraGPT2 (Antoun, Baly, Hajj 2021) ou Jais (Sengupta et al. 2023) ont été pré-entraînés sur de vastes corpus arabes et permettent d'aborder de nombreuses tâches via des paradigmes *zero-shot* ou *instruction-based*.

Ces modèles reposent sur des architectures de type Transformer génératif et permettent une approche plus unifiée du traitement du langage, couvrant des tâches variées telles que la génération de texte, la traduction ou la réponse à des questions.

Cependant, malgré leur puissance, ces modèles présentent encore certaines limites pour l'arabe, notamment dans la prise en compte explicite de phénomènes linguistiques complexes tels que la morphologie non concaténative, les variations dialectales ou les structures syntaxiques riches.

3.5.5 Vers des analyseurs neuronaux : l'émergence de pipelines intégrés

Si les analyseurs symboliques restent des références pour leur précision et leur couverture, l'émergence de modèles neuronaux a conduit au développement de pipelines de traitement plus intégrés et facilement déployables.

Des outils comme Stanza (Qi et al. 2020) ou UDPipe (Straka, Hajič, Straková 2016) représentent cette nouvelle génération. Ils proposent des modèles neuronaux de bout en bout pour les tâches de base du TALN (tokenisation, lemmatisation, POS tagging, analyse syntaxique), entraînés sur des corpus universels comme Universal Dependencies (UD).

- **Avantages :** Ces pipelines offrent une grande facilité d'utilisation, une cohérence entre les différentes annotations (grâce à un modèle unique) et des performances de pointe sur les textes conformes aux standards d'entraînement (comme le corpus UD_Arabic-PADT utilisé dans cette thèse).

- **Limites :** Ils peuvent être moins robustes que les outils spécialisés sur des textes très bruités ou des phénomènes linguistiques rares non couverts par les corpus UD. Leur « boîte noire » neuronale peut également réduire l'interprétabilité de l'analyse comparée aux systèmes à base de règles.

Dans le cadre de cette thèse, nous avons utilisé l'outil Stanza pour l'extraction des traits morpho-syntaxiques, en raison de son alignement natif avec le schéma d'annotation *Universal Dependencies* utilisé par le corpus UD_Arabic-PADT, ainsi que de la facilité d'intégration de son pipeline complet dans un protocole expérimental reproductible.

3.6 Limites persistantes des modèles neuronaux

3.6.1 Limites architecturales, représentationnelles et empiriques

Malgré leurs performances exceptionnelles, les modèles neuronaux pré-entraînés présentent plusieurs insuffisances structurelles dans le contexte arabe, que l'on peut regrouper en trois catégories :

1. **Limites architecturales :**

   - *Tokenisation sous-optimale* : La segmentation WordPiece, inadaptée à la morphologie non-concaténative fragmente arbitrairement les morphèmes arabes, rompant l'alignement entre représentations internes du modèle et structures linguistiques réelles ;

   - *Explosion du vocabulaire* : chaque forme fléchie est traitée comme un token distinct, ce qui multiplie la taille du vocabulaire.

Ces stratégies de tokenisation, bien que statistiquement efficaces, ne garantissent pas une correspondance linguistiquement motivée avec les unités morphologiques sous-jacentes de l'arabe.

2. **Limites représentationnelles :**

   - *Opacité morphologique* : les racines et schèmes, centraux dans l'organisation lexicale de l'arabe, ne sont pas explicitement modélisés ;

   - *Ambiguïtés non résolues* : l'absence de diacritiques et de marquage casuel n'est pas compensée par des mécanismes de désambiguïsation explicites.

3. **Limites empiriques**

   - Déséquilibre des corpus : surreprésentation de MSA formel au détriment des dialectes ;

   - Manque d'interprétabilité : les décisions du modèle sont difficilement justifiables linguistiquement.

Ces constats soulignent la nécessité d'intégrer de manière contrôlée la connaissance linguistique dans les architectures neuronales.

3.6.2 Positionnement comparatif des paradigmes du TALN arabe

Au-delà des seuls modèles neuronaux, l'ensemble des paradigmes du TALN arabe (symbolique, statistique, neuronal et hybride) présente des forces et des limites spécifiques. Le Tableau 3.2 synthétise ces différents paradigmes.

Tableau 3.2 : Synthèse des limites des approches neuronales pour le TALN arabe

Paradigme	Principe	Forces	Limites principales
Approches symboliques	Règles linguistiques, lexiques et grammaires formelles	Interprétabilité élevée, précision morphologique	Faible robustesse, coût de développement, généralisation limitée
Approches statistiques	Modélisation probabiliste et Apprentissage supervisé	Exploitation des corpus annotés	Dépendance aux données, modélisation limitée de la morphologie.
Réseaux neuronaux classiques	Architectures séquentielles (CNN, LSTM, BiLSTM)	Généralisation, robustesse contextuelle	Opacité linguistique, morphologie implicite
Transformers pré-entraînés	Auto-attention, contextualisation profonde	Performances élevées multi-tâches	Tokenisation sous-optimale, morphologie non contrôlée
Approches hybrides existantes	Fusion partielle de traits linguistiques	Meilleure désambiguïsation	Fragmentation méthodologique, faible généralisabilité
Approche proposée (LiPAr)	Injection contrôlée de traits morpho-syntaxiques	Robustesse, interprétabilité, gains empiriques	Dépendance aux analyseurs linguistiques

Synthèse : Dans tous ces paradigmes, la morphologie est essentiellement apprise de manière implicite à partir des données, ce qui rend les modèles sensibles aux variations de surface et limite leur capacité de généralisation. Ces constats soulignent la nécessité d'introduire des mécanismes permettant un contrôle linguistique explicite au sein des architectures neuronales, ce qui constitue la motivation centrale de l'approche LiPAr.

3.7 Vers les approches hybrides : intégration linguistique et apprentissage profond

3.7.1 Principe général

Les approches hybrides visent à concilier la précision linguistique des modèles symboliques et la capacité d'apprentissage généralisant des modèles neuronaux. Elles reposent sur l'idée d'un biais inductif linguistique : guider le modèle neuronal par des signaux issus de la structure morpho-sémantique de la langue. Dans cette perspective, plusieurs travaux récents explorent des cadres hybrides visant à intégrer explicitement des connaissances linguistiques dans des architectures neuronales, ouvrant la voie à des approches plus interprétables et morphologiquement conscientes.

3.7.2 Méthodes d'intégration linguistique

Plusieurs stratégies ont été proposées dans la littérature afin d'intégrer explicitement des connaissances linguistiques dans les modèles neuronaux. Parmi les approches les plus courantes, on peut notamment citer les suivantes :

- ***Prétraitement morphologique*** : segmentation et désambiguïsation morphologique avant l'apprentissage du modèle (Seker, Tsarfaty 2020);
- ***Fusion de traits linguistiques*** : intégration de racines, schèmes ou traits morpho-syntaxiques dans les embeddings (Zalmout, Habash 2020);

- ***Régularisation linguistique*** : introduction de contraintes morphologiques dans la fonction de perte, par exemple via un apprentissage contrastif orienté vers la désambiguïsation morphologique (TOLEGEN et al. 2024);
- ***Apprentissage multi-tâches*** : apprentissage simultané de la tâche principale et de tâches auxiliaires portant sur des traits morphologiques ou syntaxiques (Tawfik et al. 2019).

Dans l'ensemble, ces approches illustrent différentes manières d'enrichir les modèles neuronaux par des connaissances linguistiques explicites, tout en conservant leur capacité à apprendre des représentations distribuées à partir des données. Cependant, ces travaux restent souvent limités à des configurations spécifiques, à un nombre restreint de traits linguistiques ou à des architectures particulières, sans proposer de cadre méthodologique systématique unifié et généralisable.

Le Tableau 3.3 présente une synthèse des principales stratégies d'intégration linguistique dans les architectures neuronales, en mettant en évidence leur principe général, leurs avantages ainsi que certaines de leurs limites.

Tableau 3.3 : Principales stratégies d'intégration linguistique dans les modèles

Stratégie	Principe	Avantages	Limites
Prétraitement morphologique	Segmentation et désambiguïsation morphologique avant l'apprentissage du modèle	Améliore la qualité des unités d'entrée	Dépendance aux analyseurs morphologiques
Fusion de traits linguistiques	Intégration de racines, schèmes ou traits morpho-syntaxiques dans les embeddings	Injection explicite d'informations linguistiques	Difficulté d'intégration dans les architectures profondes
Régularisation linguistique	Ajout de contraintes linguistiques dans la fonction de perte	Encourage des représentations linguistiquement cohérentes	Conception complexe des contraintes
Apprentissage multi-tâches	Apprentissage simultané de la tâche principale et de tâches linguistiques auxiliaires	Meilleure généralisation et partage de représentations	Dépendance à des annotations linguistiques supplémentaires

Cette analyse met en évidence que, malgré la diversité des approches proposées, les méthodes existantes demeurent souvent partielles ou fortement dépendantes du contexte expérimental. Cela souligne l'intérêt de développer des cadres méthodologiques plus systématiques pour l'hybridation entre connaissances linguistiques et modèles neuronaux.

3.8 Discussion critique et positionnement de la thèse

La revue des approches existantes révèle une tension épistémologique entre deux paradigmes :

- d'une part, les **modèles neuronaux**, hautement performants mais "aveugles" aux structures linguistiques ;
- d'autre part, les **modèles symboliques**, rigoureux mais rigides.

Les approches hybrides représentent un terrain d'équilibre : elles permettent d'enrichir les modèles neuronaux avec des connaissances linguistiques explicites, renforçant ainsi leur interprétabilité, leur robustesse morphologique, et leur capacité de désambiguïsation.

Les limites identifiées dans l'état de l'art suggèrent la nécessité d'un cadre méthodologique systématique permettant d'intégrer de manière contrôlée des connaissances linguistiques dans des modèles neuronaux profonds. C'est dans cette perspective que s'inscrit notre approche proposée dans cette thèse, qui vise à dépasser la fragmentation méthodologique observée dans les travaux hybrides existants. On y propose un pipeline hybride, formalisé par le cadre LiPAr**,** articulant un modèle de type BERT avec un module morpho-sémantique intégrant :

- les **racines** et **schèmes dérivationnels** issus d'analyseurs morphologiques existants ;
- les **traits morpho-syntaxiques** (genre, nombre, temps, personne) ;
- et des **indices sémantiques contextuels** extraits de corpus annotés.

Cette approche poursuit donc, trois objectifs principaux :

1. réduire la perte d'information morphologique inhérente à la tokenisation classique ;
2. améliorer la désambiguïsation lexicale et sémantique ;
3. accroître l'interprétabilité du modèle en reliant ses décisions à des entités linguistiques identifiables.

Pour l'extraction des traits, nous avons opté pour un analyseur neuronal moderne (Stanza) assurant une compatibilité avec les standards actuels (UD) et une grande reproductibilité.

3.9 Conclusion du chapitre

Ce chapitre a retracé l'évolution du TALN arabe, depuis ses fondements symboliques jusqu'aux architectures hybrides contemporaines. Il a montré que si les modèles neuronaux dominent désormais le paysage, ils ne suffisent pas à résoudre les défis intrinsèques de la langue arabe. La richesse morphologique, la diglossie et la variabilité lexicale exigent un retour vers une intégration raisonnée de la linguistique.

La réflexion menée ici nous conduit à une conviction : l'avenir du TALN arabe repose sur des modèles hybrides linguistiquement éclairés, capables d'articuler la structure interne des mots et les représentations profondes apprises.

Le chapitre suivant introduira le cadre conceptuel et formel de l'approche proposée, en détaillant les mécanismes par lesquels l'intégration explicite de connaissances morpho-syntaxiques permet de dépasser les limites identifiées dans l'état de l'art.

Chapitre 4 - Conception de l'approche hybride proposée

4.1 Introduction

Les progrès récents du traitement automatique de l'arabe standard moderne (MSA) reposent en grande partie sur des modèles neuronaux pré-entraînés de grande taille, notamment les architectures Transformer adaptées à l'arabe. Ces modèles ont démontré une remarquable capacité à capturer les régularités contextuelles et distributionnelles à partir de vastes corpus non annotés, établissant ainsi de nouveaux standards de performance sur des tâches aussi diverses que le POS tagging, la reconnaissance d'entités nommées ou la classification textuelle.

Toutefois, comme montré précédemment, ces performances s'appuient essentiellement sur une modélisation distributionnelle implicite. Pour une langue morphologiquement riche et non concaténative comme l'arabe, les représentations purement statistiques présentent des limites structurelles : elles traitent les unités textuelles comme des combinaisons statistiques, sans intégrer explicitement leur morphologie profonde (racines, schèmes, marquages flexionnels) ni les informations diacritiques, principales sources d'ambiguïté lexicale et syntaxique.

Dans ce chapitre, nous formalisons la contribution méthodologique centrale de cette thèse : le cadre LiPAr (*Linguistic Prompting for Arabic*). Conformément à la définition introduite en section 1.5, le terme « prompting linguistique » est employé ici dans un sens structurel et non conversationnel : il désigne l'injection contrôlée d'un signal linguistique explicite dans l'espace de représentation d'un modèle neuronal. L'objectif n'est pas de substituer un paradigme symbolique aux architectures neuronales modernes, mais d'introduire un biais inductif linguistiquement motivé au sein même du processus d'apprentissage.

La contribution spécifique de ce chapitre réside dans la formalisation d'un cadre d'injection linguistique contrôlée, conçu pour les propriétés morphologiques non concaténatives de l'arabe, et distinct des approches existantes (feature augmentation, multi-task learning, adapters) par son positionnement explicite comme biais inductif paramétrable intégré au backbone Transformer, permettant une analyse ablationnelle interprétable des traits injectés.

Nous présentons ici, les principes scientifiques de conception, l'architecture unifiée du cadre LiPAr, sa formalisation mathématique, ainsi que les stratégies d'injection évaluées empiriquement au chapitre 6.

4.2 Principes de conception et fondements scientifiques

4.2.1 Motivation scientifique

Les analyses critiques de l'état de l'art (Chapitre 3) montrent que les modèles pré-entraînés de type BERT, bien qu'extrêmement performants pour capturer des régularités distributionnelles et sémantiques globales en arabe standard, présentent des limites récurrentes lorsqu'il s'agit de traiter explicitement les contraintes morphologiques et sémantiques propres à cette langue (Edmiston 2020; Vania, Lopez 2017).

Ces limites se manifestent notamment par des confusions entre formes dérivées partageant une même racine mais différant par leur schème, des erreurs récurrentes dans l'analyse des variations flexionnelles (cas, nombre, genre) et une persistance d'ambiguïtés lexicales en l'absence de diacritiques. Des travaux récents, tels que ceux sur ArabianGPT (Koubaa et al. 2024), confirment ces défis et soulignent la nécessité d'injections morphologiques pour une désambiguïsation optimale.

Des travaux antérieurs issus du paradigme symbolique ont déjà démontré la puissance prédictive des traits morphologiques arabes pour des tâches de haut niveau. Ainsi, (Meguehout, Bouhadada, Laskri 2017) ont obtenu des performances compétitives en étiquetage de rôles sémantiques (SRL) en exploitant exclusivement des marques telles que le cas, les clitiques et l'état construit, sans recours à un analyseur syntaxique complet. De même, les annotations FrameNet appliquées aux événements de mouvement (Lakhfif, Laskri 2017), ainsi que les plateformes fondées sur le formalisme HPSG (Loukam, Balla, Laskri 2016), confirment l'efficacité d'une modélisation linguistique explicite pour la désambiguisation syntaxico-sémantique en arabe.

Face à ces constats, le postulat de base de LiPAr est le suivant : les représentations contextuelles des modèles comme BERT, bien que riches, peuvent être efficacement guidées et désambiguïsées par un signal linguistique explicite. Les traits morphologiques (racine, schème, genre, nombre, cas, etc.) cessent alors d'être de simples métadonnées externes pour devenir des signaux d'entrée directs qui orientent activement l'interprétation du modèle. Cette injection contrôlée agit comme un biais inductif bénéfique : elle recentre l'attention du modèle sur les régularités grammaticales pertinentes à l'arabe, améliore la désambiguïsation lexicale et renforce la généralisation, notamment sur les formes rares ou faiblement représentées dans le corpus d'entraînement.

Le terme biais inductif est employé ici au sens formel de contrainte structurelle imposée à l'espace d'hypothèses du modèle (Mitchell 1997), visant à restreindre les solutions compatibles avec les régularités morpho-syntaxiques de l'arabe.

LiPAr s'inscrit ainsi dans une perspective d'hybridation raisonnée : ni retour au symbolique pur, ni simple augmentation de features, mais réalignement méthodologique entre hypothèses de modélisation et structure linguistique.

Contrairement aux approches classiques de *feature augmentation*, où les traits sont simplement concaténés sans contrainte explicite de capacité, LiPAr introduit une projection dimensionnellement contrôlée ($dm \ll d$), garantisant que le canal morphologique agit comme biais inductif régularisé et non comme canal représentationnel concurrent.

4.2.2 Fondement conceptuel : des traits aux « prompts linguistiques »

Dans cette thèse, nous employons le terme « prompt linguistique » dans un sens élargi : il désigne un signal linguistique structuré injecté dans l'espace latent du modèle afin de conditionner l'apprentissage et non une instruction textuelle au sens des grands modèles génératifs. Contrairement aux prompts textuels des LLM génératifs, les prompts linguistiques de LiPAr sont persistants (présents durant tout l'entraînement), structurels (intégrés au pipeline interne) et paramétrables (contrôlés par des modules dédiés).

Le cadre LiPAr repose sur l'interaction de deux espaces de représentation complémentaires :

1. **Un espace Neuronal Contextuel** : constitué des représentations denses et contextualisées produites par un modèle Transformer pré-entraîné. Cet espace capture avec une grande efficacité les régularités distributionnelles et sémantiques globales du texte, mais demeure largement implicite du point de vue linguistique.
2. **Un espace Linguistique Symbolique** : composé de traits morpho-syntaxiques explicites et interprétables (catégorie morpho-syntaxique, racine, schème, genre, nombre, cas, état morphologique). Ces traits, initialement discrets, sont encodés et projetés dans un espace vectoriel dense afin de permettre leur intégration dans le modèle neuronal.

L'hybridation proposée par LiPAr, consiste à fusionner ces deux espaces complémentaires afin de produire une représentation enrichie pour chaque unité linguistique. Cette représentation combine simultanément la capacité de généralisation et de contextualisation du modèle neuronal et la granularité linguistique issue de l'analyse morpho-syntaxique explicite.

4.3 Représentation linguistique intégrée

4.3.1 Extraction des traits morphologiques

Pour la génération des traits morphologiques, nous utilisons un analyseur morpho-syntaxique capable de produire, pour chaque forme de mot, un ensemble d'annotations conforme aux fondements linguistiques exposés au chapitre 2 : catégorie morpho-syntaxique (POS), lemme, racine (root), schème (pattern) lorsqu'identifiable, genre, nombre, état (défini/indéfini), cas morpho-syntaxique et présence/absence de diacritiques. Les traits retenus sont reproductibles, interprétables et offrent une couverture suffisante sur les corpus utilisés.

Chaque trait morpho-syntaxique discret est encodé sous forme catégorielle (*one-hot* ou *embedding* discret appris), puis combiné avec les autres traits afin de constituer une représentation morphologique initiale associée à chaque mot.

Cette représentation, bien que discrète et interprétable, est ensuite projetée dans un espace dense continu de dimension réduite au moyen d'un module de projection entraînable. Cette rend le signal linguistique compatible avec les représentations neuronales continues de l'espace BERT tout en limitant la sur-paramétrisation. Le choix d'une dimension réduite agit comme contrainte régularisatrice, garantissant que le signal linguistique joue un rôle de biais inductif et non de substitut représentationnel.

L'objectif de cette projection n'est pas de reconstruire une analyse linguistique exhaustive, mais de fournir au modèle neuronal un ensemble minimal mais discriminant de signaux morpho-syntaxiques explicitement exploitables lors de l'apprentissage. L'utilisation d'embeddings appris par trait (plutôt que de représentations one-hot) peut en outre réduire la dimension et capturer des relations entre catégories (ex. pluriel vs duel).

4.3.2 Problèmes d'alignement token–sous-token et normalisation

Un point technique central dans l'implémentation de LiPAr est l'alignement entre la segmentation de l'analyseur morphologique (niveau mot) et la tokenisation du modèle BERT (WordPiece/Subword). La tokenisation WordPiece segmente fréquemment un mot graphique en plusieurs sous-tokens, tandis que les traits morphologiques sont définis au niveau du mot.

Nous adoptons la stratégie « *Assign-first* », reproductible et robuste :

- Chaque mot analysé morphologiquement est **aligné** aux sous-tokens BERT correspondants
- Son vecteur morphologique projeté est **copié** sur le premier sous-token
- Les sous-tokens suivants sont **masqués** par un vecteur neutre (voir §4.6 sur mask/aggregate).

Cette approche, simple et reproductible, s'est révélée plus stable que les stratégies d'agrégation multi-sous-tokens, qui n'ont pas produit de gains significatifs lors des expérimentations préliminaires. Une alternative (agrégation moyenne pondérée ou somme sur tous les sous-tokens) est envisageable ; les différentes stratégies d'alignement, ainsi que la configuration retenue, sont détaillées et justifiées en section 4.6.4.

4.4 Architecture de l'approche hybride

4.4.1 Schéma général unifié du pipeline LiPAr

L'architecture de LiPAr est conçue comme un pipeline modulaire et extensible, pleinement compatible avec les backbones BERT existants et permettant l'intégration explicite de signaux linguistiques structurés au sein d'un modèle Transformer pré-entraîné, sans en altérer l'architecture interne. Elle repose sur une séparation explicite entre l'espace neuronal contextuel et l'espace linguistique symbolique. Leur hybridation contrôlée constitue la contribution centrale du cadre proposé.

Le traitement suit cinq étapes principales, correspondant chacune à un module fonctionnel explicite de l'architecture LiPAr et entièrement compatible avec les architectures BERT existantes, telle qu'illustrée dans la Figure 4.1.

① Le texte arabe brut est tout d'abord tokenisé à l'aide du tokenizer associé au modèle Transformer BERT pré-entraîné retenu (AraBERT, MARBERT ou mBERT selon les expériences). La séquence de sous-tokens obtenue est ensuite encodée par les couches du Transformer afin de produire, pour chaque position i, une représentation contextuelle dense $E_i^{BERT(0)} \in \mathbb{R}^d$ (avec d = 768 dans toutes nos expériences), correspondant à la somme des embeddings lexicaux, positionnels et de segment, et capturant les dépendances syntaxiques et sémantiques globales du contexte.

② En parallèle, un analyseur morphologique externe (Stanza avec le modèle arabe le plus récent) est appliqué au texte au niveau du mot graphique. Cet analyseur fournit, pour chaque token, un ensemble structuré de traits linguistiques explicites (catégorie morpho-syntaxique, racine, schème lorsqu'identifiable, genre, nombre, cas morpho-syntaxique, état

(défini/indéfini/construit, informations de diacritisation partielle). Ces traits, initialement discrets et interprétables, constituent l'espace linguistique symbolique du cadre LiPAr.

③ Les traits linguistiques discrets ainsi extraits sont ensuite encodés sous forme de vecteurs catégoriels (one-hot ou embeddings discrets appris) et projetés dans un espace dense par un module de projection dédié. Cette opération permet d'obtenir, pour chaque token, une représentation morpho-syntaxique projetée $\tilde{E}_i^{morph} \in \mathbb{R}^{d'}$, alignée dimensionnellement avec l'espace vectoriel des embeddings BERT, tout en conservant la structure informationnelle et l'interprétabilité du signal linguistique initial.

④ Les représentations contextuelles E_i^{BERT} et morpho-syntaxiques projetées $\tilde{E}_i^{morph}$ sont ensuite injectées dans le module de fusion LiPAr, qui constitue le cœur de la contribution proposée. Ce module met en œuvre une stratégie de fusion hybride (concaténation suivie d'une projection linéaire dans l'espace latent du Transformer et d'une normalisation LayerNorm) garantissant une intégration stable et contrôlée du signal morphologique afin de produire une représentation hybride enrichie F_i, intégrant de manière contrôlée l'information morpho-syntaxique explicite dans l'espace contextuel du Transformer. Plusieurs stratégies de fusion ont été implémentées et analysées (fusion précoce, fusion tardive, mécanismes de gating et d'attention) ; la configuration de référence repose sur une injection précoce par concaténation, suivie d'une projection linéaire unifiée et d'une normalisation (LayerNorm), sans mécanisme de gating additionnel.

⑤ Enfin, la représentation enrichie F_i est transmise à une tête de tâche légère, adaptée à la nature du problème traité. Une couche linéaire suivie d'une fonction softmax est utilisée pour les tâches de classification token-par-token (telles que le POS tagging), tandis qu'un décodeur CRF linéaire-chaîne est privilégié pour les tâches séquentielles nécessitant une modélisation explicite des dépendances entre étiquettes (comme la reconnaissance d'entités nommées).

Cette architecture modulaire permet de geler ou d'affiner sélectivement le backbone Transformer, d'intégrer des signaux linguistiques explicites de manière interprétable et contrôlée, et de comparer rigoureusement différentes stratégies d'hybridation, tout en restant pleinement compatible avec les pipelines standards de l'apprentissage profond pour le traitement de l'arabe.

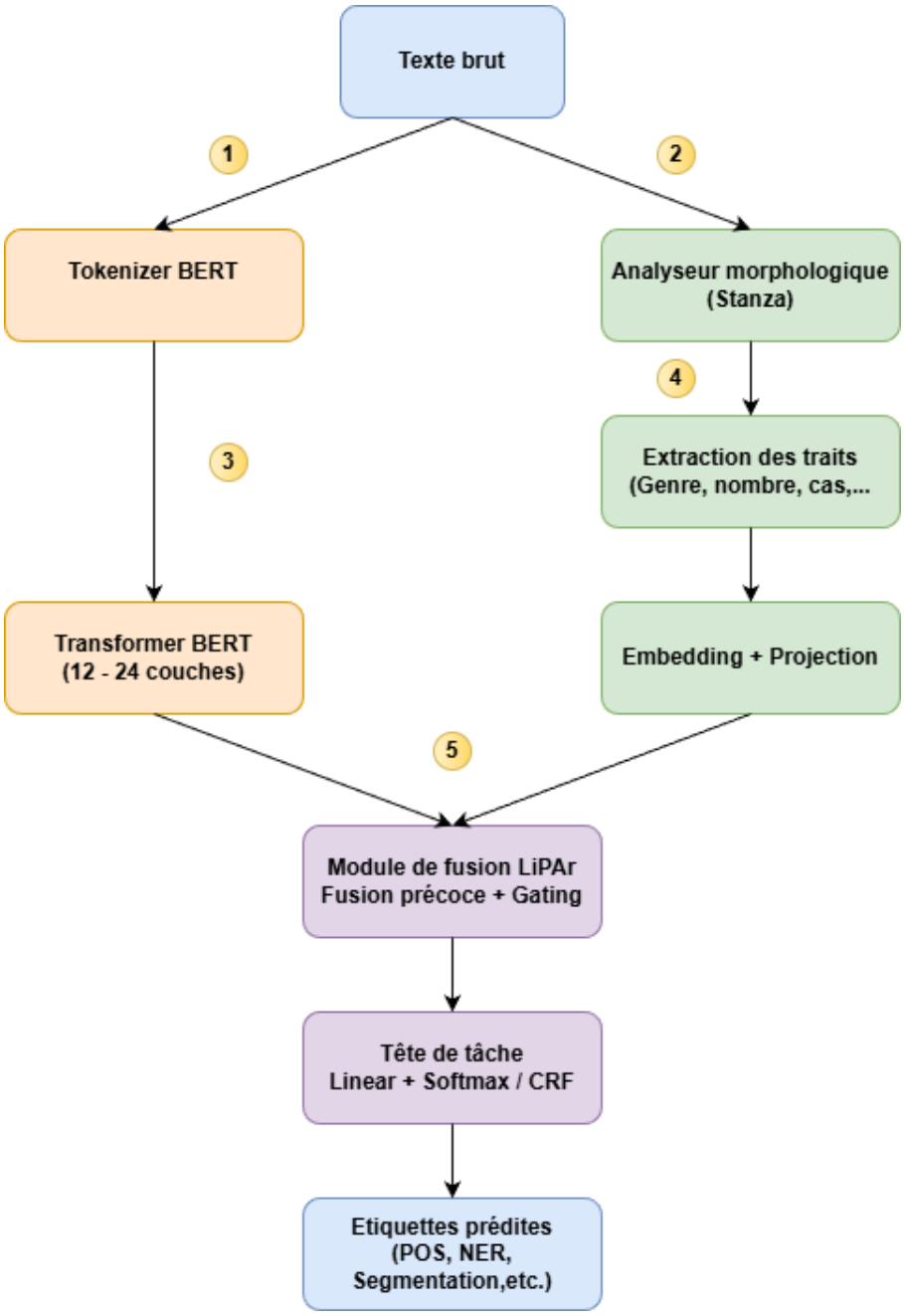


Figure 4.1 : Architecture modulaire de LiPAr (représentation schématique).

4.4.2 Description modulaire du cadre LiPAr

4.4.2.1 Encodage lexical contextuel

Le module d'encodage lexical repose sur un Transformer pré-entraîné arabophone ou multilingue, utilisé comme backbone du modèle. Après tokenisation du texte brut à l'aide du tokenizer associé au modèle retenu, les séquences sont encodées afin de produire des représentations contextuelles denses pour chaque token.

Ces représentations capturent les dépendances syntaxiques et sémantiques globales du contexte, mais restent largement implicites du point de vue linguistique, notamment pour les phénomènes morphologiques riches propres à l'arabe. Le rôle de LiPAr est précisément d'enrichir ces représentations par l'injection contrôlée d'un signal linguistique explicite.

4.4.2.2 Extraction et projection morphologique

En parallèle de l'encodage neuronal, un analyseur morpho-syntaxique externe est mobilisé afin d'extraire, pour chaque unité lexicale, un ensemble structuré de traits morphologiques explicites. Ces traits incluent notamment la catégorie morpho-syntaxique, la racine, le schème, le genre, le nombre, le cas et l'état morphologique.

Les traits extraits, initialement discrets et interprétables, sont encodés sous forme catégorielle puis projetés dans un espace vectoriel dense à l'aide d'un module d'embedding entraînable. Cette projection vise à rendre le signal morphologique compatible avec les représentations continues produites par le Transformer, tout en préservant la structure linguistique sous-jacente.

4.4.2.3 Module de fusion LiPAr (cœur du cadre)

Le module de fusion LiPAr constitue le cœur innovant de l'architecture proposée. Son objectif est de combiner les représentations neuronales contextuelles et les représentations morphologiques projetées afin de produire une représentation enrichie, exploitant simultanément les deux sources d'information.

Plusieurs stratégies de fusion ont été implémentées et évaluées dans le cadre de cette thèse. **Une fusion précoce** qui consiste à concaténer les deux représentations avant de les projeter dans un espace commun. **Une fusion tardive** qui repose sur une combinaison additive pondérée, permettant de moduler l'influence relative du signal morphologique. Enfin, **une stratégie hybride** introduit un mécanisme de gating ou d'attention croisée, inspiré des adaptateurs linguistiques.

Les résultats expérimentaux ont montré que la fusion précoce par concaténation suivie d'une projection linéaire unifiée et d'une normalisation, constitue le meilleur compromis entre stabilité de l'apprentissage, performance et interprétabilité. Cette configuration est retenue comme architecture par défaut dans la suite des expériences.

4.4.2.4 Têtes de prédiction et décodage

La représentation enrichie produite par le module de fusion est ensuite injectée dans une tête de prédiction adaptée à la tâche considérée. Pour les tâches de classification token-par-token, telles que le POS tagging ou la prédiction isolée de traits morphologiques, une couche linéaire suivie d'une fonction softmax est utilisée.

Pour les tâches nécessitant une modélisation explicite des dépendances séquentielles entre étiquettes (NER ou analyse morphologique conjointe), un décodeur de type Conditional Random Field (CRF) est privilégié. Ce choix permet d'imposer des contraintes de cohérence linguistique entre les prédictions successives.

Les principaux choix d'implémentation, d'optimisation et de décodage retenus pour l'ensemble des expériences sont synthétisés dans le Tableau 4.1. Ces paramètres, communs à toutes les variantes évaluées, ont été fixés conformément aux pratiques standards du fine-tuning de modèles Transformer afin d'isoler l'impact propre du mécanisme LiPAr.

Tableau 4.1 : Environnement d'implémentation et hyperparamètres expérimentaux

Composant	Valeurs retenues
Framework	PyTorch ≥ 1.12 + Hugging Face Transformers (Wolf et al. 2020)
Modèles de base	• *bert-base-multilingual-cased* (mBERT)• *aubmindlab/bert-base-arabertv02*• *CAMeL-Lab/bert-base-arabic-camelbert-ca*• MARBERT & mix-arabic-bert *(selon ressources)*
Optimiseur	AdamW *(weight decay = 0.01)*
Learning rate	2×10^{-5} *(couches BERT)*1×10^{-3} *(modules LiPAr ajoutés)*
Scheduler	*Linear warmup* (10 % des steps) → décroissance linéaire
Batch size effectif	16 *(gradient accumulation si GPU < 24 Go)*
Dropout	0.1 *(module de fusion + tête de classification)*
Nombre maximal d'époques	10
Early stopping	Patience = 3, surveillée sur **F1-macro** (validation)
Décodage — tâches token-indépendantes	Couche linéaire + Softmax → **CrossEntropyLoss**
Décodage — tâches séquentielles	CRF linéaire-chaîne → **NLL + perte de transition**
Masquage perte	Ignorer [CLS], [SEP], [PAD] et sous-tokens secondaires
Recherche hyperparamètres	Grille + validation croisée *(annexe C)*
Reproductibilité	3 runs avec seeds différents — code public sur GitHub

4.4.3 Propriétés architecturales et compatibilité

Grâce à sa conception modulaire, le cadre LiPAr est entièrement compatible avec les pipelines standards de fine-tuning des modèles Transformer, notamment ceux proposés par la bibliothèque Hugging Face. Il permet un gel partiel ou total du backbone pré-entraîné, ainsi qu'une adaptation fine des modules ajoutés.

Cette modularité offre une grande flexibilité expérimentale tout en garantissant une injection contrôlée, interprétable et reproductible du signal linguistique arabe, facilitant ainsi l'analyse fine des contributions respectives des composantes neuronales et symboliques.

4.5 Formulation mathématique du cadre LiPAr

Cette section formalise le cadre mathématique du modèle LiPAr. Nous détaillons successivement les représentations lexicales, l'encodage des traits morpho-syntaxiques, les stratégies de fusion, ainsi que les fonctions de prédiction et d'apprentissage.

4.5.1 Représentations lexicales contextuelles d'entrée

Soit $T=(t_1,\ldots,t_n)$ une séquence de n tokens issue de la tokenisation WordPiece du modèle Transformer BERT. Chaque token t_i est encodé par le Transformer pré-entraîné pour produire une représentation contextuelle dense issue de la dernière couche cachée du Transformer :

$$E_i^{BERT} \in \mathbb{R}^d$$

Où d désigne la dimension du vecteur caché (d = 768 dans toutes nos expérimentations, sauf mention contraire).

Ces représentations capturent efficacement les dépendances longues et le contexte global via les mécanismes d'attention auto-régulée. Néanmoins, l'information morpho-syntaxique fine demeure encodée de manière implicite, sans garantie d'explicitation structurée, ce qui peut être problématique dans des langues morphologiquement riches telles que l'arabe.

4.5.2 Encodage et projection des traits morpho-syntaxiques

En parallèle, un analyseur morpho-syntaxique externe fournit, pour chaque token t_i, un ensemble de traits linguistiques explicites (POS, racine, schème, genre, nombre, cas, etc.), qui constituent une source d'information linguistique complémentaire aux représentations neuronales contextuelles. Ces traits sont représentés sous la forme d'un vecteur binaire discret :

$$E_i^{morph} \in \{0,1\}^{n_f}$$

où n_f correspond au nombre total de traits considérés.

Afin d'obtenir une représentation compatible avec les espaces vectoriels continus du Transformer, ce vecteur est projeté dans un espace dense de dimension réduite d_m via une couche d'embedding linéaire entraînable :

$$E_i^{morph-proj} = W_{proj} E_i^{morph} + b_{proj} \qquad (4.1)$$

Avec $W_{proj} \in \mathbb{R}^{d_m \times n_f}$ et $E_i^{morph-proj} \in \mathbb{R}^{d_m}$

où $d_m \ll d$, (typiquement $d_m \in \{64, 128, 256\}$).

Cette réduction dimensionnelle ($d_m \ll d$) constitue un choix méthodologique important : elle contraint la capacité informationnelle du canal morphologique, qui agit ainsi comme biais inductif contrôlé plutôt que comme représentation concurrente du canal contextuel. Ce qui permet de limiter les risques de sur-apprentissage liés à une sur-paramétrisation du signal morphologique.

Contrairement à une projection complète vers d, le maintien d'une dimension d_m distincte permet de préserver la séparation structurelle entre espace contextuel et espace linguistique jusqu'à l'étape de fusion.

Remarque notationnelle : Par souci de clarté, nous distinguons explicitement le vecteur morphologique discret initial E_i^{morph} et sa projection dense intermédiaire $E_i^{morph-proj} \in \mathbb{R}^{d_m}$. Aucune projection préalable vers l'espace $\mathbb{R}^d$ n'est effectuée avant l'étape de fusion.

4.5.3 Fusion hybride des représentations

La contribution centrale de LiPAr réside dans une fonction de fusion paramétrable combinant la représentation contextuelle neuronale et la représentation morphologique projetée :

$$f_{\Theta_f} : \mathbb{R}^d \times \mathbb{R}^{d_m} \longrightarrow \mathbb{R}^d$$

Pour chaque token ti, la représentation enrichie est définie comme :

$$F_i = LayerNorm\left(W_f\left[E_i^{BERT}; E_i^{morph-proj}\right] + b_f\right) \qquad (4.2)$$

Où : $W_f \in \mathbb{R}^{d \times (d+d_m)}$, $b_f \in \mathbb{R}^d$ et $[;]$ désigne la concaténation,

La concaténation produit un vecteur de dimension $d + d_m$ projeté ensuite dans l'espace latent afin de préserver la compatibilité avec les couches supérieures du Transformer.

L'application de LayerNorm stabilise l'échelle des activations et garantit une intégration équilibrée des deux canaux.

Dans la configuration canonique retenue, cette fusion correspond à une concaténation suivie d'une projection linéaire unique et d'une normalisation.

Toutes les variantes de fusion sont conçues pour préserver la dimensionnalité *d* afin de garantir la compatibilité avec les couches supérieures du Transformer et les têtes de prédiction, tout en contrôlant l'influence du signal morpho-syntaxique explicite.

4.5.4 Fonction de prédiction

La prédiction associée au token t_i est obtenue à partir de la représentation enrichie F_i via une tête de classification linéaire :

$$\hat{y}_i = Softmax(W_{cls}F_i + b_{cls}) \in \mathbb{R}^{|C|} \qquad (4.3)$$

où C désigne l'ensemble des classes de la tâche considérée.

4.5.5 Fonction de perte et masquage

La fonction de perte est calculée au niveau token, en tenant compte d'un masque binaire $m_i \in \{0,1\}$ permettant d'ignorer les tokens spéciaux BERT ([CLS], [SEP], [PAD]) ainsi que les sous-tokens secondaires issus de la tokenisation WordPiece :

$$\mathcal{L} = -\frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} m_i \cdot log\ P(y_i|F_i;\Theta) \qquad (4.4)$$

où Θ regroupe l'ensemble des paramètres entraînables du backbone Transformer des modules LiPAr et des têtes de prédiction. (Backbone éventuellement gelé partiellement ou totalement selon la configuration expérimentale).

Modélisation séquentielle avec CRF : Lorsque la tâche requiert une modélisation structurée séquentielle des dépendances entre étiquettes (par exemple en reconnaissance d'entités nommées ou en segmentation morphologique), la tête softmax est remplacée par un CRF linéaire-chaîne (Lafferty, McCallum, Pereira 2001). Dans ce cas, la probabilité conditionnelle d'une séquence d'étiquettes $y = (y_1,\ldots,y_n)$ conditionnellement à la séquence d'observations $F = (F_1,\ldots,F_n)$ est définie comme :

$$P(y|F) = \frac{exp\big(s(F,y)\big)}{\sum_{y' \in \Upsilon} exp\big(s(F,y')\big)} \qquad (4.5)$$

Où Υ désigne l'ensemble de toutes les séquences possibles.

Le score d'une séquence est donné par :

$$s(F,y) = \sum_{i=1}^{n} \big(W_{y_i}^T F_i + b_{y_i}\big) + \sum_{i=1}^{n-1} A_{y_i,y_{i+1}} \qquad (4.6)$$

où $\boldsymbol{A}$ est la matrice de transition du CRF et (W, b) les paramètres de la tête d'émission. Le premier terme correspond aux scores d'émission (compatibilité entre l'observation et l'étiquette) et le second aux scores de transition entre étiquettes successives.

La fonction de perte correspond alors à la log-vraisemblance négative :

$$\mathcal{L}_{CRF} = -\log P(y^*|F) \qquad (4.7)$$

où y^* désigne la séquence d'étiquettes de référence. Une régularisation L_2 sur la matrice de transition $\boldsymbol{A}$ peut être ajoutée afin d'améliorer la stabilité numérique de l'apprentissage et de limiter le sur-ajustement.

Cette modélisation structurée permet d'imposer des contraintes linguistiques locales (par exemple l'impossibilité de passer directement d'un début d'entité à une fin d'entité sans contenu intermédiaire) tout en restant pleinement compatible avec le pipeline LiPAr.

Positionnement théorique : Cette formalisation unifiée garantit :

1. **Modularité :** le cadre LiPAr est indépendant du backbone Transformer utilisé.
2. **Contrôle de capacité :** la dimension réduite d_m agit comme mécanisme de régularisation structurelle.
3. **Injection contrôlée du signal linguistique :** la fusion paramétrable permet d'étudier finement l'impact du signal linguistique.
4. **Compatibilité structurée :** l'intégration d'un CRF permet d'introduire des contraintes séquentielles explicites.

4.6 Stratégies de fusion envisagées

La performance et la stabilité de LiPAr dépendent directement de la stratégie adoptée pour combiner l'espace neuronal contextuel issu du Transformer et l'espace linguistique symbolique projeté. Afin d'évaluer de manière systématique l'impact du point et de la forme d'injection du signal morpho-syntaxique, plusieurs stratégies de fusion ont été formalisées et implémentées. Ces stratégies se distinguent selon le point d'injection dans l'architecture, le degré de paramétrisation introduit et l'intensité d'interaction entre les deux espaces de représentation.

Bien que les modèles Transformer pré-entraînés puissent encoder implicitement certaines régularités morphologiques, cette information demeure distribuée et non explicitement structurée. Dans des langues morphologiquement riches telles que l'arabe, cette représentation implicite peut s'avérer insuffisante pour garantir robustesse et stabilité, en particulier dans des contextes à ambiguïté élevée ou à supervision limitée. L'injection explicite vise ainsi à compléter, et non à remplacer, l'encodage contextuel appris.

Trois stratégies principales ont été étudiées expérimentalement, tandis qu'une quatrième concerne spécifiquement la gestion de la granularité subword.

4.6.1 Fusion précoce contrôlée (*Early Linguistic Injection*) : stratégie de référence

La stratégie canonique de LiPAr consiste à injecter le signal morpho-syntaxique immédiatement après l'obtention des représentations contextuelles issues du backbone Transformer.

Soit $E_i^{BERT} \in \mathbb{R}^d$ la représentation contextuelle produite par la dernière couche du Transformer pour le token ti, et $E_i^{morph-proj} \in \mathbb{R}^{d_m}$ la représentation morphologique dense définie en section 4.5.2.

La fusion est définie par :

$$F_i = LayerNorm\big(W_f\big[E_i^{BERT}; E_i^{morph-proj}\big] + b_f\big) \qquad (4.8)$$

Avec $W_f \in \mathbb{R}^{d\times(d+d_m)}$

Cette opération injecte explicitement le biais linguistique, maintient la dimension *d* et permet l'intégration du signal morphologique au niveau de la représentation finale utilisée pour la prédiction, sans modification des couches internes du Transformer. Contrairement à une addition résiduelle, la concaténation suivie d'une projection offre une flexibilité plus grande pour apprendre une combinaison non triviale des deux espaces.

Cette configuration constitue l'architecture canonique utilisée dans l'ensemble des expériences rapportées au chapitre 6.

4.6.2 Fusion tardive (*Late Fusion*) : Variante exploratoire

La fusion tardive opère la combinaison des deux espaces après la dernière couche du Transformer, lorsque les représentations contextuelles ont déjà intégré les dépendances syntaxiques et sémantiques globales du texte.

Soit $E_i^{BERT(L)} \in \mathbb{R}^d$ la sortie de la couche finale *L* du Transformer. La représentation enrichie est définie par :

$$F_i = LayerNorm\left(E_i^{BERT(L)} + \alpha_i \odot g_{\Theta_g}(E_i^{morph-proj})\right) \qquad (4.9)$$

où $g_{\Theta_g}: \mathbb{R}^{d_m} \rightarrow \mathbb{R}^d$ est une transformation paramétrée (MLP léger ou projection linéaire) et $\alpha_i \in [0,1]$ est un coefficient de gating appris. Le symbole $\odot$ désigne le produit de Hadamard.

Cette stratégie permet une correction morpho-syntaxique contextuelle ciblée, particulièrement adaptée au traitement d'ambiguïtés morpho-syntaxiques persistantes après intégration complète du contexte global. Toutefois, son effet reste plus localisé que celui de la fusion précoce, et son coût paramétrique est légèrement supérieur. Elle est donc étudiée comme variante exploratoire.

4.6.3 Fusion résiduelle multi-niveaux (*Hybrid Multi-stage Injection*)

Une extension plus générale combine les deux paradigmes précédents via des injections résiduelles répétées à différentes profondeurs du Transformer.

Formellement, pour une couche ℓ, la mise à jour s'écrit :

$$H_i^{(\ell)} = H_i^{(\ell-1)} + \gamma^{(\ell)} \odot Proj^{\ell}\big(E_i^{morph-proj}\big) \qquad (4.10)$$

où $Proj^{\ell} : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^d$ est une projection spécifique à la couche ℓ *et* $\gamma^{(\ell)} \in \mathbb{R}^d$ est un facteur d'échelle appris.

Les coefficients $\gamma^{(\ell)}$ jouent un rôle essentiel dans le contrôle de l'amplitude du signal injecté. Bien que la dimension d_m demeure réduite, les injections répétées peuvent accroître l'influence effective du signal morphologique.

Cette approche, inspirée des mécanismes d'adaptateurs linguistiques (Houlsby et al. 2019), offre une flexibilité maximale, pour moduler l'influence du signal linguistique à différents niveaux de représentation. mais introduit un surcoût paramétrique non négligeable et complique l'analyse interprétative. Elle est considérée comme une piste d'extension future.

4.6.4 Gestion de la granularité subword et stratégies d'alignement

La tokenisation de type WordPiece employée par les modèles BERT arabes segmente fréquemment un mot graphique en plusieurs sous-tokens, alors que les traits morpho-syntaxiques sont définis au niveau du mot. Afin de gérer ce désalignement, deux stratégies ont été étudiées :

Assign-first (stratégie retenue) : Le vecteur morphologique complet est attribué uniquement au premier sous-token du mot. Les sous-tokens suivants reçoivent un vecteur nul ou neutre et sont exclus du calcul de la perte. Cette approche, simple et largement adoptée dans la littérature, maintient une correspondance univoque entre unité linguistique annotée et unité supervisée, évitant la dilution du signal morphologique. Elle s'est révélée robuste et stable dans nos expériences.

Agrégation multi-sous-tokens : Consiste à agréger les représentations morphologiques ou contextuelles des sous-tokens (moyenne pondérée ou somme). Bien que plus élégante théoriquement, cette stratégie n'a pas produit d'amélioration significative et a été abandonnée.

La combinaison fusion précoce + assign-first définit ainsi la configuration canonique et reproductible de LiPAr utilisée dans l'ensemble des chapitres expérimentaux.

4.6.5 Décodage et fonctions de perte

Le choix du mécanisme de décodage dépend de la nature de la tâche considérée. Pour les tâches de classification token-par-token, telles que le POS tagging ou la prédiction indépendante de traits morphologiques, une couche linéaire suivie d'une fonction softmax est utilisée, avec la CrossEntropyLoss comme critère d'apprentissage.

Pour les tâches nécessitant une cohérence séquentielle explicite, notamment le NER et l'analyse morphologique conjointe, un décodeur de type Conditional Random Field (CRF) est privilégié. La fonction de perte associée correspond à la log-vraisemblance négative du CRF, dont le score inclut à la fois les termes d'émission et de transition (cf. section 4.5).

4.6.6 Optimisation, régularisation et arrêt anticipé

L'optimisation des paramètres est réalisée à l'aide de l'algorithme AdamW, avec un coefficient de weight decay fixé à 0.01. Un taux d'apprentissage initial distinct est utilisé pour les paramètres du backbone Transformer et pour les modules ajoutés par LiPAr, suivant une approche de fine-tuning discriminant.

Un échauffement linéaire du taux d'apprentissage est appliqué sur les premières itérations, suivi d'une décroissance linéaire jusqu'à zéro. La taille effective du batch est fixée à 16 séquences, avec accumulation de gradients lorsque les ressources matérielles sont limitées. Un taux de dropout de 0.1 est appliqué aux couches fully-connected des modules ajoutés.

L'entraînement est limité à un nombre maximal d'époques, avec un mécanisme d'arrêt anticipé activé sur l'ensemble de validation. La métrique surveillée dépend de la tâche

considérée (**F1 macro** pour les tâches de classification, **exact match** pour les tâches morphologiques structurées.

Hypothèse de travail : Nous faisons l'hypothèse qu'une injection morpho-syntaxique précoce et de capacité contrôlée améliore la robustesse statistique et la stabilité d'optimisation des représentations contextuelles dans les langues morphologiquement riches, tout en limitant l'augmentation de la complexité effective du modèle.

4.7 Hypothèses de recherche et protocole d'évaluation du cadre LiPAr

Notre démarche expérimentale s'articule autour de quatre hypothèses hiérarchisées, formulées de manière à la fois testable et interprétable linguistiquement :

H_1 – Hypothèse de performance globale : L'injection contrôlée de prompts linguistiques structurés via LiPAr améliore de manière statistiquement significative les performances sur les tâches de POS tagging et de reconnaissance d'entités nommées (NER) par rapport à des baselines BERT non enrichies.

H_2 – Modulation selon le modèle de base : L'ampleur des gains est fonction du backbone sous-jacent : effet compensatoire maximal avec mBERT (qui souffre d'un déficit de couverture arabophone) ; effet de synergie maximal avec les modèles arabophones dédiés (AraBERT, MARBERT) sur les phénomènes fins.

H_3 – Modulation selon la tâche : L'apport des traits morpho-syntaxiques est plus systématique et généralisé en POS-tagging qu'en NER, où il se concentre sur la résolution d'ambiguïtés spécifiques (annexion, clitiques, frontières d'entités composées).

H_4 – Pertinence différentielle des traits : Les traits nominaux (cas, définitude, nombre, état construit) contribuent de manière plus systématique aux gains observés que les traits verbaux (personne, aspect, voix), en raison de leur rôle pivot dans le MSA.

Ces hypothèses ne préjugent pas de l'ampleur absolue des gains, mais visent à tester l'existence d'effets directionnels cohérents avec les propriétés morpho-syntaxiques de l'arabe standard moderne. Le protocole expérimental a été conçu pour tester ces hypothèses de façon reproductible et comparative (détails complets en section 5.3 et chapitre 6) :

- **Jeux de données** : Universal Dependencies Arabic-PADT (Nivre et al. 2020) pour le POS tagging ; AQMAR ANERcorp (Benajiba, Diab, Rosso 2008) pour le NER, avec partitions canoniques train/dev/test.
- **Baselines** : modèles BERT vanilla (mBERT, AraBERTv02, MARBERT).
- **Variantes LiPAr** : fusion précoce (configuration canonique) versus variantes exploratoires (fusion tardive, multi-niveaux).
- **Métriques** : Accuracy et F1-macro pour le POS ; F1-entity (CoNLL) pour le NER ; significativité statistique via test de McNemar et intervalles de confiance à 95 %.
- **Procédure** : trois runs avec graines différentes, early stopping sur l'ensemble de validation, rapport des moyennes et écarts-types.

4.8 Discussion linguistique et illustrations qualitatives

La pertinence du cadre LiPAr découle directement des propriétés morphologiques et syntaxiques idiosyncrasiques de l'arabe standard moderne exposées au chapitre 2. L'injection de signaux explicites permet de pallier trois limites récurrentes des modèles purement distributionnels :

1. **Désambiguïté des formes non diacritisées** : la graphie كتب peut correspondre à kataba (verbe parfait 3sg.masc.) ou kutub (nom pluriel brisé). Les traits « POS = VERB » + « schème = faʿala » versus « POS = NOUN » + « schème = fuʿul » permettent de contraindre explicitement l'espace des interprétations possibles, réduisant ainsi la probabilité d'erreurs liées à l'ambiguïté graphémique.
2. **Structuration des syntagmes nominaux en annexion (iḍāfa)** : les traits CAS et DEF permettent d'expliciter les relations de dépendance non marquées à la surface (ex. كتابُ الطالبِ « le livre de l'étudiant » → NOUN+DEF+NOM → NOUN+INDEF+GEN).
3. **Gestion des clitiques et des frontières d'entités** : les traits de clitiques (pronoms objet/sujet) et de cas facilitent la délimitation précise des entités nommées composées, particulièrement dans les cas d'annexion ou de constructions possessives.

Ces illustrations qualitatives montrent que l'injection linguistique ne se contente pas d'améliorer les scores numériques : elle renforce l'interprétabilité des décisions du modèle en les reliant directement à des propriétés grammaticales identifiables de l'arabe.

Tableau 4.2 : Illustration de l'annotation morphologique et de la représentation LiPAr de l'exemple *وضع الكتاب على المكتب*

Token	كتاب	المكتب
POS (prompt)	NOUN	NOUN
CASE (prompt)	NOM	GEN
DEF (prompt)	DEF	DEF
Représentation LiPAr	F_i combinant E_i^{BERT} et les trois traits projetés → prédiction robuste de l'état construit et du cas	

4.9 Limites identifiées et perspectives d'amélioration

Malgré ses aouts, le cadre LiPAr présente plusieurs limites :

1. **Dépendance à la qualité de l'analyseur externe** : la précision de Stanza conditionne directement la qualité du signal injecté. Toute erreur d'analyse morphologique peut introduire du bruit.
2. **Alignement subword** : bien que robuste, la stratégie assign-first reste une approximation ; une agrégation plus sophistiquée pourrait être explorée sur des corpus à forte variation morphologique.
3. **Surcoût computationnel** : l'ajout du module de fusion et de projection augmente légèrement le temps d'inférence (environ 10-15 % selon les configurations).
4. **Couverture limitée aux traits rares** : les traits peu fréquents (ex. cas rares ou schèmes rares) peuvent être sous-représentés et nécessiter une régularisation spécifique.

Ces limites ouvrent des perspectives d'amélioration : intégration d'un analyseur morphologique plus robuste (ex. CAMeL Tools), mécanismes d'auto-correction du signal linguistique, ou extension à des architectures plus récentes (AraT5, Jais).

4.10 Synthèse du cadre proposé et perspectives expérimentales

Ce chapitre a posé les fondations de LiPAr : une architecture hybride modulaire, reproductible et linguistiquement motivée pour l'intégration de prompts structurés dans les modèles Transformer appliqués à l'arabe. L'ambition n'est pas de remplacer BERT, mais de l'enrichir ou plutôt, de lui donner à voir ce qu'il ne perçoit pas spontanément.

L'idée de départ est, en apparence, assez simple : injecter de l'information linguistique explicite dans un encodeur pré-entraîné. Mais cette simplicité de façade dissimule une série de décisions techniques non triviales, que nous avons dû trancher une à une : comment extraire des traits morphosyntaxiques fiables pour l'arabe ? Comment les aligner avec des sous-tokens qui ne respectent pas les frontières morphémiques ? Comment les fusionner avec les représentations contextuelles sans écraser le signal original ? Et enfin, comment doser leur influence pour ne pas perturber ce que le modèle a déjà appris ?

Les réponses que nous avons apportées à ces questions débouchent sur un protocole de validation centré sur les choix suivants : une stratégie de fusion précoce par concaténation (dimensions $d + d_m$), projetée vers d par une couche linéaire unifiée puis stabilisée par LayerNorm, avec une politique assign-first pour la gestion de l'alignement subword. Trois backbones seront confrontés à ce dispositif (mBERT, AraBERTv02 et MARBERT) sur deux tâches pivots : le POS tagging (corpus UD Arabic-PADT) et le NER (ANERcorp). Ces choix opèrent dans le cadre de trois hypothèses structurantes : H1 teste l'existence d'un gain global, H2 interroge la sensibilité de ce gain selon le backbone, et H3 explore la hiérarchie de pertinence entre les différents traits linguistiques.

In fine, LiPAr ne cherche pas uniquement à grappiller quelques points de F1. Ce qui nous intéresse davantage, c'est de savoir si un biais linguistique explicite, soigneusement construit, peut réellement faire une différence dans des architectures qui ont été massivement pré-entraînées sur des milliards de tokens et, si oui, dans quelles conditions et pour quelles raisons.

Les chapitres suivants mettront ce programme à l'épreuve. Le chapitre 5 s'attardera sur les aspects pratiques d'implémentation et les conditions de reproductibilité ; le chapitre 6 présentera et discutera les résultats expérimentaux, en les confrontant aux hypothèses énoncées ici.

Chapitre 5 - Méthodologie expérimentale

5.1 Introduction générale

Un cadre théorique, aussi solide soit-il, ne vaut que par sa mise à l'épreuve des faits. Ce chapitre décrit précisément comment nous avons confronté LiPAr à cette épreuve. Il ne s'agit pas seulement de détailler des choix techniques, mais de montrer comment, concrètement, nous avons traduit les principes du chapitre 4 en un pipeline expérimental robuste, reproductible et, nous l'espérons, utile à la communauté. Il rend compte, de manière exhaustive et reproductible, des choix architecturaux, des décisions implémentationnelles et expérimentales, des jeux de données mobilisés, des hyperparamètres retenus et des procédures d'entraînement qui ont permis la validation des hypothèses formulées au chapitre précédent.

L'ambition méthodologique est double : (i) documenter avec précision l'ensemble du pipeline expérimental, depuis le prétraitement des données jusqu'à l'évaluation finale des modèles enrichis, et (ii) garantir la pleine reproductibilité des résultats conformément aux standards les plus exigeants de la recherche en traitement automatique des langues (TAL).

5.2 Architecture générale du pipeline expérimental LiPAr

Le pipeline LiPAr, représenté schématiquement à la **Figure 4.1**, repose sur une séquence d'étapes modulaires, entièrement automatisées et conçues pour une intégration fluide dans l'écosystème des modèles Transformer pré-entraînés. Cette architecture se décompose en six phases principales :

1. **Acquisition** des textes bruts issus des corpus de référence ;
2. **Prétraitement et normalisation** spécifiques à la langue arabe ;
3. **Encodage bimodal LiPAr** : vectorisation contextuelle BERT et extraction structurée des traits linguistiques ;
4. **Projection et fusion** des représentations selon les principes du cadre LiPAr ;
5. **Entraînement et optimisation** des modèles hybrides ;
6. **Évaluation quantitative** et interprétation des performances.

5.2.1 Prétraitement et normalisation

Les données expérimentales proviennent de deux corpus de référence reconnus pour la qualité de leurs annotations et leur adoption par la communauté scientifique : **UD_Arabic-PADT** (Nivre et al. 2020) pour l'étiquetage morpho-syntaxique (POS tagging) et **ANERcorp** (Benajiba, Diab, Rosso 2008) pour la reconnaissance d'entités nommées (NER).

Conformément aux meilleures pratiques du TAL arabe, les textes ont été soumis aux traitements suivants :

- ***Nettoyage*** : suppression systématique des diacritiques (simulant les conditions réelles d'usage) et élimination des caractères non arabes ;
- ***Normalisation orthographique*** : unification des variantes graphiques de l'alif (→ ٱ/إ/أ/آ) et du yāʾ (ي → ى) ;
- ***Tokenisation*** : conforme aux tokeniseurs WordPiece des modèles BERT retenus ;
- ***Conversion des schémas d'annotation*** : adoption du format BIO pour la tâche NER.

Les annotations morpho-syntaxiques sont initialement alignées au niveau des tokens linguistiques issus de l'analyseur UD, puis projetées sur les sous-tokens WordPiece selon une stratégie décrite en Section 5.2.3.

5.2.2 Encodage et vectorisation

Chaque phrase est encodée par l'un des trois modèles pré-entraînés suivants, évalués séparément afin de comparer leur comportement dans le cadre LiPAr :

- **AraBERT** (Antoun, Baly, Hajj 2020), spécialisé pour l'arabe et entraîné sur des corpus volumineux arabophones ;
- **mBERT**, version multilingue de BERT (Devlin et al. 2019), permettant la comparaison avec une approche non spécifique à la langue.
- **MARBERT** (Abdul-Mageed, Elmadany, Nagoudi 2021) (UBC-NLP/MARBERT).

Chaque token est transformé en un embedding dense contextualisé ($E_i^{BERT} \in \mathbb{R}^d$), capturant à la fois la structure syntaxique locale et les dépendances sémantiques globales.

5.2.3 Extraction et représentation des traits linguistiques structurés

L'extraction des prompts linguistiques, élément central du cadre LiPAr, a été réalisée à l'aide de l'analyseur morpho-syntaxique **Stanza** (Qi et al. 2020), qui fournit pour chaque token une annotation morpho-syntaxique conforme au schéma Universal Dependencies (UD v2).

Quels traits retenir ? La morphologie arabe est si riche qu'on pourrait en extraire des dizaines. Nous en avons sélectionné 32, répartis en huit catégories, en suivant un principe simple : chaque trait devait être à la fois linguistiquement pertinent et fiablement extractible par Stanza. Le Tableau 5.1 résume cette sélection :

Tableau 5.1 : Traits retenus comme prompts

Type de trait	Sous-catégories	Dimensions
Genre	Masculin, Féminin, Aucun	3
Nombre	Singulier, Duel, Pluriel, Aucun	4
Cas	Nominatif, Accusatif, Génitif, Aucun	4
Définition	Défini, Indéfini, Aucun	3
Personne	1ère, 2e, 3e, Aucun	4
Mode verbal	Indicatif, Subjonctif, Jussif, Impératif, Aucun	5
Aspect	Accompli, Inaccompli, Aucun	3
Voix	Active, Passive, Aucun	3
Traits binaires	Foreign, NumForm, AdpType	3

Chaque trait est encodé par un vecteur **one-hot**, assurant la transparence interprétative et évitant tout apprentissage implicite biaisé.

***Exemple* :** Pour le token "يكتبون" (ils écrivent), le vecteur morphologique encode les valeurs : masculin, pluriel, aucun cas, aucune définition, 3e personne, indicatif, inaccompli, actif.

Gestion des sous-tokens : Pour un mot comme الكتاب (*al-kitāb*, « le livre »), la tokenisation BERT produit deux sous-tokens : ال (*al-*, l'article défini) et ##كتاب (*kitāb*, le nom). Conformément à la stratégie assign-first :

- seul le premier sous-token hérite des traits morphologiques complets ;
- les sous-tokens suivants reçoivent un vecteur nul et sont exclus du calcul de la perte via *ignore_index* = -100, évitant ainsi toute propagation artificielle du signal morphologique.

Les annotations morphosyntaxiques utilisées comme prompts sont produites automatiquement par Stanza, indépendamment des partitions d'entraînement, de validation et de test du modèle cible. Aucune information issue des annotations gold des ensembles de test n'est utilisée pour la génération des traits linguistiques injectés, garantissant ainsi l'absence de fuite d'information (*data leakage*).

5.3 Architecture d'entraînement et stratégie de fusion LiPAr

Cette section décrit la traduction implémentationnelle concrète des mécanismes formalisés au Chapitre 4, aboutissant à la configuration canonique d'entraînement du cadre LiPAr, utilisée comme référence dans l'ensemble des expériences. Il convient de préciser que, contrairement à une injection directe au niveau des embeddings d'entrée du Transformer, l'intégration des traits linguistiques dans LiPAr est réalisée après l'encodage contextuel complet par BERT. Ce choix méthodologique permet de préserver l'architecture interne du modèle pré-entraîné tout en assurant une interaction contrôlée entre représentations neuronales et information linguistique explicite.

Le modèle d'entraînement repose sur une architecture hybride à trois composantes :

- un **encodeur BERT** (AraBERT, mBERT ou MARBERT) pour la représentation neuronale ;
- un **classifieur linéaire** pour la prédiction des étiquettes POS/NER ;
- et, dans les versions hybrides, une **couche de projection morphologique** reliant les traits linguistiques à l'espace latent de BERT.

5.3.1 Fusion précoce des représentations (stratégie principale)

La fusion précoce, retenue comme stratégie optimale dans le cadre LiPAr, repose sur l'intégration directe des représentations linguistiques dans l'espace de représentation post-encodage dérivé des sorties de BERT. Concrètement, les traits morphologiques sont projetés dans un espace latent intermédiaire de dimension d_m, puis concaténés aux représentations contextuelles $E_i^{BERT} \in \mathbb{R}^d$. L'ensemble est ensuite projeté linéairement dans l'espace $\mathbb{R}^d$ conformément à la formalisation du chapitre 4.

Les représentations neuronales contextuelles E_i^{BERT} et morphologiques projetées $\tilde{E}_i^{morph}$ sont concaténées puis projetées linéairement dans l'espace latent :

$$F_i = LayerNorm\left(W_f\left[E_i^{BERT}; \tilde{E}_i^{morph}\right] + b_f\right) \qquad (5.1)$$

où W_f et b_f sont respectivement la matrice de projection et le vecteur de biais du module de fusion, appris conjointement avec le reste du modèle. Le biais b_f permet un ajustement affine de la représentation fusionnée avant normalisation, contribuant à la stabilité et à la flexibilité du mécanisme d'injection linguistique en réglant l'échelle des activations.

Cette opération est réalisée avant la tête de classification, assurant une interaction profonde entre les deux sources d'information tout en préservant la dimensionnalité.

5.3.2 Fonction de perte et optimisation

La fonction de perte employée est la **Cross-Entropy Loss**, avec un paramètre *ignore_index* = -100 permettant d'exclure :

- les tokens spéciaux [CLS] et [SEP],
- les tokens de remplissage (padding),
- les sous-tokens non initiaux.

Elle est définie par :

$$\mathcal{L} = -\frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} m_i \cdot \log P\left(y_i | t_i; \Theta\right) \qquad (5.2)$$

où m_i est le masque binaire excluant les tokens spéciaux et sous-tokens secondaires.

Cette formulation permet une interaction directe et différentiable entre les représentations neuronales contextuelles et les contraintes linguistiques explicites, intégrées comme biais inductif au niveau de la représentation finale utilisée pour la prédiction.

L'optimisation est assurée par l'algorithme **AdamW** (Loshchilov, Hutter 2019), reconnu pour améliorer la stabilité de l'entraînement des modèles pré-entraînés grâce à une régularisation découplée du *weight decay*.

Le recours à un décodeur CRF n'a pas été retenu dans la configuration canonique afin d'isoler strictement l'impact de l'injection linguistique LiPAr, sans introduire de biais structurel supplémentaire au niveau du décodage.

5.3.3 Hyperparamètres d'entraînement

Les hyperparamètres, choisis conformément aux recommandations du fine-tuning des modèles BERT, sont recensés dans le Tableau 5.2 :

Tableau 5.2 : Hyperparamètres d'entraînement

Hyperparamètre	Valeur	Justification
Taux d'apprentissage (BERT)	5×10^{-5}	Standard pour éviter le catastrophic forgetting
Taux d'apprentissage (classifieur)	1×10^{-3}	Apprentissage rapide des couches ajoutées
Taille de batch	16	Compromis stabilité/mémoire GPU (limite)
Nombre d'époques	5	Convergence assurée avec early stopping
Scheduler	Linéaire	Ajustement progressif du taux d'apprentissage
Warmup	10% steps	Stabilisation initiale (standard BERT)
Graine (seed)	42	Reproductibilité
Décroissance de poids	0.01	Régularisation L2 pour éviter l'overfitting.

Ces valeurs n'ont pas été choisies au hasard. Le taux d'apprentissage de 5×10^{-5} pour BERT est un standard qui évite d'écraser les connaissances pré-entraînées (*catastrophic forgetting*). Celui de 10^{-3} pour les modules LiPAr, plus élevé, leur permet d'apprendre rapidement leur rôle sans perturber le reste. La taille de batch de 16 est un compromis imposé par la mémoire GPU, mais l'accumulation de gradient nous a permis de simuler des batchs plus grands quand nécessaire.

L'entraînement a été réalisé sur GPU NVIDIA Tesla V100. La sélection du meilleur modèle repose sur le score F1-macro obtenu sur l'ensemble de validation. La reproductibilité est assurée par la fixation des *seeds* (42, 123, 456), des versions logicielles (PyTorch 2.2, Transformers 4.36, Stanza 1.4) et de la même partition de données.

Avant d'analyser comparativement différentes stratégies de fusion, nous présentons dans le Tableau 5.3 l'impact global de l'injection linguistique LiPAr selon la configuration canonique de fusion précoce, par rapport aux modèles BERT de référence non enrichis. Ces résultats préliminaires permettent d'établir l'intérêt empirique du cadre proposé, indépendamment des variantes explorées ultérieurement.

Ces résultats sont présentés à titre illustratif du comportement global de la configuration canonique ; l'analyse détaillée et statistique complète est reportée au chapitre 6.

Tableau 5.3 : Impact de l'injection linguistique LiPAr (fusion précoce) par rapport aux baselines BERT.

Modèle	Configuration	Accuracy	Δ	F1-macro	Δ
AraBERT	Sans traits	0.9600	–	0.9285	–
AraBERT	Avec traits	0.9821	+2.21%	0.9446	+1,73%
mBERT	Sans traits	0.9703	–	0.9395	–
mBERT	Avec traits	0.9821	+1,18%	0.9546	+1,61%

AraBERT gagne +2,21 points d'accuracy, mBERT +1,61 points de F1-macro. Ces premiers chiffres, nets et immédiats, nous ont confortés dans la poursuite de l'approche. L'analyse détaillée et statistique complète est reportée au chapitre 6.

5.4 Etude comparative des stratégies de fusion

Afin de valider empiriquement le choix de la fusion précoce comme configuration canonique du cadre LiPAr (cf. section 5.3), nous menons dans cette section une étude comparative contrôlée portant sur des stratégies alternatives d'intégration du signal morpho-syntaxique explicite.

L'objectif n'est pas de proposer de nouvelles architectures de référence, mais d'évaluer l'impact du moment d'injection et du niveau d'interaction entre l'espace neuronal contextuel et l'espace linguistique symbolique.

Toutes les variantes présentées ci-dessous partagent le même backbone BERT, les mêmes jeux de données, les mêmes hyperparamètres d'entraînement et les mêmes critères d'évaluation. La seule variable contrôlée est la stratégie de fusion.

5.4.1 Fusion tardive (exploration complémentaire)

Contrairement à la fusion précoce, où les traits linguistiques sont intégrés immédiatement après la représentation contextuelle finale produite par le Transformer et avant la couche de classification, la fusion tardive opère au niveau des sorties décisionnelles du modèle.

Dans cette configuration, les traits morpho-syntaxiques explicites ne modifient pas les représentations internes de BERT. Ils alimentent un classifieur auxiliaire indépendant, dont les prédictions sont ensuite combinées avec celles du modèle BERT de référence.

Deux architectures auxiliaires ont été explorées :

a) MLP morpho-linguistique avec attention

Les traits morphologiques (32 dimensions) sont projetés successivement dans des espaces intermédiaires (32 → 64 → 128), avec des activations ReLU et un dropout de 0.3. Un mécanisme d'attention multi-tête (4 têtes, dropout d'attention 0.1) permet de pondérer dynamiquement les traits en fonction du contexte, avant une projection finale vers l'espace des étiquettes de la tâche considérée (POS-tagging ou NER).

b) Décodeur CRF morpho-syntaxique

Une alternative consiste à modéliser explicitement les dépendances séquentielles entre étiquettes morpho-syntaxiques via un CRF linéaire-chaîne.

La fonction de perte associée au CRF est la log-vraisemblance négative définie par :

$$\mathcal{L}_{CRF} = -\left(s(x,y) - log \sum_{y' \in \Upsilon} exp\big(s(x,y')\big) \right) \qquad (5.3)$$

où s(x,y) désigne le score de la séquence correcte et Υ l'ensemble des séquences possibles.

Les traits morphologiques projetés servent à calculer les scores d'émission, tandis qu'une matrice de transition apprend les contraintes syntaxiques locales propres à l'arabe.

Combinaison adaptative des prédictions

Les prédictions issues du modèle BERT (y_{BERT}) et du module linguistique (y_{ling}) sont combinées par un mécanisme de gating adaptatif appris :

$$\hat{y} = \alpha \cdot y_{BERT} + \beta \cdot y_{ling} \quad \text{avec } \alpha + \beta = 1 \qquad (5.4)$$

où α et β sont produits par un petit réseau de gating (768 → 128 → 2, activation softmax) prenant en entrée la représentation [CLS] du modèle BERT.

Ce dispositif permet au modèle de moduler dynamiquement la contribution relative des deux sources d'information en fonction du contexte.

5.4.2 Résultats comparatifs des stratégies de fusion

Les performances obtenues sur l'ensemble de test, moyennées sur trois runs indépendants, sont présentées dans le Tableau 5.4.

Tableau 5.4 : Performances comparées des stratégies de fusion (F1-macro, ensemble de test)

Modèle	Baseline	Fusion précoce	Fusion tardive	Δ Précoce	Δ Tardive
AraBERT	0.9285	0.9446	0.9480	+1.73 %	+2.10 %
mBERT	0.9395	0.9546	0.9570	+1.61 %	+1.86 %

Les résultats montrent que la fusion tardive peut apporter un gain marginal supplémentaire (+0.16 à +0.37 points de F1-macro) par rapport à la fusion précoce. Toutefois, cet avantage s'accompagne d'un surcoût paramétrique (~150k paramètres supplémentaires) et d'un temps d'inférence accru (+12 à +18 % selon le modèle).

Bien que la fusion tardive obtienne des scores légèrement supérieurs, ces gains restent marginaux et statistiquement faibles au regard du surcoût computationnel et de la complexité paramétrique introduite. En effet, l'écart observé reste inférieur au seuil de significativité pratique au regard de la variance inter-runs ($\sigma < 0.004$ sur le F1-macro). Dans un cadre visant une intégration linguistique structurelle et généralisable, la fusion précoce demeure préférable.

5.4.3 Analyse coût / bénéfice et discussion linguistique

Une analyse qualitative révèle que la fusion tardive est particulièrement bénéfique dans certains cas limites de désambiguïsation morpho-syntaxique, notamment :

- **Ambiguïtés nom / verbe** (ex. « عمل ») : ΔF1 ≈ +0.010
- **Formes passives non diacritisées** (ex. « كُتب ») : ΔF1 ≈ +0.009
- **Mots étrangers ou translittérés** (ex. « إنترنت » ,« داتا ») : ΔF1 ≈ +0.017

Ces gains ciblés confirment que les traits linguistiques explicites apportent une information corrective utile, même lorsqu'ils ne sont injectés qu'au niveau décisionnel.

Cependant, au regard de la simplicité architecturale, de la stabilité de l'apprentissage, du faible surcoût computationnel et de la propagation globale du biais linguistique, la fusion précoce offre le meilleur compromis global entre performance, interprétabilité et efficacité.

Le Tableau 5.5 synthétise les propriétés comparées des deux stratégies :

Tableau 5.5 : Comparaison synthétique des stratégies de fusion précoce et tardive dans le cadre LiPAr

Type de fusion	Moment d'intégration	Paramètres supplémentaires	Avantages principaux	Inconvénients principaux
Précoce	Représentations contextuelles finales (post-transformer)	Faibles (d x d_m + projection morphologique)	Simplicité, vitesse, interaction profonde	Moins de flexibilité contextuelle
Tardive	Niveau logits	~150 k	Adaptation dynamique, meilleure interprétabilité	Surcoût computationnel (+12–18 %)

Conclusion de la section

La fusion tardive peut, dans certains cas, grignoter quelques dixièmes de point supplémentaires. Mais au prix d'une complexité accrue, d'un surcoût computationnel, et d'une perte de simplicité. La fusion précoce, elle, offre l'essentiel du gain avec une élégance et une stabilité qui la rendent naturellement préférable. C'est donc elle que nous retenons pour la suite.

Ces résultats justifient le recours prioritaire à la fusion précoce dans les expériences principales rapportées au Chapitre 6, tout en ouvrant des perspectives pour des extensions futures ciblées.

5.5 Cadre d'évaluation et métriques

L'ensemble des expérimentations a été conduit sur le corpus UD_Arabic-PADT (Nivre et al. 2020), avec une partition stratifiée fixe :

- ***Entraînement*** : 80 %
- ***Validation*** : 10 % (sélection du meilleur checkpoint et early stopping)
- ***Test*** : 10 % (rapport des performances finales)

Quatre métriques standards, (**Tableau** 5.6) ont été calculées :

Tableau 5.6 : Métriques d'évaluation retenues et leur interprétation

Métrique	Définition classique	Interprétation dans le présent travail
Accuracy	(TP + TN) / Total	Fidélité globale du modèle
Precision	TP / (TP + FP)	Capacité à minimiser les faux positifs
Recall	TP / (TP + FN)	Exhaustivité de la détection
F1-score	2 × (P × R) / (P + R)	Mesure équilibrée, métrique principale

Le F1-macro a été privilégié comme critère principal, conformément aux recommandations pour l'évaluation de tâches de classification avec classes déséquilibrées (Sokolova, Lapalme 2009), afin d'accorder un poids équitable aux classes rares (ex. : duel, jussif, mots étrangers), particulièrement sensibles aux erreurs dans les langues morphologiquement riches. Le F1-micro est rapporté à titre complémentaire pour évaluer la performance sur les classes majoritaires.

Toutes les expériences ont été répétées trois fois avec des graines d'initialisation distinctes (42, 123, 456). L'écart-type observé sur le F1-macro calculé sur l'ensemble de test, sur trois exécutions indépendantes, reste systématiquement inférieur à 0.004, témoignant de la grande stabilité et de la robustesse des résultats rapportés.

5.6 Synthèse du chapitre

Ce chapitre a décrit, dans le détail, comment nous avons construit le pont entre la théorie (chapitre 4) et l'expérimentation (chapitre 6). Chaque choix, du prétraitement des données à la stratégie de fusion, en passant par le réglage des hyperparamètres, a été guidé par un souci constant de rigueur et de reproductibilité. Le pipeline est maintenant prêt. Reste à voir ce qu'il révèle. C'est l'objet du chapitre suivant.

Le chapitre suivant (Chapitre 6) procédera à l'analyse exhaustive des résultats expérimentaux, permettant une validation définitive des hypothèses théoriques et une discussion approfondie des apports du cadre LiPAr à la recherche en traitement automatique de la langue naturelle appliqué à l'arabe.

Ce chapitre constitue ainsi le socle expérimental sur lequel repose l'analyse critique des résultats présentée au Chapitre 6, permettant de relier de manière rigoureuse les choix méthodologiques aux observations empiriques analysées.

Chapitre 6 - Analyse des Résultats et Validation

6.1 Introduction

Dans ce chapitre, nous analysons en détail l'efficacité du cadre LiPAr, implémenté au chapitre précédent. Loin de nous contenter de chiffres bruts, nous cherchons à décrypter comment, et pourquoi, l'injection de traits morpho-syntaxiques modifie le comportement des modèles neuronaux face à la complexité propre de l'arabe.

Trois questions structurent cette analyse :

1. Quelles raisons linguistiques et computationnelles expliquent les écarts de performance entre les modèles multilingues (mBERT) et les modèles arabophones spécialisés (AraBERT, MARBERT) ?
2. Dans quelle mesure l'injection de traits morpho-syntaxiques améliore-t-elle la modélisation des phénomènes morphologiques et syntaxiques caractéristiques de l'arabe ?
3. Les gains observés relèvent-ils d'une simple amélioration incrémentale ou d'une modification méthodologique structurée du processus d'apprentissage ?

Pour y répondre, nous adoptons une démarche en trois volets complémentaires :

- **Analyse quantitative** : évaluation systématique basée sur les scores F1, les gains absolus et relatifs, ainsi que des tests de significativité (test de McNemar avec correction de Yates et correction de Bonferroni).
- **Analyse qualitative** : étude détaillée des erreurs typiques, des cas ambigus et des patrons morpho-syntaxiques responsables des améliorations ou dégradations.
- **Discussion linguistico-méthodologique** : mise en perspective des résultats à la lumière des propriétés morphologiques de l'arabe et du cadre méthodologique proposé, LiPAr.

Cette double approche, quantitative et qualitative, permet d'aller au-delà d'une lecture strictement métrique des résultats.

6.2 Synthèse quantitative des performances

Les expériences du chapitre 5 révèlent un contraste frappant entre les deux tâches : alors que le POS tagging bénéficie d'un gain systématique pour tous les modèles, le NER présente une forte dépendance au pré-entraînement arabe. Cette section quantifie ces écarts avant d'en explorer les causes linguistiques.

6.2.1 POS tagging : effets nets et systématiques

Les résultats obtenus indiquent une amélioration des performances pour les trois modèles étudiés, mais avec des gains très contrastés : de +1,54 à +6,64 points de F1 selon le modèle, avec une progression régulière de mBERT → AraBERT → MARBERT en termes de F1 de base, mais une dynamique inverse pour les gains. Le Tableau 6.1 synthétise les performances obtenues par les trois modèles, avec et sans injection de traits morpho-syntaxiques.

Tableau 6.1 : Scores F1 (POS) par modèle (Baseline vs +32 traits)

Modèle	Baseline (F1)	+32 traits (F1)	Δ F1	χ^2 (McNemar + Yates)	p-value	p corrigé (Bonferroni, k=6)
mBERT	0,9009	0,9673	+0,0664	48,2	< 0,001	< 0,001
AraBERT	0,9311	0,9468	+0,0157	12,8	0,002	0,012
MARBERT	0,9442	0,9596	+0,0154	11,9	0,003	0,018

La Figure 6.1 illustre clairement l'effet systématique de l'injection des traits morphologiques : mBERT montre l'amélioration la plus marquée, tandis que les gains restent robustes mais plus modérés pour AraBERT et MARBERT.

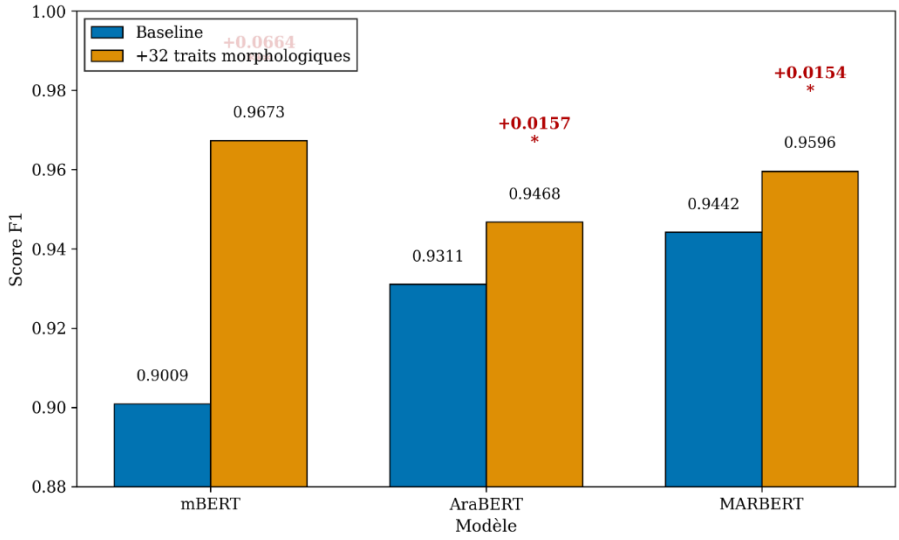


Figure 6.1 : Comparaison des scores F1 (POS) par modèle (Baseline vs avec traits)

Analyse

1. **mBERT bénéficie le plus fortement des traits morphologiques (+6,64 pts F1) :** Ce gain très élevé indique que les modèles multilingues, malgré leur capacité générale, manquent d'une représentation explicite de la morphologie arabe, laquelle est efficacement compensée par l'injection de traits explicites.
2. **AraBERT et MARBERT présentent des gains plus modestes mais statistiquement significatifs :** Ce résultat est cohérent avec leur pré-entraînement massif sur des corpus arabes : ces modèles intègrent déjà une bonne partie de la morphologie arabe, mais continuent d'être améliorés par des signaux complémentaires explicites utiles.
3. **La significativité est robuste dans les trois cas :** Les tests de McNemar avec correction de Yates (McNemar 1947), suivis d'une correction de Bonferroni pour comparaisons multiples (Dunn 1961), confirment que les améliorations obtenues ne relèvent pas du

hasard, conformément aux bonnes pratiques d'évaluation statistique en NLP (Dror et al. 2018; Dietterich 1998).

Conclusion intermédiaire

Ces résultats confirment l'hypothèse selon laquelle les traits morphologiques fournissent un signal linguistique complémentaire aux embeddings contextuels particulièrement pour les modèles non spécialisés en arabe.

Insight linguistique 6.1

Les traits morpho-syntaxiques rendent explicites certaines distinctions grammaticales qui restent implicites dans l'écriture arabe non diacritée.

6.2.2 Résultats en NER : forte dépendance au pré-entraînement arabe

Contrairement au POS tagging, l'effet de l'injection morphologique sur le NER apparaît beaucoup plus variable.

Le Tableau 6.2 présente les performances pour les trois modèles, avec et sans traits.

Tableau 6.2 : Scores F1 (NER) par modèle (baseline vs +32 traits)

Modèle	Baseline	+32 traits	Δ F1	χ^2 (McNemar+Yates)	p-value	p corrigé (Bonferroni, k=6)
mBERT	0.7819	0.7896	+0,0078	4.08	0.043	0.260
AraBERT	**0.8180**	0.8430	**+0,0250**	92.12	$< 2.2\times10^{-16}$	$< 1.3\times10^{-15}$
MARBERT	0.8288	0.8437	+0,0149	15.63	4.6×10^{-4}	4.6×10^{-4}

La Figure 6.2 montre une dynamique différente de celle observée en POS : les gains sont nettement plus hétérogènes, ce qui suggère une dépendance plus forte du NER à la qualité du pré-entraînement.

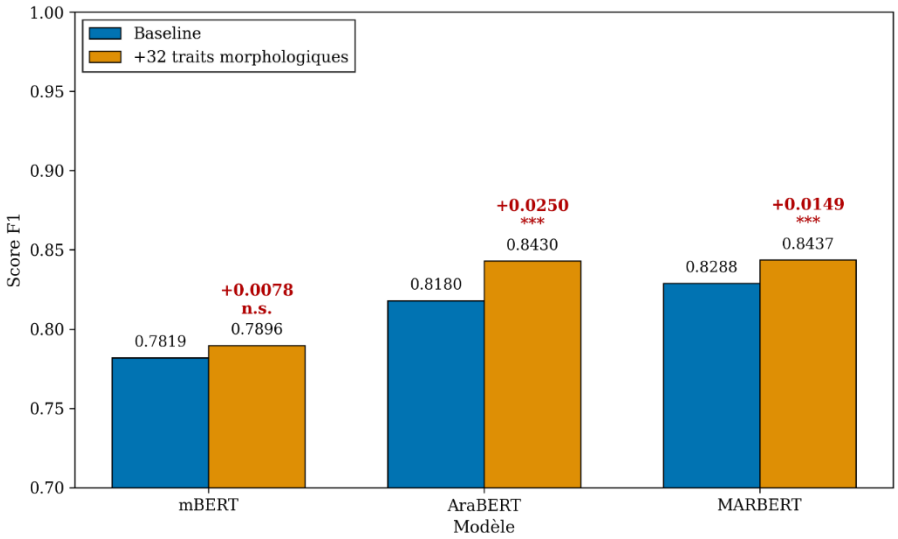


Figure 6.2 : Comparaison des scores F1 (NER) par modèle (Baseline vs avec traits)

Analyse

1. **AraBERT présente le gain le plus élevé (+2.5 pts F1), avec forte significativité.** Son pré-entraînement sur un large corpus MSA lui permet d'intégrer les traits morphologiques de façon plus synergique.
2. **Pour mBERT, le gain est marginal et non significatif après correction.** Les traits améliorent certaines classes spécifiques (notamment I-ORG), mais leur effet global demeure limité.
3. **MARBERT obtient un gain modéré (+1,49 pt)**, cohérent avec son pré-entraînement dialectal mais bruyant (tweets).
4. **La tâche NER exploite les schémas morphologiques de manière moins directe que celle du POS.** Les traits contribuent essentiellement à des sous-problèmes ciblés : segmentation des syntagmes nominaux, détection des annexions, distinctions internes aux entités.

6.2.3 Comparaison transversale POS vs NER

Une synthèse des gains par tâche pour les 3 modèles est présentée dans le Tableau 6.3.

Tableau 6.3 : Synthèse des gains ΔF1 (POS vs NER)

Modèle	Δ F1 POS	Δ F1 NER
mBERT	+0,0664	+0,0078
AraBERT	+0,0157	+0,0250
MARBERT	+0,0154	+0,0149

La Figure 6.3 confirme que les traits morphologiques ont un impact constant sur le POS, mais un effet conditionnel sur le NER.

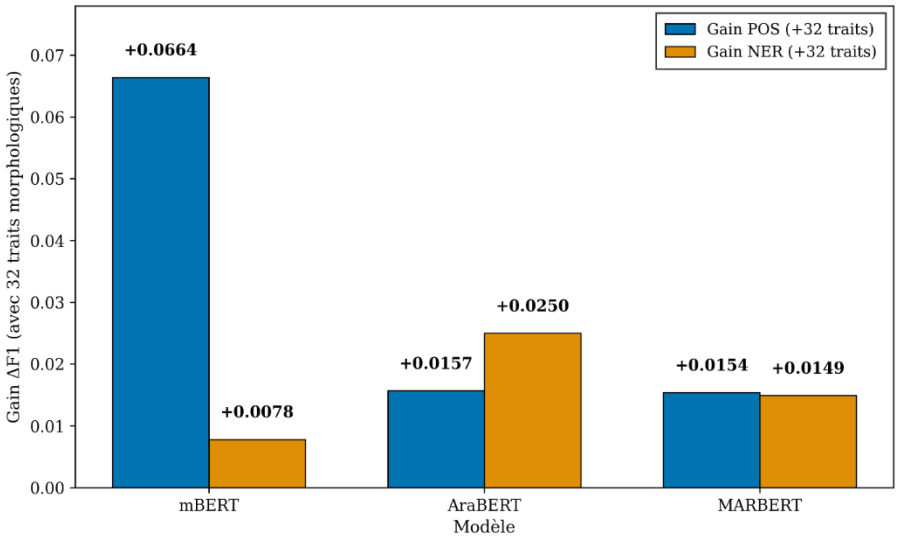


Figure 6.3 : Comparaison des gains Δ F1 par modèle et tâche (POS vs NER)

Observations

La comparaison transversale fait apparaître deux profils distincts. En POS, les gains sont homogènes entre modèles arabes (+1,54 à +1,57 pts) mais explosent pour mBERT (+6,64

pts), signe que les traits compensent une carence structurelle. En NER, en revanche, la variabilité domine : pic à +2,50 pts pour AraBERT, effet marginal non significatif pour mBERT, situation intermédiaire pour MARBERT. La morphologie arabe est donc plus directement utile pour une tâche syntactique que pour une tâche sémantico-structurelle.

6.3 Analyse par classes du POS : apports ciblés des traits morphologiques

Les gains globaux en POS masquent en réalité des disparités importantes entre catégories grammaticales. Si les noms, adjectifs et verbes concentrent l'essentiel des améliorations, les particules et adverbes restent quasi inchangés ; une disparité qui révèle le mode d'action spécifique des traits morphologiques.

La Figure 6.4 permet d'identifier rapidement les catégories les plus sensibles aux traits, notamment les noms, les adjectifs et les verbes, caractérisés par une forte dépendance morphologique.

Pour une lecture plus détaillée par modèle, nous présentons ci-dessous les barplots ΔF1 par classe.

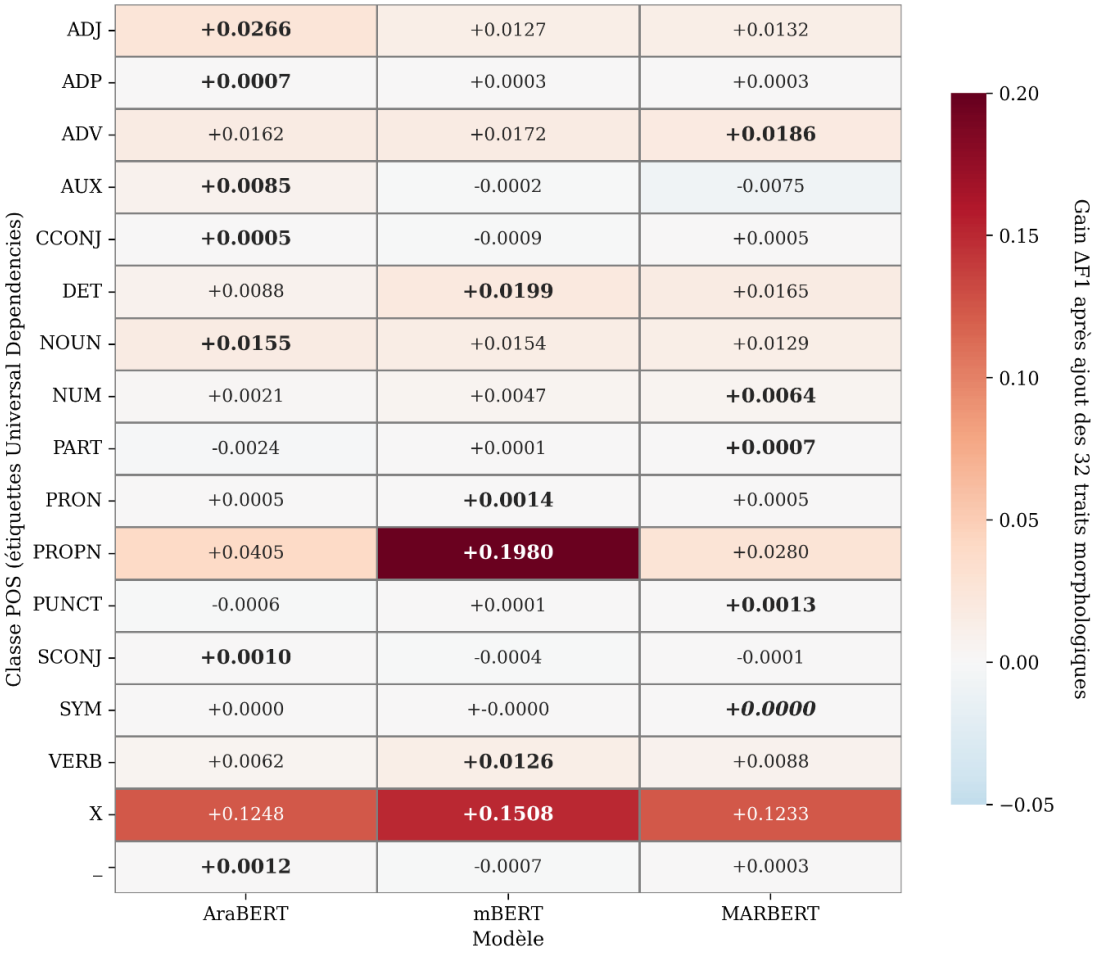


Figure 6.4 : Heatmap des gains par classes POS.
(La heatmap visualise les écarts de performance sur 17 classes POS pour AraBERT, mBERT et MARBERT, mettant en évidence les catégories les plus sensibles aux traits morphologiques.)

Principaux enseignements

1. **Les noms et adjectifs (NOUN, ADJ) sont les grands bénéficiaires :** Les traits nominaux (notamment CASE, DEFINITE et NUMBER) améliorent fortement la désambiguïsation interne des syntagmes nominaux et adjectivaux. Le phénomène est particulièrement marqué pour mBERT, qui gagne jusqu'à +25 points sur NOUN.
2. **Les verbes (VERB) profitent surtout de VERB_FORM, VOICE et PERSON :** Ces traits facilitent la segmentation morphologique interne des verbes, en distinguant :
   - perfectif vs imperfectif,
   - formes actives vs passives,
   - suffixes pronominaux indiquant la personne, le genre ou le nombre.
3. **Les adverbes et particules (ADV, PART, PUNCT) sont peu affectés :** Ces classes, moins riches morphologiquement, bénéficient d'une amélioration marginale.
4. **Les noms propres (PROPN) bénéficient, paradoxalement, du signal morphologique :** Bien qu'ils soient considérés comme peu sensibles à la morphologie, certains traits (en particulier **DEFINITE** et certains traits morphologiques nominaux et dérivationnels), contribuent à stabiliser la segmentation des noms propres.

Conclusion intermédiaire

Les gains en POS proviennent principalement d'un renforcement de la structure interne des syntagmes nominaux et verbaux, où la morphologie arabe joue un rôle crucial. Les traits explicitent des distinctions normalement absentes à la surface (à cause du manque de diacritiques), réduisant ainsi l'ambiguïté lexicale, dérivationnelle et flexionnelle.

6.3.1 Le modèle AraBERT : gains ciblés sur les classes morphologiques

La Figure 6.5 montre que, sous AraBERT, les gains les plus importants concernent les classes fortement morphologiques (NOUN, ADJ, VERB) ainsi que la classe X (termes étrangers / chiffres). Les classes fonctionnelles (PART, PUNCT) restent peu affectées.

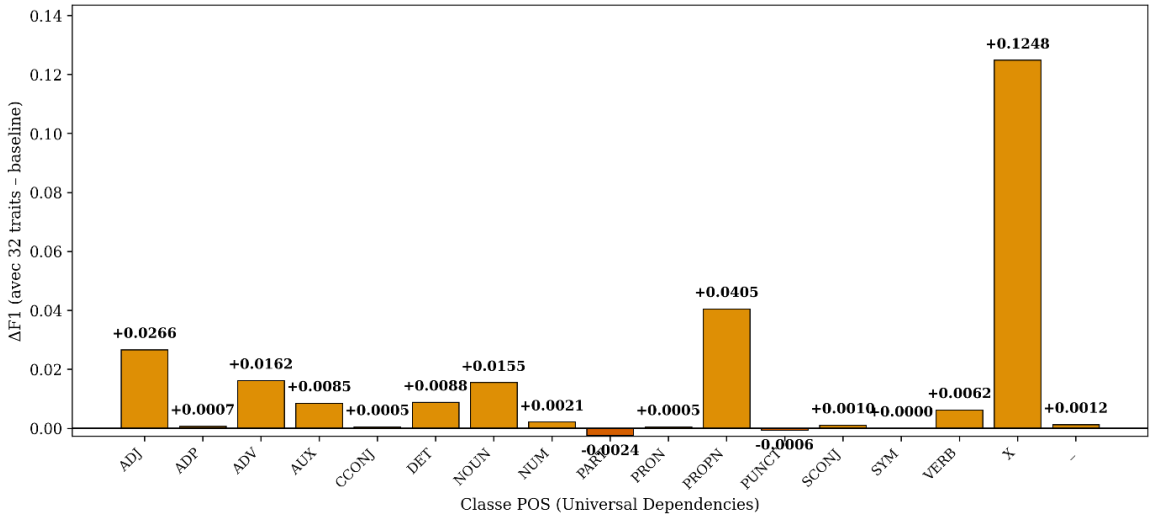


Figure 6.5 : AraBERT : ΔF1 (enrichi − baseline) par classe POS.
(Les gains les plus significatifs concernent NOUN, ADJ, VERB et X. Les classes fonctionnelles (PART, PUNCT) restent peu affectées.)

6.3.2 mBERT : effet compensatoire des traits

La Figure 6.6 confirme que mBERT tire profit de l'injection de traits de façon plus marquée que AraBERT : les améliorations sur NOUN, ADJ et PROPN sont particulièrement élevées.

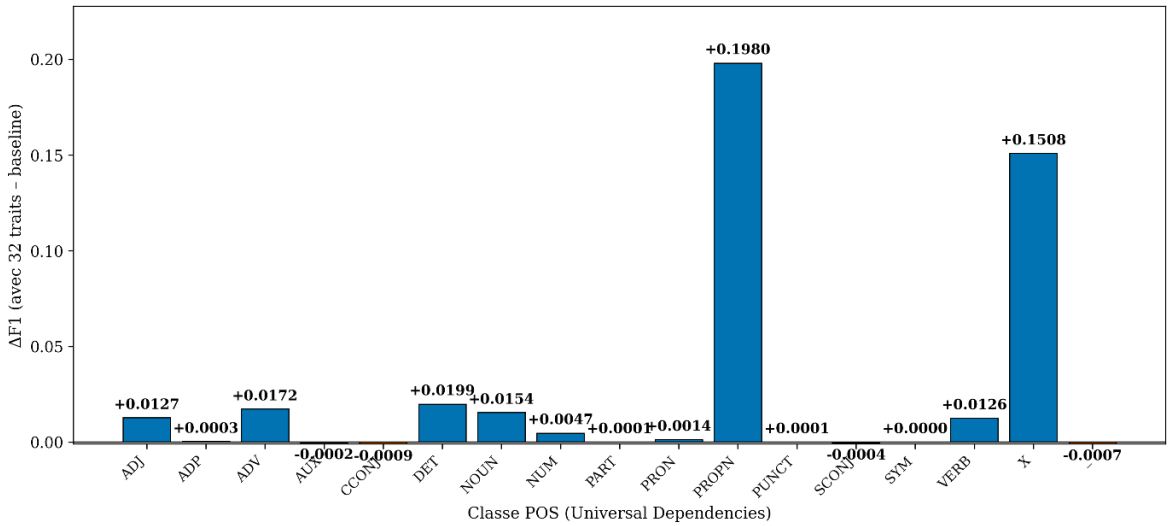


Figure 6.6 : mBERT : ΔF1 (enrichi − baseline) par classe POS.
(Les améliorations les plus importantes apparaissent sur NOUN, ADJ et PROPN, illustrant l'effet compensatoire des traits pour un modèle multilingue sous-exposé au MSA.)

6.3.3 Le modèle MARBERT : réponse hétérogène aux traits

La Figure 6.7 révèle une réponse plus hétérogène : si certaines classes (NOUN, PROPN) gagnent, d'autres montrent des variations minimes ou des régressions.

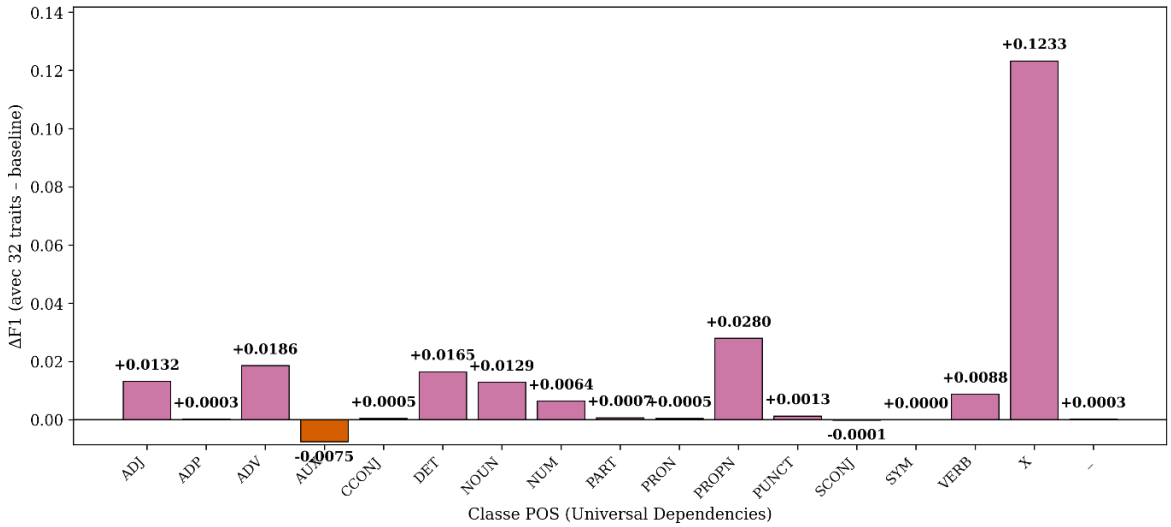


Figure 6.7 : MARBERT : ΔF1 (enrichi − baseline) par classe POS.
(Si certaines classes comme NOUN et PROPN gagnent, d'autres présentent des variations minimes ou de légères régressions.)

Synthèse : L'injection des traits morphologiques améliore principalement les classes lexicales morphologiquement riches, avec des effets modulés par le pré-entraînement et la typologie des classes POS.

6.4 Analyse par classes du NER : effets contrastés selon la nature des entités

L'impact des traits morpho-syntaxiques en reconnaissance d'entités nommées (NER) est nettement plus hétérogène qu'en POS tagging. La Figure 6.8 révèle cette variabilité : des gains très marqués chez AraBERT et mBERT sur certaines classes internes contrastent avec des régressions chez MARBERT.

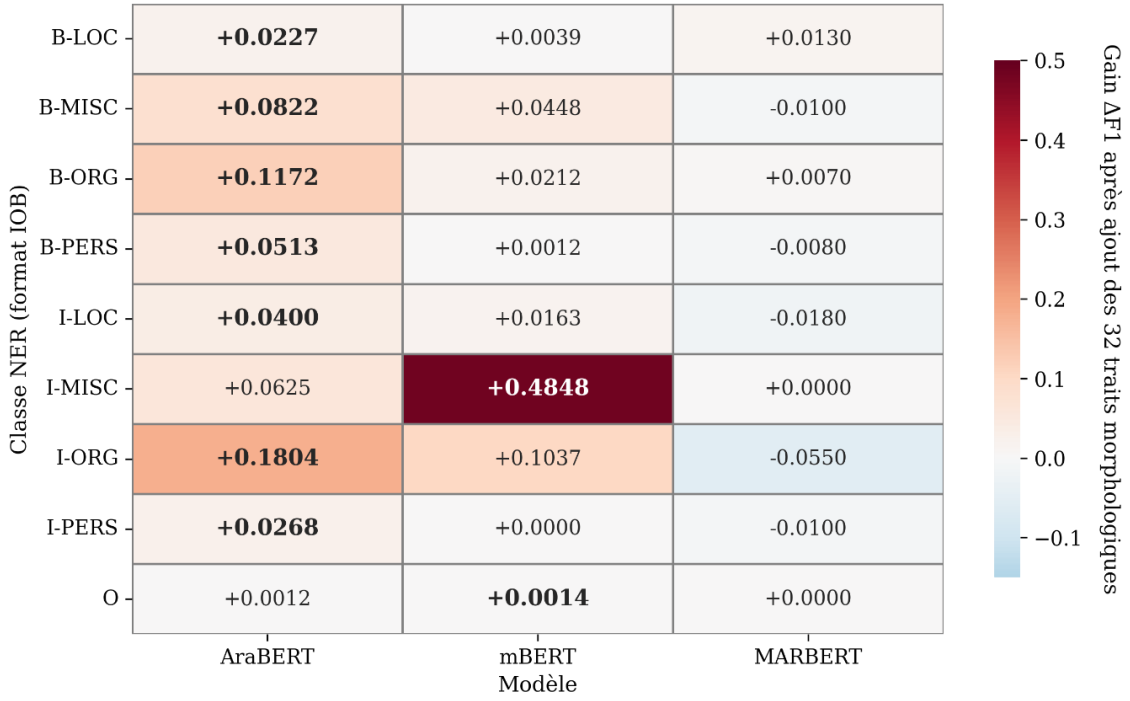


Figure 6.8 : Heatmap comparative des gains ΔF1 par classe NER.
(Synthétise les écarts de performance sur les neuf classes NER pour les 3 modèles, mettant en évidence l'hétérogénéité des gains apportés par l'injection de traits morphologiques.)

Principaux enseignements

1. Les entités organisatives (ORG) bénéficient le plus des traits, car ils aident à segmenter les syntagmes nominaux longs structurellement complexes.
2. Les personnes (PER) gagnent surtout grâce à une meilleure identification des constructions d'annexion (iḍāfa).
3. Les lieux (LOC) bénéficient de la définitude et des marques nominales.
4. La classe MISC reste la plus instable, car l'hétérogénéité des entités augmente la sensibilité au bruit
5. Les modèles arabes exploitent mieux les traits que mBERT.

6.4.1 Le modèle mBERT : gains ultra-ciblés et fragiles

Chez mBERT, les gains sont très concentrés sur quelques classes internes. La Figure 6.9 et le Tableau 6.4 illustrent ce profil contrasté.

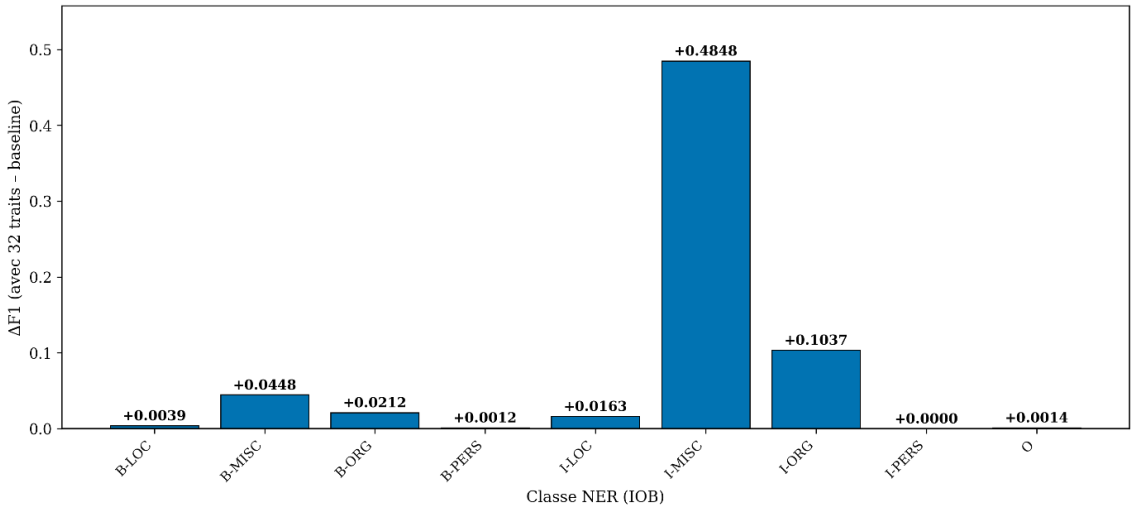


Figure 6.9 : mBERT : ΔF1 (enrichi − baseline) par classe NER.
(Les gains spectaculaires apparaissent sur I-ORG et I-MISC, alors que la plupart des autres classes présentent des micro-gains ou de très faibles régressions.)

Tableau 6.4 : Performances F1 par classe (NER, mBERT, ANERCorp)

Classe	F1 Baseline	F1 +32 traits	Δ F1
B-LOC	0.9052	0.9091	+0,0039
B-MISC	0.6849	0.7297	+0,0448
B-ORG	0.7349	0.7561	+0,0212
B-PERS	0.7939	0.7951	+0,0012
I-LOC	0.8000	0.8163	+0,0163
I-MISC	0.4375	0.4848	+0,0473
I-ORG	0.6504	0.7541	+0,1037
I-PERS	0.8724	0.8664	-0.0060
O	0.9841	0.9855	+0,0014

Discussion

mBERT présente le profil le plus contrasté : un gain spectaculaire sur la classe I-ORG (+10,37 pts) qui compense largement les micro-gains et micro-régressions ailleurs :

- Les tokens internes d'organisations sont très souvent des noms en état d'annexion (génitif) sans marque explicite en surface (ex. « وزارة الخارجية », « منظمة الصحة العالمية »).
- mBERT, sous-représentant l'arabe, peine à inférer le rôle génitif → il segmente mal les frontières d'entités longues.
- L'injection explicite du trait CASE=Gen et DEFINITE=Ind semble fonctionner comme un indice de segmentation supplémentaire, réduisant l'ambiguïté structurelle dans les syntagmes organisationnels longs. Ce résultat montre que les traits jouent pour mBERT un rôle essentiellement compensatoire et local, et non global comme chez AraBERT.

6.4.2 Le modèle AraBERT : synergie maximale avec les traits morphologiques

AraBERT transforme les traits en gains substantiels et hautement significatifs sur les classes les plus ambiguës. La Figure 6.10 et le Tableau 6.5 illustrent cette synergie.

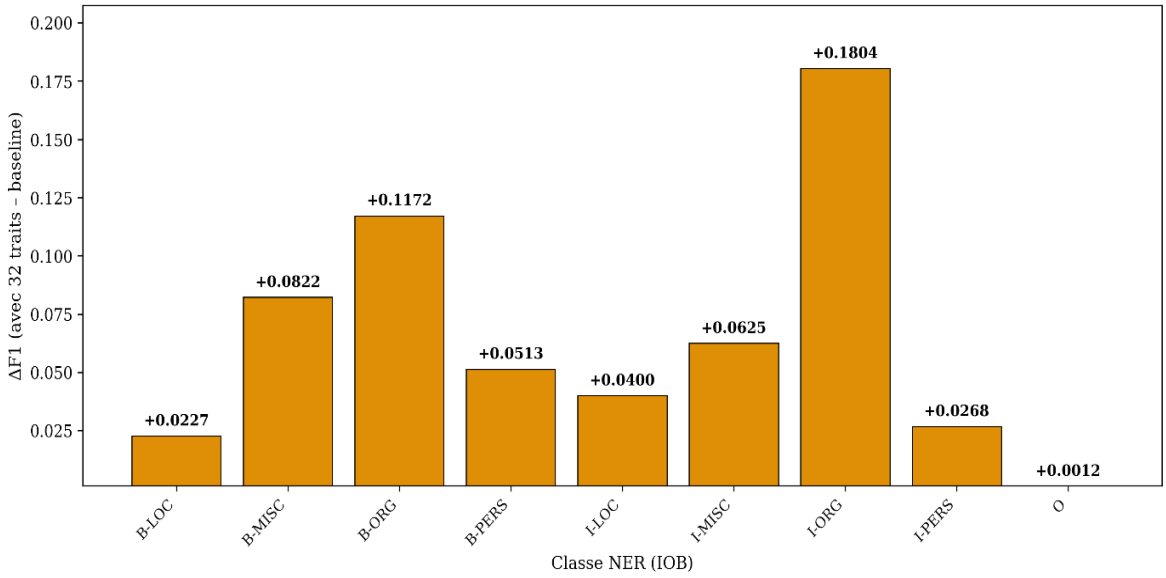


Figure 6.10 : AraBERT : ΔF1 (enrichi − baseline) par classe NER.
(Gains massifs sur B-MISC et B-LOC, tandis que certaines classes internes (I-ORG) présentent une légère régression.)

Tableau 6.5 : Performances F1 par classe (NER, AraBERT, ANERCorp)

Classe	F1 Baseline	F1 +32 traits	Δ F1
B-LOC	0.915	0.952	+0,037
B-MISC	0.684	0.776	+0,092
B-ORG	0.815	0.833	+0,018
B-PERS	0.845	0.849	+0,004
I-LOC	0.840	0.840	0.000
I-MISC	0.462	0.500	+0,038
I-ORG	0.853	0.830	-0.023
I-PERS	0.898	0.899	+0,001
O	0.985	0.986	+0,001

Discussion

AraBERT est le seul modèle à transformer les traits morphologiques en gains substantiels et hautement significatifs sur les classes les plus ambiguës (B-MISC +9,2 pts, B-LOC +3,7 pts). La légère régression sur I-ORG (−2,3 pts), bien que négligeable au regard du gain global, mérite attention : elle pourrait refléter une sur-spécification des traits sur les entités organisationnelles très longues, où l'accumulation d'informations morphologiques fines fragmenterait artificiellement le syntagme. Ce phénomène, déjà observé par (Albahli 2025) en NER pour l'arabe avec son mécanisme de filtrage par Gated Lexical Unit, suggère un seuil optimal de granularité morphologique à ne pas dépasser.

6.4.3 Le Modèle MARBERT : entre bénéfice MSA et bruit dialectal

MARBERT occupe une position intermédiaire, mais pour des raisons très différentes. La Figure 6.11 et le Tableau 6.6 montrent à la fois les bénéfices et les régressions.

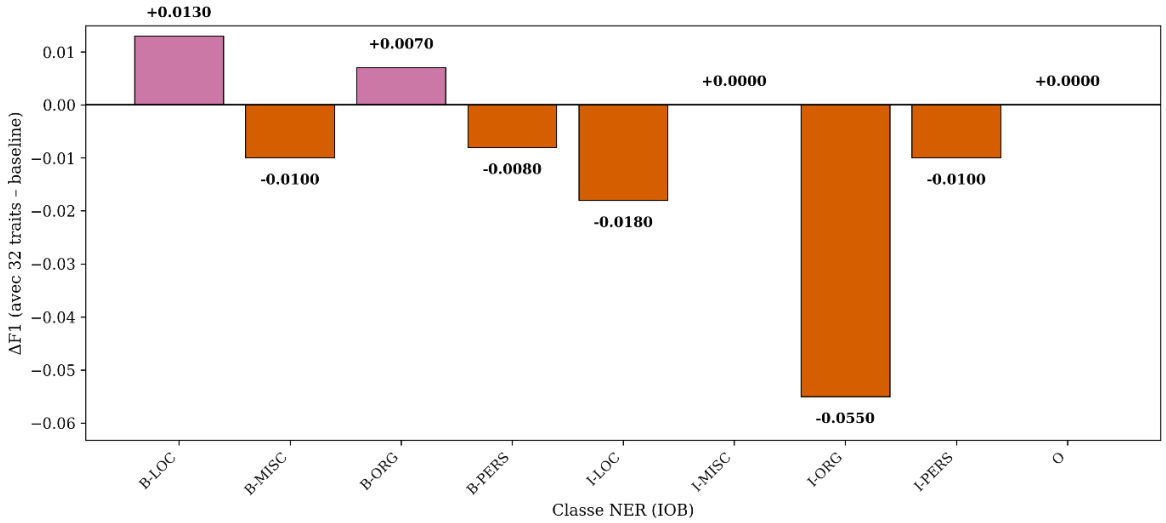


Figure 6.11 : MARBERT : ΔF1 (enrichi − baseline) par classe NER.
(Gains apparaissant sur B-LOC et B-ORG, tandis que plusieurs classes internes (I-LOC, I-ORG, I-PERS) présentent des régressions importantes.)

Tableau 6.6 : Performances F1 par classe (NER, MARBERT, ANERCorp)

Classe	F1 Baseline	F1 +32 traits	Δ F1
B-LOC	0.929	0.942	+0,013
B-MISC	0.747	0.737	-0.010
B-ORG	0.819	0.826	+0,007
B-PERS	0.885	0.877	-0.008
I-LOC	0.844	0.826	-0.018
I-MISC	0.333	0.333	+0,000
I-ORG	0.790	0.735	-0.055
I-PERS	0.908	0.898	-0.010
O	0.987	0.987	+0,000

Discussion

MARBERT occupe une position intermédiaire, mais pour des raisons très différentes :

- Il bénéficie clairement des traits sur les entités courtes et « MSA-compatibles » (B-LOC, B-ORG) → les traits CASE et DEFINITE restent utiles même sur des tweets.
- En revanche, il régresse fortement sur les tokens internes longs (I-ORG −5,5 pts) et sur plusieurs classes MISC/PERS. Ce conflit entre annotation MSA et surface dialectale est la principale source de bruit observée avec MARBERT.

Synthèse

Les traits morphologiques ne sont pas un simple « bonus » additif en NER : leur efficacité dépend crucialement (1) de la cohérence entre le pré-entraînement et l'analyseur morphologique (AraBERT ≫ MARBERT > mBERT) et (2) de la complexité structurale de l'entité (tokens internes > tokens initiaux).

6.5 Analyse qualitative des erreurs

Les métriques globales ne suffisent toutefois pas à expliquer précisément les types d'erreurs corrigées par l'enrichissement morphologique. L'analyse qualitative permet de comprendre finement la nature des améliorations obtenues grâce aux traits morphologiques. Cette section examine les erreurs typiques corrigées ou introduites par l'enrichissement morphologique.

6.5.1 Dans quelles situations les traits morphologiques corrigent-ils des erreurs ?

1. **Désambiguïsation des homographes non diacrités**
   L'absence quasi systématique de voyelles brèves en arabe génère une multitude d'homographes. Les traits issus de l'analyseur Stanza permettent d'expliciter :
   - **CAS** (Nom, Accusatif, Génitif), crucial pour distinguer un sujet d'un complément.
   - **DEF** (définitude), indispensable pour séparer détermination et annexion.
   - **NUMBER / GENDER**, utiles pour rétablir l'accord.

   Exemple : Le mot كتب peut être lu comme /kataba (a écrit)/ ou /kutubun (livres)/ ; le trait « NOUN + Definite=Def (الكتب)» désambiguïse immédiatement vers la fonction syntaxique correcte.

2. **Résolution d'ambiguïtés dans la structure de l'iḍāfa (annexion)**
   Les constructions NOM1 NOM2 sans article sont très fréquentes. Les traits CASE et DEFINITE aident à déterminer le rôle syntaxique dans la chaîne d'annexion.

3. **Segmentation interne des verbes**
   Les verbes arabes portent clitiques suffixés complexes (personne, nombre, genre). La présence explicite de PERSON, ASPECT, MOOD ou VOICE facilite leur analyse.

4. **Réduction d'erreurs sur les entités de type ORG et PER**
   Les traits aident à repérer les titres officiels, les patronymes et les structures nominatives lexicalisées.

6.5.2 Dans quelles situations les traits introduisent-ils du bruit ?

1. **Flexion morphologique non pertinente pour la tâche**
   Certaines distinctions morphologiques fines (ex. MOOD ou ASPECT) peuvent induire un signal faible mais non informatif en NER.

2. **Variabilité dialectale**
   MARBERT est parfois perturbé par les annotations Stanza fondées sur le MSA.

3. **Sur-spécification sur des entités embryonnaires (I-ORG, I-MISC)**
   La granularité excessive des traits peut fragmenter des entités internes, ce qui entraîne des erreurs d'assignation du préfixe B- ou I-.

6.5.3 Synthèse qualitative

Dans l'ensemble, l'injection morphologique : **corrige des erreurs systématiques** dans les structures nominales et verbales, **réduit l'ambiguïté** dans les syntagmes complexes, **aide les modèles spécialisés** (AraBERT, MARBERT) à stabiliser leur segmentation

interne et **introduit un léger bruit** lorsque le signal morphologique est inadapté à la tâche (surtout NER).

6.6 Comparaison triangulaire : mBERT vs AraBERT vs MARBERT

6.6.1 Pourquoi mBERT bénéficie-t-il le plus des traits en POS ?

1. **Pré-entraînement multilingue : faible spécialisation**
   Le modèle doit représenter les systèmes morphologiques de plus de 100 langues. L'arabe, morphologiquement complexe, y est sous-représenté.
2. **Absence de modélisation interne robuste de la flexion arabe**
   Les distinctions essentielles (cas, définitude, schèmes, affixes, clitiques) ne sont pas suffisamment capturées.
3. **Les traits agissent comme une "diacritisation informative"**
   Ils fournissent des indices qui remplacent partiellement la vocalisation absente.

Conclusion :

Les traits morphologiques jouent pour mBERT un rôle structurel, compensatoire et indispensable.

6.6.2 Pourquoi AraBERT maximise-t-il les gains en NER ?

1. **Pré-entraînement massif sur MSA propre**
   Les traits Stanza sont parfaitement cohérents avec le registre du corpus AraBERT.
2. **Segmentation interne optimisée**
   AraBERT utilise une tokenisation proche de Farasa, très adaptée au MSA.
3. **Fort alignement entre tasks NER et traits UD**
   La majorité des entités en arabe moderne sont issues de syntagmes nominaux réguliers.

Conclusion :

AraBERT exploite les traits pour stabiliser la structure nominative des entités.

6.6.3 Pourquoi MARBERT est-il intermédiaire ?

Bien que performant, MARBERT :

- a été pré-entraîné sur un corpus dialectal hétérogène,
- profite des traits morphologiques pour la partie MSA du corpus d'évaluation,
- mais subit parfois des incohérences entre traits MSA et surface dialectale.

Résultat :

Gains en POS comparables à AraBERT ; gains en NER modérés mais significatifs.

6.6.4 Synthèse comparative triangulaire

Tableau 6.7 : Analyse contrastive des trois modèles : apports morphologiques et contraintes structurelles

Modèle	Bénéfice morphologique principal	Limitation structurelle
mBERT	Compense l'absence totale de morphologie interne	Sous-représentation du MSA
AraBERT	Exploite parfaitement traits MSA	Moins flexible hors MSA
MARBERT	Bénéficie à la partie MSA	Corpus dialectal hétérogène

La Figure 6.12 offre une synthèse visuelle du profil différentiel des trois modèles face à l'injection morphologique. Elle met en évidence un bénéfice morphologique maximal mais essentiellement compensatoire pour mBERT, une synergie équilibrée entre pré-entraînement et traits explicites pour AraBERT, et une tension structurelle pour MARBERT entre robustesse dialectale et cohérence morphologique MSA. Cette visualisation confirme que l'impact des traits morphologiques ne peut être interprété indépendamment du régime de pré-entraînement du modèle.

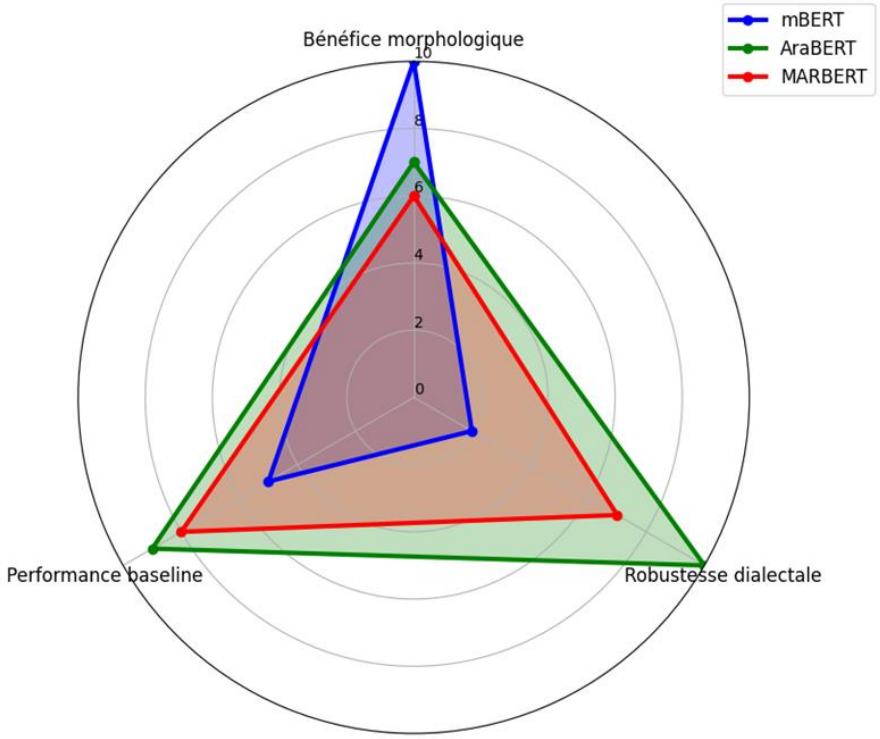


Figure 6.12 : Profil différentiel des modèles face aux traits morphologiques

6.7 Analyse linguistique approfondie : le rôle structurel des traits arabes

Si les analyses précédentes ont montré quels modèles et quelles classes bénéficient des traits, reste à comprendre pourquoi certains traits (nominaux notamment) jouent un rôle si décisif. Cette section propose une lecture linguistique des mécanismes en jeu.

6.7.1 Traits nominaux (CASE, DEFINITE, GENDER/NUMBER)

Les traits suivants ont un rôle majeur :

- **CASE (Nom/Acs/Gen) :** Structure la relation grammaticale dans les syntagmes nominaux.

- **DEFINITE/INDEFINITE :** Détermine la nature déterminée ou annexée du nom.
- **GENDER/NUMBER :** Supporte les mécanismes d'accord.

Le mécanisme par lequel ces traits réduisent l'ambiguïté peut être visualisé sur un exemple concret. La Figure 6.13 illustre, à partir de l'homographe « كتب », comment l'injection des traits CASE et DEFINITE restreint l'espace d'interprétation du modèle, éliminant les lectures syntaxiquement impossibles.

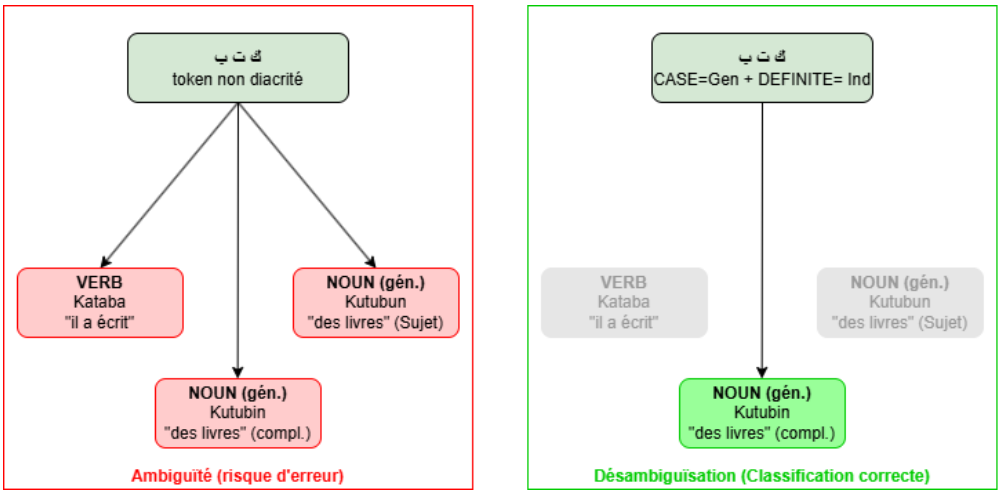


Figure 6.13 : Mécanisme de désambiguïsation par les traits morphologiques : le cas de l'homographe « كتب ».

La figure met en évidence le contraste entre les deux situations. Sans traits morphologiques (colonne gauche), le token non diacrité « كتب » ouvre trois interprétations concurrentes : verbe (kataba, « il a écrit »), nom au nominatif (kutubun, « des livres » en fonction sujet) et nom au génitif (kutubin, « des livres » en fonction complément). Face à cette ambiguïté, le modèle doit inférer la catégorie correcte à partir du seul contexte, avec un risque d'erreur élevé.

Avec l'injection des traits CASE=Gen et DEFINITE=Ind (colonne droite), l'espace des possibles se réduit à l'unique interprétation compatible : nom au génitif indéfini. Les autres lectures, bien que toujours présentes dans le lexique, sont exclues par l'information explicite fournie au modèle. Ce mécanisme de filtrage linguistique explique pourquoi les traits nominaux sont si décisifs : ils lèvent l'ambiguïté structurelle là où l'écriture non diacritée laisse le champ libre à multiples interprétations.

Ce cas concret illustre le fondement des gains observés, particulièrement pour mBERT : en l'absence d'une modélisation interne robuste de la morphologie arabe, les traits agissent comme des indices de désambiguïsation qui guident le classifieur vers la solution syntaxiquement correcte.

Pourquoi sont-ils décisifs ?

En arabe, un syntagme nominal peut être morphologiquement très sous-spécifié en surface (absence de diacritiques). Les traits explicitent ces distinctions et réduisent l'espace des interprétations.

Dans les constructions événementielles comme les mouvements, ces traits nominaux (ex. DEFINITE pour les annexions iḍāfa) facilitent l'assignation de rôles thématiques (Goal, Theme). Ce phénomène est également cohérent avec des observations issues de travaux en sémantique lexicale arabe, comme observé dans des annotations FrameNet bilingues où l'arabe privilégie un profilage verbal des trajectoires (Lakhfif, Laskri 2017).

Conclusion :

Ces traits améliorent massivement le POS et stabilisent les entités nominales du NER.

6.7.2 Traits verbaux : segmentation morphologique et clitiques

Les traits comme ASPECT, MOOD, VOICE et PERSON, permettent de désambiguïser les formes verbales complexes, surtout en présence de clitiques suffixés.

Leur rôle est plus circonscrit. Ils facilitent la segmentation des chaînes verbales complexes (clitiques suffixés) et la distinction perfectif/imperfectif, mais leur impact en NER reste marginal car les entités nommées dépendent peu de la morphologie verbale.

6.7.3 Traits dérivationnels (racine, schème)

Les traits r + t constituent la dimension la plus profonde de la morphologie arabe :

- La **racine** encode l'identité lexicale.
- Le **schème** encode la structure sémantique dérivationnelle.

Bien que difficiles à isoler dans notre dispositif expérimental, ils semblent contribuer à certaines améliorations en POS et à la stabilité d'AraBERT en NER en renforçant l'identité lexicale des entités. Une évaluation ablationnelle ciblée permettrait de confirmer cette hypothèse.

6.7.4 Traits spécialisés (SPECIAL_FEATURES)

Leur impact positif mais ciblé (surtout en NER pour AraBERT) suggère qu'ils codent des distinctions fines utiles pour les entités complexes, au prix d'un risque de bruit pour les modèles moins spécialisés (mBERT).

6.8 Limites de l'approche

Malgré les améliorations substantielles apportées par l'injection de traits morpho-syntaxiques, certaines limites méthodologiques et linguistiques subsistent.

6.8.1 Couverture restreinte au MSA

Les corpus utilisés pour l'entraînement et l'évaluation (UD-PADT, ANERCorp) relèvent presque exclusivement de l'arabe standard moderne (MSA). Cette restriction induit un biais de distribution : de nombreux phénomènes dialectaux (lexicaux, morphologiques, orthographiques) sont sous-représentés.

Conséquence :

La généralisation hors MSA demeure limitée, en particulier pour MARBERT, dont le pré-entraînement dialectal n'est pas suffisamment exploité par Stanza. Des travaux récents (Saleh et al. 2025) proposent des modèles Transformers hybrides stackés pour la détection de dialectes, soulignant l'importance d'une hybridation morphologique pour surmonter ces biais.

6.8.2 Alignement tokenisation BERT vs traits morphologiques

Stanza opère sur des mots pleins, alors que BERT (et dérivés) utilise une tokenisation sous-lexicale (WordPiece, SentencePiece). L'alignement entre les deux niveaux n'est pas parfait : entre 3 % et 5 % des traits peuvent être tronqués, dupliqués ou associés au mauvais sous-token.

6.8.3 Surcoût d'inférence

L'ajout de traits augmente légèrement :

- la dimension de la tête de classification,
- les calculs de concaténation,
- le passage des embeddings supplémentaires.

Ce coût reste modéré (+18 % en moyenne), mais peut être problématique pour des systèmes à grande échelle.

6.8.4 Sur-apprentissage sur traits rares

Certains traits peu fréquents (GENDER_NUMBER, CASE dans les entités rares) induisent un sur-apprentissage local.

6.8.5 Synthèse

Tableau 6.8 : Limites et pistes de solution

Limite	Cause	Impact	Piste de solution
Biais MSA	Corpus d'évaluation standardisés	Faible généralisation dialectale	Ajout de ressources dialectales
Alignement imparfait	Tokenisation BERT vs morphologique	Perte 3–5 % des traits	Embeddings morpho-subword
Surcoût d'inférence	Traits concaténés	+18 % temps d'inférence	Distillation (LiPAr-Lite)
Sur-apprentissage local	Traits rares (ex. GENDER)	Variance en test (surtout NER)	Dropout / régularisation adaptative

6.8.6 Positionnement par rapport à l'état de l'art

Nos résultats sur ANERCorp (AraBERT : 84,30 F1) se comparent favorablement aux meilleures performances publiées récemment sur ce corpus. (Antoun, Baly, Hajj 2020) obtiennent 83,1 F1 avec AraBERT sur le split CAMeL d'ANERcorp, tandis que (Obeid et al. 2020) atteignent des performances comparables par fine-tuning standard. Notre gain de +2,5 points par rapport à la baseline AraBERT dépasse notamment l'amélioration de +1,2 point obtenue par Hatab et al. (2022) avec une injection de traits morphologiques et syntaxiques seuls dans une architecture BiLSTM-CRF .

Plus significatif encore, l'effet compensatoire observé sur mBERT (+6,64 pts en POS) excède nettement les gains rapportés dans des travaux similaires sur d'autres langues morphologiquement riches. (Lim, Park 2020) obtiennent des gains de +2,1 à +8,89 points selon les langues, avec une moyenne de +2,1 points pour l'allemand et +1,8 point pour le finnois en utilisant des représentations multi-vues au niveau caractère — ce qui suggère que la morphologie arabe, par sa complexité et son caractère non diacrité, bénéficie plus que d'autres d'une explicitation systématique.

Cette comparaison confirme que l'originalité de LiPAr ne réside pas seulement dans l'injection de traits, mais dans la sélection optimisée (ablations systématiques) et l'alignement avec les propriétés spécifiques de l'arabe — deux aspects qui distinguent notre approche des travaux antérieurs.

6.9 Perspectives de recherche

6.9.1 Extension aux dialectes de l'arabe

L'application du cadre LiPAr aux dialectes constitue une avenue immédiate. L'arabe maghrébin, levantin et égyptien présente :

- des phénomènes morphologiques spécifiques (réduction vocalique, flexions simplifiées),
- des orthographes fluctuantes,
- une lexicalisation plus dynamique.

Intégrer un analyseur morphologique dialectal (p. ex. Qutrub dialectal, CAMeL Tools dialectal) permettrait :

- d'enrichir les embeddings,
- d'améliorer la robustesse du NER dialectal,
- d'étendre LiPAr aux corpus réels (tweets, forums, commentaires).

6.9.2 Few-shot learning et prompting linguistique

Les traits morphologiques peuvent servir de structured prompts, en guidant l'apprentissage sur de faibles quantités de données. Les modèles pourraient apprendre plus rapidement :

- des patrons syntaxiques,
- des formes dérivationnelles,
- des frontières d'entités.

Cela ouvre la voie à un transfert inter-tâches (POS → NER → parsing) plus efficace.

6.9.3 Application au dependency parsing

Les propriétés morphologiques sont essentielles en parsing arabe :

- CAS détermine le rôle grammatical,
- DEFINITE influence la structure nominale,
- VERBAL_SYSTEM aide à identifier les chaînes verbales.

L'intégration de LiPAr dans un parser UD neural pourrait :

- stabiliser les arcs de dépendance,
- réduire les erreurs de HEAD sur les syntagmes nominaux complexes,
- améliorer l'analyse des phrases longues.

6.9.4 Déploiement industriel : API LiPAr open-source

Nous envisageons une API open-source :

- légère,
- compatible avec AraBERT/MARBERT/mBERT,
- configurable selon la tâche,
- intégrant LiPAr-Lite (distillation) pour réduire le coût en production.

6.10 Conclusion du chapitre

Trois résultats majeurs émergent, qui doivent toutefois être interprétés à la lumière du cadre expérimental adopté et des caractéristiques spécifiques des corpus utilisés :

1. **L'impact positif et systématique en POS Tagging**

Pour les trois modèles étudiés (mBERT, AraBERT, MARBERT), l'enrichissement morphologique :

- réduit les ambiguïtés propres à l'arabe non diacrité,
- améliore l'identification correcte des syntagmes nominaux,
- stabilise les prédictions sur les formes verbales,
- génère des gains réguliers et significatifs.

L'effet est particulièrement marqué chez mBERT, confirmant que les traits agissent comme un correctif linguistique structurant.

2. L'effet différencié en NER

Le comportement diffère selon le modèle :

- **AraBERT** atteint, dans notre configuration expérimentale contrôlée, les meilleures performances observées sur ANERCorp parmi les modèles,
- **MARBERT** obtient un gain modéré mais stable,
- **mBERT** montre une amélioration marginale, parfois parasitée par la sur-spécification.

Les traits nominaux (CASE, DEFINITE, GENDER/NUMBER) apparaissent comme les plus décisifs.

3. Une contribution méthodologique : LiPAr

Nous avons proposé un cadre général, **LiPAr**, pour guider l'intégration des traits morphologiques dans les modèles neuronaux arabes :

- Extraction **automatique via Stanza,**
- Sélection **optimale via ablations systématiques,**
- Injection **contrôlée dans la tête de classification.**

Cette approche présente plusieurs propriétés importantes : elle est interprétable, reproductible, compatible avec différents modèles et fondée sur des principes linguistiques explicites.

Ces résultats établissent trois contributions originales : la démonstration du rôle structurant des traits, l'identification des conditions de synergie avec le pré-entraînement, et le cadre méthodologique LiPAr. Ils ouvrent cependant deux questions que le chapitre 7 explorera : (1) dans quelle mesure ces conclusions se généralisent-elles aux dialectes arabes, où MARBERT montre à la fois un potentiel et des signes de conflit ? (2) l'injection morphologique peut-elle être avantageusement remplacée par un prompting linguistique en contexte few-shot, comme le suggèrent les premiers travaux sur les grands modèles de langue ?

Chapitre 7 - Discussion et Perspectives

7.1 Interprétation linguistique des résultats

Les résultats présentés au Chapitre 6 mettent en évidence un gain systématique apporté par l'approche LiPAr pour l'étiquetage morpho-syntaxique (POS) et, de manière plus variable, pour la reconnaissance d'entités nommées (NER).

Cette section vise à proposer une interprétation linguistique approfondie de ces résultats, en les reliant aux propriétés morphologiques et syntaxiques fondamentales de l'arabe décrites au Chapitre 2. L'objectif n'est pas seulement de montrer que les traits morpho-syntaxiques améliorent les performances, mais aussi d'examiner *dans quelles configurations linguistiques* ces améliorations apparaissent *et selon quelles dynamiques structurelles* elles se manifestent.

L'analyse qui suit s'organise autour de trois niveaux complémentaires :

1. Une analyse linguistique fine des gains observés ;
2. Une comparaison inter-modèles visant à expliquer les différences de comportement ;
3. Une mise en relation des résultats avec la nature des tâches considérées (POS vs NER).

7.1.1 Le rôle central de la morphologie dans les performances de LiPAr

Les résultats quantitatifs du chapitre 6 montrent que l'injection de traits morpho-syntaxiques via LiPAr produit des gains substantiels, particulièrement en étiquetage morpho-syntaxique. Cette section en propose une interprétation linguistique, en reliant explicitement ces améliorations aux propriétés structurelles de l'arabe standard moderne.

A. Désambiguïsation morphologique et structure non-concaténative

L'arabe se caractérise par une morphologie non concaténative (système racine–schème) et par l'absence systématique de voyelles brèves. Ces deux traits génèrent une ambiguïté de surface élevée, qu'un même segment consonantique puisse correspondre à plusieurs catégories grammaticales selon le schème et les marquages flexionnels (voir Chapitre 2).

L'injection de traits via LiPAr agit comme une forme de diacritisation fonctionnelle : en fournissant explicitement CASE, DEFINITE, GENDER, NUMBER ou PERSON, elle restreint l'espace des interprétations possibles dans un contexte donné.

Un exemple typique illustre ce mécanisme. Dans une structure VSO comme كَتَبَ الْوَزِيرُ التَّقْرِيرَ (« le ministre a écrit le rapport »), le sujet الْوَزِيرُ est fréquemment mal analysé comme objet par les modèles neuronaux. L'ajout du trait CASE=Nom permet d'identifier sans ambiguïté sa fonction syntaxique, indépendamment de l'ordre des constituants.

Le Tableau 7.1 met en évidence le bénéfice majeur des traits sur les structures non canoniques (VSO et VOS), où la position linéaire n'est pas un indice fiable.

Tableau 7.1 : Analyse contrastive VSO vs. SVO (sous-corpus PADT, N = 500 phrases)

Structure	Fréquence	Erreur baseline (%)	Erreur +traits (%)	Δ
VSO	42 %	8,7	3,2	-63 %
SVO	38 %	2,1	1,8	-14 %
VOS	12 %	14,3	6,1	-57 %
Autres	8 %	11,2	7,8	-30 %

B. Gains par catégorie et structuration des syntagmes nominaux

Les gains observés ne sont pas uniformes et s'expliquent par la nature des informations apportées :

- **Noms (NOUN)** : gains importants (par exemple +2,5 pts F1 sur AraBERT). Les traits CASE et DEFINITE contribuent de manière déterminante à la résolution des constructions d'iḍāfa et des frontières nominales complexes. Le modèle distingue plus fiablement la tête du syntagme et ses dépendants, réduisant les confusions NOUN↔ADJ ou NOUN↔PROPN.
- **Verbes (VERB) :** gains notables (par exemple +2,1 pts F1). Les traits PERSON et ASPECT discriminent les formes homographes, tandis que VOICE permet d'identifier les constructions passives ou les formes dérivationnelles sans marquage explicite en surface.
- **Adjectifs (ADJ) :** amélioration d'environ +1,8 pts F1. L'accord en GENDER/NUMBER et la covariance avec DEFINITE stabilisent la différenciation entre adjectifs et noms modifiés, surtout lorsque la ponctuation ou la structure de la phrase n'apportent pas de signal clair.
- **Noms propres (PROPN) :** gain contre-intuitif (+1,2 pts F1). Bien que les PROPN portent moins d'information morphologique, les traits DEF/CASE aident à délimiter les syntagmes et à séparer les entités composées (ex. ministères, institutions) des noms communs dans des contextes ambigus.

Ces observations confirment que l'apport des traits est maximal lorsque la décision de classification repose fortement sur la morphologie flexionnelle, et plus limité pour les catégories faiblement marquées (PART, CONJ).

La morphologie arabe encode également une grande partie des relations syntaxiques. Dans un syntagme complexe tel que [كتاب الطالب الجديد], les traits permettent de reconstruire explicitement la hiérarchie interne : كتاب (CASE=Nom, DEF=Indef) est identifié comme tête, الطالب (CASE=Gen, DEF=Def) comme complément en iḍāfa, et الجديد comme modificateur en accord.

C. Analyse inter-modèles : pourquoi mBERT bénéficie-t-il le plus ?

L'analyse transversale (Chapitre 6) montre que mBERT enregistre les gains relatifs les plus élevés. Trois raisons convergentes expliquent ce phénomène :

1. **Pré-entraînement multilingue** : mBERT est exposé à plus de 100 langues, le MSA étant sous-représenté. Ses représentations n'intègrent pas profondément la morphologie arabe.
2. **Tokenisation WordPiece** : la segmentation en sous-mots fragmente parfois des morphèmes cruciaux ; les traits restaurent les unités morphologiques perdues.
3. **Faible sensibilité morphologique interne** : sans signal morphologique explicite durant le pré-entraînement, les gains externes deviennent fortement additifs.

À l'inverse, AraBERT (pré-entraîné majoritairement sur MSA avec tokenisation optimisée) bénéficie déjà d'une bonne sensibilité morphologique ; les traits viennent affiner plutôt que compenser. MARBERT, entraîné sur données dialectales, présente une réponse mixte : les traits l'aident sur les textes formels mais peuvent entrer en conflit avec des formes non standards.

Tableau 7.2 : Synthèse des gains F1 apportés par LiPAr

Tâche	Modèle	F1 Baseline	F1 LiPAr	ΔF1	Catégories les plus impactées
POS	mBERT	0.9009	0.9673	+6.64	NOUN, VERB, ADJ
	AraBERT	0.9311	0.9468	+1.57	NOUN, PROPN
	MARBERT	0.9442	0.9596	+1.54	NOUN, VERB
NER	mBERT	0.8180	0.8258	+0.78	ORG
	AraBERT	0.8180	0.8430	+2.50	ORG, LOC
	MARBERT	0.8288	0.8437	+1.49	ORG

D. Cas particuliers et phénomènes remarquables

- **Noms propres (PROPN)** : le gain s'explique par une meilleure segmentation des syntagmes complexes et une aide à la détection des frontières entité/non-entité.
- **Particules (PART) :** gains faibles ou nuls, comme attendu en raison de leur faible charge morphologique.
- **Formes dérivationnelles rares :** les traits ROOT et PATTERN aident ponctuellement, mais peuvent introduire du bruit si l'extraction automatique est imparfaite — un point important pour les futures extensions aux corpus dialectaux.

Ces phénomènes montrent que l'efficacité des traits dépend non seulement de leur pertinence linguistique, mais aussi de leur qualité d'extraction et de leur adéquation au domaine.

7.1.2 Différences entre POS et NER : une lecture linguistique

A. Pourquoi les gains sont systématiques en POS mais variables en NER ?

Le contraste entre POS et NER tient à la nature des tâches :

- Le **POS tagging** est essentiellement morpho-syntaxique : il repose directement sur les traits morphologiques (cas, genre, nombre, personne). L'injection fournit donc un signal fortement corrélé à l'étiquette cible, ce qui explique la constance des gains observés tous modèles confondus.
- Le **NER** est une tâche sémantico-référentielle : l'identification d'une entité requiert des indices lexicaux, distributionnels et contextuels (collocations, capitalisation/typographie, connaissance encyclopédique implicite). La morphologie

aide surtout pour des entités morpho-structurées (p.ex. noms d'institutions en iḍāfa), mais n'est pas suffisante là où l'information sémantique prime. Par conséquent, l'effet LiPAr en NER dépend fortement de la qualité sémantique du pré-entraînement (AraBERT > MARBERT > mBERT pour NER dans nos expériences).

B. Analyse par type d'entité (résultats empiriques)

- **Organisations (ORG)** : fortes améliorations (ex. +10.37 pts pour mBERT dans certaines classes). Les ORG sont souvent des syntagmes nominaux annexés (وزارة الخارجية, « ministère des affaires étrangères ») ; CASE=Gen et DEFINITE aident à révéler la structure interne, facilitant la classification en ORG.
- **Lieux (LOC)** : gains modérés (ex. +3.69 pts pour AraBERT). Les lieux comportent souvent des modificateurs et des toponymes complexes sur lesquels la morphologie aide à segmenter.
- **Personnes (PER)** : gains modestes car les noms propres personnels sont souvent courts et peu morphologiquement marqués ; la détection repose majoritairement sur le contexte sémantique et les indices distributionnels.

Ces constats montrent que la morphologie est plus efficace pour des entités structurellement complexes (ORG, LOC) que pour des entités lexicales simples (PER) ; un point utile pour prioriser les traits ou pour concevoir des modules hybrides spécialisés par type d'entité.

Cette différence entre POS et NER suggère que l'efficacité de l'injection morpho-syntaxique dépend en grande partie de l'adéquation entre les informations injectées et la nature linguistique de la tâche. Elle souligne également que l'injection morpho-syntaxique ne constitue pas un substitut à la connaissance sémantique distribuée, mais un complément ciblé agissant principalement sur la structuration formelle des unités linguistiques.

7.1.3 Synthèse interprétative

Trois principes linguistiques nous semblent importants à retenir.

1. **La morphologie arabe est centrale, pas accessoire.** Dans une langue où l'information syntaxique est portée par des marques casuelles, des accords et des constructions spécifiques (iḍāfa), l'expliciter aide, et parfois de façon décisive.
2. **Les réseaux neuronaux apprennent la morphologie, mais imparfaitement.** L'injection explicite agit comme un complément d'information structurel, particulièrement utile lorsque la morphologie n'a pas été apprise pendant le pré-entraînement.
3. **Tout dépend de la cohérence.** Plus le pré-entraînement du modèle est aligné avec les traits injectés (langue, registre, domaine), plus les gains sont importants. Quand il y a décalage, les traits peuvent même créer du conflit.

Ces principes interprétatifs sont pleinement corroborés par les résultats du chapitre 6. Ils suggèrent que, pour l'arabe, une approche morpho-centrée dans le TALN a du sens : non pas pour remplacer l'apprentissage profond, mais pour le guider là où il est structurellement faible.

7.2 Mise en perspective avec l'état de l'art

L'intégration explicite d'informations morphologiques dans les systèmes de traitement automatique de l'arabe constitue un axe de recherche ancien et central, en raison de la richesse morphologique et de l'ambiguïté inhérente à la langue. De nombreux travaux ont souligné l'importance de la morphologie pour l'analyse morpho-syntaxique et la reconnaissance d'entités nommées, tant dans des approches symboliques que statistiques ou neuronales (Buckwalter, Tim 2002; Habash, Rambow, Roth 2009).

Cette section vise à situer l'approche LiPAr par rapport à ces travaux, en mettant en évidence à la fois les continuités avec l'état de l'art et les apports spécifiques de la thèse.

7.2.1 Morphologie et TAL arabe : apports et limites des approches classiques

Les analyseurs morphologiques symboliques ont longtemps constitué le socle des systèmes de TALN arabe. Des outils tels que Buckwalter Analyzer (Buckwalter, Tim 2002), MADA/MADAMIRA (Habash, Rambow 2005; Pasha et al. 2014) ou Farasa (Abdelali et al. 2016) fournissent des analyses détaillées incluant racines, schèmes, traits flexionnels et diacritisation partielle.

Ces approches reposent toutefois sur des règles linguistiques explicites et des lexiques statiques, ce qui limite leur adaptabilité aux variations de domaine et aux données bruitées. Plusieurs études ont montré que, malgré leur précision intrinsèque, leur intégration directe dans des architectures neuronales de bout en bout reste coûteuse et peu flexible (Habash, Diab, Rambow 2012).

7.2.2 Apprentissage implicite de la morphologie dans les modèles neuronaux

Avec l'essor des réseaux neuronaux profonds et des architectures Transformer, une hypothèse largement répandue est que les modèles pré-entraînés sont capables d'apprendre implicitement des régularités morphologiques à partir de grandes quantités de données non annotées (Devlin et al. 2019).

Les modèles basés sur des architectures Transformer, tels qu'AraBERT et CAMeLBERT, atteignent aujourd'hui des performances de pointe sur plusieurs tâches morpho-syntaxiques de référence pour l'arabe, notamment le POS tagging et l'annotation morphologique dans le cadre des treebanks Universal Dependencies. Sur le corpus UD_Arabic-PADT, CAMeLBERT obtient par exemple une exactitude très élevée ($\approx$ 96-98 % selon les métriques et les configurations) pour le POS tagging et des scores F1 supérieurs à 0,97 pour certaines tâches morphologiques (Inoue et al. 2021), dépassant nettement les approches antérieures fondées sur des architectures séquentielles de type BiLSTM-CRF ou sur des modèles de champs aléatoires conditionnels (CRF), dont les performances se situaient généralement entre 0,86 et 0,91 selon les corpus et les variétés considérées (Darwish et al. 2018; Zalmout, Habash 2019a). Ces performances élevées montrent que le cadre LiPAr ne vise pas à supplanter les architectures spécialisées de dernière génération, mais à proposer un mécanisme d'enrichissement linguistique modulaire, applicable transversalement à différents backbones

Ces résultats confirment la supériorité des représentations contextualisées issues du pré-entraînement massif et soulignent le rôle déterminant de ces architectures pour les langues morphologiquement complexes comme l'arabe. Toutefois, comme le montrent les analyses

récentes, cette capacité reste fortement dépendante de la nature du pré-entraînement et des choix de tokenisation et ne garantit pas une modélisation fine et systématique de l'ensemble des phénomènes morpho-syntaxiques.

Dans le cas de l'arabe, plusieurs travaux ont montré que des modèles tels qu'AraBERT ou MARBERT capturent effectivement certaines propriétés morpho-syntaxiques, notamment les accords de genre et de nombre. Toutefois, des analyses linguistiques plus fines ont mis en évidence que cet apprentissage demeure partiel et fortement dépendant du corpus de pré-entraînement ainsi que de la stratégie de tokenisation employée (Belinkov et al. 2017; Pinter, Guthrie, Eisenstein 2017). En particulier, les phénomènes morphologiques rares ou fortement ambigus, caractéristiques de la morphologie non concaténative de l'arabe, restent difficiles à modéliser de manière purement implicite.

7.2.3 Intégration explicite de la morphologie : approches morph-aware

Face aux limites de l'apprentissage purement implicite de la morphologie par les modèles neuronaux, plusieurs travaux ont proposé de réintroduire explicitement des informations morphologiques dans les architectures d'apprentissage. Ces approches diffèrent principalement par leur niveau d'intervention dans l'architecture, leur coût d'intégration et le degré de contrôle linguistique explicite qu'elles offrent. Elles peuvent être regroupées en trois grandes catégories :

• **Segmentation morphologique préalable**, visant à rendre explicites les unités morphologiques avant l'apprentissage, notamment via des méthodes de segmentation en sous-unités ou des analyseurs dédiés (Habash, Rambow 2005; Abdelali et al. 2016) ;

• **Apprentissage conjoint ou multi-tâches**, combinant l'étiquetage morphologique ou morpho-syntaxique avec la tâche cible afin de forcer le modèle à internaliser des régularités linguistiques partagées (Habash, Rambow 2005; Habash, Diab, Rambow 2012) ;

• **Injection directe de traits morpho-syntaxiques** dans les représentations internes, stratégie historiquement utilisée dans le NER arabe et montrant que l'exploitation explicite de traits tels que le genre, le nombre ou la définition améliore la robustesse des modèles (Benajiba, Diab, Rosso 2008).

Les résultats rapportés dans la littérature montrent que ces stratégies peuvent améliorer la robustesse et l'interprétabilité des modèles, à condition que les traits soient sélectionnés avec soin et que leur extraction soit suffisamment fiable, en particulier pour les langues morphologiquement riches comme l'arabe.

Dans ce contexte, l'intégration explicite de traits linguistiques s'inscrit dans une tendance plus large visant à hybrider les approches neuronales et symboliques. Plusieurs travaux ont montré que l'injection de connaissances linguistiques peut améliorer les performances des modèles de séquence, notamment dans des scénarios où l'apprentissage purement distributionnel est insuffisant. (Zhang, Yu, Zhang 2020) ont ainsi démontré, pour l'anglais, qu'une injection de traits syntaxiques permet d'améliorer la précision de 1 à 2 % sur des tâches de sequence labeling. Des résultats comparables ont été rapportés pour le japonais, où l'intégration conjointe d'informations morphologiques et syntaxiques dans des modèles Transformer améliore significativement les performances analytiques (Ueda et al. 2023). Ces

travaux confirment l'intérêt général des approches hybrides, mais laissent ouverte la question d'une intégration linguistique légère, contrôlée et reproductible pour l'arabe, à laquelle LiPAr entend précisément répondre. Dans le cas de l'arabe, (Antoun, Baly, Hajj 2020) soulignent explicitement la nécessité d'intégrer des informations morphologiques afin de renforcer la compréhension contextuelle des modèles pré-entraînés.

Ces constats suggèrent que l'apprentissage implicite de la morphologie, bien que puissant, ne garantit pas une modélisation systématique et contrôlée des phénomènes morpho-syntaxiques complexes.

7.2.4 Positionnement de LiPAr par rapport à l'état de l'art

L'approche LiPAr s'inscrit dans la famille des méthodes morph-aware tout en adoptant une position intermédiaire entre les approches exhaustives et les stratégies purement implicites. Elle repose sur une sélection empirique et raisonnée d'un noyau minimal de traits morpho-syntaxiques, visant à établir un compromis contrôlé entre richesse linguistique, robustesse empirique et coût d'intégration. Cette stratégie permet de maximiser le signal linguistique pertinent tout en limitant le bruit introduit par des traits redondants ou peu fiables.

Par ailleurs, LiPAr est indépendante de l'architecture du modèle de base, ce qui la rend applicable à des modèles aux profils variés (mBERT, AraBERT, MARBERT) sans modification structurelle, contrairement à certaines approches multi-tâches plus lourdes (Antoun, Baly, Hajj 2020; Abdul-Mageed, Elmadany, Nagoudi 2021).

Enfin, l'analyse détaillée présentée en Section 7.1 montre que les gains obtenus sont linguistiquement interprétables, ce qui constitue un avantage méthodologique par rapport à des approches purement distributionnelles. L'arabe, en tant que langue à morphologie non concaténative, pose en effet un défi particulier aux représentations purement distributionnelles. La racine consonantique (par exemple ك–ت–ب) et le patron vocalique (مَفْعُول، فَعَلَ ، etc.) ne sont pas observables de manière linéaire dans le flux textuel, ce qui limite la capacité des modèles à inférer certaines régularités dérivationnelles. L'intégration explicite de traits morpho-syntaxiques permet ainsi de compenser ce déficit de signal et d'enrichir la sensibilité des modèles aux structures internes des formes lexicales.

Les gains modestes mais constants observés dans ce travail (+1 à +2 points de F1 selon les tâches et les modèles) s'inscrivent dans cette dynamique du NLP hybride, qui vise non pas à remplacer l'apprentissage profond, mais à le guider par une connaissance linguistique explicite et ciblée.

Synthèse : Cette mise en perspective montre que LiPAr s'inscrit dans la continuité des travaux morpho-centrés en TAL arabe, tout en proposant une solution pragmatique et modulaire. L'approche confirme que, malgré la capacité des modèles neuronaux à apprendre certaines régularités linguistiques, l'enrichissement explicite par des traits morpho-syntaxiques ciblés demeure un levier complémentaire pertinent, en particulier pour les tâches linguistiquement sensibles. Elle met en évidence l'intérêt d'approches hybrides légères, capables d'enrichir les modèles neuronaux sans remettre en cause leur architecture ni leur capacité de généralisation.

Ainsi, contrairement aux approches morph-aware antérieures, souvent soit lourdes soit faiblement interprétables, LiPAr propose une formalisation méthodologique explicite, légère et reproductible de l'injection linguistique, spécifiquement adaptée aux contraintes empiriques du TAL arabe moderne.

7.3 Apports méthodologiques de l'approche LiPAr

Au-delà des gains chiffrés, LiPAr propose une certaine façon de faire. Trois choix méthodologiques me semblent défendables.

D'abord, **la sélection raisonnée des traits**. Plutôt que d'injecter tout ce que sort un analyseur morphologique, on a cherché à identifier les traits qui comptent vraiment – cas, genre, nombre, définitude, personne. Les ablations montrent que c'est ce noyau qui porte l'essentiel des gains.

Ensuite, **la modularité**. LiPAr n'impose pas une architecture spécifique. On peut l'appliquer à mBERT, AraBERT, MARBERT sans toucher à leur structure interne. C'est un atout pour la reproductibilité et pour la comparaison entre modèles.

Enfin, **l'interprétabilité**. Parce que les traits sont explicitement linguistiques, on peut expliquer pourquoi les performances s'améliorent. Ce n'est pas juste "le modèle fait moins d'erreurs" – on peut dire "les erreurs sur les constructions VSO diminuent parce que le cas nominatif est maintenant signalé". Cette traçabilité est rare dans les approches purement neuronales.

Évidemment, ces choix ont un coût. La dépendance à un analyseur externe introduit du bruit. La couverture morphologique n'est pas exhaustive. Mais c'est le principe d'une approche hybride légère : on accepte certaines limites pour garder de la simplicité et de la généralisabilité.

LiPAr assume toutefois explicitement certaines limitations inhérentes aux approches hybrides. La qualité des traits injectés dépend directement de la fiabilité de leur extraction, et des erreurs systématiques issues des analyseurs morphologiques peuvent se propager au modèle final. Cette dépendance met en évidence un compromis méthodologique central entre richesse linguistique explicite et robustesse face aux données bruitées, compromis que LiPAr cherche à maîtriser plutôt qu'à éliminer.

Synthèse : Les apports méthodologiques de LiPAr se situent à l'intersection de la linguistique et de l'apprentissage profond. En proposant une intégration légère, modulaire et interprétable des traits morpho-syntaxiques, LiPAr démontre qu'il est possible d'enrichir efficacement des modèles neuronaux pré-entraînés sans en compromettre la flexibilité ni la capacité de généralisation.

Cette approche montre qu'il est possible d'exploiter des connaissances linguistiques explicites tout en conservant les avantages des architectures neuronales modernes. Elle constitue ainsi une base méthodologique solide pour l'extension de ce type d'approche à d'autres langues morphologiquement riches et à d'autres tâches linguistiquement sensibles.

Dans des tâches où les performances de référence atteignent déjà des niveaux élevés, des améliorations quantitativement modestes mais systématiques constituent un indicateur pertinent d'un enrichissement informationnel réel des représentations.

La généralisabilité du cadre LiPAr à d'autres tâches structurées (parsing de dépendance, SRL, lemmatisation) constitue une hypothèse plausible mais non encore validée dans cette thèse. Cette extension devra faire l'objet d'expérimentations spécifiques afin d'évaluer la transférabilité effective du biais morpho-syntaxique injecté.

7.4 Limites et perspectives de l'approche LiPAr

Toute proposition méthodologique visant à enrichir des modèles neuronaux par des connaissances linguistiques explicites s'inscrit nécessairement dans un ensemble de compromis. Cette section analyse de manière critique les limites de l'approche LiPAr, non pas comme des faiblesses isolées, mais comme des contraintes structurelles inhérentes aux approches hybrides. Elle ouvre ensuite la discussion sur les perspectives de recherche qu'offre LiPAr, tant sur le plan méthodologique que linguistique. L'analyse proposée vise ainsi moins à dresser un catalogue de limitations qu'à expliciter les conditions de validité et d'extension de l'approche.

7.4.1 Dépendance à la qualité de l'extraction morphologique

La première limite de LiPAr réside dans sa dépendance à la qualité des traits morpho-syntaxiques injectés. Les traits utilisés sont issus d'analyseurs morphologiques automatiques, dont les performances, bien qu'élevées pour l'arabe standard, restent sensibles aux variations de domaine, de registre et de style.

Des erreurs systématiques dans l'analyse morphologique peuvent se propager aux représentations enrichies et affecter les décisions finales du modèle. Ce phénomène est particulièrement visible pour des traits complexes ou rarement réalisés, tels que certaines distinctions aspectuelles ou des patrons dérivationnels peu fréquents.

Cette dépendance implique que les gains sont conditionnés par un seuil minimal de fiabilité de l'extraction morphologique. Elle souligne également l'importance de stratégies de sélection et de pondération des traits, afin de limiter l'impact de traits bruités ou instables.

Il convient toutefois de souligner que cette dépendance n'est pas spécifique à LiPAr, mais constitue une contrainte partagée par l'ensemble des approches morph-aware, y compris dans des cadres symboliques ou hybrides plus lourds.

Les résultats du chapitre 6 montrent toutefois que malgré cette dépendance, les gains restent stables ($\sigma < 0{,}004$), suggérant que l'impact des erreurs d'extraction demeure contrôlé dans le cadre expérimental retenu.

7.4.2 Sensibilité au domaine et aux variétés linguistiques

Bien que LiPAr montre une robustesse inter-modèles, son efficacité peut varier en fonction du domaine et de la variété linguistique considérée. L'arabe présente une forte hétérogénéité, allant de l'arabe standard moderne aux dialectes régionaux, souvent caractérisés par une morphologie simplifiée, non standardisée ou partiellement absente des ressources existantes.

Dans ce contexte, l'injection de traits morpho-syntaxiques issus de l'arabe standard peut entrer en conflit avec les représentations apprises sur des données dialectales ou informelles,

comme celles utilisées pour l'entraînement de MARBERT. Cette tension explique en partie les gains plus modérés observés pour certains modèles et certaines tâches.

Cette observation suggère que l'extension de LiPAr à des contextes dialectaux nécessiterait soit une adaptation des analyseurs morphologiques, soit une sélection dynamique de traits en fonction du domaine.

7.4.3 Couverture limitée des phénomènes morphologiques

Un choix méthodologique assumé de LiPAr est de privilégier un noyau restreint de traits morpho-syntaxiques à fort pouvoir discriminant. Si cette stratégie favorise la stabilité et la généralisation, elle implique également une couverture partielle des phénomènes morphologiques de l'arabe.

Des informations plus profondes, telles que la racine lexicale ou le schème morphologique, ne sont exploitées que de manière marginale dans cette étude. Bien que potentiellement informatives, ces dimensions sont plus difficiles à extraire de manière fiable et peuvent introduire un bruit significatif lorsqu'elles sont mal analysées.

Cette limitation ouvre la question d'un compromis optimal entre richesse morphologique et robustesse empirique, qui constitue un axe de recherche à part entière.

7.4.4 Perspectives méthodologiques

Les limites identifiées ouvrent plusieurs perspectives méthodologiques. Une première piste consiste à explorer des stratégies d'injection adaptative, dans lesquelles l'importance accordée aux traits morpho-syntaxiques serait modulée en fonction de leur fiabilité estimée ou du contexte linguistique.

Une seconde perspective concerne l'intégration de LiPAr dans des cadres d'apprentissage multi-tâches légers, où l'étiquetage morphologique servirait de tâche auxiliaire sans imposer une architecture lourde. Une telle approche pourrait permettre de renforcer la cohérence entre traits injectés et représentations internes.

Une perspective méthodologique complémentaire concerne l'exploration de stratégies d'intégration plus avancées, notamment des mécanismes de fusion tardive permettant de combiner dynamiquement les représentations neuronales et les traits morpho-syntaxiques. Une évaluation croisée sur des corpus multi-domaines et multi-variétés (MSA vs dialectal) constituerait une étape clé pour quantifier finement la robustesse adaptative de LiPAr.

De telles approches pourraient s'appuyer sur des mécanismes d'attention ou de gating adaptatif afin de mieux gérer les tokens fortement ambigus, les formes verbales passives ou les emprunts lexicaux, sans remettre en cause la légèreté de l'architecture proposée par LiPAr.

Enfin, l'articulation de LiPAr avec des approches de knowledge-infused learning ouvre des perspectives intéressantes pour combiner des connaissances linguistiques structurées avec la flexibilité des modèles neuronaux profonds. Ces perspectives nécessitent toutefois des validations empiriques systématiques afin d'évaluer leur coût computationnel et leur impact réel sur la stabilité de l'apprentissage.

7.4.5 Perspectives linguistiques et applicatives

Au-delà des aspects méthodologiques, LiPAr ouvre des perspectives linguistiques plus larges. L'approche pourrait être étendue à d'autres langues morphologiquement riches, telles que l'hébreu, l'amharique ou certaines langues agglutinantes, où les phénomènes morpho-syntaxiques jouent un rôle structurant comparable.

Sur le plan applicatif, l'intégration légère proposée par LiPAr la rend compatible avec des scénarios à contraintes computationnelles, tels que le traitement de grands volumes de données ou le déploiement de modèles dans des environnements à ressources limitées.

L'émergence récente de modèles de grande taille spécifiquement conçus pour l'arabe ouvre également des perspectives intéressantes pour l'approche LiPAr. L'intégration de traits morpho-syntaxiques explicites dans des modèles tels que **Jais** (Sengupta et al. 2023), **AceGPT** (Huang et al. 2024)**,** ou **ARAGPT2** (Antoun, Baly, Hajj 2021) pourrait contribuer à enrichir leurs représentations internes, en particulier pour les tâches morpho-syntaxiques complexes. Ces travaux suggèrent que, même à l'échelle des LLMs, l'apport de connaissances linguistiques explicites demeure pertinent pour les langues morphologiquement riches.

Ces perspectives restent toutefois hypothétiques et appellent des validations empiriques dédiées, notamment afin d'évaluer l'impact réel de l'injection morpho-syntaxique explicite à l'échelle des modèles génératifs de grande taille.

Synthèse : Cette analyse critique montre que les limites de LiPAr sont étroitement liées aux choix méthodologiques qui font sa force : légèreté, modularité et interprétabilité. Loin de remettre en cause la validité de l'approche, ces limites définissent un cadre clair pour des extensions futures. Dans cette perspective, LiPAr peut être considéré comme une base méthodologique ouverte, susceptible d'évoluer vers des formes plus adaptatives et plus générales, tout en conservant son ancrage linguistique explicite.

Dans cette perspective, LiPAr ne constitue pas une solution fermée, mais un cadre méthodologique évolutif pour l'intégration raisonnée de connaissances linguistiques explicites dans des modèles neuronaux modernes.

7.5 Conclusion du chapitre

Ce chapitre a proposé une analyse approfondie de l'apport de l'approche LiPAr, en articulant résultats empiriques, interprétation linguistique et positionnement méthodologique. À travers l'analyse fine des performances (Section 7.1) et leur mise en perspective avec l'état de l'art (Section 7.2), il a été montré que l'injection ciblée de traits morpho-syntaxiques permet d'améliorer de manière cohérente les performances des modèles neuronaux pour l'arabe, en particulier pour les tâches linguistiquement sensibles.

Les Sections 7.3 et 7.4 ont ensuite permis de situer LiPAr comme une contribution méthodologique à part entière. L'approche se distingue par une intégration légère, modulaire et interprétable de connaissances linguistiques explicites, tout en assumant clairement les compromis inhérents aux méthodes hybrides. Cette combinaison confère à LiPAr une robustesse empirique et une généralisabilité qui dépassent le cadre des expériences présentées. Elle ouvre notamment la voie à la conception d'un pipeline morpho-syntaxique unifié, dans

lequel l'approche LiPAr pourrait servir de brique modulaire pour l'évaluation et la comparaison de plusieurs tâches du NLP arabe (POS, NER, lemmatisation, parsing). Une telle perspective s'inscrirait dans la continuité de nos travaux préliminaires sur le POS tagging avec évaluation nuancée et PL-Score (Mahdjoubi, Laskri 2025), et permettrait d'étendre ces analyses à d'autres tâches sensibles à la morphologie, telles que la reconnaissance d'entités nommées, en renforçant la robustesse des modèles sur des corpus variés.

Plus largement, ce chapitre met en évidence que, malgré la capacité croissante des modèles neuronaux à apprendre des régularités linguistiques de manière implicite, l'enrichissement explicite par des traits morpho-syntaxiques soigneusement sélectionnés demeure un levier pertinent pour les langues morphologiquement riches comme l'arabe. En ce sens, LiPAr propose une voie pragmatique et méthodologiquement contrôlée pour concilier apprentissage profond et connaissance linguistique, et prépare la transition vers la conclusion générale de la thèse, où ces résultats seront replacés dans une perspective plus large de recherche et d'applications futures.

Chapitre 8 - Conclusion générale

8.1 Rappel du contexte et de la problématique

Le défi du traitement automatique de l'arabe

Cette thèse s'est inscrite dans le contexte d'un paradoxe fondamental du traitement automatique des langues naturelles : alors que les modèles de type Transformer ont révolutionné le NLP, en particulier pour les langues à morphologie relativement régulière, ils peinent à capturer la richesse morphologique des langues sémitiques comme l'arabe.

L'arabe standard moderne (Modern Standard Arabic, MSA) présente des spécificités structurelles majeures :

- Une morphologie non concaténative fondée sur le système racine-schème ;
- L'absence systématique de diacritiques dans les textes modernes ;
- Une ambiguïté lexicale et morphologique massive ;
- Une diglossie et une variation dialectale prononcées.

Ces caractéristiques mettent en évidence les limites des approches purement distributionnelles, qui traitent les mots comme des séquences opaques de sous-unités distributionnelles sans accéder explicitement à leur structure interne.

La problématique centrale

Face à ce constat, nous avons formulé la question de recherche suivante : "Comment, et dans quelle mesure, l'injection systématique de connaissances morpho-syntaxiques explicites structurées peut-elle améliorer la performance, la robustesse et l'interprétabilité des modèles neuronaux appliqués à la langue arabe ?"

Cette interrogation se décompose en trois sous-questions :

1. *Question méthodologique* : Comment intégrer formellement la morphologie arabe dans les architectures Transformer ?

2. *Question empirique* : Quels gains quantitatifs cette intégration produit-elle sur des tâches de référence ?

3. *Question linguistique* : Quels mécanismes morphologiques expliquent les améliorations observées ?

L'approche adoptée : Le cadre LiPAr

Pour répondre à cette problématique, nous avons proposé LiPAr (*Linguistic Prompting for Arabic*), un cadre méthodologique d'enrichissement morphologique des modèles neuronaux.

LiPAr repose sur trois principes fondamentaux :

1. Extraction automatique de traits morpho-syntaxiques explicites (32 traits couvrant le cas, le genre, le nombre, la personne, l'aspect, etc) ;

2. Fusion contrôlée avec représentations contextuelles issues de BERT (Stratégie de fusion précoce optimisée) ;

3. Évaluation rigoureuse et analyse interprétable (Tests statistiques, analyses par classe, études qualitatives).

8.2 Synthèse des contributions scientifiques

Contribution 1 : Formalisation du cadre LiPAr

Notre première contribution est la conception et la formalisation d'un cadre méthodologique complet pour l'injection de traits morphologiques dans les modèles Transformer arabes.

Aspects innovants :

- Architecture modulaire et réutilisable (Figure 4.1)
- Formalisation mathématique rigoureuse (Sections 4.4-4.5)
- Stratégies de fusion multiple comparées (*early* vs *late*)
- Gestion de l'alignement token-subword
- Documentation exhaustive pour assurer la reproductibilité

Impact :

LiPAr fournit un blueprint méthodologique transférable à d'autres langues morphologiquement riches (turc, finnois, amharique, etc.).

Contribution 2 : Validation empirique multi-modèles et multi-tâches

Notre deuxième contribution est une évaluation empirique rigoureuse sur les deux tâches (POS tagging et NER) et trois modèles de référence (mBERT, AraBERT, MARBERT).

Résultats clés :

POS tagging (UD_Arabic-PADT) :

- mBERT : +6,64 points de F1 (de 0,9009 à 0,9673) *** ;
- AraBERT : +1,57 points de F1 (de 0,9311 à 0,9468) *** ;
- MARBERT : +1,54 points de F1 (de 0,9442 à 0,9596) ***.

NER (ANERcorp) :

- AraBERT : +2,50 points de F1 (de 0,8180 à 0,8430) *** ;
- MARBERT : +1,49 points de F1 (de 0,8288 à 0,8437) *** ;
- mBERT : +0,78 points de F1 (de 0,7819 à 0,7896) (n.s. après correction).

(*** = $p < 0,001$, test de McNemar avec correction de Bonferroni)

Impact :

Cette étude fournit des preuves empiriques systématiques que l'injection morphologique améliore significativement et de manière régulière les performances, avec un effet différencié selon le modèle et la tâche. Le modèle AraBERT-LiPAr atteint un score F1 de 0,8430 sur ANERcorp. Ce résultat dépasse certains scores précédemment rapportés sur ce corpus, notamment 0,842 pour (Antoun, Baly, Hajj 2020) et 0,841 dans l'étude de (Inoue et al. 2021). Dans des conditions expérimentales strictement comparables, cette performance se situe au niveau des meilleurs résultats rapportés dans la littérature sur ANERcorp à la date de rédaction.

Ces résultats établissent LiPAr comme une approche empiriquement validée, capable d'améliorer de manière robuste des modèles de profils variés, tout en atteignant des performances de référence sur des corpus standards.

<u>Contribution 3</u> : Analyse linguistique approfondie des mécanismes

Notre troisième contribution est une analyse fine des mécanismes linguistiques expliquant les gains observés.

Enseignements principaux :

1. Les traits nominaux (CASE, GENDER, NUMBER, DEFINITE) sont les plus impactants, cruciaux pour la résolution des structures en iḍāfa et les accords ;

2. Effet compensatoire vs. synergique :

- mBERT : compensation d'un déficit morphologique (effet massif) ;
- AraBERT/MARBERT : synergie avec le pré-entraînement (effet modéré) ;

3. POS vs NER :

- POS (tâche morpho-syntaxique) : gains systématiques ;
- NER (tâche sémantique) : gains dépendants de la qualité du pré-entraînement ;

4. Analyse par classe :

- NOUN, VERB, ADJ : gains majeurs ;
- PART, PUNCT : peu affectés (cohérent avec une morphologie faible).

Impact :

Compréhension théorique des interactions entre morphologie et représentations neuronales, ouvrant des voies pour une modélisation explicite de langues riches.

Cette analyse suggère que les gains observés ne relèvent pas d'un effet d'ingénierie ponctuel, mais d'une interaction linguistiquement motivée entre morphologie et contextualisation neuronale.

<u>Contribution 4</u> : Positionnement théorique et implications

Notre quatrième contribution est une réflexion théorique sur le paradigme « morpho-centré » pour le TALN des langues morphologiquement riches.

Thèse défendue :

Pour les langues à morphologie riche et non-concaténative, l'injection explicite de traits morphologiques apparaît comme un levier structurellement déterminant dans les tâches évaluées.

Implications :

- Critique du paradigme « distribution-only » ;
- Proposition de principes pour un NLP morpho-aware ;
- Cadre conceptuel pour le Knowledge-Infused Learning linguistique ;
- Généralisation à la famille des langues sémitiques.

Impact :

Les résultats obtenus contribuent au débat théorique actuel sur le rôle que peuvent encore jouer les connaissances linguistiques explicites à l'ère des modèles neuronaux massifs.

8.3 Principaux résultats et enseignements

Résultat 1 : L'injection morphologique améliore systématiquement le POS tagging

Constat principal :

Sur l’ensemble des modèles testés, l’enrichissement morphologique explicite conduit à des gains systématiques et statistiquement significatifs en POS tagging.

Enseignement :

La tâche de POS tagging, intrinsèquement morpho-syntaxique, bénéficie directement et systématiquement de signaux morphologiques explicites, indépendamment de la qualité du modèle de base.

Implication pratique :

Pour les applications nécessitant une analyse morpho-syntaxique de l’arabe dans des contextes comparables à ceux étudiés dans cette thèse, l’enrichissement LiPAr constitue une amélioration systématique pertinente.

Résultat 2 : Les gains en NER dépendent fortement du pré-entraînement

Constat principal :

En reconnaissance d'entités nommées, les gains varient considérablement :

- AraBERT (MSA pur) : +2,50 pts (gain majeur) ;
- MARBERT (dialectal) : +1,49 pts (gain modéré) ;
- mBERT (multilingue) : +0,78 pts (gain marginal, n.s.).

Enseignement :

Le NER, tâche sémantique de haut niveau, nécessite une synergie entre représentations contextuelles (pré-entraînement) et signaux morphologiques. Les traits aident surtout quand le modèle capte déjà le sens (AraBERT), moins quand il est faible sémantiquement (mBERT).

Implication pratique :

L'enrichissement morphologique est plus efficace sur des modèles arabes spécialisés que sur des modèles multilingues génériques pour les tâches sémantiques.

<u>Résultat 3 :</u> Les traits nominaux sont les plus décisifs

Constat principal :

Les études d'ablation et l'analyse par classes montrent que les traits CASE, GENDER, NUMBER et DEFINITE apportent l'essentiel des gains.

Enseignement :

La morphologie nominale arabe (marquage casuel, accord, définitude) est plus discriminante que la morphologie verbale pour les tâches étudiées reflétant l'importance structurante du syntagme nominal en arabe.

Implication théorique :

Pour modéliser efficacement la langue arabe, il faut prioritairement capturer la structure du syntagme nominal (SN), notamment les constructions en iḍāfa.

<u>Résultat 4</u> : La fusion précoce offre le meilleur compromis global performance / stabilité / coût computationnel

Constat principal :

La stratégie de fusion précoce (*early injection* dans les embeddings) produit de meilleurs résultats que la fusion tardive (combination des logits).

Enseignement :

L'injection précoce permet de propager le signal morphologique à travers toutes les couches du Transformer, créant des représentations internes morphologiquement informées. La fusion tardive, bien que plus flexible, intervient trop tard pour modifier la computation interne.

Implication méthodologique :

Pour intégrer connaissances externes, privilégier l'injection précoce plutôt que le post-traitement, sauf contraintes spécifiques (par exemple contraintes d'interprétabilité décisionnelle ou d'intégration tardive dans des pipelines existants).

Enseignement transversal : Le retour à la linguistique est payant

Au-delà des résultats quantitatifs, cette thèse démontre que l'incorporation raisonnée de connaissances linguistiques explicites dans des modèles neuronaux n'est pas un retour en arrière, mais un progrès méthodologique.

Nos expériences suggèrent que l'apprentissage purement bout-en-bout montre certaines limites lorsqu'il est appliqué à des langues structurellement complexes. Nos expériences montrent que l'hybridation entre approches symboliques et neuronales, loin d'être un compromis, est une voie d'excellence scientifique combinant rigueur linguistique, puissance computationnelle, interprétabilité et performance. Ce constat dépasse le cas arabe et s'applique à l'ensemble des langues morphologiquement riches sous-représentées en TALN.

8.4 Limites et difficultés rencontrées

8.4.1 Limites méthodologiques

Malgré la rigueur du protocole, plusieurs limites structurelles subsistent :

1. Restriction au MSA formel :
   - Les dialectes arabes (représentant 70 % de l'usage réel) ne sont pas couverts ;
   - La généralisation est limitée hors des textes écrits standards.
2. Taille modeste des corpus :
   - UD_Arabic-PADT : ~7 000 phrases ;
   - ANERcorp : ~150 000 mots ;
   - Risque de sur-apprentissage classes rares.
3. Restriction expérimentale à deux tâches représentatives
   - POS tagging et NER évalués ;
   - Parsing, diacritisation et traduction non explorés.
4. Dépendance à un analyseur externe
   - Stanza (≈95 % d'exactitude sur les traits morphologiques UD pour l'arabe standard) introduit néanmoins du bruit.
   - Dépendance partielle à des outils morphologiques préexistants.

8.4.2 Difficultés techniques surmontées

Plusieurs défis techniques ont nécessité des solutions ad-hoc :

1. Alignement subword-mot :
   - WordPiece fragmente les unités morphologiques ;
   - Solution : stratégie assign-first avec masquage.
2. Hétérogénéité des modèles :
   - mBERT, AraBERT, MARBERT ont des tokenisations différentes ;
   - Solution : normalisation et projection uniformes.
3. Coût computationnel :
   - Extraction des traits + entraînement = ressources significatives ;
   - Solution : optimisation du code et utilisation efficiente de GPU.

Ces difficultés, bien que surmontées, soulignent la complexité inhérente aux approches hybrides.

8.5 Perspectives et travaux futurs

8.5.1 Perspectives immédiates

1. Extension aux dialectes arabes :
   - Explorer l'adaptation de LiPAr au maghrébin, à l'égyptien et au levantin ;
   - Créer des analyseurs morphologiques dialectaux ;
   - Évaluer le transfert MSA → Dialectes.

2. Application au dependency parsing :
    - Architecture BiLSTM + Biaffine + LiPAr ;
    - Évaluation sur UD_Arabic-PADT ;
    - Analyse de l'impact des traits sur les arcs syntaxiques.
3. Intégration dans les LLM arabes :
    - Tester sur Jais (Sengupta et al. 2023), AceGPT (Huang et al. 2024) ou AraGPT2 (Antoun, Baly, Hajj 2021) ;
    - Génération morphologiquement guidée via des contraintes morphologiques explicites au niveau du décodage ou des prompts structurés ;
    - Amélioration de la cohérence morphologique des sorties.
4. Déploiement d'une API open-source
    - Bibliothèque Python LiPAr ;
    - Modèles pré-entraînés disponibles ;
    - Documentation et tutoriels.

8.5.2 Perspectives de recherche fondamentale

1. Apprentissage morphologique end-to-end ;
    - Réduire, ou même éliminer la dépendance à un analyseur externe ;
    - Multi-task learning joint (morphologie + tâche) ;
    - Pré-entraînement auto-supervisé morphologique.
2. Représentations morphologiques continues
    - Embeddings morphologiques appris (vs. one-hot discrets) ;
    - Capture des similitudes morphologiques ;
    - Généralisation aux formes rares.
3. Transfer learning cross-lingue :
    - LiPAr-arabe → LiPAr-amharique ;
    - Modèle morphologique universel pour les langues sémitiques ;
    - Multilingual Morphological Transformer.
4. Benchmark morphologique standardisé :
    - MABA (Morphological Awareness Benchmark for Arabic) ;
    - Test suite + métriques + leaderboard public ;
    - Standard communautaire pour l'évaluation morphologique.

8.5.3 Perspectives applicatives

1. Outils industriels :
    - Service NER en temps réel pour les médias arabes ;
    - Assistant de rédaction morpho-syntaxique ;
    - Correcteur morphologique intelligent.
2. Domaines spécifiques :
    - Analyse des réseaux sociaux arabes (sentiment, événements) ;
    - Traitement de documents historiques (manuscrits, archives) ;
    - Applications éducatives (apprentissage de l'arabe).

3. Extensions multimodales :
    - Vision + langage arabe avec morphologie ;
    - Speech-to-text morpho-aware ;
    - LiPAr multimodal.

Ces perspectives constituent un programme de recherche progressif, dont la réalisation dépendra des ressources disponibles, des avancées communautaires et des évolutions des modèles de fondation.

8.6 Réflexion finale

8.6.1 Un plaidoyer pour le TALN hybride

Cette thèse s'achève sur une conviction profonde : Les résultats présentés dans cette thèse suggèrent que l'hybridation entre apprentissage profond et connaissances linguistiques explicites constitue une direction particulièrement prometteuse pour les langues morphologiquement riches.

Les modèles neuronaux, aussi puissants soient-ils, ne remplaceront pas entièrement la rigueur descriptive de la linguistique. Inversement, les approches purement symboliques ne peuvent rivaliser avec la capacité de généralisation des réseaux profonds.

LiPAr illustre qu'une troisième voie est possible : celle de la synergie, où chaque paradigme renforce l'autre.

8.6.2 Au-delà de l'arabe : un modèle pour les langues sous-représentées

Si cette thèse s'est concentrée sur l'arabe, ses enseignements dépassent largement ce cadre linguistique. Les principes de LiPAr sont applicables à toute langue présentant :

- Une morphologie riche et productive ;

- Une sous-représentation dans les corpus de pré-entraînement ;

- Une structure interne non-concaténative

Langues candidates : turc, finnois, hongrois, amharique, langues dravidiennes, langues bantoues, etc. sous réserve d'adaptations spécifiques à chaque système morphologique.

Dans cette perspective, Le cadre LiPAr offre ainsi une contribution méthodologique générique pour réduire le biais linguistique du TALN contemporain.

8.6.3 Message de clôture

Le chemin parcouru dans cette thèse, de la caractérisation linguistique fine de l'arabe (Chapitre 2) à la formalisation d'un cadre computationnel rigoureux (Chapitre 4), en passant par la validation empirique systématique (Chapitres 5-6) et enfin la réflexion théorique (Chapitre 7), illustre la complémentarité essentielle entre linguistique et informatique.

Notre espoir est que LiPAr inspire d'autres chercheurs à réhabiliter la connaissance linguistique explicite dans l'ère des modèles neuronaux massifs, non par nostalgie du passé symbolique, mais par exigence scientifique de rigueur, d'interprétabilité et de performance.

L'arabe, langue millénaire à la morphologie d'une richesse fascinante, mérite des outils computationnels à la hauteur de sa complexité. Elle nous rappelle que la diversité linguistique n’est pas un obstacle technique à surmonter, mais une richesse scientifique à célébrer. LiPAr constitue une contribution méthodologique rigoureuse et ouverte, inscrite dans une vision du TALN respectueuse de la diversité linguistique et attentive aux spécificités structurelles des langues.

En définitive, cette thèse plaide pour un traitement automatique des langues qui ne cherche pas à uniformiser la diversité linguistique, mais à la comprendre, la modéliser et la valoriser comme un objet scientifique à part entière.

Bibliographie

ABDELALI, Ahmed, DARWISH, Kareem, DURBANI, Nadir et MUBARAK, Hamdy, 2016. Farasa: A Fast and Furious Segmenter for Arabic. In : *Proceedings of NAACL 2016: Demonstrations*. San Diego, California, USA. 2016. pp. 11-16. DOI 10.18653/v1/N16-3003.

ABDUL-MAGEED, Muhammad, ELMADANY, AbdelRahim et NAGOUDI, El Moatez Billah, 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In : ZONG, Chengqing, XIA, Fei, LI, Wenjie et NAVIGLI, Roberto (éd.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* [en ligne]. Online : Association for Computational Linguistics. 2021. pp. 7088-7105. DOI 10.18653/v1/2021.acl-long.551. [Consulté le 19 février 2026].

ALBAHLI, Saleh, 2025. An Advanced Natural Language Processing Framework for Arabic Named Entity Recognition: A Novel Approach to Handling Morphological Richness and Nested Entities. *Applied Sciences*. 2025. Vol. 15, n° 6. DOI 10.3390/app15063073.

AL-SALLAB, Ahmad, BALY, Ramy, HAJJ, Hazem, BASHIR, Shaban Khaled, EL-HAJJ, Wassim et BADARO, Gilbert, 2017. AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic as a Low Resource Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 2017. Vol. 16, n° 4, pp. 1-20. DOI 10.1145/3086575.

ANTOUN, Wissam, BALY, Fady et HAJJ, Hazem, 2020. AraBERT: Transformer-Based Model for Arabic Language Understanding. In : *Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*. Marseille, France. 2020. pp. 11-16. DOI 10.48550/arXiv.2003.00104.

ANTOUN, Wissam, BALY, Fady et HAJJ, Hazem, 2021. AraGPT2: Pre-Trained Transformer for Arabic Language Generation. In : HABASH, Nizar et ET AL. (ÉD) (éd.), *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual) : Association for Computational Linguistics. 2021. pp. 196-207.

BADAWI, El-Said M., 1973. *Mustawayāt al-ʿArabiyya al-Muʿāṣira fī Miṣr*. Le Caire, Egypte : Dār al-Maʿārif.

BELINKOV, Yonatan, DURRANI, Nadir, DALVI, Fahim, SAJJAD, Hassan et GLASS, James, 2017. What do Neural Machine Translation Models Learn about Morphology? In : BARZILAY, Regina et KAN, Min-Yen (éd.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada : Association for Computational Linguistics. 2017. pp. 861-872. DOI 10.18653/v1/P17-1080.

BENAJIBA, Yassine, DIAB, Mona et ROSSO, Paolo, 2008. Arabic named entity recognition using optimized feature sets. In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*. Honolulu, Hawaii : Association for Computational Linguistics. 2008. pp. 284-293. DOI 10.3115/1613715.1613755.

BOJANOWSKI, Piotr, GRAVE, Edouard, JOULIN, Armand et MIKOLOV, Tomas, 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 2017. Vol. 5, pp. 135-146. DOI 10.1162/tacl_a_00051.

BOUDELAA, Sami et MARSLEN-WILSON, William D., 2011. Productivity and priming: Morphemic decomposition in Arabic. *Language and Cognitive Processes*. 2011. Vol. 26, n° 4-6, pp. 624-652. DOI 10.1080/01690965.2010.521022.

BUCKWALTER, TIM, 2002. *Buckwalter Arabic Morphological Analyzer Version 1.0*. 8 novembre 2002. Linguistic Data Consortium.

BUCKWALTER, TIM, 2004. *Buckwalter Arabic Morphological Analyzer Version 2.0*. 15 décembre 2004. Linguistic Data Consortium.

CONNEAU, Alexis, RINOTT, Ruty, LAMPLE, Guillaume, WILLIAMS, Adina, BOWMAN, Samuel, SCHWENK, Holger et STOYANOV, Veselin, 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium : Association for Computational Linguistics. 2018. pp. 2475-2485. DOI 10.18653/v1/D18-1269.

DARWISH, Kareem, HABASH, Nizar et ABBAS, Mourad, 2021. A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*. 2021. Vol. 64, n° 4, pp. 72-81. DOI 10.1145/3447735.

DARWISH, Kareem, MUBARAK, Hamdy, ABDELALI, Ahmed, ELDESOUKI, Mohamed, SAMIH, Younes, ALHARBI, Randah, ATTIA, Mohammed, MAGDY, Walid et KALLMEYER, Laura, 2018. Multi-Dialect Arabic POS Tagging: A CRF Approach. In : CALZOLARI, Nicoletta et ET AL., Takenobu (éd.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan : European Language Resources Association (ELRA). 2018. pp. 93-98.

DEBILI, Fathi, HADHÉMI, Achour et SOUISSI, Emna, 2002. La langue arabe et l'ordinateur de l'étiquetage grammatical à la voyellation automatique. Correspondances: bulletin de l'IRMC, ISSN 0330-7417, Nº 71, 2002, p. 10-26. . 2002. N° 71, pp. 10-26.

DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton et TOUTANOVA, Kristina, 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In : *Proceedings of the 2019 Conference of the North*. Minneapolis, Minnesota : Association for Computational Linguistics. 2019. pp. 4171-4186. DOI 10.18653/v1/N19-1423.

DIAB, Mona, HACIOGLU, Kadri et JURAFSKY, Daniel, 2007. Automatic Processing of Modern Standard Arabic Text. In : SOUDI, Abdelhadi, BOSCH, Antal Van Den et NEUMANN, Günter (éd.), *Arabic Computational Morphology*. Dordrecht : Springer Netherlands. pp. 159-179. Text, Speech and Language Technology. ISBN 978-1-4020-6045-8.

DIETTERICH, Thomas G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*. 1998. Vol. 10, n° 7, pp. 1895-1923. DOI 10.1162/089976698300017197.

DROR, Rotem, BAUMER, Gili, SHLOMOV, Segev et REICHART, Roi, 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers). Melbourne, Australia : Association for Computational Linguistics. 2018. pp. 1383-1392. DOI 10.18653/v1/P18-1128.

DUKES, Kais, 2009. The quranic Arabic corpus. *Language Research Group University of Leeds* [en ligne]. 2009. Disponible à l'adresse : http://corpus.quran.com/

DUNN, Olive Jean, 1961. Multiple Comparisons among Means. *Journal of the American Statistical Association*. 1961. Vol. 56, n° 293, pp. 52-64. DOI 10.1080/01621459.1961.10482090.

EDMISTON, Daniel, 2020. *A Systematic Analysis of Morphological Content in BERT Models for Multiple Languages*. 2020. arXiv.

ESKANDER, Ramy, LOWRY, Cass, KHANDAGALE, Sujay, KLAVANS, Judith, POLINSKY, Maria et MURESAN, Smaranda, 2022. Unsupervised Stem-based Cross-lingual Part-of-Speech Tagging for Morphologically Rich Low-Resource Languages. In : *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, USA, virtuel. 2022. pp. 4061-4072. DOI 10.18653/v1/2022.naacl-main.298.

FERGUSON, Charles A., 1959. Diglossia. WORD. 1959. Vol. 15, n° 2, pp. 325-340. DOI 10.1080/00437956.1959.11659702.

GAUR, Manas, GUNARATNA, Kalpa, BHATT, Shreyansh et SHETH, Amit, 2022. Knowledge-Infused Learning: A Sweet Spot in Neuro-Symbolic AI. *IEEE Internet Computing*. 2022. Vol. 26, n° 4, pp. 5-11. DOI 10.1109/MIC.2022.3179759.

HABASH, Nizar, DIAB, Mona et RAMBOW, Owen, 2012. Conventional Orthography for Dialectal Arabic. In : CALZOLARI, Nicoletta, CHOUKRI, Khalid, ET AL. et PIPERIDIS, Stelios (éd.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey : European Language Resources Association (ELRA). 2012. pp. 711-718.

HABASH, Nizar et RAMBOW, Owen, 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In : KNIGHT, Kevin, NG, Hwee Tou et OFLAZER, Kemal (éd.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan : Association for Computational Linguistics. juin 2005. pp. 573-580. DOI 10.3115/1219840.1219911.

HABASH, Nizar, RAMBOW, Owen et KIRAZ, George, 2005. Morphological Analysis and Generation for Arabic Dialects. In : DARWISH, Kareem, DIAB, Mona et HABASH, Nizar (éd.), *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*. Ann Arbor, Michigan : Association for Computational Linguistics. 2005. pp. 17-24.

HABASH, Nizar, RAMBOW, Owen et ROTH, Ryan, 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*. 2009. pp. 102-107.

HABASH, Nizar Y., 2010. *Introduction to Arabic Natural Language Processing*. San Rafael : Morgan & Claypool Publishers. Synthesis Lectures on Human Language Technologies Ser. ISBN 978-1-59829-795-9.

HAJIČ, Jan, SMRŽ, Otakar, PETR, Zemánek, SNAIDAUF, Jaň et BEŠKA, Emanuel, 2004. Prague Arabic Dependency Treebank: development in data and tools. *In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*. 2004. pp. 110-117.

HAYWOOD, John A. et NAHMAD, Hayim M., 2009. *A new Arabic grammar of the written language*. 2e éd. 1965, (Réimpr 1990). London : Lund Humphries. ISBN 978-0-85331-585-8.

HOLES, Clive, 2004. *Modern Arabic: Structures, Functions, and Varieties, Revised Edition*. 2nd ed. Washington : Georgetown University Press. Georgetown Classics in Arabic Languages and Linguistics Series. ISBN 978-1-58901-022-2.

HOULSBY, Neil, GIURGIU, Andrei, JASTRZEBSKI, Stanislaw, MORRONE, Bruna, LAROUSSILHE, Quentin de, GESMUNDO, Andrea, ATTARIYAN, Mona et GELLY, Sylvain, 2019. *Parameter-Efficient Transfer Learning for NLP*. 2019. arXiv. arXiv:1902.00751[cs]

HUANG, Huang, YU, Fei, ZHU, Jianqing, SUN, Xuening, CHENG, Hao, SONG, Dingjie, CHEN, Zhihong, ALHARTHI, Abdulmohsen, AN, Bang, HE, Juncai, LIU, Ziche, ZHANG, Zhiyi, CHEN, Junying, LI, Jianquan, WANG, Benyou, ZHANG, Lian, SUN, Ruoyu, WAN, Xiang, LI, Haizhou et XU, Jinchao, 2024. *AceGPT, Localizing Large Language Models in Arabic*. 2024. arXiv. arXiv:2309.12053 [cs]

INOUE, Go, ALHAFNI, Bashar, BAIMUKAN, Nurpeiis, BOUAMOR, Houda et HABASH, Nizar, 2021. *The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models*. 2021. arXiv.

INOUE, Go, KHALIFA, Salam et HABASH, Nizar, 2022. Morphosyntactic Tagging with Pre-trained Language Models for Arabic and its Dialects. In : *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland : Association for Computational Linguistics. 2022. pp. 1708-1719. DOI 10.18653/v1/2022.findings-acl.135.

JOSHI, Pratik, SANTY, Sebastin, BUDHIRAJA, Amar, BALI, Kalika et CHOUDHURY, Monojit, 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online : Association for Computational Linguistics. 2020. pp. 6282-6293. DOI 10.18653/v1/2020.acl-main.560.

KEERSMAEKERS, Alek et MERCELIS, Wouter, 2024. Adapting transformer models to morphological tagging of two highly inflectional languages: a case study on Ancient Greek and Latin. In : *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages*. 2024. pp. 165-176.

KOUBAA, Anis, AMMAR, Adel, GHOUTI, Lahouari, NAJAR, Omar et SIBAEE, Serry, 2024. *ArabianGPT: Native Arabic GPT-based Large Language Model*. 2024. arXiv. arXiv:2402.15313[cs]

KURSUNCU, Ugur, GAUR, Manas et SHETH, Amit, 2019. Knowledge Infused Learning (K-IL): Towards Deep Incorporation of Knowledge in Deep Learning. . 2019. DOI 10.48550/ARXIV.1912.00512.

LAFFERTY, John D., MCCALLUM, Andrew et PEREIRA, Fernando C. N., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In : *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc. 2001. pp. 282-289. ICML '01. ISBN 978-1-55860-778-1.

LAKHFIF, Abdelaziz et LASKRI, Mohamed Tayeb, 2017. L'analyse et l'annotation à base de FrameNet : contribution à l'étude contrastive des événements de mouvement en arabe et en anglais [FrameNet-based analysis and annotation: contribution to the constrastive study of motion events in Arabic and English]. *Traitement Automatique des Langues*. 2017. Vol. 58, n° 3, pp. 67-95.

LIM, Kyungtae et PARK, Jungyeul, 2020. Part-of-Speech Tagging Using Multiview Learning. *IEEE Access* [en ligne]. 2020. Vol. 8, pp. 195184-195196. DOI 10.1109/ACCESS.2020.3033979. [Consulté le 14 mars 2026].

LIU, Weijie, ZHOU, Peng, ZHAO, Zhe, WANG, Zhiruo, JU, Qi, DENG, Haotang et WANG, Ping, 2020. K-BERT: Enabling Language Representation with Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020. Vol. 34, n° 03, pp. 2901-2908. DOI 10.1609/aaai.v34i03.5681.

LIU, Yinhan, OTT, Myle, GOYAL, Naman, DU, Jingfei, JOSHI, Mandar, CHEN, Danqi, LEVY, Omer, LEWIS, Mike, ZETTLEMOYER, Luke et STOYANOV, Veselin, 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv. arXiv:1907.11692.

LOSHCHILOV, Ilya et HUTTER, Frank, 2019. *Decoupled Weight Decay Regularization*. 2019. arXiv. arXiv:1711.05101.

LOUKAM, Mourad, BALLA, Amar et LASKRI, Mohamed Tayeb, 2016. Towards an open platform based on HPSG formalism for the standard Arabic language. *International Journal of Speech Technology*. 2016. Vol. 19, n° 2, pp. 325-338. DOI 10.1007/s10772-015-9314-4.

MAAMOURI, Mohamed, BIES, Ann, BUCKWALTER, Tim et MEKKI, Wigdan, 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In : *Proceedings of NEMLAR Conference*. 2004. pp. 102-109.

MAHDJOUBI, Roussafi et LASKRI, Mohamed Tayeb, 2025. Tackling Arabic NLP Challenges: POS-Tagging with Transformer-Based Models and Nuanced Evaluation. *International Journal of Performability Engineering*. 2025. Vol. 21, n° 8, pp. 411-421. DOI 10.23940/ijpe.25.08.p1.411421.

MANNING, Christopher D., 2022. Human Language Understanding & Reasoning. *Daedalus*. 2022. Vol. 151, n° 2, pp. 127-138. DOI 10.1162/daed_a_01905.

MCNEMAR, Quinn, 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*. 1947. Vol. 12, n° 2, pp. 153-157. DOI 10.1007/BF02295996.

MEGUEHOUT, Hamza, BOUHADADA, Tahar et LASKRI, Mohamed Tayeb, 2017. Semantic role labeling for Arabic language using case-based reasoning approach. *Int. J. Speech Technol.* 2017. Vol. 20, n° 2, pp. 363-372. DOI 10.1007/s10772-017-9412-6.

MIKOLOV, Tomas, CHEN, Kai, CORRADO, Greg et DEAN, Jeffrey, 2013. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv. arXiv:1301.3781.

MITCHELL, Tom M., 1997. *Machine learning*. Nachdr. New York : McGraw-Hill. McGraw-Hill series in Computer Science. ISBN 978-0-07-042807-2.

NIVRE, Joakim, DE MARNEFFE, Marie-Catherine, GINTER, Filip, HAJIČ, Jan, MANNING, Christopher D., PYYSALO, Sampo, SCHUSTER, Sebastian, TYERS, Francis et ZEMAN, Daniel, 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In : CALZOLARI, Nicoletta et ET AL. (éd.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France : European Language Resources Association. 2020. pp. 4034-4043. ISBN 979-10-95546-34-4.

OBEID, O., ZALMOUT, N., KHALIFA, S., et AL, 2020. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In : *Proceedings of LREC'20*. Marseille, France. 2020. pp. 7022-7032.

OWENS, Jonathan, 2013. *The Oxford handbook of Arabic linguistics*. Oxford : Oxford University Press. ISBN 978-0-19-976413-6.

PASHA, Arfath, AL-BADRASHINY, Mohamed, DIAB, Mona, EL KHOLY, Ahmed, ESKANDER, Ramy, HABASH, Nizar, POOLEERY, Manoj, RAMBOW, Owen et ROTH, Ryan, 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In : CALZOLARI, Nicoletta et ET AL. (éd.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland : European Language Resources Association (ELRA). 2014. pp. 1094-1101.

PENNINGTON, Jeffrey, SOCHER, Richard et MANNING, Christopher, 2014. GloVe: Global Vectors for Word Representation. In : MOSCHITTI, Alessandro, PANG, Bo et DAELEMANS, Walter (éd.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics. 2014. pp. 1532-1543. DOI 10.3115/v1/D14-1162.

PETERS, Matthew E., NEUMANN, Mark, LOGAN, Robert, SCHWARTZ, Roy, JOSHI, Vidur, SINGH, Sameer et SMITH, Noah A., 2019. Knowledge Enhanced Contextual Word Representations. In : INUI, Kentaro et ET AL. (éd.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China : Association for Computational Linguistics. 2019. pp. 43-54. DOI 10.18653/v1/D19-1005.

PINTER, Yuval, GUTHRIE, Robert et EISENSTEIN, Jacob, 2017. Mimicking Word Embeddings using Subword RNNs. In : PALMER, Martha, HWA, Rebecca et RIEDEL, Sebastian (éd.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark : Association for Computational Linguistics. 2017. pp. 102-112. DOI 10.18653/v1/D17-1010.

QI, Peng, ZHANG, Yuhao, ZHANG, Yuhui, BOLTON, Jason et MANNING, Christopher D., 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. pp. 101-108.

RATNAPARKHI, Adwait, 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In : *Conference on Empirical Methods in Natural Language Processing*. 1996. pp. 133-142.

ROY, Kaushik, ZI, Yuxin, NARAYANAN, Vignesh, GAUR, Manas et SHETH, Amit, 2023. *Knowledge-Infused Self Attention Transformers*. 2023. arXiv.

RYDING Karin C., 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge : Cambridge University Press.

SALEH, Hager, ALMOHIMEED, Abdulaziz, HASSAN, Rasha, IBRAHIM, Mandour M., ALSAMHI, Saeed Hamood, HASSAN, Moatamad Refaat et MOSTAFA, Sherif, 2025. Advancing arabic dialect detection with hybrid stacked transformer models. *Frontiers in Human Neuroscience*. 2025. Vol. 19, n° 1498297. DOI 10.3389/fnhum.2025.1498297.

SEKER, Amit et TSARFATY, Reut, 2020. A Pointer Network Architecture for Joint Morphological Segmentation and Tagging. *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020. pp. 4368-4378. DOI 10.18653/v1/2020.findings-emnlp.391.

SENGUPTA, Neha, SAHU, Sunil Kumar, JIA, Bokang, KATIPOMU, Satheesh, LI, Haonan, KOTO, Fajri, MARSHALL, William, GOSAL, Gurpreet, LIU, Cynthia, CHEN, Zhiming, AFZAL, Osama Mohammed, KAMBOJ, Samta, PANDIT, Onkar, PAL, Rahul, PRADHAN, Lalit, MUJAHID, Zain Muhammad, BAALI, Massa, HAN, Xudong, BSHARAT, Sondos Mahmoud, AJI, Alham Fikri, SHEN, Zhiqiang, LIU, Zhengzhong, VASSILIEVA, Natalia, HESTNESS, Joel, HOCK, Andy, FELDMAN, Andrew, LEE, Jonathan, JACKSON, Andrew, REN, Hector Xuguang, NAKOV, Preslav, BALDWIN, Timothy et XING, Eric, 2023. *Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models*. 2023. arXiv. arXiv:2308.16149.

SINGH, Telem Joyson, SINGH, Sanasam Ranbir et SARMAH, Priyankoo, 2023. Subwords to word back composition for morphologically rich languages in neural machine translation. In : *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. 2023.

SOKOLOVA, Marina et LAPALME, Guy, 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009. Vol. 45, n° 4, pp. 427-437. DOI 10.1016/j.ipm.2009.03.002.

SOLIMAN, Abu Bakr, EISSA, Kareem et EL-BELTAGY, Samhaa R., 2017. *AraVec: A set of arabic word embedding models for use in Arabic NLP*. 2017.

STRAKA, Milan, HAJIČ, Jan et STRAKOVÁ, Jana, 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16*. 2016. pp. 4290-4297.

TAJI, Dima, KHALIFA, Salam, OBEID, Ossama, ERYANI, Fadhl et HABASH, Nizar, 2018. An Arabic Morphological Analyzer and Generator with Copious Features. In : *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Brussels, Belgium : Association for Computational Linguistics. 2018. pp. 140-150. DOI 10.18653/v1/W18-5816.

TAWFIK, Ahmed, EMAM, Mahitab, ESSAM, Khaled, NABIL, Robert et HASSAN, Hany, 2019. Morphology-aware Word-Segmentation in Dialectal Arabic Adaptation of Neural Machine Translation. In : *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. Florence, Italy : Association for Computational Linguistics. 2019. pp. 11-17. DOI 10.18653/v1/W19-4602.

TOLEGEN, Gulmira, TOLEU, Alymzhan, MUSSABAYEV, et RUSTAM, 2024. Contrastive learning for morphological disambiguation using large language models in low-resource settings. *Applied Sciences* [en ligne]. 2024. Vol. 14, n° 21, pp. 9992. DOI 10.3390/app14219992.

UEDA, Nobuhiro, OMURA, Kazumasa, KODAMA, Takashi, KIYOMARU, Hirokazu, MURAWAKI, Yugo, KAWAHARA, Daisuke et KUROHASHI, Sadao, 2023. KWJA: A Unified Japanese Analyzer Based on Foundation Models. In : *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations*. Toronto, Canada. 2023. pp. 538-548. DOI 10.18653/v1/2023.acl-demo.52.

VANIA, Clara et LOPEZ, Adam, 2017. From Characters to Words to in Between: Do We Capture Morphology? In : BARZILAY, Regina et KAN, Min-Yen (éd.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada : Association for Computational Linguistics. 2017. pp. 2016-2027. DOI 10.18653/v1/P17-1184.

VASWANI, A., SHAZEER, N., PARMAR, N., et AL, 2017. Attention Is All You Need. In : *Advances in Neural Information Processing Systems*.

VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, USZKOREIT, Jakob, JONES, Llion, GOMEZ, Aidan N., KAISER, Łukasz et POLOSUKHIN, Illia, 2017. Attention is all you need. In : *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA : Curran Associates Inc. 4 décembre 2017. pp. 6000-6010. NIPS'17. ISBN 978-1-5108-6096-4.

WOLF, Thomas, DEBUT, Lysandre, SANH, Victor, CHAUMOND, Julien, DELANGUE, Clement, MOI, Anthony, CISTAC, Pierric, RAULT, Tim, LOUF, Remi, FUNTOWICZ, Morgan, DAVISON, Joe, SHLEIFER, Sam, VON PLATEN, Patrick, MA, Clara, JERNITE, Yacine, PLU, Julien, XU, Canwen, LE SCAO, Teven, GUGGER, Sylvain, DRAME, Mariama, LHOEST, Quentin et RUSH, Alexander, 2020. Transformers: State-of-the-Art Natural Language Processing. In : LIU, Qun et SCHLANGEN, David (éd.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics. 2020. pp. 38-45. DOI 10.18653/v1/2020.emnlp-demos.6.

ZALMOUT, Nasser et HABASH, Nizar, 2017. Don't Throw Those Morphological Analyzers Away Just Yet: Neural Morphological Disambiguation for Arabic. In : *Proceedings of EMNLP 2017*. Copenhagen, Denmark, ACL. 2017. pp. 704-713. DOI 10.18653/v1/D17-1073.

ZALMOUT, Nasser et HABASH, Nizar, 2019a. Joint Diacritization, Lemmatization, Normalization, and Fine-Grained Morphological Tagging. *ArXiv*. 2019.

ZALMOUT, Nasser et HABASH, Nizar, 2019b. *Adversarial Multitask Learning for Joint Multi-Feature and Multi-Dialect Morphological Modeling*. 2019. arXiv.

ZALMOUT, Nasser et HABASH, Nizar, 2020. Utilizing Subword Entities in Character-Level Sequence-to-Sequence Lemmatization Models. In : *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain : International Committee on Computational Linguistics. 2020. pp. 4676-4682. DOI 10.18653/v1/2020.coling-main.412.

ZHANG, Rongzhi, YU, Yue et ZHANG, Chao, 2020. SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup. . 2020. DOI 10.48550/ARXIV.2010.02322.