Analyse morpho-syntaxique massivement multilingue à l'aide de ressources typologiques, d'annotations universelles et de plongements de mots multilingues

Manon Scholivet

Aix*Marseille université
Socialement engagée

NNT/NL : 2021AIXM0306/013ED184

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 15 octobre 2021 par

Manon SCHOLIVET

Analyse morpho-syntaxique massivement multilingue à l'aide de ressources typologiques, d'annotations universelles et de plongements de mots multilingues

Discipline
Informatique

Spécialité
Traitement Automatique des Langues

École doctorale
ED184

Laboratoire
Laboratoire d'Informatique & Systèmes (LIS)

Composition du jury

Benoît CRABBÉ LLF	Rapporteur
Anne-Laure LIGOZAT LISN	Rapporteure
Laurent BESACIER LIG	Examinateur
Miryam DE LHONEUX LIIR	Examinatrice
Cécile CAPPONI LIS	Examinatrice
Alexis NASR LIS	Directeur de thèse
Carlos RAMISCH LIS	Co-Directeur de thèse
Benoit FAVRE LIS	Invité

LIS

TALEP

Je soussignée, Manon Scholivet, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique d'Alexis Nasr et de Carlos Ramisch, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 25/02/21

“N’oublie jamais, celui qui croit savoir n’apprend plus.”

Pierre Bottero

Résumé

L'annotation de données est un problème majeur dans toutes les tâches d'apprentissage automatique. Dans le domaine du Traitement Automatique des Langues (TAL), ce problème est multiplié par le nombre de langues existantes.

De nombreuses langues se retrouvent sans annotations, et sont alors mises à l'écart des systèmes de TAL. Une solution possible pour intégrer ces langues dans les systèmes est de tenter d'exploiter les langues disposant de nombreuses annotations, d'apprendre des informations sur ces langues bien dotées, et de transférer ce savoir vers les langues peu dotées.

Pour cela, il est possible de se reposer sur des initiatives comme les *Universal Dependencies*, qui proposent un schéma d'annotation universel entre les langues. L'utilisation de plongements de mots multilingues et de traits typologiques issus de ressources comme le *World Atlas of Language Structures* (WALS) sont des solutions permettant un partage de connaissances entre les langues.

Ces pistes sont étudiées dans le cadre de cette thèse, à travers la prédiction de l'analyse syntaxique, de la morphologie et des parties du discours sur 41 langues au total. Nous montrons que l'impact du WALS peut être positif dans un cadre multilingue, mais que son utilité n'est pas systématique dans une configuration d'apprentissage *zero-shot*. D'autres représentations des langues peuvent être apprises sur les données, et donnent de meilleurs résultats que le WALS, mais ont l'inconvénient de ne pas fonctionner dans un cadre de *zero-shot*. Nous mettons également en évidence l'importance de la présence d'une langue proche lors de l'apprentissage des modèles, ainsi que les problèmes liés à l'utilisation d'un modèle de caractère pour les langues isolées.

Mots clés : multilingue, étiquetage morpho-syntaxique, analyse syntaxique, zero-shot, traits typologiques

Abstract

Data annotation is a major problem in all machine learning tasks. In the field of NLP, this problem is multiplied by the number of existing languages.

Many languages do not have any annotations, and are therefore excluded from NLP systems. One possible solution to integrate these languages into the systems is to try to leverage the languages having many annotations, and to try to learn information about these resource-rich languages, and to transfer this knowledge to the low-resources languages.

It is possible to rely on initiatives such as Universal Dependencies, which propose a universal annotation scheme between languages. The use of multilingual word embeddings and typological features from resources such as the WALS are solutions allowing knowledge sharing between languages.

These tracks are studied in the framework of this thesis, through the prediction of parsing, morphology and parts of speech on 41 languages in total. We show that the impact of the WALS can be positive in a multilingual setting, but that its usefulness is not systematic in a zero-shot learning setting. Other language representations can be learned from the data, and perform better than the WALS, but have the downside of not working in a zero-shot setting. We also highlight the importance of the presence of a nearby language when learning patterns, as well as the problems associated with using a character pattern for isolated languages.

Keywords: multilingual, tagging, parsing, zero-shot, typological features

Remerciements

Après ces 4 années bien difficiles, il y a énormément de gens que je souhaite remercier, pour leur aide, leur patience, leurs conseils, et leur soutien.

J'aimerais commencer par remercier les membres du jury. Merci aux deux rapporteur·e·s, **Benoît Crabbé** et **Anne-Laure Ligozat** d'avoir pris le temps de lire cette thèse, et d'avoir fait des retours aussi justes que pertinents. Merci à **Laurent Besacier** d'avoir accepté de faire partie du jury, mais aussi de mon comité de suivi de thèse. Les trois comités réalisés avec lui et **Yann Vaxès** m'ont permis d'avancer dans la bonne direction. Merci également à **Miryam de Lhoneux**, pour son intérêt pour mes travaux et pour sa gentillesse lors de NAACL 2019. Enfin, merci à **Cécile Capponi**, que j'ai souvent eu l'occasion de croiser au sein du LIS. Merci pour nos échanges, toujours intéressants et emplis de bienveillance.

Parmi les membres du jury, j'ai bien évidemment une pensée particulière pour mes directeurs de thèse. Pour commencer, même s'il n'est pas officiellement sur les papiers, j'aimerais remercier **Benoit** pour son implication pendant ces 4 années. Merci de toujours avoir des idées aux moments de creux, d'être un incroyable puits de connaissances que tu partages avec tous les intéressés ! Merci de toujours me pousser à faire les choses, surtout quand je m'en crois pourtant incapable.

J'aimerais ensuite remercier **Carlos**, pour m'avoir prise en stage lors de mon M1, de m'avoir fait découvrir le monde magique des expressions polylexicales et du TAL en général. Merci de m'avoir dit un jour qu'ouvrir une porte ne voulait pas nécessairement dire qu'on ferme toutes les autres. Merci pour les nombreuses belles leçons qui m'ont aidée à grandir, autant scientifiquement qu'humainement.

Merci enfin à **Alexis**, avec qui j'ai probablement le plus travaillé pendant ces 4 années. Ta patience et ta bienveillance m'ont permis d'avancer, d'apprendre à identifier où se situaient mes intérêts scientifiques, à prendre de l'indépendance et à mieux m'affirmer professionnellement. Merci de m'avoir accompagnée jusqu'au bout, et d'avoir continué à croire en moi, même dans les moments les plus difficiles.

Ma gratitude va vers énormément de personnes qui m'ont aidée dans mon parcours, que ce soit des enseignants, des collègues (je pense à **Olivier**, pour m'avoir aidée à m'intégrer au LIF, et à **Marie-Hélène** pour avoir été la première à me dire "et pourquoi pas une thèse?"), des supers stagiaires (surtout **Thibault Roux**, qui même si il a découvert plein de problèmes dans mes travaux, a réalisé un super travail) ou des amis. Je dis toujours que j'aime pas les gens, et que j'ai du mal à me faire des amis, mais finalement, la liste des amis à remercier va être longue.

Merci à **Park**, **Baki**, **Darkkis**, **Chichine**, et **Hichem** pour m'avoir aidée à me détendre (ou à rager, au choix) le soir, quand je feed sur lol après les longues journées de travail.

Un merci particulier à toi **Baki**, toi qui es un chercheur pour qui j'ai tant d'admi-

ration. Merci de m'avoir dit que tu me respectais en tant que chercheuse, merci de m'avoir aidée à prendre confiance en moi. Merci pour tout.

Victoria, **Manon**, **Ivannah**, **Julien**, **Bastien**. Merci pour tout nos crousti moments, merci d'être là quand j'ai besoin de rire, ou quand j'ai besoin de soutien. Merci Victoria de m'avoir sautée dessus il y a maintenant 9 ans pour me proposer des photocopies. Dire qu'à l'époque, je suis arrivée en me disant que je ne voulais pas d'amis, que j'étais juste là pour bosser et c'est tout... Si j'avais su où on en serait aujourd'hui ! Merci infiniment pour votre présence si précieuse.

Un immense merci à mes bros, mes compagnons de thèse (ou presque). **Cindy**, **Franck**, **Raphael**, **Jeremy** et **Alexie**, merci d'avoir été là quand plus rien n'allait, de m'avoir aidée avec les cours, la thèse (merci Franck, best stagiaire ever) et tant d'autres choses. Je vous dois tellement de cookies pour vous remercier ! On se donne rendez-vous à New^n Moux pour continuer à faire la fête tous les étés.

Jérémy, toi qui as fait la majeure partie des graphiques et autres dessins de cette thèse (et de la soutenance !), toi qui as accepté de relire mes chapitres avant que je les envoie à mes chefs parce que je doutais de moi, toi qui m'as aidée à prendre en compte les retours les plus compliqués, toi qui as pris de ton temps pour ma thèse et moi, les mots me manquent pour te dire à quel point ton aide m'a été précieuse. Merci de me faire rire, de m'écouter pleurer, et de supporter la boule d'émotions que je suis. Merci d'être toujours là.

Enfin, les derniers mais pas des moindres, j'aimerais remercier ma famille. **Papa**, même si je t'avais demandé d'arrêter de me demander tout le temps comment allait ma thèse, sache que maintenant, tu peux m'en parler autant que tu veux ! Parce que je suis enfin fière de moi. Merci pour ton amour inconditionnel. **Maman**, j'aimerais te remercier pour les mille petites (et grosses !) choses que tu fais pour moi. Merci d'avoir relu cette thèse la veille de son envoi pour vérifier que je n'avais pas écrit n'importe comment. Merci pour tes nombreuses petites attentions du quotidien pour lesquelles je ne te remercie pas assez. Merci pour ton écoute les jours où j'ai besoin de parler, et pour les jours où j'ai juste envie de papoter sans m'arrêter ! **Lyne**, ma grande sœur que j'aime fort. Merci d'être là pour prendre les rênes quand je suis perdue et de m'aider à prendre les bonnes décisions. Merci d'être là depuis toujours, et de toujours être là. Merci d'être un phare de ma vie, un de mes deux Tiboulens. Et enfin, **Anaïs**. Ma sœur, ma colocataire, peut être ma meilleure amie. Mon deuxième Tiboulen. Aucun mot ne pourra résumer l'admiration et la gratitude que j'éprouve pour toi. Ma thèse est un peu la tienne, de même que ton mémoire était un peu le mien. Merci d'exister.

Je finirai ces remerciements sur une deuxième citation de Pierre Bottero, grand écrivain et poète de ce siècle, dont de nombreux mots sont gravés pour toujours en moi. Merci encore à tou·te·s de m'avoir aidée à aller de l'avant.

"Le doute est une force. Une vraie et belle force. Veille simplement qu'elle te pousse toujours en avant."

Table des matières

Résumé 4

Abstract 5

Remerciements 6

Table des matières 8

Liste des acronymes 10

Introduction générale 12

1 Concepts et notions 18
- 1.1 Notions linguistiques . . . 18
- 1.2 Apprentissage automatique . . . 24
- 1.3 Concepts liés au TAL . . . 28

2 Analyse syntaxique multilingue et délexicalisée 33
- 2.1 Introduction . . . 33
- 2.2 État de l'art . . . 34
- 2.3 Les différentes représentations de la langue . . . 36
- 2.4 Analyseur . . . 39
- 2.5 Cadre expérimental . . . 41
- 2.6 Résultats et analyse . . . 43
- 2.7 Comment le vecteur W est-il pris en compte par l'analyseur? . . . 48
- 2.8 Conclusions et perspectives . . . 50

3 Lexicalisation 52
- 3.1 Introduction . . . 52
- 3.2 MUSE . . . 54
- 3.3 PanLex . . . 55
- 3.4 Harmonisation . . . 58
- 3.5 Évaluations monolingues . . . 59
  - 3.5.1 Méthodes . . . 59
  - 3.5.2 Résultats . . . 62
- 3.6 Évaluations multilingues . . . 66
  - 3.6.1 Méthodes . . . 66
  - 3.6.2 Résultats . . . 70

3.7 Le problème de l'harmonisation . . . 74
3.8 Lexicalisation de l'analyseur syntaxique . . . 78
3.8.1 Lexicalisation . . . 83
3.9 Conclusions et perspectives . . . 87

4 Étiquetage en parties du discours **89**
4.1 Introduction . . . 89
4.2 État de l'art . . . 91
4.3 Étiqueteur . . . 92
4.4 Résultats et analyses . . . 96
4.4.1 Monolingue . . . 97
4.4.2 Multilingue . . . 98
4.4.3 Zero-shot . . . 103
4.4.4 Variabilité en zero-shot . . . 109
4.4.5 Familles de langues . . . 116
4.5 Conclusions . . . 121

5 Système de prédictions complètes **123**
5.1 Introduction . . . 123
5.2 État de l'art . . . 124
5.3 Tagparser . . . 126
5.4 Résultats et analyses . . . 131
5.4.1 Morphologie . . . 132
5.4.2 POS et traits morphologiques de références VS prédits . . . 135
5.4.3 WALS . . . 137
5.5 Conclusions et perspectives . . . 140

Bibliographie **146**

ANNEXES **155**
A Annexes chapitre 2 . . . 156
B Annexes chapitre 3 . . . 160
B.1 Ré-entraînement des analyseurs syntaxiques . . . 160
B.2 Modèle de caractères . . . 160
C Annexes chapitre 4 . . . 170
D Annexes chapitre 5 . . . 179
D.1 Comparaison avec l'état de l'art . . . 179

Liste des acronymes

AP

moyenne des précisions – ou *Average Precision* –. 68

BLI

induction de lexique bilingue – ou *Bilingual Lexicon Induction* –. 66, 67

CI

indice de connectivité – ou *Connectedness Index* –. 113–115

LPP

langue la plus proche. 111–113, 115, 116, 121

MAP

moyenne des précisions moyennes – ou *Mean Average Precision* –. 68, 70–73, 76, 77

MRR

rang réciproque moyen – ou *Mean Reciprocal Rank* –. 67, 68, 76

OOO

OddOneOut. 61, 62, 64–66

POS

partie du discours – ou *Part Of Speech* –. 12–17, 20–22, 24–30, 35, 36, 39, 42, 51, 52, 74, 77, 78, 80, 81, 87–94, 96–101, 103, 104, 108, 109, 122–126, 128–132, 134–144, 169, 173, 174, 179–181

RR

rang réciproque. 67

TAL

Traitement Automatique des Langues. 4, 12, 13, 18, 20, 28, 33, 123, 144

UD

Universal Dependencies. 14–16, 18, 22, 23, 33–35, 37, 40, 41, 50, 52, 89, 125, 131, 142

WALS

World Atlas of Language Structures. 4, 14–18, 23, 34–39, 42, 43, 45–48, 51, 82, 90–92, 95, 96, 100–103, 106, 107, 109, 113–115, 117, 122–124, 126, 130–134, 137, 139–145, 159

Introduction générale

La moitié des langues parlées aujourd'hui sont en danger de disparation, et on estime que 90% des langues risquent de disparaître au cours du siècle (MOSELEY 2010). Parmi les langues existantes (entre 3 000 et 8 000 selon la définition de ce qu'est une langue), toutes n'ont pas de système d'écriture. Le site de l'Ethnologue [1] recense seulement 4 065 langues écrites pour un total de 7 139 langues, basées sur un peu plus d'une trentaine de systèmes d'écriture. Sur l'ensemble des langues écrites, l'immense majorité ne dispose d'aucune annotation. Les annotations sont des informations associées à des entités (comme les mots ou les phrases), et peuvent par exemple renseigner la nature d'un mot, sa partie du discours – ou *Part Of Speech* – (POS), ou encore le sujet d'un verbe donné. Sur les 2 485 langues pour lesquelles JOSHI et al. 2020 ont répertorié des données, plus de 2 200 n'ont aucune donnée annotée. Cette absence d'annotations implique des limitations pour le développement d'outils informatiques pour un grand nombre de langues.

Le Traitement Automatique des Langues (TAL) est un champ d'étude permettant de créer des outils de traitement des langues pour diverses applications, comme la traduction automatique ou la conception de chatbots, tandis que l'apprentissage automatique est un ensemble de méthodes permettant de détecter des régularités dans les données. Les méthodes d'apprentissage automatique sont fréquemment utilisées dans les tâches de TAL. L'amélioration continue des systèmes de TAL est très liée à l'annotation de données, car les systèmes d'apprentissage automatique nécessitent des annotations [2] afin de fournir des exemples aux systèmes pour qu'ils puissent apprendre. De nombreuses langues se retrouvent pénalisées à cause de cela. Afin de pouvoir traiter plus de langues à l'aide de systèmes automatiques, il est nécessaire de trouver des solutions pour s'émanciper le plus possible du besoin de données annotées. Depuis des années, des chercheur·e·s s'intéressent à trouver des solutions possibles au problème des langues dites *peu dotées*, c'est-à-dire qui n'ont que peu d'annotations disponibles. La création de système de TAL pour toutes les langues pourrait permettre d'aider à réduire la division technologique associée aux applications du TAL pour éviter la situation d'un monde à deux vitesses, où les langues peu dotées se retrouvent mises de côté, faute d'utilisation dans les systèmes de TAL au quotidien.

AGIĆ et al. 2016 proposent par exemple d'utiliser des systèmes pour des langues bien dotées, et de les adapter à de nouvelles langues pour lesquelles on ne dispose

1. www.ethnologue.com
2. les systèmes d'apprentissage automatique *supervisé*, qui sont les systèmes les plus utilisés en TAL.

pas d'annotations, à l'aide de textes traduits dans la langue cible et la langue source. DUONG et al. 2015 montrent qu'en disposant d'un tout petit corpus annoté en syntaxe de la langue cible, il est possible de pré-apprendre un système d'apprentissage automatique sur une langue bien dotée puis de finir l'apprentissage sur le corpus de la langue cible. La plupart des initiatives se tournent vers des systèmes multilingues afin d'exploiter les données déjà à notre disposition pour en faire bénéficier des langues peu dotées. CHOMSKY et LASNIK 2008 ont émis l'hypothèse que toutes les langues contiennent des structures et des règles similaires, et qu'il existe une grammaire universelle innée à tous les êtres humains. Les systèmes multilingues pourraient donc être la clef permettant de développer des systèmes de TAL pour traiter des langues sans données.

Le but que nous cherchons à atteindre dans cette thèse est de créer un système d'analyse linguistique du texte sur les données annotées déjà existantes des différentes langues à notre disposition. Ce système serait capable de généraliser les connaissances apprises à de nouvelles langues. Il serait alors possible d'annoter un texte de n'importe quelle langue grâce à cet unique système. Dans cette thèse, plusieurs tâches de prédiction seront réalisées :

— la tâche d'étiquetage des partie du discours – ou *Part Of Speech* – (POS), qui consiste à associer à chaque mot d'une phrase sa POS, qui est la nature du mot (le mot est-il un nom, un verbe, un déterminant, ...)

— la tâche d'analyse syntaxique, qui consiste à retrouver les dépendances entre les mots de la phrase : qui est le sujet, qui est l'objet, ...

— la tâche prédisant les POS, les traits morphologiques et l'arbre syntaxique de façon jointe. Cette tâche permet de prédire les POS et l'analyse syntaxique d'un seul coup, et prédit également les traits morphologiques. Les traits morphologiques peuvent correspondre, par exemple, aux informations de genre, de nombre d'un mot, ou bien encore sa conjugaison.

Les POS, les traits morphologiques et l'arbre syntaxique d'une phrase sont souvent utilisés dans d'autres systèmes de TAL, comme l'extraction d'informations ou l'analyse sémantique. Ce sont des informations de base qui permettent de faire abstraction de la variabilité du lexique, permettant d'obtenir des représentations plus proches du sens des énoncés. Un tel système semble cependant difficile à obtenir. La généralisation des connaissances apprises à une nouvelle langue est confrontée à de nombreuses difficultés, entre autres :

— la cohérence des annotations d'une langue à l'autre. Les initiatives se chargeant de créer de nouveaux corpus ont également créé leurs propres guides d'annotation, très différents les uns des autres, résultant en des corpus annotés de façon très diverse.

— les différences de systèmes d'écriture (alphabet latin, cyrillique, ...)

— la différence des lexiques entre les langues (le mot *poulpe* en français s'écrit *octopus* en anglais)

— les différences entre langues, que ce soit au niveau phonologique, morphologique, syntaxique, ...

Pour tendre vers un tel système, un certain de nombre de ressources pourront être utiles. Pour commencer, il est primordial de disposer de schémas d'annotation qui soient communs à toutes les langues. Il existe aujourd'hui des initiatives comme les *Universal Dependencies* (UD) (NIVRE, MARNEFFE et al. 2016) ayant pour but de créer des corpus annotés de façon consistante entre les langues. UD propose un schéma d'annotation, un guide d'annotation, des outils, une communauté, et des corpus annotés. L'idée n'est pas d'obtenir une annotation unique pour toutes les langues, mais que les phénomènes identiques à travers les langues soient représentés de manière similaire. Le projet reste suffisamment ouvert pour que les phénomènes propres à chaque langue puissent être annotés, sans les forcer à rentrer dans une case trop stricte, sauf quand cela a déjà été annoté dans d'autres langues. Ces annotations universelles comprennent, entre autres, des annotations des POS, des traits morphologiques et des dépendances syntaxiques.

L'hétérogénéité des annotations n'est cependant pas le seul défi, car les langues utilisent des lexiques différents, c'est-à-dire que les mots n'ont pas les mêmes formes, même quand ils font référence à des concepts ou entités identiques. Les annotations proposées par UD ne permettent cependant pas de représenter le lexique de manière universelle. L'idée d'un lexique universel serait de pouvoir recenser les mots de toutes les langues, et les représenter d'une manière commune. Les plongements de mots permettent de façon générale de représenter le lexique dans des systèmes monolingues. Les plongements de mots sont une manière de représenter les mots sous forme de vecteur. Deux vecteurs proches indiquent que les deux mots qui y sont associés ont un sens proche. Or, il a été montré que les espaces vectoriels des plongements de mots ont des structures similaires d'une langue à l'autre (MIKOLOV et al. 2013), permettant d'aligner les espaces vectoriels des différentes langues dans un espace commun. Une représentation universelle des mots pourrait ainsi être obtenue, pouvant être utile pour des systèmes multilingues.

Les langues ne sont pas isolées, et elles partagent des caractéristiques communes entre elles, souvent issues d'une phylogénie liée à leur évolution. Les points communs et les différences entre les langues ont largement été étudiés et des ressources regroupent certaines de ces informations. Une question centrale des travaux que nous présentons est la suivante : est-il possible de tirer profit des points communs entre les langues pour aider le partage de connaissances dans des systèmes multilingues ? Des ressources comme le WALS (DRYER et HASPELMATH 2013) ou Glottolog (HAMMARSTRÖM et al. 2021) peuvent être d'excellentes sources d'information pour obtenir des connaissances sur la typologie de plus de 2 500 langues. Ces ressources sont des ensembles de caractéristiques (par exemple concernant l'ordre des mots) renseignées pour chaque langue décrite. Autrement dit, chaque langue est représentée par un ensemble de valeurs pour chacune de ces caractéristiques. Des langues partageant beaucoup de valeurs pourront être considérées proches, et vice-versa. Un exemple d'information contenue dans le WALS sera l'ordre du sujet, du verbe et de l'objet d'une langue. Le français, par exemple, est majoritairement SVO (Sujet-Verbe-

Objet), alors que le japonais est SOV (Sujet-Objet-Verbe). Les ressources comme le WALS permettent ainsi de décrire une langue, d'obtenir une représentation de cette langue, qui, bien qu'imparfaite puisque partielle, permet tout de même d'expliciter comment les langues peuvent être reliées les unes aux autres.

Les travaux proposés ici mettent en lien l'utilisation des UD, du WALS et de plongements de mots multilingues pour une quarantaine de langues afin de se rapprocher de la création d'un système *universel* pouvant réaliser des prédictions pour toutes les langues. Pour cela, nous entraînerons des modèles massivement multilingues, ainsi que des modèles de zero-shot. Un système *zero-shot* est capable de faire des prédictions d'annotations pour des données très différentes de celles qu'il a vu à l'entraînement, par exemple, les données d'une nouvelle langue. Nous explorerons tout au long de cette thèse la question suivante : l'utilisation de ces ressources permet-elle de tendre vers un meilleur partage de connaissances entre les langues ? Dans un cadre d'analyse syntaxique, d'étiquetage de POS mais aussi pour une chaîne de traitement réalisant ces prédictions et celle des traits morphologiques, nous verrons comment le WALS peut aider à rapprocher ou éloigner certaines langues, en particulier dans un cadre de zero-shot.

Un papier ayant eu un rôle important dans la direction prise par cette thèse est celui de AMMAR, MULCAIRE, BALLESTEROS et al. 2016. Les auteurs de cet article se sont déjà intéressés à la création de systèmes multilingues basés sur les UD et utilisant le WALS pour de l'analyse syntaxique. Leurs modèles étaient appris sur un ensemble de 7 langues indo-européennes, contrairement aux 38 [3] langues de nos systèmes, qui sont diversifiées, et issues de familles de langues très différentes. Les auteurs n'ont pas constaté de gain lié à l'utilisation du WALS dans leurs travaux. Cependant, le petit nombre de langues et la faible diversité des valeurs du WALS pour les langues choisies ont probablement limité la pleine exploitation de cette ressource. La campagne d'évaluation [4] CoNLL 2017 (ZEMAN, POPEL et al. 2017) a également largement impacté l'orientation de nos travaux. Dans le cadre de cette campagne d'évaluation, de nombreux systèmes de bout-en-bout ont été mis en place pour prédire l'analyse syntaxique d'une phrase à partir de texte brut. C'est dans cet esprit que cette thèse a commencé, avec comme objectif de créer des systèmes massivement multilingues permettant de prédire l'analyse syntaxique, ainsi que les tâches intermédiaires d'étiquetage des POS et des traits morphologiques. Les corpus utilisés lors de cette campagne d'évaluation sont ceux que nous utilisons tout au long de nos travaux. Une deuxième édition de cette campagne d'évaluation a eu lieu en 2018 (ZEMAN, HAJIČ et al. 2018) avec plus de langues.

Bien que nous ne traitons que peu de langues réellement peu dotées dans ces travaux, nos questions et hypothèses sont surtout tournées autour du zero-shot. Nous

3. 37 langues ont été utilisées pour les apprentissages de l'analyse syntaxique délexicalisée. Pour les chapitres suivants, une langue (le norvégien) a été séparée en deux langues : le bokmål et le nynorsk, résultant en l'apprentissage de 38 langues dans la suite.

4. Une campagne d'évaluation consiste à soumettre un problème à la communauté scientifique. Les chercheur·e·s proposent des solutions possibles, et le potentiel de chaque méthode peut être ainsi comparé et évalué.

simulerons des conditions de langues sans données annotées, et nous verrons ce que le WALS et les plongements de mots multilingues peuvent apporter dans une telle situation. Nous verrons cependant que les méthodes de zero-shot existent cependant à différents degrés. Bien que dans nos travaux, nous n'utilisons pas de données annotées, les systèmes proposés requièrent tout de même un minimum de données afin de les développer. En particulier, pour nos systèmes, de grands corpus de textes non annotés sont nécessaires pour entraîner les plongements de mots, ainsi que des (petits) dictionnaires bilingues pour les aligner dans un espace commun. L'annotation des traits du WALS est également nécessaire, cette annotation demande cependant peu d'effort, puisque la plupart des locuteurs de la langue peuvent aider à rapidement combler les informations manquantes.

Une promesse de l'apprentissage automatique est qu'il devrait être possible de concaténer les annotations issues de toutes les langues, et tout devrait bien se passer. Nous verrons qu'en pratique, c'est rarement le cas. Une grande partie de cette thèse consistera à se placer dans des conditions contrôlées afin d'identifier ce qui ne fonctionne pas.

Les 5 contributions principales de cette thèse consistent en des analyses de :

— l'utilité des représentations de la langue, en particulier du WALS, pour aider au partage de connaissances entre langues en multilingue et en zero-shot

— la lexicalisation, à l'aide de plongements de mots multilingues, de systèmes monolingues, multilingues et zero-shot

— l'impact de la présence d'une langue proche de la langue cible dans l'ensemble d'entraînement dans un cadre d'apprentissage zero-shot

— la restriction de l'apprentissage à des familles de langues

— l'impact de l'utilisation des caractères des mots

Ces contributions ont toutes lieu dans le cadre de l'analyse automatique de textes comprenant la prédiction des POS, des traits morphologiques ou des dépendances syntaxiques.

Pour identifier quelles peuvent être les sources de partage de connaissance entre les langues, nous commencerons par explorer l'utilisation des UD et de représentations de la langue issues de ressources telles que le WALS dans un cadre d'analyse syntaxique délexicalisée dans le chapitre 2. Ce chapitre sera notre point de départ pour aborder les problématiques liées au multilinguisme, et les questions sur la manière d'utiliser le WALS. Nous montrerons que le vecteur du WALS classiquement utilisé dans la littérature (AMMAR, MULCAIRE, BALLESTEROS et al. 2016; TÄCKSTRÖM et al. 2013; ZHANG et BARZILAY 2015) issu de NASEEM et al. 2012 peut ne pas être suffisant pour tirer pleinement profit du WALS, c'est la raison pour laquelle nous proposerons une version étendue de ce vecteur.

Afin de lexicaliser les modèles, c'est-à-dire, donner accès aux mots à nos systèmes, nous utiliserons des plongements de mots multilingues issus de FastText, que nous alignerons dans un espace vectoriel commun afin de tendre vers un lexique "universel". Le chapitre 3 décrit et évalue ces plongements de mots, de manière intrinsèque et extrinsèque. Il est important de noter qu'au moment de la réalisation de ces travaux, les

plongements de mots contextuels comme BERT (DEVLIN et al. 2019) n'étaient pas encore la norme. C'est pourquoi les travaux de cette thèse se basent sur des plongements de mots non-contextuels.

L'utilisation de plongements de mots multilingues permettront de nous intéresser au problème de prédiction des POS dans la phrase. Dans le chapitre 4, nous constaterons qu'une fois encore, le WALS peut se montrer utile pour les langues isolées. L'impact de l'utilisation des caractères des mots sera également testé, et nous verrons que bien que les caractères puissent être utiles pour certaines langues, ils peuvent également être très néfastes pour d'autres, en particulier dans un cadre de zero-shot. Nous verrons cependant que le plus important en zero-shot est la présence d'une langue proche dans l'ensemble d'entraînement. La question de l'apprentissage sur des sous-ensembles de langues sera également abordée, et nous constaterons là encore qu'une langue proche est l'élément déterminant pour l'obtention de bonnes prédictions en zero-shot.

Enfin, dans le chapitre 5, nous ferons les prédictions complètes des POS, des traits morphologiques, et de l'analyse syntaxique de la phrase dans une tâche jointe à partir de texte tokenisé. Nous verrons là encore que le WALS peut se montrer utile, en particulier en zero-shot, même si cela dépend grandement des langues.

L'état de l'art de cette thèse sera présenté au sein même des chapitres de contributions afin de situer nos travaux par rapport à chaque tâche abordée. Nous avons en effet considéré qu'il était plus intéressant d'expliquer le contexte dans lequel se situe nos travaux au moment où cela serait nécessaire afin que ces éléments soient gardés à l'esprit au moment de la lecture de chaque chapitre. Les notions nécessaires à la compréhension de cette thèse seront définies dans le chapitre 1.

Chapitre 1.

Concepts et notions

Nous allons à présent proposer des définitions aux concepts utilisés dans cette thèse. Des définitions propres à la linguistique, aux définitions liées à l'apprentissage automatique en passant par les concepts liés au TAL, nous tenterons d'expliquer de la manière la plus accessible possible les notions nécessaires à la compréhension de cette thèse. Deux ressources jouent un rôle fondamental dans cette thèse et vont être, entre autres, présentées ici : le WALS et les UD.

1.1. Notions linguistiques

Le domaine du TAL est intrinsèquement lié à la linguistique. Cette thèse, bien que traitant d'informatique, manipule de nombreuses notions linguistiques que nous allons brièvement présenter dans cette section.

Token

Dans nos travaux, nous nous intéresserons uniquement aux textes écrits, et pas à l'oral. Les *tokens* vont nous permettre de représenter les mots, c'est un outil pratique : un token est une chaîne de caractères. Le *mot* correspond au concept, et le *token* est la représentation d'une occurrence d'un mot.

La plupart des tokens représentent des mots de la langue. Tous les mots peuvent être représentés à l'aide d'un token, mais tous les tokens ne sont pas des mots. Chaque virgule, point ou parenthèse est également un token. Autrement dit, la ponctuation est également une unité à part entière de la phrase. Il arrive également qu'un mot compte comme deux tokens ou plus. Par exemple, dans la phrase *Le poulpe mange du crabe*, le mot *du* est la contraction des tokens *de le*.

Ainsi, la *tokenisation* est le processus permettant de transformer un texte brut en une suite de tokens bien séparés. La tokenisation consiste donc à segmenter le texte en phrases et en mots.

Morphologie

La morphologie est l'étude de la forme des mots. De nombreux traits morphologiques peuvent décrire un mot, comme le genre, qui se décline au masculin, féminin

Les	poulpes	mangeront	une	huître
Nombre=Plur Genre=Masc Défini=Def	Nombre=Plur Genre=Masc	Nombre=Plur Mode=Indicatif Personne=3 Temps=Futur	Nombre=Sing Genre=Fem Défini=Indef	Nombre=Sing Genre=Fem

TABLEAU 1.1. – Exemples de marqueurs morphologiques pour la phrase *Les poulpes mangeront une huître.*

ou neutre. Le temps, qui peut être au présent, au passé, etc., le mode, comme l'indicatif, le conditionnel et bien d'autres. Le paradigme d'un mot correspond aux différentes façons d'écrire un même mot en fonction de ses marqueurs morphologiques.

La morphologie n'a pas la même importance selon les langues. Dans les langues flexionnelles comme le français, où les mots changent de forme selon leur rapport grammatical aux autres mots, la morphologie est plus ou moins marquée. Au contraire, les marqueurs morphologiques sont très peu nombreux dans les langues isolantes comme le chinois, pour lesquelles chaque mot a une unique écriture ne changeant jamais. La morphologie nous permettra d'aider à retrouver la syntaxe d'une phrase, et sera également prédite dans la fin de nos travaux.

Forme de surface et lemme

La forme de surface du mot est sa forme telle qu'on la trouve dans le texte. Cependant, lorsque l'on cherche dans un dictionnaire par exemple, c'est une forme canonique du mot qu'on cherche. Le choix de cette forme simplifiée est assez arbitraire. Pour la phrase *Les poulpes mangent des crabes*, on cherchera le mot *manger*, qui est l'infinitif du verbe *mangent*. Cette forme simplifiée du mot s'appelle le lemme. Le plus souvent, cela correspond au mot une fois qu'on lui retire tous les marqueurs morphologiques.

Homographe

Des mots homographes sont des mots s'écrivant de façon identique. Les homographes peuvent se prononcer de la même façon comme dans la phrase *Un **avocat** mange un **avocat***. Mais il est possible d'avoir des homographes ayant des prononciations différentes : *Le poulpe **est** né à l'**est** de l'océan*. Ces homographes ne sont alors pas des *homophones*.

Il est possible d'avoir des homographes entre langues, ce qui sera important dans la suite de cette thèse. Parfois, les homographes auront le même sens, comme le mot *excellent* en français et en anglais, mais peuvent également n'avoir aucun lien, comme le mot *pain*, en français, qui désigne un excellent aliment de base de nombreuses cuisines dans le monde, alors qu'en anglais, le mot *pain* signifie *douleur*, soit un sens

presque opposé au sens français.

Certains homographes entre les langues peuvent avoir des rôles syntaxiques très différents, par exemple le mot *a* en français est un verbe conjugué, le verbe *avoir*, alors qu'en anglais, *a* est un déterminant.

Diacritiques

Un diacritique est un symbole ajouté à une lettre ou un caractère d'un alphabet, afin d'en changer le sens ou la prononciation. Ce sont les accents, le tréma, la cédille, etc... Chaque alphabet dispose de ses propres diacritiques. Les diacritiques seront évoqués à quelques reprises dans ces travaux, par exemple pour expliquer certains phénomènes ayant lieu lors du traitement du vietnamien, qui utilise l'alphabet latin comme de nombreuses autres langues sur lesquelles nous travaillons. Mais pour le vietnamien, des phénomènes proches de langues ayant un alphabet unique ont lieu lors des expériences. Une explication possible est que le vietnamien utilise de nombreux diacritiques, présents uniquement dans cette langue.

Lexique

Le lexique d'une langue correspond à l'ensemble des mots qui la compose. Le lexique n'est pas figé et évolue au fil du temps. De nouveaux mots apparaissent, par exemple *Covid-19*, alors que d'autres disparaissent, tombés en désuétude. Connaissiez-vous le verbe *Paperasser* ? Il est fort probable que la réponse soit négative, ce mot ayant tendance à disparaître. La taille du lexique d'une langue varie ainsi en fonction du temps.

En TAL nous n'avons accès qu'à une partie du lexique seulement. Les mots inconnus lors d'une tâche peuvent cependant être représentés à l'aide d'un token. Par exemple dans la phrase *Le poulpe nage près du croin*, même si nous ne connaissons pas le mot *croin*, il nous a été possible de le représenter avec un token.

Partie du discours

Les mots peuvent être regroupés en fonction de leur proximité au niveau de leur comportement syntaxique. Les partie du discours – ou *Part Of Speech* – (POS) correspondent à la catégorie à laquelle appartiennent les mots. Par exemple, les verbes, les noms, ou encore les adjectifs. Souvent, les mots appartenant à une même catégorie peuvent se substituer à un autre mot de la même nature. Par exemple, dans la phrase *Le poulpe mange le crabe*, crabe est un nom, et peut donc être remplacé par un autre nom : *Le poulpe mange le poisson*, ou encore *Le poulpe mange le train*. La phrase peut ne plus avoir de sens, mais elle reste grammaticalement correcte. Cependant, on ne peut généralement pas remplacer un nom par un adverbe. La phrase *Le poulpe mange le hier* n'est plus syntaxiquement acceptable, le mot *hier* n'étant pas un nom.

Les POS sont essentielles à ce travail, la tâche de prédiction des POS étant centrale. Les POS servent également d'entrées aux analyseurs syntaxiques.

Syntaxe

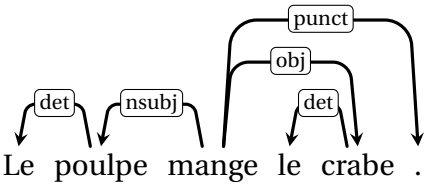
La syntaxe régit l'ordre des mots dans la phrase. Autrement dit, c'est le respect des règles de construction d'une phrase dans une langue donnée. On s'intéressera alors à plusieurs phénomènes relatifs à la syntaxe, comme l'importance de l'ordre des mots (*Le poulpe mange le crabe* ne veut pas dire la même chose que *Le crabe mange le poulpe*) ou encore aux fonctions grammaticales des mots (déterminant, sujet, objet, ...)

Arbre syntaxique

Classiquement, la structure syntaxique des phrases est représentée sous forme d'arbre. Deux types d'arbres sont couramment utilisés : les structures syntagmatiques, et les structures de dépendances. Dans ces travaux, nous ne nous intéresserons qu'aux structures de dépendances. Ces structures consistent à définir les relations de dépendances entre deux mots, et créer un lien entre ceux-ci. Les nœuds de l'arbre sont des mots, et les dépendances ont une étiquette.

Chaque dépendance est constituée d'une *tête* et d'un *dépendant*. Chaque mot de la phrase est le dépendant d'exactement une dépendance, mais les mots peuvent être la tête de plusieurs dépendances. Un seul mot n'est dépendant d'aucune dépendance : c'est la racine de la phrase.

Les étiquettes des dépendances correspondent à la relation syntaxique existant entre la tête et le dépendant, pouvant être une relation de *sujet* ou d'*objet* par exemple.



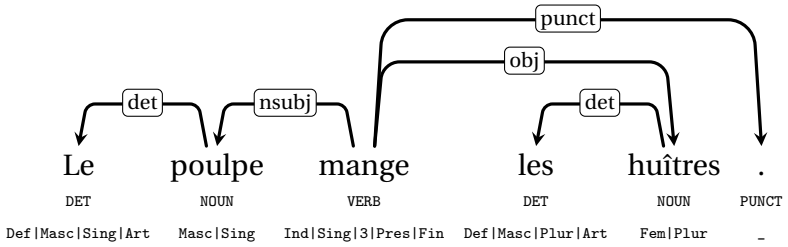
Dans cet exemple, *mange* est la racine de la phrase, *poulpe* est le sujet, et *crabe* est l'objet. Les déterminants sont reliés au nom qu'ils déterminent.

Annotations

Les annotations sont des informations associées à des entités (mots, phrases ou corpus) permettant de relier chaque entité à une information la concernant. Les annotations peuvent concerner les POS, les lemmes, les traits morphologiques ou encore l'arbre syntaxique de la phrase.

Des exemples d'annotations pour la phrase *Le poulpe mange les huîtres.* sont disponibles ci-dessus. On peut observer les annotations de la syntaxe en dépendance en haut, les traits morphologiques tout en bas, et les POS juste au-dessus.

Cependant, cette représentation n'est pas pratique. Dans un corpus, cette phrase serait représentée comme dans l'exemple ci-dessous. Ce format, où les annotations



sont rangées par colonnes [1] et chaque ligne représente un mot de la phrase, est appelé le format CoNLL.

ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
1	Le	le	DET	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	2	det
2	poulpe	poulpe	NOUN	Gender=Masc|Number=Sing	3	nsubj
3	mange	manger	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root
4	les	le	DET	Definite=Def|Gender=Masc|Number=Plur|PronType=Art	5	det
5	huîtres	huître	NOUN	Gender=Fem|Number=Plur	3	obj
6	.	.	PUNCT	_	3	punct

Lorsque les annotations sont créées à la main par des annotateurs humains, on parle d'annotations de référence. Le terme anglais *gold* est souvent employé pour qualifier ces annotations. Ces annotations permettent de fournir des exemples aux systèmes d'apprentissage. Elles peuvent aussi servir de base de comparaison pour évaluer des modèles afin de vérifier si leurs prédictions sont identiques à celles de références, afin d'évaluer les systèmes.

Corpus

Un corpus est une collection de textes ayant une certaine unité. Un corpus est souvent la base du travail des chercheur·e·s, que ce soit pour extraire des statistiques sur les données, ou pour entraîner des systèmes d'apprentissage automatique. Les corpus peuvent être accompagnés d'annotations, par exemple des annotations des POS, des traits morphologiques, ou de l'arbre syntaxique de la phrase. Lorsqu'ils ne le sont pas, on parle de corpus bruts. Un *corpus arboré* est un corpus annoté à l'aide d'arbres syntaxiques.

Universal Dependencies

La cohérence de l'annotation des données entre les langues est un problème majeur pour les tâches d'analyse multilingue, la plupart des corpus étant annotés en utilisant différents guides d'annotation et jeux d'étiquettes. L'initiative des *Universal Dependencies* (UD) (NIVRE, MARNEFFE et al. 2016) a pour but de créer des corpus arborés consistants entre les langues, facilitant ainsi les analyses lors de travaux sur la langue.Des annotations universelles sont donc fournies pour les POS, les traits

1. En général, identifiant du mot dans la phrase, puis forme de surface, lemme, POS, traits morphologiques, tête de la dépendance du mot courant, et étiquette de cette dépendance

Français	Le poulpe	mange	le crabe
Anglais	The octopus	eats	the crab
Vietnamien	Con bạch	tuộc ăn	con cua
Ordre	S	V	O

Japonais	Tako wa	kani	o tabemasu
Basque	Olagarroak	karramarroa	jaten du
Traduction	*Le poulpe*	*le crabe*	*mange*
Ordre	S	O	V

morphologiques et les dépendances syntaxiques. Ces annotations sont une vraie aide aux travaux de recherches en lien avec les problématiques du multilingue.

Dans ce travail, nous utilisons la version 2.0 de UD pour nos corpus d'entraînement de développement [2], et les corpus de tests de la campagne d'évaluation CoNLL 2017 [3]. 64 corpus arborés de UD en 45 langues sont disponibles pour l'entraînement et le développement [4]. Cependant, ces corpus sont de taille très variable (de 529 mots pour le kazakh à 1 1842 867 mots pour le tchèque).

Les corpus de tests contiennent au minimum 10 000 mots par langue et sont disponibles pour 49 langues. 4 langues n'ont pas de corpus d'entraînement correspondant.

Traits typologiques

La typologie linguistique permet de classifier les langues selon des caractéristiques structurelles communes. De nombreux traits typologiques, aussi appelés caractéristiques typologiques, existent pour rapprocher des langues, un des plus connus étant celui de l'ordre des mots du sujet (S), du verbe (V) et de l'objet (O) dans une phrase. Le français est une langue de type SVO, tout comme le vietnamien, alors que le basque ou le japonais respecte l'ordre SOV.

Il existe plusieurs bases de données recensant les traits typologiques des langues, comme *Glottolog* (HAMMARSTRÖM et al. 2021) ou le *World Atlas of Language Structures* (WALS) (DRYER et HASPELMATH 2013).

Le WALS est une base de données de propriétés structurelles (phonologiques, grammaticales et lexicales) réunie par 55 auteurs à partir de matériel descriptif par exemple des grammaires de référence. Cette base de données nous a permis d'associer à chaque langue des corpus de UD un ensemble de traits décrivant des propriétés pertinentes pour l'analyse syntaxique. Le WALS décrit 2 676 langues grâce à un ensemble de 192 traits, répartis entre 11 familles :

— Phonologie : 20 traits (i.e. 10A Vowel Nasalization)

2. http ://hdl.handle.net/11234/1-1983
3. http ://hdl.handle.net/11234/1-2184
4. Ces corpus ont été fixés au début de la thèse. La version 2.8 de UD est cependant disponible depuis le 21 mai 2021, avec 202 corpus arborés pour 114 langues.

— Morphologie : 12 traits (i.e. 26A Préfixage vs suffixage en morphologie flexionnelle)

— Catégories Nominales : 29 traits (i.e. 30A Nombre de genres)

— Syntaxe Nominale : 8 traits (i.e. 58B Nombre de noms possessifs)

— Catégories Verbales : 17 traits (i.e. 70A L'impératif morphologique)

— Ordre des mots : 56 traits (i.e. 81A Ordre du Sujet de l'Objet et du Verbe)

— Clauses Simples : 26 traits (i.e. 112A Morphèmes négatifs)

— Complex Sentences : 7 traits (i.e. 122A Relativisation sur des sujets)

— Lexique : 13 traits (i.e. 131A Bases numériques)

— Langages des signes : 2 traits, incluant les particules de question en langues des signes

— Autres : 2 traits, incluant le Système d'écriture

Bien que tous les traits ne soient pas renseignés pour toutes les langues, cette ressource sera notre source principale de traits typologiques pour les expériences qui suivront. La version utilisée est celle datant de 2014 [5], même si des mises à jour ont été faites depuis 2020. Notre utilisation de cette ressource sera abordée dans la section 2.3.

1.2. Apprentissage automatique

L'apprentissage automatique est un ensemble de méthodes permettant de détecter automatiquement des régularités dans des données. Ces régularités peuvent être utilisées pour mieux décrire ces données ou prédire certaines de leurs caractéristiques. De nombreuses prédictions peuvent être générées à l'aide de l'apprentissage automatique, comme les étiquettes de POS, les traits morphologiques, ou bien l'analyse syntaxique d'une phrase.

Il existe deux grands types de familles d'apprentissage : l'apprentissage supervisé, et l'apprentissage non-supervisé. En apprentissage supervisé, on dispose d'un couple (X,Y), où X est la donnée et Y son annotation. La tâche est d'entraîner le système à prédire correctement Y à partir de X. Pour cela, on donne X au système, et le système génère une prédiction. Si cette prédiction est différente de Y, les paramètres du système sont mis à jour pour se rapprocher de la solution. Pour les méthodes non-supervisés, des exemples X lui sont montrés, mais le système n'a pas accès à Y, la "bonne réponse". Il doit apprendre par lui-même à regrouper les exemples qui se ressemblent, et créer ainsi des *clusters* (des regroupements), des hiérarchies, ou juste identifier des similarités.

De nombreux outils sont utilisés pour faire de l'apprentissage automatique, et nous définirons dans cette section ceux exploités dans le cadre de cette thèse.

5. https://zenodo.org/record/3607439#.YPk-_3UzbmE

Modèle

Un *modèle* correspond au produit du processus d'apprentissage : c'est l'outil à qui l'on donne des données en entrée et qui effectue les prédictions de la tâche en sortie. Un modèle est une fonction qui associe des prédictions à des données d'entrées (par exemple, en associant une POS au mot donné). Les paramètres de cette fonction sont trouvés à l'aide d'algorithmes d'apprentissage qui utilisent des données. Un modèle est construit à partir de deux éléments essentiels : l'algorithme d'apprentissage, et les exemples fournis en entrée. Le changement de l'un ou l'autre de ces éléments générera un nouveau modèle.

Réseaux de neurones

Les réseaux de neurones sont des modèles. Deux types de réseaux de neurones nous intéressent pour cette thèse : les *perceptrons multicouches* (MLP) et les bi-LSTMs.

Les **MLP** sont parmi les réseaux de neurones les plus simples. Ils consistent en un empilement de couches linéaires. Une couche linéaire est un produit matriciel entre un vecteur d'entrée et une matrice de poids (qui paramètre cette couche), résultant en un vecteur de sortie. Ces poids seront trouvés lors de l'apprentissage. Un MLP est une alternance de couches linéaires et de fonctions non linéaires. Les MLP peuvent servir à la classification, par exemple si on utilise la représentation d'un mot ou d'un contexte comme vecteur d'entrée, alors le MLP produit en vecteur de sortie une distribution de probabilités sur les POS possibles pour ce mot.

Les réseaux de neurones nous permettront de réaliser des tâches de classification, par exemple la prédiction de la POS d'un mot. Si on donne le mot *poulpe* au réseau, on souhaite qu'il apprenne à classifier ce mot dans la catégorie `NOM`. Un autre cas intéressant est la prédiction de la POS des mots dans le contexte d'une phrase, qui est notre cas d'utilisation principale.

Afin de s'adapter au mieux à ce type de tâche, les réseaux de neurones peuvent être déclinés en des versions plus élaborées. Les réseaux de neurones **récurrents** par exemple permettent de garder l'historique des prédictions et de mieux encoder la séquence d'entrée qu'une simple concaténation de ces informations.

Le	poulpe	mange	le	crabe	.
`DET`	`NOM`	`VERB`	`DET`	`NOM`	`PONCT`

Imaginons que dans l'exemple ci-dessus, le réseau a prédit la POS `DET` pour le mot "Le". Les déterminants sont très souvent suivis de `NOM`. Ce contexte peut donc être utile au système. On utilisera des réseaux récurrents (LSTM, *Long Short-Term Memory* (HOCHREITER et SCHMIDHUBER 1997)) pour encoder les séquences de mots et de POS qui forment le contexte du mot courant.

Il peut également être utile de voir dans le "futur", autrement dit, de voir les mots qui viennent plus loin dans la phrase. Pour cela, on peut utiliser des réseaux de neurones

récurrents **bidirectionnels**, qui consistent en un réseau de neurones récurrent lisant la phrase dans le sens de lecture, et un deuxième item lisant dans le sens inverse au sens de lecture, permettant ainsi d'avoir un aperçu du futur de la séquence. Par exemple, lorsqu'on cherche à prédire la POS du mot *le*, si ce mot est suivi de *crabe*, qu'on sait être souvent précédé d'un `DET`, l'information de la présence du mot *crabe* pourra aider la prédiction. Nous utiliserons des bi-LSTMs dans ces travaux.

Représentations

Il existe plusieurs méthodes pour représenter des données en entrée d'un réseau de neurones.

One-hot Un one-hot est une représentation très simple : s'il y a N valeurs possibles à représenter, le one-hot sera un vecteur de taille N. Toutes les valeurs de ce vecteur sont à 0, sauf une qui est à 1. Imaginons par exemple que nous avons une liste de 4 noms d'animaux :

— poulpe

— chat

— crabe

— licorne

Le vecteur [0,0,0,1] représentera le mot *poulpe*, [0,0,1,0] le mot *chat*, [0,1,0,0] le mot *crabe* et [1,0,0,0] le mot *licorne*. Cette représentation peut être intéressante lorsqu'il y a peu de données à représenter, et qu'elles sont à distance égale les une des autres.

Plongements de mots Bien que les one-hot soient très pratiques et simples d'utilisation, ils ont le défaut de devenir très grands dans le cas où les valeurs à représenter sont nombreuses, par exemple pour un lexique, qui peut facilement contenir un million de valeurs différentes. De plus, les mots *poulpe*, *pieuvre* et *caillou* seront à égales distances alors que les deux premiers mots sont plus proches entre eux que du troisième. Les *plongements de mots* sont une façon de représenter les données sous forme vectorielle.

L'hypothèse de départ est que les mots qui apparaissent dans un même contexte ont le même sens.

— "Le *XXX* vit sous la mer."

— "Je vais commander la salade de *XXX*."

— "Les *XXX* se délectent de crabes."

— "Avec ses huit tentacules, le *XXX* est capable d'attraper des objets."

Vous avez beau n'avoir jamais vu le mot *XXX*, vous pouvez en déduire son sens, qui semble être proche de celui du mot *poulpe*. C'est ce qu'on appelle l'hypothèse distributionnelle. FIRTH 1957 résume bien cette idée dans sa célèbre citation *You shall*

know a word by the company it keeps, pouvant être traduite de façon non-littéral par *Dis moi qui le mot fréquente et je te dirais qui il est*.

Il est ainsi possible d'entraîner des plongements de mots à l'aide de grands corpus de textes bruts. Les vecteurs obtenus pour chaque mot permettront ainsi d'avoir une représentation proche dans l'espace pour des mots aux sens proches.

Depuis quelques années, l'émergence de plongements de mots contextuels est constatée, et sont à leur apogée depuis la création de BERT, un modèle de *transformers* (DEVLIN et al. 2019). Les plongements de mots contextuels permettent en particulier de distinguer des homographes en fonction de leur contexte. Les plongements des mots *avocat* dans ces deux phrases seront alors très différents :

— L'avocat est au tribunal

— Il y a de l'avocat dans cette salade

Modèle de caractère

L'utilisation du terme *modèle de caractère* désigne l'apprentissage de représentations pour les caractères des mots, de la même manière que des représentations des mots ont été construites pour le lexique avec les plongements de mots. Pour nous, les représentations des caractères des mots seront construites uniquement à partir des préfixes et des suffixes des mots.

Apprentissage Zero-shot

Certaines situations ne permettent pas d'utiliser des données d'entraînement correspondant aux données de test qui seront utilisées pour réaliser les prédictions, par exemple dans le cas où une langue ne dispose pas de données d'entraînement. Les méthodes d'apprentissage sans exemples sont classiquement appelées des méthodes d'apprentissage *zero-shot* lorsqu'il n'existe aucun exemple des annotations à générer pour une instance. L'apprentissage *few-short* correspond à la situation où quelques exemples sont tout de même disponibles.

Nous nous intéresserons dans cette thèse à différents types d'apprentissage, en commençant par celui ou de nombreuses ressources sont disponibles, comme des annotations en POS ou les arbres syntaxiques, puis en réduisant les ressources utilisées pour arriver à une configuration de zero-shot. Cependant, le zero-shot permet tout de même d'exploiter un certain nombre de sources d'informations connexes qui donnent au système une chance de faire la tâche cible correctement. Par exemple, on peut entraîner un modèle sur une langue et l'appliquer sur une autre langue. On peut aussi exploiter des données non annotées ou des ressources linguistiques qui décrivent la langue cible.

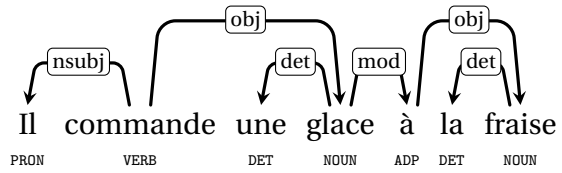
1.3. Concepts liés au TAL

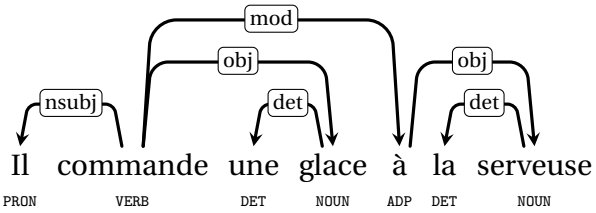
Nous allons ici définir les outils et ressources propres au Traitement Automatique des Langues (TAL) utilisés dans le cadre de cette thèse.

Analyse syntaxique (*Parsing*)

La tâche d'analyse syntaxique consiste à prédire la structure syntaxique d'une phrase. Pour cela, on utilise des analyseurs syntaxiques (des *parsers* en anglais). Ces derniers prennent en entrée une phrase et produisent une structure syntaxique pour celle-ci.

La prédiction de l'arbre syntaxique de la phrase est une tâche pouvant être plus ou moins difficile selon les langues. En effet, une langue avec des dépendances très éloignées [6] sera généralement plus dure à analyser. La difficulté réside également dans les ambiguïtés possibles d'une phrase.





Dans cet exemple, les deux phrases ont la même séquence de POS, et pourtant l'analyse syntaxique est différente. Dans la phrase *Il commande une glace à la fraise*, le mot *à* est rattaché à *glace*, car c'est la glace qui est à la fraise. Cependant, dans la phrase *Il commande une glace à la serveuse*, le mot *à* est rattaché à *commande*, car c'est à la serveuse que la commande a été passée.

Les analyseurs utilisés dans cette thèse sont des analyseurs en transition *arc-eager* (NIVRE 2004) reposant sur un MLP. Un analyseur est composé de trois structures :

— un buffer, à partir duquel les tokens sont lus en entrée

— une pile, contenant une liste de mots déjà rencontrés dont l'analyseur peut encore avoir besoin

— une liste de transitions précédemment réalisées par l'analyseur

L'état de ces trois structures à chaque instant représente la *configuration* de l'analyseur. Les transitions réalisables par un analyseur *arc-eager* sont :

6. autrement dit, lorsqu'il y a beaucoup de mots entre la tête et le dépendant de la dépendance

— `SHIFT` : place le mot courant au sommet de la pile

— `REDUCE` : supprime le mot en sommet de pile

— `LEFT` : crée une dépendance syntaxique entre le mot courant du buffer et le sommet de la pile, puis supprime le mot en sommet de pile

— `RIGHT` : crée une dépendance syntaxique entre le sommet de la pile et le mot courant du buffer, puis place le mot courant au sommet de la pile

Les analyseurs en transitions reposent sur un classifieur, comme un MLP, dont le rôle est de choisir la prochaine transition à réaliser en fonction de l'historique, du mot courant, des précédents, de l'état de la pile etc. Le système utilisé est *glouton*, c'est-à-dire que la transition choisie sera toujours celle ayant la probabilité maximale.

Étiquetage (*Tagging*)

Une tâche d'étiquetage consiste à assigner, à chaque unité [7] d'une séquence, une étiquette. L'étiquetage en POS assigne ainsi à chaque mot une étiquette correspondant à la partie du discours de ce mot. Il en va de même pour la morphologie.

Ces deux tâches sont considérées comme des tâches relativement simples, les résultats pour l'étiquetage en POS étant en moyenne supérieurs à 90% d'exactitude, et supérieurs à 85% pour la prédiction de la morphologie (SMITH et al. 2018).

Il existe bien évidemment des cas ambigus, certains mots pouvant par exemple être à la fois un nom ou un adjectif selon le contexte.

— *Les poulpes ne mangent pas d'oranges.*

— *Les poulpes sont parfois oranges.*

Dans cet exemple, le mot *oranges* est d'abord un nom, faisant référence aux fruits, puis dans la seconde phrase, c'est un adjectif qui se réfère à la couleur.

Les étiqueteurs utilisés dans cette thèse sont similaires aux analyseurs présentés ci-dessus. Un étiqueteur est composé de deux structures :

— un buffer, à partir duquel les tokens sont lus en entrée

— une liste de transitions précédemment réalisées par l'étiqueteur

Le nombre de transitions réalisables par les étiqueteurs est égal au nombre d'étiquettes [8] qu'il est possible de prédire. Chacune de ces transitions consiste alors à affecter cette étiquette au mot courant du buffer, puis de passer au mot suivant.

Le classifieur permettant de choisir la prochaine transition à réaliser est similaire à celui de l'analyseur. En fonction de l'historique, du mot courant, des précédents, etc., le système choisira la transition ayant la probabilité maximale.

7. Ces unités peuvent être les mots, les caractères, les syllabes,...

8. Selon la tâche en cours, ces étiquettes peuvent par exemple être les POS ou les traits morphologiques

Métriques

Les métriques évaluent la qualité des prédictions effectuées par un modèle, permettant de comparer des expériences. Nous allons définir les différentes métriques que nous utiliserons dans nos travaux.

Exactitude L'exactitude est le pourcentage d'étiquettes correctement prédites. Formellement, l'exactitude est :

$$\frac{nb\_étiquettes\_bien\_prédites * 100}{nb\_étiquettes\_totales}$$

Cette mesure nous permettra d'évaluer nos résultats d'étiquetage en POS, ainsi que l'étiquetage des traits morphologiques. Imaginons que nous ayons la phrase suivante :

	Le	poulpe	mange	le	crabe	.
Réf.	DET	NOM	VERB	DET	NOM	PONCT
Pred.	VERB	NOM	VERB	NOM	NOM	PONCT

Sur les 6 étiquettes, seulement 4 sont correctes. L'exactitude sera alors de $\frac{4*100}{6} = 66.67$. Autrement dit, 66.67% des étiquettes ont bien été prédites.

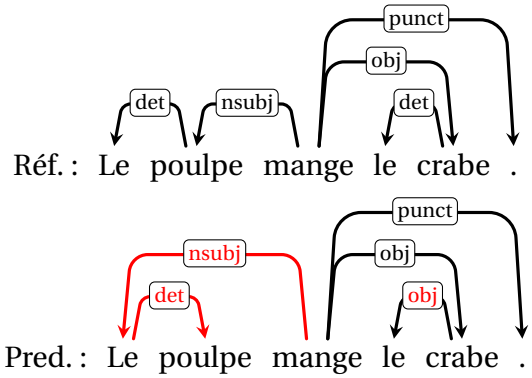
UAS, LAS Le score d'attachement - ou *Unlabeled Attachment Score* - (UAS) et le score d'attachement étiqueté - ou *Labeled Attachment Score* - (LAS) sont deux métriques permettant d'évaluer la tâche de prédiction de l'analyse syntaxique d'une phrase.

Étant donné une phrase et son analyse syntaxique, le UAS est la proportion de dépendances bien prédites (tête et dépendant corrects) dans la phrase. Le LAS est la proportion de dépendances bien prédites et bien étiquetées (tête, dépendant et étiquette corrects).

$$UAS = \frac{nb\_dépendances\_bien\_prédites * 100}{nb\_dépendances\_totales}$$

$$LAS = \frac{nb\_dépendances\_et\_étiquettes\_bien\_prédites * 100}{nb\_dépendances\_totales}$$

Le UAS n'évalue que si la structure de l'arbre syntaxique est correcte, le LAS permet également d'évaluer les étiquettes des dépendances. Un score UAS est toujours supérieur ou égal au LAS, puisque les erreurs du UAS sont également des erreurs pour le



LAS, mais les erreurs sur les prédictions des étiquettes affectent le LAS, mais pas le UAS.

Dans l'exemple ci-dessus, le UAS serait de $\frac{3}{5}$, puisque la dépendance *nsubj* et la dépendance *det* sont fausses. Le LAS quant à lui est de $\frac{2}{5}$, puisqu'en plus des deux dépendances mal prédites, l'étiquette de la dépendance entre *crabe* et *le* est également fausse.

Moyenne De nombreuses langues sont traitées dans cette thèse. Afin d'avoir une vue d'ensemble, de nos résultats sur toutes les langues, nous avons deux possibilités : soit calculer le score sur un corpus de test contenant les tests de toutes les langues, soit faire la moyenne des scores [9] de chaque corpus de test calculé individuellement. Ce deuxième choix consiste à calculer ce qu'on appelle la *macro moyenne*, elle a l'avantage d'être indépendante de la taille du corpus de test de chaque langue, et n'introduit pas de biais en faveur des langues sur-représentées dans l'ensemble de test.

Corrélation Le calcul de corrélation entre deux listes de valeurs permet d'établir une mesure de l'indépendance de ces deux listes. Il existe plusieurs moyens de calculer des corrélations, nous en verrons deux dans ces travaux. La première, la corrélation de Pearson, permet de mesurer l'existence d'une relation linéaire entre les deux listes de valeurs. La corrélation de Spearman mesure l'existence d'une relation monotone entre les listes de valeurs. Cette mesure se base sur les rangs pour calculer la corrélation.

Le calcul d'une corrélation est associé à une *p-value*. Cette p-value indique la probabilité que la corrélation soit due au hasard, une p-value de 0.6 indiquant qu'il y a 60% de chance que la corrélation soit l'effet du hasard. Cela est souvent dû à un trop faible nombre de données dans les listes de valeurs comparées. Nous considérerons qu'un score de corrélation est *significatif* si la p-value est inférieure à 0.05, c'est-à-dire s'il y a moins de 5% de chance que la corrélation soit due au hasard.

De nombreuses corrélations vont être étudiées dans ces travaux, pour vérifier la non-existence ou la possibilité d'une relation causale entre l'utilisation de certaines ressources ou méthodes avec les résultats obtenus. Une corrélation peut être le résultat

9. qui sont eux même des moyennes

d'une causalité. Par exemple, plus le nombre de brownies consommés à la fin du repas est grand, plus le risque de douleur au ventre est élevé. Ces évènements ne sont pas seulement corrélés, l'un est la cause de l'autre. Mais une corrélation ne signifie pas toujours qu'il y ait causalité. Un exemple célèbre est celui de la quantité de chocolat consommé par pays, qui est très corrélé au nombre de prix Nobel obtenus dans le dit pays. Cette corrélation ne signifie pas que la consommation de chocolat augmente la probabilité d'obtenir un prix Nobel. Il existe une troisième variable qui est dite *cachée* : la richesse d'un pays. Plus un pays est riche, plus les habitants peuvent consommer de chocolat, qui est une denrée chère. Mais la richesse d'un pays influe aussi les moyens investis dans la recherche, limitant ou favorisant la probabilité d'obtenir un prix Nobel dans le pays.

Une relation de causalité se traduira par une relation de corrélation, mais l'absence de corrélation implique l'absence de causalité. Cependant, les scores de Pearson et Spearman ne permettent pas de capturer tous les types de corrélations, seulement les corrélations linéaires et monotones. L'interprétation des résultats est à faire avec précaution. De plus, ces scores ne sont pas binaires : les résultats peuvent être compris entre -1 et 1, et plus les résultats sont proches de 0, moins il y a de corrélation. L'interprétation de ces scores n'est pas évidente. Un score de 0.89 ou de -0.96 indique une forte corrélation. Un score de 0.3 est plus difficilement interprétable.

Les notions nécessaires à la compréhension de cette thèse ont été définies. Nous allons à présent passer aux chapitres de contributions de nos travaux.

Chapitre 2.

Analyse syntaxique multilingue et délexicalisée

Sommaire

2.1 Introduction 33
2.2 État de l'art 34
2.3 Les différentes représentations de la langue 36
2.4 Analyseur 39
2.5 Cadre expérimental 41
2.6 Résultats et analyse 43
2.7 Comment le vecteur W est-il pris en compte par l'analyseur ? 48
2.8 Conclusions et perspectives 50

2.1. Introduction

Depuis qu'elle repose sur l'apprentissage automatique, l'analyse syntaxique en dépendances requiert, comme de nombreuses autres tâches en TAL, de grandes quantités de données pour apprendre la syntaxe de la langue. Certaines langues comme l'anglais bénéficient d'une grande quantité de telles données. En revanche, pour beaucoup de langues, il n'en existe actuellement pas. L'annotation de données étant une tâche coûtant extrêmement chère, certaines langues comme le hausa, une langue africaine parlée par plus 80 millions de locuteurs natifs (WIKIPEDIA 2021), sont des langues pour lesquelles aucun corpus annoté n'existe.

L'analyse syntaxique multilingue est une solution possible pour tenter de tirer profit des langues bien dotées et des caractéristiques communes à plusieurs langues. Deux problèmes se posent alors : comment représenter les données, et comment représenter la langue ?

Avec l'apparition des UD [1], l'utilisation d'annotations universelles pour la syntaxe permet en partie de répondre au premier problème. En effet, les UD définissent un ensemble de relations de dépendances, de parties du discours et de traits morphologiques qui se veulent communs à toutes les langues (NIVRE, MARNEFFE et al. 2016).

1. http://universaldependencies.org

Les corpus arborés de UD sont disponibles pour de nombreuses langues et sont basés sur un guide d'annotation commun.

Bien que les UD proposent un cadre unique pour l'annotation de différentes langues, ils ne définissent cependant pas de lexique commun. C'est la raison pour laquelle l'utilisation d'un analyseur délexicalisé (ZEMAN et RESNIK 2008) a été fait pour ce travail. Cette technique consiste à ignorer le lexique lors de l'entraînement de l'analyseur. Cet appauvrissement des données conduit à des analyseurs moins précis, mais offre une solution simple au problème du lexique. Une façon d'éviter la délexicalisation est l'utilisation de plongements de mots multilingues (AMMAR, MULCAIRE, TSVETKOV et al. 2016). Cette alternative sera étudiée dans le chapitre 5.

L'hypothèse que nous souhaitons tester est celle de l'utilité d'une représentation de la langue pour la prédiction de l'arbre syntaxique de la phrase. L'idée est de donner à l'analyseur une phrase dans une langue quelconque, plus une représentation de la langue afin de lui permettre d'apprendre les similarités et les différences entre plusieurs langues.

La question de la représentation de la langue est donc au cœur de ce chapitre. Trois méthodes de représentations seront présentées. En plus d'un simple identifiant de la langue, nous verrons comment tenter d'extraire les *paramètres* syntaxiques, premièrement à partir des données, ensuite en utilisant le *World Atlas of Language Structures* (WALS) [2] (DRYER et HASPELMATH 2013).

Nos contributions consistent à comparer ces différentes méthodes et à évaluer leurs effets dans un cadre d'entraînement multilingue, où le corpus d'entraînement est composé de corpus arborés de 37 langues, puis dans un cadre de *zero-shot*, c'est-à-dire pour une langue n'ayant pas de données d'entraînement (AMMAR, MULCAIRE, BALLESTEROS et al. 2016 ; GUO et al. 2015).

2.2. État de l'art

Ce travail est à l'intersection de trois tendances dans la littérature sur l'analyse syntaxique en dépendances multilingue. La première est le *transfer parsing*, qui consiste à apprendre un analyseur sur une langue (ou un ensemble de langues) puis à le tester sur une autre. L'avantage de cette méthode est qu'elle peut s'appliquer à des langues avec peu ou pas de ressources, puisqu'il suffit d'avoir des données d'entraînement pour la langue source. La seconde est l'*analyse syntaxique délexicalisée*, qui a pour but d'ignorer le lexique. Cette deuxième méthode à l'avantage de neutraliser les biais de genre textuel ou de domaine, qui sont fortement marqués dans le vocabulaire des corpus. La troisième et dernière tendance est l'utilisation de *ressources typologiques* telles que le WALS, qui offrent de nombreux renseignements sur les langues. Ces ressources sont disponibles pour de très nombreuses langues, pour lesquelles il n'existe pas toujours de données annotées.

Le transfert de modèles (*transfer parsing*) est une solution efficace lorsqu'il s'agit de traiter des langues avec peu de ressources. MCDONALD et al. 2011 décrivent deux types

2. https://wals.info/

de transfert : le premier se base sur des corpus parallèles dont une des deux langues n'a pas, ou pas assez, de données d'entraînement pour apprendre un analyseur, alors que l'autre langue en possède. Le deuxième, les approches par transfert direct, repose sur les similarités entre les langues et ne nécessitent pas de corpus parallèle. Par exemple, LYNN et al. 2014 proposent un analyseur pour l'irlandais entraîné d'abord sur une autre langue, puis appliqué à l'irlandais. Ces méthodes donnent parfois des résultats inattendus : bien qu'elles n'appartiennent pas à la même famille de langues, c'est l'indonésien qui donne les meilleurs résultats dans cette approche. L'hypothèse des auteurs serait que les dépendances très distantes sont mieux représentées dans l'indonésien que dans les autres langues testées.

Les langues peu dotées peuvent parfois avoir une petite quantité de données d'entraînement disponible. En concaténant les corpus d'entraînement de deux langues, on peut obtenir un analyseur bilingue pour vérifier si une amélioration est possible comparé à un analyseur monolingue (VILARES et al. 2015). Les méthodes de transfert direct et les analyseurs bilingues sont proches des méthodes utilisées dans ce chapitre, puisqu'on retrouve cette idée de concaténation des corpus d'entraînement.

La combinaison de corpus de multiples langues pour l'entraînement d'un analyseur est facilitée par les récentes avancées sur les standards multilingues et les ressources disponibles, en particulier grâce aux *Universal Dependencies* pour la syntaxe en dépendances (NIVRE, MARNEFFE et al. 2016). La recherche sur l'analyse syntaxique multilingue est fortement encouragée par des initiatives comme les campagnes d'évaluation CoNLL 2017 et 2018, sur l'analyse syntaxique en dépendances multilingue à partir de textes bruts (ZEMAN, HAJIČ et al. 2018 ; ZEMAN, POPEL et al. 2017).

Comme nous l'avons mentionné ci-dessus, les analyseurs délexicalisés ignorent la forme superficielle et les lemmes des mots lors de l'analyse d'une phrase, utilisant des traits plus abstraits comme les étiquettes de POS. Le partage du lexique entre langues étant généralement faible, l'utilisation d'analyseurs délexicalisés est particulièrement pertinente pour l'apprentissage d'analyseurs multilingues. L'approche proposée par ZEMAN et RESNIK 2008 consiste à adapter un analyseur pour une nouvelle langue en utilisant soit un corpus parallèle, soit un analyseur délexicalisé. Cette méthode peut être utilisée pour construire rapidement un analyseur si la langue source et la langue cible sont suffisamment proches. MCDONALD et al. 2011 montrent que les analyseurs délexicalisés peuvent être facilement transférés entre les langues, donnant de meilleurs résultats que des analyseurs non supervisés.

Une autre source d'information pouvant être utilisée pour réaliser le transfert est une description des traits typologiques des langues, tel que ceux présents dans le WALS (DRYER et HASPELMATH 2013) ou Glottolog[3] (HAMMARSTRÖM et al. 2021) qui donnent des informations sur la structure des langues. Ces traits pourraient être utiles pour guider un analyseur multilingue, l'informant sur les paramètres du modèle qui pourraient être partagés parmi les langues qui ont des caractéristiques communes. NASEEM et al. 2012 et ZHANG et BARZILAY 2015 utilisent l'ensemble des traits issus de la catégorie de l'Ordre des mots du WALS qui sont disponibles pour leurs langues.[4]

3. Plus de ressources du genre peuvent être trouvées sur `https://clld.org/`.
4. Ces traits, que nous définirons et utiliserons également par la suite, sont les suivants : 81A, 85A,

PONTI et al. 2018 utilisent plutôt les traits qu'ils jugent pertinents dans les diverses catégories (pas uniquement celle concernant l'ordre des mots). SHI et al. 2017 ont obtenu les meilleurs résultats sur les langues surprises de la campagne d'évaluation CoNLL 2017, et se sont basé sur le WALS pour estimer quelle était la langue la plus proche de chaque langue surprise afin d'entraîner un modèle délexicalisé dessus.

TÄCKSTRÖM et al. 2013 utilisent une méthode de transfert multilingue délexicalisée, montrant en quoi le partage de paramètres basés sur des traits typologiques et l'appartenance à une famille de langues peut être utilisé dans un analyseur en dépendances discriminant. Les traits typologiques qu'ils choisissent sont basés sur ceux utilisés par NASEEM et al. 2012, en retirant deux traits qu'ils ne considèrent pas utiles (88A : *Ordre du démonstratif et du nom* et 89A : *Ordre du Nombre et du Nom*).

AMMAR, MULCAIRE, BALLESTEROS et al. 2016 concatènent des corpus pour entraîner un analyseur multilingue. Les auteurs utilisent un analyseur à transitions entraîné sur un ensemble de traits lexicaux incluant des plongements de mots multilingues, des clusters de Brown, et des étiquettes POS détaillées. Ils utilisent également un vecteur de représentation de la langue encodé en one-hot, l'ensemble de 6 traits issus du travail de NASEEM et al. 2012, et testent en plus l'utilisation de l'intégralité des traits du WALS. Leurs expériences ont été réalisées sur sept langues richement dotées. Bien qu'ils aient montré que, dans un cadre lexicalisé, la concaténation de corpus peut donner des résultats similaires à des analyseurs syntaxiques monolingues, les origines et les limites de ces gains restent floues.

Une méthode de représentation des langues basée sur des statistiques des langues a été explorée dans WANG et EISNER 2018, qui se basent eux même sur les travaux de LIU 2010. Les auteurs ont pu mettre en évidence l'impact positif de ce type de représentation dans un cadre d'analyse syntaxique délexicalisé. FISCH et al. 2019 ont montré que ce type de représentation pouvait même obtenir de meilleurs résultats qu'un vecteur issu du WALS.

2.3. Les différentes représentations de la langue

Dans ce travail, nous utilisons trois méthodes de représentation de la langue. La première (que nous appellerons ID) consiste en un simple identifiant de la langue, sous la forme d'un vecteur one-hot de dimension 37 (pour les 37 langues de notre corpus d'entraînement, voir section 2.5) en entrée de l'analyseur. Pour chaque mot passé en entrée, on concatène le vecteur one-hot correspondant à la langue du mot. Les deux autres méthodes sont décrites ci-dessous.

Apprentissage à partir des données Une méthode pour représenter la langue consiste à apprendre un vecteur à partir de nos données d'entraînement, utilisant des statistiques sur les dépendances vues dans les corpus d'entraînement. Deux vecteurs seront appris de cette façon.

86A, 87A, 88A et 89A

Le premier, que nous appellerons W_d, consiste à encoder, pour chaque étiquette syntaxique, la distance moyenne qui sépare les deux éléments de la relation dans une langue. Chaque dépendance d sera encodée sur deux composantes du vecteur W_d : une pour le cas où d est une dépendance droite et l'autre pour les dépendances gauches. Ce type de vecteur est comparable à ceux utilisés par WANG et EISNER 2018.

La deuxième représentation, appelée W_{df}, ajoute à W_d l'information sur la fréquence de chacune de ces dépendances. Ainsi, s'il existe n étiquettes de dépendance, W_d sera de taille $2*n$ et W_{df} de taille $4*n$. Un exemple de W_d et W_{df} est donné, dans la Figure 2.1.

6,00	3,36	12,80	12,59	...
G_acl	D_acl	G_advcl	D_advcl	...

6,00	3,36	0,00005	0,02	12,80	...
G_acl	D_acl	freq_G_acl	freq_D_acl	G_advcl	...

FIGURE 2.1. – Exemple de W_d (en haut) et W_{df} (en bas). G_acl représente les dépendances gauches étiquetées acl, dont la distance moyenne entre les éléments est de 6. D_acl représente la même chose, mais pour une dépendance droite. freq_G_acl est la fréquence des dépendances acl gauches dans la langue, et freq_D_acl est la fréquence des dépendances acl droites.

Ces vecteurs sont appris sur notre corpus d'entraînement (section. 2.5) et ne sont donc pas disponibles pour les langues n'ayant pas de corpus d'entraînement. Ils ne conviennent donc pas aux tests de type zero-shot, contrairement aux vecteurs issus du WALS présentés ci-dessous.

World Atlas of Language Structures Le WALS (DRYER et HASPELMATH 2013) est une ressource décrivant 2 676 langues grâce à un ensemble de 192 traits. Ces traits se répartissent en 11 familles (p.ex. Phonologie, Morphologie, Ordre des mots...).

Le WALS peut ainsi être représenté via une matrice W de 2 676 lignes et 192 colonnes, où chaque cellule $W(l, f)$ donne la valeur du trait f pour la langue l. Chaque ligne $W(l)$ est le vecteur de trait de la langue l.

Cette matrice a été épurée puis complétée comme expliqué ci-dessous pour correspondre à nos conditions expérimentales.

Pour commencer, nous n'avons gardé que les lignes correspondant aux 49 langues de notre corpus de test. Inversement, les langues mortes de UD (le Vieux Slave (cu), le Gotique (got), le Grec Ancien (grc), et le Latin (la)) n'apparaissent pas dans le WALS et ont donc été mises de côté. On obtient alors une version réduite de W contenant 45 lignes.

Deux représentations de la langue ont été extraites à partir du WALS. La première, que l'on appellera W_N, est basée sur les travaux de NASEEM et al. 2012, qui ont sélectionné les 6 traits de la famille de l'*Ordre des mots* qui étaient entièrement renseignés pour leurs 17 langues cibles. Ces traits couvrent des phénomènes tels que l'ordre verbe-objet ou adjectif-nom, et ont été largement discutés dans la littérature AMMAR,

MULCAIRE, BALLESTEROS et al. 2016; TÄCKSTRÖM et al. 2013; ZHANG et BARZILAY 2015. La matrice qui en résulte possède 45 lignes (langues) pour 6 colonnes (traits). Cependant, le WALS est une matrice creuse, puisque certains traits ne sont pas renseignés pour certaines langues [5]. C'est pourquoi nous avons fait le choix de ne garder que les langues pour lesquelles au moins la moitié du vecteur est renseignée, éliminant ainsi 5 langues de plus : le galicien (gl), le haut sorabe (hsb), le kazakh (kk), le slovaque (sk), et le ouïghour (ug). Nos tests sont effectués sur cet ensemble de 40 langues [6]. Cependant, 3 de ces langues (bxr, kmr, sme) ne disposent pas de corpus d'entraînement et seront traitées légèrement différemment des autres.

La deuxième représentation de la langue extraite du WALS, que l'on appellera W_{22}, est une version plus étendue de W_N. Alors que NASEEM et al. 2012 se restreignent à 6 traits, nous incluons dans W_{22} tous les traits renseignés pour au moins 80% de nos 40 langues. 22 traits résultent de cette sélection.

En plus des traits de la famille de l'*Ordre des mots*, nous avons également inclus ceux de la famille des *Propositions simples*. La famille des *Phrases Complexes* a également été considérée, mais aucun trait ne dépassait le seuil des 80%. Il en résulte une matrice de 40 lignes et 22 colonnes, correspondant à 3 traits de la famille des *Propositions simples* et 19 traits de celle de l'*Ordre des mots*. Les différents traits et leurs valeurs sont détaillés en annexe dans le tableau .1.

Les matrices W_N et W_{22} obtenues ne sont cependant pas complètes : elles contiennent respectivement 4 et 35 valeurs non-renseignées. Nous avons renseigné automatiquement ces valeurs grâce la méthode décrite ci-dessous.

Chaque matrice W (W_N et W_{22}) permet de comparer deux langues l_1 et l_2 en utilisant la distance de Hamming (Le nombre de dimensions pour lesquelles les valeurs diffèrent) entre leurs vecteurs $W(l_1)$ et $W(l_2)$, noté $d(l_1, l_2)$. Pour remplacer les valeurs manquantes, nous avons sélectionné, pour chaque langue l_1 contenant au moins une valeur non-renseignée ('?'), la valeur correspondante dans le vecteur de la langue l_2 la plus proche qui soit entièrement renseignée, où $l_2 = \underset{l_i \mid \text{“?”} \notin W(l_i)}{\arg\min} d(l_1, l_i)$.

Les matrices W_N et W_{22} ne fournissent qu'une description partielle des langues, fortement biaisée en faveur de l'analyse syntaxique et ignorant les autres aspects (i.e. la phonologie). Néanmoins, il est pertinent de comparer les distances de langues appartenant aux mêmes familles typologiques dans ce mode de représentation. Pour cela, nous nous sommes intéressés à 3 familles présentes dans notre ensemble de 40 langues : les langues Romanes (6 langues : catalan, espagnol, français, italien, portugais et roumain), les langues Germaniques (6 langues : danois, allemand, anglais, néerlandais, norvégien et suédois) et les langues Slaves (7 langues : bulgare, tchèque, croate, polonais, russe, slovaque, ukrainien). Nous avons alors calculé la proximité des vecteurs de ces langues. Nous définissons la distance interne moyenne (MID) d'un ensemble de langues $L = \{l_1, \ldots, l_n\}$, comme la moyenne des distances de chaque paire

5. Lorsqu'un trait n'est pas renseigné, cela signifie qu'il n'y a pas eu d'étude pour ce trait pour cette langue. Il existe généralement une valeur pour les langues non-concernées par le trait.

6. Les codes de ces langues sont : ar, bg, bxr, ca, cs, da, de, el, en, es, et, eu, fa, fi, fr, ga, he, hi, hr, hu, id, it, ja, kmr, ko, lv, nl, no, pl, pt, ro, ru, sl, sme, sv, tr, uk, ur, vi, zh

	Romane	Germanique	Slave	Aléatoire
W_N	0.33	1.33	0.67	2.41
W_{22}	4.13	4.47	4.19	10.15

TABLEAU 2.1. – MID de chaque famille de langues comparée à l'aléatoire. Aléatoire est la moyenne de 50 000 familles de 6 langues. Plus la valeur est petite, plus les langues de l'ensemble ont des valeurs proches.

dans L :

$$MID(L) = \frac{1}{n^2 - n} \sum_{\substack{(l_i, l_j) \in L \times L \\ i \neq j}} d(l_i, l_j)$$

Nous avons calculé le MID de chaque famille de langues, ainsi que la moyenne du MID de 50 000 familles de 6 langues aléatoires (ce qui correspond au nombre de langues des familles Germanique et Romane) et l'avons comparé au MID d'ensembles aléatoires de 6 langues. Les résultats du tableau 2.1 montrent clairement que les vecteurs du WALS permettent de capturer des similarités au sein d'une famille de langues, puisque le MID des vecteurs de langues d'une même famille est nettement inférieur au simple hasard.

2.4. Analyseur

L'analyseur utilisé dans les expériences de ce chapitre est celui du logiciel Macaon [7]. Cet analyseur par transition de type *arc-eager* NIVRE 2008 est entraîné avec un oracle dynamique GOLDBERG et NIVRE 2012. La prédiction des transitions est faite par un perceptron multi-couches (MLP) similaire au système de CHEN et MANNING 2014, consistant en une couche d'entrée, une couche cachée et une couche de sortie. Deux ensembles de traits entièrement délexicalisés ont été définis pour la prédiction : BASIQUE et ÉTENDU. BASIQUE est un ensemble classique composé de 9 traits liés au POS, 7 traits syntaxiques, 32 traits morphologiques et un trait de distance (la distance entre la tête et le dépendant). L'ensemble ÉTENDU ajoute à BASIQUE de nouveaux traits correspondant aux vecteurs du WALS (W_N ou W_{22}), ou l'identifiant de la langue. Chaque trait est associé à un plongement de taille 3, initialisé à zéro. L'identifiant de la langue (issu de la représentation ID) est un vecteur one-hot de dimension 37 (correspondant aux 37 langues du corpus d'entraînement). La couche d'entrée du MLP correspond à la concaténation des plongements des différents traits, dont les dimensions varient de 396 à 465 selon la configuration (avec ou sans vecteurs de langue W_N et W_{22}, ou de l'identifiant de la langue ID). La couche de sortie est composée de 263 neurones, correspondant au nombre de transitions que l'analyseur peut prédire. La couche cachée est de taille 1 000, avec un *dropout* durant l'entraînement de 0.4, le nombre d'itérations est égal à 10, la fonction d'activation est la fonction ReLu, la fonction objectif est un softmax de vraisemblance négative, et l'algorithme d'appren-

7. https://gitlab.lis-lab.fr/franck.dary/macaon

tissage est AMSgrad, utilisant les paramètres par défaut de Dynet NEUBIG et al. 2017. Les valeurs des hyperparamètres ont été définies dans des conditions similaires à *Multi* W_N, cf. section 2.5).

À chaque étape du processus d'analyse syntaxique, l'analyseur prédit une action à réaliser, pouvant aboutir à la création d'une nouvelle dépendance entre deux mots de la phrase. La prédiction des actions est basée sur la valeur des traits fournis en entrée au MLP. Dans la configuration BASIQUE, ces traits décrivent différents aspects du gouverneur, du dépendant, et du contexte. Par exemple, si la tête est un verbe et que le dépendant est un nom situé avant le verbe, une dépendance sujet aura une forte probabilité d'être prédite pour les langues utilisant majoritairement l'ordre sujet-verbe (SV) (voir figure 2.2).

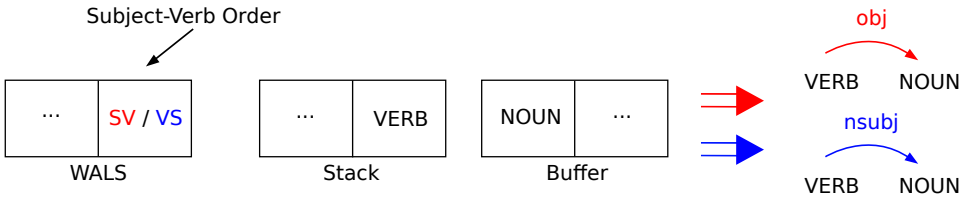


FIGURE 2.2. – Les caractéristiques typologiques pourraient guider les analyseurs pour partager des informations entre des langues similaires au-delà de la simple identification de la langue.

Avec l'ensemble de traits ÉTENDU, l'utilisation de l'ordre SV pour une langue est explicite. Le MLP a donc la possibilité de combiner une *configuration phrastique* (i.e. un verbe avant un nom) avec une *configuration de la langue* (i.e., la langue est SV) quand il prédit une action. Les langues partageant un trait dans W auront la possibilité de générer la même prédiction pour une configuration phrastique correspondant à ce trait (i.e. le nom précédant le verbe et la langue est de type SV).

Cet analyseur ne peut pas prédire d'arbres non-projectifs. La présence d'une dépendance non-projective génère systématiquement une erreur d'analyse lors de la phase de test. Le taux moyen de non-projectivité du corpus de test est égal à 1%, avec un écart-type de 1% pour les 40 langues. 14 corpus (pour 20 langues) ont un taux de non-projectivité inférieur à 1%, et le taux maximum est de 8% pour le corpus du néerlandais (corpus Lassysmal). Nous avons fait des tests utilisant une transformation en arbre pseudo-projectif NIVRE et NILSSON 2005, mais l'impact sur les résultats étant négligeable, nous avons choisi de garder l'algorithme projectif original.

Les analyseurs délexicalisés, entraînés à partir des corpus de UD, prennent en entrée les parties du discours universelles (UPOS) et les traits morphologiques (FEAT) et prédisent les étiquettes de l'arbre de dépendances (incluant des sous-relations syntaxiques pouvant être spécifique au langage (p. ex., *acl :relcl*)). Les traits morphologiques sont renseignés pour presque tous les corpus, mais présentent de grandes différences. Par conséquent, nous avons fait le choix de ne garder que les traits les plus fréquents qui apparaissent dans au moins 28 langues. De plus, les traits morphologiques sont représentés comme une liste de paires *(clef, valeur)* que nous avons

séparés, afin que chaque paire soit considérée indépendamment, produisant ainsi un ensemble de 16 traits morphologiques par mot[8].

2.5. Cadre expérimental

Corpus Nos expériences ont été réalisées sur les données de la campagne d'évaluation CoNLL 2017 ZEMAN, POPEL et al. 2017, en utilisant la tokenisation de référence et en ignorant les contractions (i.e. *du=de+la*). Nos modèles sont évalués individuellement sur chacune des 40 langues pour lesquelles nous disposons d'un vecteur $W(l)$ (section. 2.3), en utilisant les corpus de test de la campagne d'évaluation pour faciliter la comparaison avec les travaux similaires. Le corpus de test pour chaque langue est obtenu par la concaténation de tous les corpus de test disponibles pour cette langue, et contient au minimum 10 000 tokens pour chaque langue.

Trois langues n'ont pas de corpus d'entraînement ou de développement (le bouriate (bxr), le kurmandji (kmr) et le same du Nord (sme)). L'entraînement et le développement se font sur des corpus multilingues (ML) dérivés des 37 langues de UD restantes, que l'on nommera TRAIN-ML et DEV-ML. La taille des corpus de UD peut fortement varier d'une langue à l'autre, allant de 529 mots pour le Kazakh (kk) à 1 842 867 pour le Tchèque (cs).

Ainsi, concaténer tous les corpus pour constituer TRAIN-ML et DEV-ML surreprésenterait certaines langues et introduirait un biais en leur faveur. C'est pourquoi nous avons décidé d'équilibrer le nombre de tokens de TRAIN-ML et DEV-ML entre les langues. L'évaluation de nos modèles se fait en décodant chaque corpus de test séparément, donnant un score par langue, puis la moyenne de ces scores est effectuée (c'est la macro moyenne), donnant autant de poids à chaque langue lors de l'évaluation. C'est pourquoi aucun équilibrage n'a été fait sur les corpus de test.

La construction des corpus de développement et d'entraînement se fait en deux étapes. Tout d'abord la totalité des données des corpus disponibles (entraînement et développement) est séparée en deux : de 10% pour un sous corpus de développement et 90% pour l'entraînement. Ensuite, pour chacun des sous-corpus précédents, on sélectionne de façon aléatoire (avec remise) des phrases pour chaque langue jusqu'à ce que les corpus finaux atteignent 2 000 tokens pour le DEV et 20 000 tokens pour le TRAIN). Enfin les données ainsi sélectionnées sont mélangées (pour chacun des corpus) afin d'éviter de conserver l'ordre des langues. Avec cette procédure, une même phrase peut apparaître plusieurs fois. Néanmoins, cette approche garantit une représentation équilibrée de chaque langue dans TRAIN-ML et DEV-ML.

Métrique La qualité des arbres prédits est évaluée par une mesure standard pour l'analyse syntaxique en dépendance : le score d'attachement étiqueté (LAS). Le score UAS a été omis, car le UAS et le LAS sont étroitement corrélés ($r = 0.98$). Le script d'évaluation de la campagne d'évaluation CoNLL 2017 a été utilisé. On donne le LAS

8. Ces traits morphologiques sont : Number, Case, VerbForm, Person, Mood, Tense, PronType, NumType, Polarity, Gender, Voice, Degree, Reflex, Poss, Definite et Aspect

par langue, ainsi que le MACRO-LAS, qui est la macro-moyenne du LAS de toutes les langues qui ont un corpus d'entraînement. Cette mesure est indépendante de la taille du corpus de test de chaque langue, et n'introduit pas de biais en faveur des langues sur-représentées dans l'ensemble de test.

Configuration d'entraînement Nos expériences sur plusieurs paires ⟨corpus d'entraînement, représentation de la langue⟩ sont désignées par les codes suivants :

Mono : Corpus monolingue.
Le corpus d'entraînement de la langue l consiste simplement à prendre toutes les phrases de la langue l dans TRAIN-ML. 37 analyseurs avec la configuration BASIQUE ont été entraînés, un pour chaque langue. Cette configuration correspond à la situation standard des expériences en analyse syntaxique : entraînement et test sur une même langue.

Multi : Corpus multilingue.
Un analyseur est entraîné sur l'intégralité de TRAIN-ML, sans indication de la langue. Le modèle est délexicalisé, le corpus ne contient donc que les étiquettes POS de référence, les traits morphologiques de référence, et les relations syntaxiques à apprendre.

Multi ID : Corpus multilingue + identifiant de la langue.
Un analyseur entraîné avec la configuration ÉTENDU sur TRAIN-ML, en utilisant l'identifiant de la langue attaché à chaque mot.

Multi W_N, *Multi* W_{22} : Corpus + WALS.
Deux analyseurs entraînés avec l'ensemble de traits ÉTENDU, sur TRAIN-ML. L'entraînement de *Multi* W_N (resp. W_{22}) ajoute à chaque mot de *Multi* un vecteur $W(l)$ issu du WALS, qui correspond à la langue du mot. Un seul modèle est entraîné pour *Multi* W_N (resp. W_{22}).

Multi W_d, *Multi* W_{df} : Corpus + vecteurs de données.
Deux analyseurs entraînés avec l'ensemble de traits ÉTENDU, sur TRAIN-ML, avec W_d (respectivement W_{df}) attaché à chaque mot. Cet ajout est fait de la même façon que pour les tests *Multi* W_N et *Multi* W_{22}.

ZS **:** *Zero-shot.*
Dans cette configuration, 37 corpus d'entraînements sont dérivés de TRAIN-ML : pour chaque langue l, un corpus d'entraînement est construit à partir de toutes les phrases de TRAIN-ML, sauf celles appartenant à la langue l. Cette configuration représente la situation où un analyseur dans la configuration BASIQUE est entraîné pour une langue pour laquelle aucune donnée d'entraînement n'est disponible, comme dans les méthodes de transfert direct.

ZS W_{22} **:**

Ajoute à *ZS* l'indication de la langue de chaque mot à travers son vecteur du WALS, de la même manière que *Multi* W_{22} faisait avec *Multi*.

2.6. Résultats et analyse

Nos expériences ont été réalisées dans les configurations décrites ci-dessus. Le LAS est donné pour chaque langue, ainsi que la macro-moyenne du LAS (MACRO), dans le tableau 2.2. Nous commentons ci-dessous les résultats pour *Mono*, et comparons les résultats de certaines expériences (voir tableau 2.3).

***Mono* :** Les résultats de l'expérience *Mono*, qui consiste à entraîner des modèles monolingues ne partageant pas d'informations entre les langues, montrent une importante variation des performances selon les langues. Le LAS varie de 46,78 pour le turc (tr) à 81,44 pour l'italien (it). Plus d'expériences seraient nécessaires pour expliquer les raisons d'une telle variabilité, mais n'entrent pas dans le cadre de ce travail. Nous pouvons tout de même émettre certaines hypothèses. Tout d'abord, certaines particularités spécifiques aux langues, comme l'équilibre entre les marqueurs morphologiques et syntaxiques (i.e. les langues morphologiquement riches sont probablement favorisées dans notre configuration, puisque l'analyse morphologique est donnée en entrée de l'analyseur). D'autres sont spécifiques au genre textuel. Bien que la délexicalisation permette de neutraliser certains biais de genre, ce dernier peut aussi influencer la syntaxe, notamment la longueur des phrases, qui sont généralement plus dures à analyser plus elles sont longues, ou encore la proportion de constructions difficiles, comme les rattachements prépositionnels ou les coordinations ambiguës. Enfin, l'hétérogénéité de la qualité des annotations selon les langues peut également expliquer la variabilité du LAS.

***Mono* vs *Multi* :** Alors que *Mono* entraîne des modèles monolingues ne partageant pas d'informations entre les langues, *Multi*, au contraire, s'entraîne sur toutes les langues disponibles, permettant un éventuel partage d'informations entre les langues. Une chute attendue des performances est observée lorsque l'on passe de *Mono* à *Multi*. Le MACRO LAS chute de 5,68 points. L'hypothèse principale pour expliquer cette chute est le bruit qui est introduit par le mélange des langues. Concaténer toutes les phrases de plusieurs langues sans préciser l'identité de la langue introduit du bruit dans l'analyseur. Par exemple, la configuration phrastique associée à une dépendance sujet/verbe dans une langue SV ou VS sera très différente, et l'analyseur n'est pas capable de faire la distinction entre ces langues et voit donc des contradictions. La chute du LAS est cependant très variable selon les langues, allant même jusqu'à une augmentation dans le cas de l'espagnol (es) (+0,51 points). Nous n'avons pas actuellement d'explication pour ce résultat, tout au plus une intuition qui serait que *Multi* apprendrait implicitement une langue moyenne (bruitée) qui serait plus proche de l'espagnol (es) que du chinois (zh) par exemple (dont le LAS chute de 14,37 points), les langues composant *Multi* étant plus proches de l'espagnol (es) que du chinois (zh) en moyenne.

Mono	*Multi*	*Multi* ID	*Multi* W_N	*Multi* W_{22}	*Multi* W_d	*Multi* W_{df}	*ZS*	*ZS* W_{22}	Lang.	
65,89	60,59	64,04	63,15	64,38	64,08	63,29	27,50	34,34	ar	
78,59	74,32	79,25	76,26	77,47	78,28	78,94	67,05	63,73	bg	S
77,18	72,76	76,63	73,03	76,27	76,90	77,17	70,48	68,88	ca	R
68,92	68,01	69,41	68,72	69,61	69,75	69,67	62,08	59,47	cs	S
73,62	67,38	70,26	70,19	70,25	70,57	72,19	61,56	63,87	da	G
71,07	63,76	70,36	69,18	69,22	70,37	70,44	59,65	60,51	de	G
77,11	71,26	76,16	73,29	75,84	76,11	76,90	63,74	65,96	el	
70,05	66,02	71,17	69,91	70,19	70,78	71,21	60,87	62,11	en	G
71,47	71,98	73,83	72,29	73,22	73,89	74,12	71,40	70,76	es	R
66,98	63,76	68,41	65,75	67,79	67,94	69,08	58,52	58,77	et	
63,26	55,76	60,11	60,22	59,39	60,00	60,31	33,50	31,68	eu	
72,85	66,02	69,50	69,63	70,00	69,67	68,02	31,25	34,27	fa	
60,97	56,29	59,65	57,37	59,28	59,99	59,92	50,69	48,72	fi	
75,74	74,25	76,16	74,79	75,82	75,80	76,50	72,68	71,97	fr	R
66,55	60,41	66,86	64,68	65,96	66,02	66,30	43,51	42,75	ga	
70,21	63,03	67,97	66,09	67,45	67,19	68,33	51,36	52,98	he	
78,91	73,86	74,86	75,77	74,45	76,66	76,52	54,23	57,33	hi	
71,03	67,49	71,37	70,00	70,40	71,42	71,28	62,88	65,71	hr	S
67,08	62,55	66,84	67,19	67,51	67,79	68,14	48,24	49,62	hu	
68,64	58,38	63,98	62,61	64,57	64,31	64,57	45,57	43,03	id	
81,44	76,45	80,56	76,97	79,83	80,45	81,73	75,82	77,04	it	R
78,26	68,22	75,81	74,85	75,56	76,59	76,89	7,64	28,53	ja	
47,68	37,17	38,61	38,07	39,66	41,11	39,36	17,99	23,18	ko	
59,89	54,11	59,98	58,23	60,17	59,78	60,86	44,41	47,38	lv	
62,56	57,21	59,28	58,13	58,59	59,11	60,00	51,11	48,98	nl	G
74,59	73,19	76,95	73,51	75,93	77,67	76,75	53,09	55,49	no	G
81,24	74,24	80,52	74,78	79,02	80,44	80,06	69,15	70,72	pl	S
72,00	65,74	69,83	68,74	69,86	69,88	69,96	62,85	65,86	pt	R
70,99	67,72	71,47	70,38	70,61	71,26	71,19	55,84	61,01	ro	R
74,06	61,35	74,72	68,65	74,45	74,92	75,06	55,09	55,50	ru	S
67,10	64,12	66,44	64,75	66,36	66,63	66,69	60,52	63,26	sl	S
72,05	69,97	72,55	70,71	71,88	73,21	73,25	64,80	67,05	sv	G
46,78	41,01	43,16	43,26	41,62	45,73	43,57	29,46	30,33	tr	
71,60	69,40	75,30	69,77	72,81	74,05	74,56	67,08	64,92	uk	S
74,35	69,15	70,93	70,71	70,76	72,41	71,47	58,93	59,41	ur	
54,40	42,42	53,75	51,94	51,72	52,81	52,42	25,86	41,07	vi	
59,83	45,46	58,48	53,42	54,87	56,95	59,19	22,24	24,48	zh	
69,32	63,64	68,25	66,41	67,64	68,39	68,54	51,86	53,80	MACRO	
-	33,32	34,10	30,37	28,49	-	-	-	-	bxr	
-	40,34	37,20	41,41	44,04	-	-	-	-	kmr	
-	47,34	45,60	47,63	42,38	-	-	-	-	sme	

TABLEAU 2.2. – LAS pour chaque langue, et MACRO LAS, pour les 9 configurations. *Mono* correspond à la configuration classique d'apprentissage sur une langue et test sur une même langue. Les expériences de type *Multi* sont des expériences multilingues, auxquelles on ajoute des représentations des langues dans le but d'aider le partage de connaissances entre langues. Les expériences de type *ZS* sont des systèmes zero-shot, où la langue de test n'est pas vue à l'apprentissage. Les langues suivies d'un S appartiennent à la famille des langues Slaves, G appartiennent à la famille des langues Germaniques et R appartiennent à la famille des langues Romanes. bxr, kmr et sme ne disposent pas de corpus d'entraînement, les expériences multilingues correspondent donc à du zero-shot pour ces trois langues.

X	Y	$\overline{X-Y}$	σ	min	max
Mono	*Multi*	5.68	3.32	-0.51 es	14.37 zh
Mono	*Multi* ID	1,08	2,37	-3,70 uk	9,07 ko
Multi W_{22}	*Multi*	4.00	2.58	0.59 hi	13.10 ru
Multi ID	*Multi*	4,60	2,91	1,00 hi	13,37 ru
Multi W_d	*Multi*	4,75	2,57	1,55 fr	13,57 ru
Multi W_{22}	*Multi* W_N	1.24	1.45	-1.64 tr	5.80 ru
Multi ID	*Multi* W_{22}	0,61	0,91	-1,05 ko	3,61 zh
Multi W_d	*Multi* ID	0,14	0,89	-1,53 zh	2,57 tr
Multi W_{df}	*Multi* W_d	0,15	0,89	-2,16 tr	2,24 zh
Mono	*ZS*	17,47	13,57	0,07 es	70,62 ja
ZS W_{22}	*ZS*	1,95	4,60	-3,32 bg	20,89 ja

TABLEAU 2.3. – Différence entre les configurations X et Y : moyenne des différences ($\overline{X-Y}$), écart type des différences (σ), minimum et maximum avec la langue correspondante.

***Mono* vs *Multi* ID :** *Multi* ID est un modèle multilingue où le partage d'informations entre langues est possible. L'information de la langue d'origine est explicitée à l'aide d'un identifiant de la langue. *Mono* est un système monolingue, sans connaissances partagées entre langues. Lors de l'ajout de l'identifiant de la langue dans les expériences de type *Multi*, on pourrait s'attendre à ce que les résultats soient équivalents à ceux de *Mono*, puisque le modèle a accès à la même quantité de données, ainsi qu'à l'information de la langue en cours de traitement. Pourtant, on observe 1.08 points d'écart entre les deux modèles, en faveur du modèle *Mono*. Une explication serait que bien que les informations données en entrée soient similaires, la taille du réseau de neurones, c'est-à-dire sa capacité en nombre de paramètres, ne change pas. Le système a donc le même espace pour stocker 37 fois plus d'informations, et perd donc probablement en richesse à cause de cela. Il serait intéressant d'entraîner un système dont la taille reste proportionnelle à la taille des données d'entrées afin de vérifier cette hypothèse.

***Multi* vs *Multi* W_{22}, *Multi* W_d, *Multi* ID :**[9] Alors que *Multi* est un modèle multilingue sans informations sur l'origine de la langue en cours d'analyse, *Multi* W_{22}, *Multi* W_d et *Multi* ID explicitent cette information, à l'aide respectivement d'un vecteur issu du WALS, d'un vecteur appris sur des statistiques sur les dépendances de chaque langue, et d'un vecteur one-hot représentant la langue en cours de traitement. Nous obtenons ici le premier résultat important de ce chapitre : lors de l'ajout d'une représentation de la langue à l'analyseur, la MACRO LAS augmente de 4 à 4,75 points comparée à *Multi*. Le LAS augmente pour toutes les langues, pour les trois expériences. Ajouter l'information ID à *Multi* permet une augmentation des résultats MACRO de 4,60 points. Cette augmentation était attendue puisque dans ce test, les configurations phrastiques sont associées à l'ID de la langue, ce qui aide à diminuer le bruit dans

9. W_N et W_{df} sont analysés plus tard.

les données. Deux interprétations sont possibles pour *Multi* W_N : l'interprétation optimiste est que les vecteurs de représentation de la langue aident à diminuer le bruit introduit par le mélange des langues dans *Multi* en "expliquant" certaines informations contradictoires dans les données grâce à l'utilisation des traits linguistiques encodés dans le WALS pour W_{22}, et grâce aux informations de distance contenues dans W_d. L'interprétation pessimiste consiste à dire que ces vecteurs sont un simple encodage arbitraire des langues. Dans ce cas, le MLP de l'analyseur apprendrait à associer les configurations phrastiques à certaines langues spécifiquement, et apprendrait alors différents modèles pour différentes langues, ce qui reviendrait à l'expérience ID. Plus d'expériences sont menées dans la suite pour comprendre comment le MLP utilise les vecteurs W.

Multi W_N **vs** ***Multi*** W_{22} **:** *Multi* W_N et *Multi* W_{22} sont deux systèmes multilingues, entraînés sur toutes les langues disponibles. Leur différence réside dans l'utilisation d'un vecteur permettant de représenter la langue. *Multi* W_N utilise un vecteur du WALS classiquement utilisé dans l'état de l'art, alors que *Multi* W_{22} utilise un vecteur du WALS étendu proposé dans cette thèse. Les vecteurs W_N et W_{22} n'ont pas le même impact lorsqu'on les ajoute à *Multi*. L'ajout de W_{22} permet l'augmentation de 4 points de la mesure MACRO tandis que l'ajout de W_N augmente le score MACRO par rapport au MACRO obtenu avec *Multi* de 2,77 points seulement. L'analyseur est donc capable de tirer profit d'une description des langues plus riche lors de son apprentissage. Ce résultat pourrait indiquer que les résultats décevants sur l'analyse syntaxique rapportés par AMMAR, MULCAIRE, BALLESTEROS et al. 2016, qui utilisaient le vecteur W_N, pourraient venir des traits extraits du WALS qui n'étaient pas suffisamment riches pour expliquer à l'analyseur les différences syntaxiques importantes entre les langues.

Multi ID **vs** ***Multi*** W_{22} **:** *Multi* ID et *Multi* W_{22} sont deux systèmes multilingues appris sur toutes les langues disponibles, le premier utilisant un one-hot pour représenter l'identifiant de la langue de la phrase en cours d'analyse, et le deuxième un vecteur issu du WALS, proposé dans le cadre de cette thèse. Malgré de meilleurs résultats obtenus en utilisant W_{22} plutôt que W_N, ce nouveau vecteur n'est toujours pas capable de surpasser les performances de l'utilisation d'un simple identifiant de la langue. Mais l'augmentation de la mesure MACRO entre W_N et W_{22} peut laisser supposer que le choix des traits du WALS influence fortement les résultats. Plus d'expériences sont nécessaires pour confirmer cette hypothèse. On pourra également se poser la question de l'influence de la méthode de remplacement des valeurs inconnues.

Multi ID **vs** ***Multi*** W_d **:** *Multi* ID et *Multi* W_d sont deux systèmes multilingues appris sur toutes les langues disponibles, le premier utilisant un one-hot pour représenter l'identifiant de la langue de la phrase en cours d'analyse, et le deuxième un vecteur issu de statistiques faites sur les dépendances de chaque langue. Le résultat de ces expériences représente le deuxième résultat majeur de ce travail : le vecteur W_d permet de battre les résultats de *Multi* ID. Même si l'hypothèse qu'une partie du vecteur

W_d sert à représenter la langue s'avère juste, une partie du modèle a été capable d'apprendre des informations supplémentaires, partagées entre les différentes langues. Ce résultat est tout de même à nuancer, puisque le LAS de *Multi* W_d n'est supérieur que de 0,14 points à celui de *Multi* ID.

Multi W_d **vs** ***Multi*** W_{df} **:** *Multi* W_N et *Multi* W_{22} sont deux systèmes multilingues, entraînés sur toutes les langues disponibles. Leur différence réside dans l'utilisation d'un vecteur permettant de représenter la langue, tous les deux issus de statistiques faites sur les dépendances de chaque langue. W_d est basé sur des informations sur les distances entre les dépendances, et W_{df} ajoute l'information de la fréquence de ces dépendances. Les résultats de *Multi* W_d et de *Multi* W_{df} sont assez proches (+0,15 points pour la mesure MACRO de W_{df}). L'information sur la fréquence des dépendances permet donc l'apprentissage de connaissances supplémentaires pour l'analyseur. Cependant, le temps d'apprentissage de *Multi* W_{df} étant assez conséquent, nous avons fait le choix de nous concentrer sur les résultats de *Multi* W_d. Le vecteur W_{df} étant deux fois plus gros que W_d, on peut également supposer que la taille de la couche cachée du MLP de l'analyseur n'est pas suffisante pour utiliser toutes les informations du vecteur W_{df}.

Mono **vs** *ZS* **:** Les modèles *Mono* consistent en l'apprentissage de modèles monolingues où il n'existe pas de partage de connaissances entre langues. *ZS* est un modèle zero-shot, appris sur toutes les langues disponibles sauf une, qui correspond à la langue sur laquelle les tests seront effectués. *ZS* correspond à des conditions extrêmes, mais également plus réalistes. La situation simulée est celle où aucune donnée d'entraînement n'est disponible pour la langue l. La chute des performances comparée à *Mono* était prévisible, mais n'en reste pas moins dramatique : la MACRO LAS chute de 17,47 points. Cette chute varie énormément selon les langues (cela se ressent sur l'écart type qui atteint 13,57 points). Certaines langues ne sont presque pas affectées, comme l'espagnol (es) qui ne perd que 0,07 point. En revanche, le japonais (ja) perd 70,62 points de LAS, la plus grosse chute toutes expériences et langues confondues. Les langues les plus isolées sont celles qui souffrent le plus du passage à *ZS*. Les familles de langues sont moins affectées, avec une chute de "seulement" 6,65 points en moyenne pour les langues romanes par exemple.

ZS **vs** *ZS* W_{22} **:** *ZS* et *ZS* W_{22} sont deux systèmes de zero-shot. Contrairement à *ZS*, *ZS* W_{22} utilise une représentation de la langue issue du WALS, permettant éventuellement un meilleur partage de connaissances entre langue. Cette comparaison constitue notre troisième et dernier résultat majeur. Cette expérience consiste à observer si l'ajout d'un vecteur issu du WALS permet de limiter la chute des score LAS de l'expérience *ZS*. Comme on peut le voir dans les Tables 2.2 et 2.3, l'ajout de W_{22} permet d'augmenter la MACRO LAS de 1,95 points, ce qui pourrait laisser penser que W_{22} permet bien d'aider le modèle à gérer les langues inconnues. Cependant, on remarque que l'écart type est assez élevé (4,60 points). En effet, lorsque l'on regarde le détail langue par langue, 11 de nos 37 langues ont de moins bons résultats avec le

Configuration	Features	Accuracy
Multi	input	0.432
Multi W_N	input	0.678
Multi W_{22}	input	0.954
Multi	hidden	0.436
Multi W_N	hidden	0.682
Multi W_{22}	hidden	0.956

TABLEAU 2.4. – Précision de l'identification du langage pour un classificateur de régression logistique entraîné sur les activations après la couche cachée, ou en entrée pour les configurations *Multi*, *Multi* W_N et *Multi* W_{22}. Le classificateur est entraîné sur l'ensemble de développement, les résultats sont rapportés sur l'ensemble de test.

modèle *ZS* W_{22} qu'avec *ZS*. Le bulgare perd même 3,32 points de LAS. Les langues ayant des résultats extrêmement bas dans *ZS* ont réussi à tirer profit du vecteur W_{22}, comme le japonais (ja) qui, après son importante chute au passage à *ZS*, remonte de 20,89 points avec *ZS* W_{22}. Le vecteur W_{22} reste une description extrêmement partielle de la langue, et il est probable qu'il ne soit pas suffisant pour compenser l'absence de corpus d'entraînement. La question précédemment soulevée se pose à nouveau : le vecteur sert-il simplement d'identifiant de la langue ? Si l'analyseur utilise, ne serait-ce qu'en partie, W_{22} pour détecter la langue de la phrase analysée, l'absence de cette langue dans le corpus d'entraînement rend la tâche impossible.

2.7. Comment le vecteur W est-il pris en compte par l'analyseur ?

Nous allons à présent chercher à comprendre quels sont les effets des traits typologiques du WALS sur le modèle de l'analyseur.

Comme évoqué précédemment, une hypothèse pour expliquer le comportement de l'analyseur en présence de W serait qu'il utilise les traits du WALS pour identifier la langue au lieu d'apprendre des généralisations sur les phénomènes syntaxiques. Le tableau 2.4 donne la précision d'un classificateur de régression logistique entraîné pour prédire l'ID de langue en fonction soit des données d'entrée du MLP de l'analyseur, soit des activations après la couche cachée, pour *Multi*, *Multi* W_N et *Multi* W_{22}. Les résultats montrent qu'en effet, les traits du WALS, en particulier W_{22}, améliorent considérablement les prédictions du classificateur de langues, suggérant que l'analyseur peut utiliser l'information de l'identité de la langue dans ses prédictions. Le fait que ces informations soient toujours disponibles juste avant la couche de décision signifie qu'elles peuvent être utilisées pour prédire les actions de l'analyseur.

Un autre axe d'analyse consiste à comparer la répartition des activations pour deux langues. Les activations sont mesurées au niveau de la couche cachée avant

		Moyenne			Max			Min		
L1	L2	*Multi*	*Multi* W_{22}		*Multi*	*Multi* W_{22}		*Multi*	*Multi* W_{22}	
nl	de	0.860	0.854	↘	1.027	0.940	↘	0.798	0.793	↘
pt	fr	0.878	0.912	↗	1.335	3.550	↗↗	0.794	0.700	↘
bxr	ga	0.890	1.160	↗↗	1.600	4.888	↗↗	0.782	0.757	↘

TABLEAU 2.5. – Statistiques JSD au niveau du neurone entre les activations au niveau de la couche cachée des modèles d'analyseur pour des paires de langues sélectionnées.

la fonction Unité Linéaire Rectifiée (ReLU), et devraient suivre une loi normale au niveau des neurones. Nous calculons la divergence Jensen-Shannon (JSD) entre les activations d'un neurone donné pour une paire de langues. Le tableau 2.5 montre la JSD moyenne, maximum et minimum au niveau des neurones entre des paires de langues sélectionnées. Nous avons sélectionné trois paires de langues avec une distance croissante. Le néerlandais et l'allemand (nl-de) appartiennent à la même famille de langue (germanique) et ont des vecteurs W_{22} identiques. Le portugais (pt) et le français (fr) appartiennent également à la même famille (romane) mais leurs vecteurs diffèrent par six caractéristiques (par exemple 101A Expression des sujets pronomiaux, 143E Morphèmes négatifs préverbaux). À l'autre extrême, le bouriate et l'irlandais (bxr-ga) ont des vecteurs W_{22} très différents, avec seulement deux valeurs partagées sur 22.

Pour la paire néerlandais-allemand (nl, de), la différence moyenne entre les distributions d'activation dans *Multi* W_{22} (0,854) est légèrement plus faible que dans *Multi* (0,86), ce qui suggère que W_{22} pourrait aider à tirer parti de la similitude entre ces langues. Cette hypothèse est renforcée par l'augmentation du LAS en passant de *Multi* à *Multi* W_{22} (Table 2.2).

Pour la paire portuguais-français (pt, fr), l'ajout de W_{22} entraîne une augmentation de la distance moyenne entre les distributions d'activation (0,912) par rapport à *Multi* (0,878). De manière analogue, cette différence augmente également d'une plus grande marge (de 0,89 à 1,16) pour la paire (bxr, ga) qui est la paire constituée des langues les plus éloignées l'une de l'autre. Dans l'ensemble, ces observations indiquent que W_{22} renforce le partage de paramètres entre des langages similaires et augmente le contraste entre des langages différents. A titre d'exemple, la figure 2.3 montre que les distributions pour le neurone avec la JSD la plus élevée sont très similaires pour (nl, de) alors qu'elles sont différentes pour (bxr, ga).

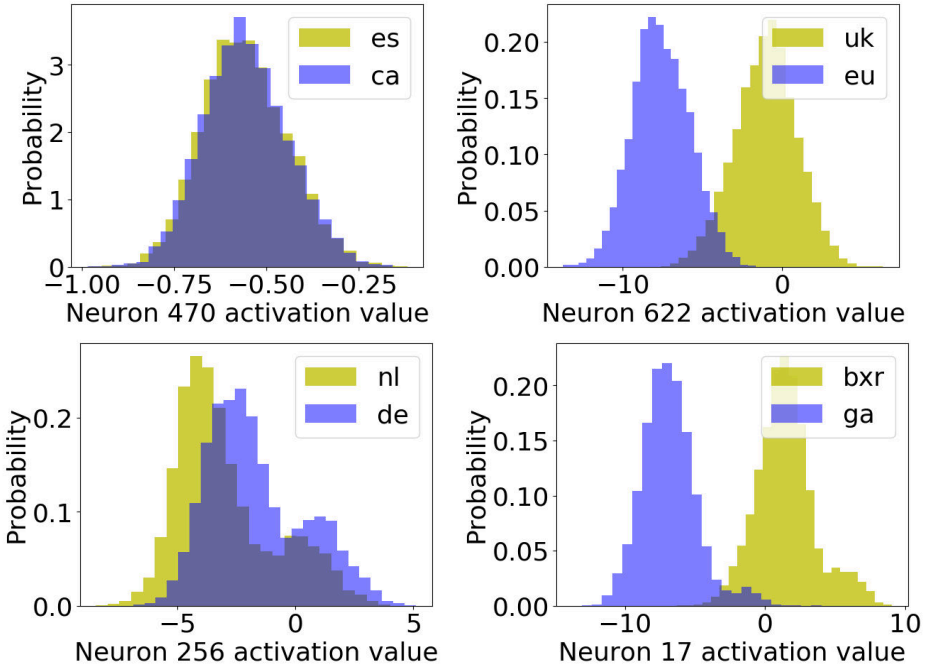


FIGURE 2.3. – Distributions d'activation pour le neurone avec la JSD la plus élevée sur *Multi* pour les paires (es, ca), (uk, eu), (nl, de) et (bxr, ga).

Il est difficile de tirer des conclusions définitives des résultats obtenus à partir de l'analyse du modèle. La première analyse a montré que l'analyseur a accès à l'identité de la langue lors de la prédiction des actions d'analyse. Cependant, nous ne savons pas à quel point il s'appuie sur ces informations. La seconde analyse a montré que l'analyseur, lorsqu'il a accès au vecteur W_{22}, tend à faire converger entre elles des langues proches, et à faire diverger des langues plus distantes. Il se peut que l'analyseur utilise l'identité du langage dans le cas où W_{22} ne contient pas suffisamment d'informations. D'autres analyses doivent être effectuées pour valider ou invalider cette hypothèse. Pour cela, il serait par exemple possible d'utiliser des méthodes d'*adversarial learning* (GANIN et al. 2016), qui permettraient de pénaliser l'analyseur s'il est capable d'identifier correctement la langue en cours d'analyse.

2.8. Conclusions et perspectives

Les deux grandes questions que nous nous sommes posées dans ce chapitre étaient : comment représenter les données, et comment représenter la langue ? Pour la représentation des données, nous avons adapté des corpus de UD, sur lesquels nous avons réalisé un équilibrage sur la représentation de chaque langue. Nous avons ensuite analysé des représentations possibles d'une langue dans le cadre d'une analyse en dépendances multilingue délexicalisé, et avons testé si ces représentations permettaient de partager des connaissances entre les langues.

Les meilleurs résultats sont obtenus avec les représentations apprises sur les don-

nées (W_d,W_{df}), bien qu'ils dépassaient de peu (+0.14 et +0.29 respectivement) la représentation par un simple identifiant (ID). Mais ces représentations ont l'inconvénient de ne pas être utilisables dans le cadre d'une langue pour laquelle il n'existe pas de données d'apprentissage, contrairement aux représentations utilisant le WALS. Cependant, des travaux (WANG et EISNER 2017) ont tenté de prédire W_d à partir de corpus non annotés en syntaxe, mais annotés en POS, et ce avec une précision raisonnable, même s'il est nécessaire d'adapter l'analyseur afin qu'il puisse tenir compte du bruit introduit par les prédictions. D'autres représentations de la langue aurait également pu être testées, par exemple un vecteur W_{de}, qui utiliserait l'écart type plutôt que la fréquence.

L'utilisation du vecteur du WALS ne permet cependant pas une amélioration nette des résultats pour toutes les langues lors d'un d'apprentissage de type *zero-shot*. Des analyses plus détaillées des comportements en fonction des langues seront menées dans les chapitres 4 et 5. Il est possible que la méthode de remplacements des valeurs non-renseignées soit peu efficace, et il pourrait être intéressant de tenter d'entraîner un réseau essayant de prédire les valeurs manquantes.

L'utilisation d'analyseurs délexicalisés a permis de mettre de côté le problème de la représentation du lexique dans des systèmes multilingues. Cependant, de nombreuses informations sont perdues lors de l'absence du lexique. La question de la création d'un lexique *universel* va ainsi être abordée dans le chapitre suivant.

Chapitre 3.

Lexicalisation

Sommaire

3.1 Introduction 52
3.2 MUSE 54
3.3 PanLex 55
3.4 Harmonisation 58
3.5 Évaluations monolingues 59
3.5.1 Méthodes 59
3.5.2 Résultats 62
3.6 Évaluations multilingues 66
3.6.1 Méthodes 66
3.6.2 Résultats 70
3.7 Le problème de l'harmonisation 74
3.8 Lexicalisation de l'analyseur syntaxique 78
3.8.1 Lexicalisation 83
3.9 Conclusions et perspectives 87

3.1. Introduction

Bien que l'utilisation de plongements de mots contextuels soit désormais la norme, en particulier depuis l'apparition de BERT (DEVLIN et al. 2019), ce n'était pas le cas au début de cette thèse. Lorsque la question de l'utilisation de plongements de mots s'est posée pour les travaux de ce chapitre, ce type de plongements de mots faisaient seulement leurs débuts, et le choix des plongements de mots non-contextuels classiquement utilisés à l'époque s'est fait. Bien qu'il soit évidemment très intéressant de tester cette nouvelle approche, il n'a pas été possible de le faire à temps.

Les expériences réalisées dans le chapitre 2 sur l'analyse syntaxique délexicalisée se plaçaient dans un contexte difficile pour l'analyseur, où les seules informations fournies en entrée étaient les POS et les traits morphologiques. Le choix de la délexicalisation était plus simple pour nous, puisqu'il suffisait d'utiliser les données des UD, déjà disponibles à un niveau de représentation universel pour toutes les langues. L'utilisation du lexique semble cependant essentielle, et nous avons fait le choix pour cela d'utiliser des plongements de mots. La création de plongements de mots semble

raisonnable, même pour des langues n'ayant pas ou peu de données annotées, les seules ressources nécessaires pour obtenir de tels plongements étant de grands corpus de textes non annotés.

Il existe plusieurs possibilités pour obtenir des plongements de mots multilingues. Tout d'abord, à la manière de AMMAR, MULCAIRE, TSVETKOV et al. 2016, il est possible de prendre des plongements de mots monolingues, puis de les projeter, les aligner dans un espace vectoriel commun, à l'aide de dictionnaires, de données parallèles ou de méthodes non supervisées. Certains plongements de mots multilingues peuvent être appris à partir de textes alignés pour la traduction automatique (HERMANN et BLUNSOM 2014). Il est également possible d'apprendre des plongements de mots multilingues directement, à la manière de M-BERT (DEVLIN et al. 2019), en apprenant des plongements de mots sur de grand corpus de textes non annotés issus de plusieurs langues.

Les espaces vectoriels des plongements de mots présentent des similarités dans leurs structures d'une langue à l'autre (MIKOLOV et al. 2013). Autrement dit, l'arrangement des mots dans l'espace vectoriel est similaire entre les langues, les distances d'un mot à l'autre étant similaires [1] : les positions relatives sont les mêmes, mais pas les positions absolues (voir figure 3.1). Le mot *chat* est proche du mot *chien*, lui-même proche du mot *loup*, mais ce dernier est plus éloigné du mot *chat*, etc. Et ces proximités restent similaires pour les traductions de ces mots dans d'autres langues.

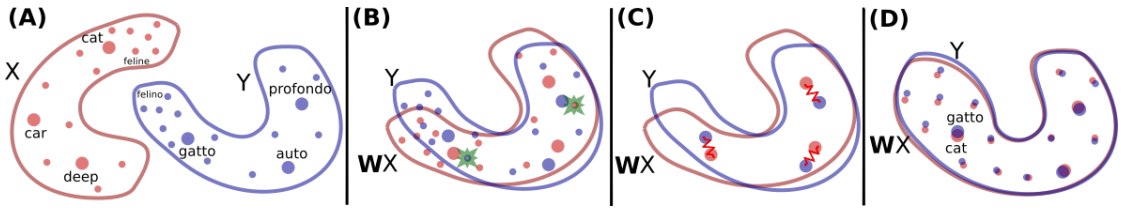


FIGURE 3.1. – Figure issue de CONNEAU et al. 2017, donnant l'intuition de la méthode d'alignement de MUSE.

Ainsi, il est possible, à partir de plongements de mots monolingues, d'aligner ces plongements de mots dans un espace commun. Dans ces travaux, cet espace commun sera celui de l'anglais (en) qui sera la langue pivot. Ce choix de l'anglais (en) est uniquement motivé par la disponibilité de plus de dictionnaires utilisant cette langue.

Pour cela, on utilisera l'outil MUSE [2] (CONNEAU et al. 2017), un logiciel permettant d'aligner deux espaces de plongements de mots de deux langues différentes, de manière supervisée ou non-supervisée, à l'aide de la méthode Procrustes.

On utilisera également les plongements de mots pré-entraînés de FastText [3] (BOJANOWSKI et al. 2017). Nous n'avons pas utilisé leurs plongements de mots déjà

1. Dans un monde parfait où la quantité de corpus d'entraînement pour les plongements de mots est infinie.

2. https://github.com/facebookresearch/MUSE

3. https://fasttext.cc/docs/en/pretrained-vectors.html. Ces plongements de mots sont de taille 300, et sont tous en minuscules.

alignés avec MUSE, car l'intégralité de nos langues n'était pas disponible pour ces plongements alignés dans un espace vectoriel commun.

De nombreux chercheurs se sont intéressés aux méthodes d'obtention et d'évaluation des plongements de mots, et des plongements de mots multilingues (AMMAR, MULCAIRE, TSVETKOV et al. 2016 ; FARUQUI et al. 2016 ; GLAVAŠ et al. 2019 ; JASTRZEBSKI et al. 2017 ; RUDER et al. 2019 ; STRINGHAM et IZBICKI 2020). Bien qu'aucune méthode d'évaluation ne soit parfaite, nous allons tenter d'exploiter les avantages de chacune afin de tirer des conclusions sur la qualité des plongements de mots obtenus.

Dans ce chapitre, nous allons donc nous intéresser au fonctionnement de MUSE pour aligner les plongements de mots (section 3.2). Nous aurons besoin de dictionnaires bilingues pour réaliser les alignements. Or, nous ne disposons pas de dictionnaires pour toutes nos langues. Nous verrons dans la section 3.3 comment mettre à profit la ressource PanLex (KAMHOLZ et al. 2014) pour en extraire des dictionnaires. Une étape d'harmonisation des vecteurs des mots homographes a été réalisée, et sera détaillée dans la section 3.4. Nous ferons ensuite le tour des différentes méthodes d'évaluations monolingues et leurs résultats sur nos données dans la section 3.5, afin de voir l'évolution des résultats avant alignement, après alignement, et après harmonisation. L'évaluation multilingue sera quant à elle abordée dans la section 3.6. Nous verrons que l'harmonisation a un fort impact négatif sur l'évaluation des plongements de mots. Nous verrons dans la section 3.7 dans quelle mesure cette harmonisation impacte la suite des expériences. Nous lexicaliserons l'analyseur syntaxique afin de vérifier l'impact des plongements de mots dans le système. Enfin, nous présenterons dans la section 3.9 les conclusions de ce chapitre, ainsi que différentes améliorations qu'il aurait été possible de mettre en place pour augmenter la qualité des plongements de mots multilingues.

3.2. MUSE

MUSE est une librairie qui vise à réaliser un alignement entre les espaces de plongements de mots de deux langues différentes. Cette librairie possède deux modes : supervisé et non supervisé. CONNEAU et al. 2017 ont montré que leurs résultats en apprentissage non-supervisé était proches, voire meilleurs que certains apprentissages supervisés pour des langues similaires. Cependant, cette conclusion est moins vraie pour les langues éloignées. Étant donné la variabilité des langues utilisées dans ces travaux, nous avons fait le choix de nous concentrer sur la méthode supervisée.

Cette méthode requiert simplement un dictionnaire de traduction entre deux langues dont on cherche à aligner les lexiques. Les auteurs ont montré qu'un dictionnaire de taille 1 000 obtenait déjà des résultats d'une qualité raisonnable, même si de meilleurs résultats sont obtenus sur des dictionnaires de taille 5 000. Les mots de ces dictionnaires serviront de base pour calculer la transformation entre les espaces. À l'issu de la transformation, tous les mots auront une version alignée, même ceux ne faisant pas partie du dictionnaire. Les auteurs fournissent les dictionnaires qu'ils ont utilisés : ces dictionnaires sont générés en utilisant les mots les plus fréquents de

la langue source et sont traduits vers la langue cible, le tout en gérant les éventuelles traductions multiples des mots. Par exemple, le mot *the* en anglais (en) aura trois traductions en français : *le*, *la* et *les*. Il y aura alors trois entrées dans le dictionnaire : *the-le*, *the-la* et *the-les*.

MUSE calcule une transformation linéaire de l'espace de plongements de mots source afin de minimiser la distance entre les vecteurs des mots traduits fournis par le dictionnaire. Ainsi, il conserve la structure initiale des espaces vectoriels, ne changeant pas les positions relatives entre les mots d'un même espace. MUSE est basé sur la méthode Procrustes.

Dans l'ensemble des lexiques produits dans ce chapitre, les paramètres par défaut de MUSE ont été utilisés.

3.3. PanLex

Panlex[4] (KAMHOLZ et al. 2014) est une base de données lexicale conçue pour gérer des traductions "panlingues", c'est à dire des traductions de toutes les langues vers toutes les autres langues. PanLex a été créé en assemblant des ressources gratuites et libres.

PanLex dispose de ressources pour plus de 5 700 langues. Cette grande variété de langues disponibles nous permet d'extraire des dictionnaires pour les langues ayant peu de données et de ressources à leur disposition. À l'aide de PanLex, des dictionnaires bilingues ont pu être créés pour toutes nos langues vers l'anglais (en).

Pour un certain nombre de langues, des dictionnaires étaient déjà fournis. Ce sont les dictionnaires utilisés pour l'alignement de leur plongement de mots en utilisant MUSE (CONNEAU et al. 2017). Ces dictionnaires ont été créés par un outil de traduction interne aux développeurs de MUSE. C'est en se calquant sur leur jeu de données de dictionnaires que nous avons fait le choix de garder 5 000 et 1 500 mots pour créer des dictionnaires d'entraînement et de test. Ainsi nous pouvions avoir des jeux de données homogènes.

Présentation Le système de PanLex, contrairement à beaucoup d'autres (Y. KIM et al. 2019), ne se base pas sur une langue pivot. Lorsque de nombreuses langues doivent être traduites, il est difficile d'apprendre un système de traduction pour chaque paire de langues, car cela nécessiterait beaucoup de modèles : si n est le nombre de langues, il faudrait $n * (n-1)$ modèles. De plus, cela suppose l'existence de données parallèles ou des dictionnaires pour chaque paire de langue, ce qui est impossible à trouver pour de nombreuses paires de langues. La solution de la langue pivot permet de simplifier ces problèmes.

L'utilisation d'une langue pivot consiste à apprendre des système de traduction pour chaque paire de langues comprenant la langue pivot, qu'on appelle P. Ainsi, pour traduire un mot de la langue $L1$ vers la langue $L2$, il suffit de traduire le mot de

4. http://panlex.org

$L1$ vers P, puis de P vers $L2$. Il suffit alors d'apprendre $2 * (n-1)$ modèles, allant de P vers toutes les langues, et de toutes les langues vers P.

Cette méthode présente cependant des défauts, en particulier celui de perdre les spécificités de certaines langues si ces spécificités sont inexistantes dans la langue pivot.

Par exemple, le plus souvent, l'anglais (en) est utilisé comme langue pivot, puisque c'est une des langues pour lesquelles le plus de données sont disponibles. Mais si on souhaite traduire le pronom français (fr) *vous* en russe (ru) (dont la bonne traduction est Вы), en passant par l'anglais (en), il est fort probable qu'on arrive au pronom ты, soit l'équivalent du pronom *tu* en français (fr). Cela vient du fait que *vous*, lors de la traduction en anglais (en), deviendra *you*. Or, *you* peut signifier à la fois *tu* et *vous*, mais c'est la forme *tu* qui est la plus fréquente. Lors de la traduction de l'anglais (en) vers le russe (ru), cette ambiguïté mènera à une erreur de traduction.

La méthode de traduction de PanLex garde cette idée de pivot, sans pour autant utiliser une unique langue pivot. Dans sa base de données, PanLex a accès à la traduction de nombreux mots entre de nombreuses langues. Pour une langue source S, une langue cible T et un mot m, les traductions de ce mot pour toutes les langues disponibles depuis la langue S vont être extraites, puis pour chacun de ces mots, une traduction vers la langue cible T va être cherchée. Ainsi, plusieurs "chemins" sont trouvés pour aller de S à T pour le mot m, et plusieurs traductions seront trouvées, formant ainsi un graphe de traduction (voir figure 3.2). Les traductions sont alors pondérées par le nombre de chemins. Il semble y avoir une pondération sur les chemins en fonction de leur fiabilité, mais il n'a pas été possible de trouver une source confirmant cette intuition. Nous allons maintenant voir comment nous allons utiliser ces traductions pour en extraire des dictionnaires.

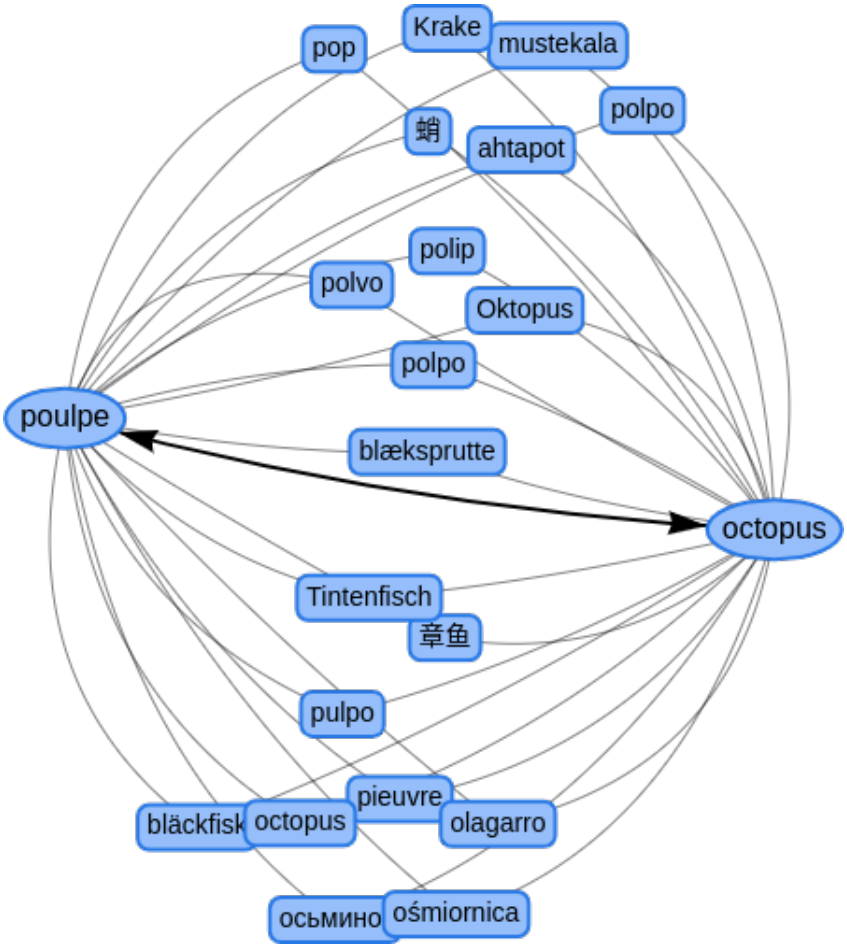


FIGURE 3.2. – Exemple de fonctionnement de PanLex pour la traduction du mot français *poulpe* vers l'anglais (figure issue de https://translate.panlex.org/?langDe=fra-000&langAl=eng-000&txt=poulpe)

Utilisation La version utilisée pour nos expériences est celle datant du 1er juillet 2019[5].

De ces graphes, des *sens* sont extraits et associés aux mots des différentes langues. Tous les mots pour une langue L sont extraits de PanLex. Chaque mot m a un (ou plusieurs) *sens*, sous forme d'un numéro d'identification, qui lui est associé. On parcourt ensuite la liste des mots de l'anglais, et si un mot t a le même *sens* que m, alors la paire t-m est ajoutée au dictionnaire $Anglais$-L. Il peut y avoir des mots apparaissant plusieurs fois pour chaque langue. Par exemple, pour le dictionnaire $Francais$-$Anglais$, on trouve :

avocat	lawyer
avocat	avocado
avocate	lawyer

5. https://db.panlex.org/panlex-20190701-csv.zip

Après cette étape, les mots sont triés par fréquence pour la langue L, si un corpus d'entraînement de cette langue est disponible. Si ce n'est pas le cas, les mots restent mélangés. On garde ensuite les 5 000 premiers mots pour créer le dictionnaire qui servira à l'alignement des plongements de mots. Les 1 500 mots suivants sont utilisés pour créer un jeu de données de test, qui permettra d'évaluer les plongements de mots que nous utilisons (voir section 3.5.2).

Les langues L ayant un dictionnaire $Anglais$-L fourni par MUSE[6] utiliseront ce dictionnaire pour l'alignement des plongements de mots. Les langues n'en n'ayant pas utiliseront les dictionnaires extraits de PanLex.

Les langues pour lesquelles nous avons généré des dictionnaires grâce à Panlex sont : le bouriate (bxr), le vieux slave (cu), l'irlandais (ga), le galicien (gl), le gothic (got), le kazahk (kk), le kurmandji (kmr), le norvégien nynorsk (nno), le norvégien bokmål (nob), le same du nord (sme), le ouïghour (ug), l'ourdu (ur) et le basque (eu).

3.4. Harmonisation

L'utilisation de MUSE permet d'obtenir des plongements de mots multilingues, alignés dans l'espace vectoriel de l'anglais pour toutes les langues. Cependant, pour des raisons techniques, nous souhaitons anonymiser l'origine de la langue, et ainsi avoir une unique entrée dans le lexique qui soit indépendante de la langue à laquelle le mot appartient. Cette étape, qu'on appelle l'étape d'*harmonisation* pose néanmoins un problème dans le cas des mots homographes, deux mots possédant la même forme dans deux langues (i.e. le mot *weekend* en français et en anglais).

Nous sommes partis de l'hypothèse que des homographes ont des sens proches. Après alignement des plongements de mots dans l'espace de l'anglais, ces mots devraient donc déjà avoir un vecteur proche. Il suffirait alors de faire la moyenne des vecteurs de ces mots pour toutes les langues dans lesquels ils apparaissent. Cette étape d'harmonisation permet ainsi de rendre les mots indépendants de la langue d'origine.

Afin de vérifier la proportion d'homographes ayant au moins un sens commun entre les deux langues, nous avons extrait la liste des homographes des corpus d'entraînement de l'anglais (en) et du français (fr). Nous avons éliminé tous les mots correspondant à de la ponctuation, à des chiffres et à des noms propres, qui biaiseraient les résultats puisque ayant évidemment un sens en commun. Nous avons également éliminé les mots d'une langue apparaissant dans l'autre langue uniquement sous forme de citation (par exemple, *"Le groupe de musique The Cranberries..."*)

Nous avons ensuite annoté manuellement les homographes selon l'existence ou non d'au moins un sens commun de l'homographe entre les deux langues. Sur les

6. Les langues ayant un dictionnaire fourni par MUSE sont : l'arabe (ar), le bulgare (bg), le catalan (ca), le chinois (zx), le croate (hr), le tchèque (cs), le danois (da), le néerlandais (nl), l'estonien (et), le finnois (fi), le français (fr), l'allemand (de), le grec (el), l'hébreu (he), l'hindi (hi), le hongrois (hu), l'indonésien (id), l'italien (it), le japonais (ja), le coréen (ko), le letton (lv), le perse (fa), le ponolais (pl), le portugais (pt), le roumain (ro), le russe (ru), le slovaque (sk), le slovène (sl), l'espagnol (es), le suédois (sv), le turc (tr), l'ukrainien (uk) et le vietnamien (vi).

225 mots de ce jeu de données, 202 ont effectivement un sens commun, contre 23 n'en n'ayant pas, soit environ 90% des homographes allant bien dans le sens de notre hypothèse pour la paire de langue fr-en.

Cependant, nous verrons dans la suite de ce chapitre que le choix de l'harmonisation des plongements de mots était préjudiciable. Certains mots comme *a*, une conjugaison du verbe avoir en français (fr), peuvent avoir un tout autre rôle syntaxique dans une autre langue, par exemple en anglais (en) où *a* est un déterminant. Toutefois, nous verrons que ce choix ne devrait pas invalider les conclusions des chapitres suivants (voir section 3.7).

3.5. Évaluations monolingues

Nous nous intéresserons aux méthodes d'évaluations monolingues afin d'évaluer la qualité des plongements de mots avant alignement, après alignement, et après harmonisation afin d'estimer l'impact de ces modifications sur la qualité intrinsèque des plongements de mots que nous utilisons.

3.5.1. Méthodes

La première étape est l'évaluation des plongements de mots monolingues issus de FastText. Autrement dit, on veut vérifier que les plongements de mots de départ permettent de rapprocher des mots dont les sens sont proches : les vecteurs des mots *poulpe* et *pieuvre* doivent être plus proches l'un de l'autre qu'ils le seraient du mot *caillou* par exemple.

Similarité de mots Une manière classique consiste à recourir à la similarité. L'utilisation de jeux de données de similarités de mots est également fréquente pour ce type d'évaluation, et pourrait être adaptée à notre problème, puisque plusieurs de jeux de données correspondant à nos langues sont disponibles. Les jeux de données de similarités sont des listes de paires de mots annotées par des humains. Pour chaque paire, les annotateurs doivent noter, selon une échelle [7], à quel point ils considèrent que les mots sont similaires. La moyenne des annotateurs est faite, et on obtient un score de similarité pour chaque paire de mots. On mesure alors la distance des vecteurs des mots de la paire pour les plongements de mots à évaluer à l'aide de leur similarité cosinus (l'angle entre les deux vecteurs), et on vérifie si cette distance est corrélée à la référence humaine.

L'évaluation de similarité consiste donc en l'utilisation de jeux de données associant un score de similarité d'après des humains, $sim(w_i)$, pour deux mots w_{i_1} et w_{i_2}. La similarité cosinus entre les vecteurs des mots w_{i_1} et w_{i_2} est calculée telle que :

$$simCos(w_i) = \frac{pdm(w_{i_1}) \cdot pdm(w_{i_2})}{\|pdm(w_{i_1})\| \|pdm(w_{i_2})\|}$$

7. Allant parfois de 0 à 6, parfois de 0 à 10.

où $pdm(w_{i_1})$ est le vecteur du mot w_{i_1}.

On peut alors calculer la corrélation de Spearman entre sim et $simCos$ pour estimer si les plongements de mots rapprochent les mêmes mots que des annotateurs humains, donnant ainsi un score à nos plongements de mots pour un jeu de données de similarité donné.

Trois jeux de données de similarité, disponibles pour plusieurs langues, sont utilisés dans ce chapitre :

— WS-353 (FINKELSTEIN et al. 2001) est un jeu de données un peu ancien pour lequel les processus d'annotations n'étaient pas encore très bien rodés. Il contient 353 paires de mots, annotés par 13 à 16 annotateurs pour chaque paire utilisant une échelle allant de 0 à 10. Il existe pour 12 langues : arabe (ar), allemand (de), anglais (en), espagnol (es), persan (fa), français (fr), italien (it), néerlandais (nl), portugais (pt), russe (ru), suédois (sv), chinois (zh)

— Simlex (HILL et al. 2015), est un corpus contenant des mots très fréquents et est surtout constitué de noms (666 paires de noms, 222 paires de verbes et 111 adjectifs). Il contient 999 paires de mots en tout et existe pour 7 langues : allemand (de), anglais (en), espagnol (es), français (fr), italien (it), néerlandais (nl), suédois (sv). Bien que le jeu de données de certaines langues soit correctement annoté par des humains (50 annotateurs pour chaque paire sur une échelle de 0 à 6), certaines langues sont de simples traductions brutes de l'anglais où le score de la paire de l'anglais a été conservé.

— Multisimlex (VULIĆ et al. 2020) est le jeu de données le plus récent que nous ayons. Les jeux de données de Simlex sont inclus dans Multisimlex. 334 paires sont issues de SemEval-17 (CAMACHO-COLLADOS et al. 2017), 67 paires de CARD-660 (PILEHVAR et al. 2018), 224 verbes proviennent de GERZ et al. 2016 et 122 paires d'adjectifs et d'adverbes sont issues de de la base de données de l'Université de South Florida (NELSON et al. 2004). En tout, il y a 1888 paires de mots pour 10 langues : arabe (ar), anglais (en), espagnol (es), estonien (et), finnois (fi), français (fr), hébreu (he), polonais (pl), russe (ru), chinois (zh). Chaque paire a été annotée par au moins 10 annotateurs avec une échelle allant de 0 à 6.

Langues peu dotées : évaluation par regroupement sémantique Un problème de taille persiste : l'évaluation des plongements de mots des langues n'ayant pas de ressources pour l'évaluation par similarité. Seulement 16 de nos langues ont pu être évaluées, car disposant d'un jeu de données de similarité. STRINGHAM et IZBICKI 2020 proposent d'utiliser la ressource Wikidata, une ressource multilingue disponible pour de nombreuses langues, définissant des relations sémantiques entre les mots. Wikidata est un graphe de connaissance multilingue (disposant de données pour 581 langues, dont 43 de nos 49 langues) et géré par la *Wikimedia Foundation*. Cette ressource contient des *objets* pouvant représenter des concepts, des objets matériels, etc. Ces objets peuvent être par exemple *poulpe*, *joie* ou encore *Marseille*. Des relations

sémantiques entre ces objets sont définies, et peuvent correspondre à l'appartenance à une catégorie, le fait que l'objet soit une instance d'une catégorie ou d'un autre objet, etc. Par exemple, l'objet *Paris* est une instance de *capitale*. Grâce à cette propriété, il est possible de générer des catégories telles que *capitale*, *religion*, *devise*, *sport*, etc., qui contiennent des listes de mots appartenant à la catégorie.

Exemples de catégories existantes dans Wikidata et d'objets appartenant à la catégorie :

— Religion : Bouddhisme, Christianisme, Hindouisme, Islam

— Émotion négative : colère, dégoût, peur, tristesse

— Sport : judo, football, rugby, handball, course à pied

Nous allons alors pouvoir créer des jeux de données de catégories sémantiques que nous utiliserons pour évaluer la qualité de nos plongements de mots monolingues. Le jeu de données de test généré pour chaque langue est une liste de 5 catégories contenant une liste de mots appartenant à la catégorie. Ces catégories sont *capitale*, *ancienne ville*, *religion*, *pays*, *élément chimique*. La raison de ce choix est que ces catégories existent dans la base de données pour la plupart des langues en quantité relativement importante. Les deux méthodes d'évaluations proposées par STRINGHAM et IZBICKI 2020, Topk et OddOneOut (OOO), permettent ainsi d'évaluer si les catégories forment des clusters, des regroupements, dans l'espace des plongements de mots.

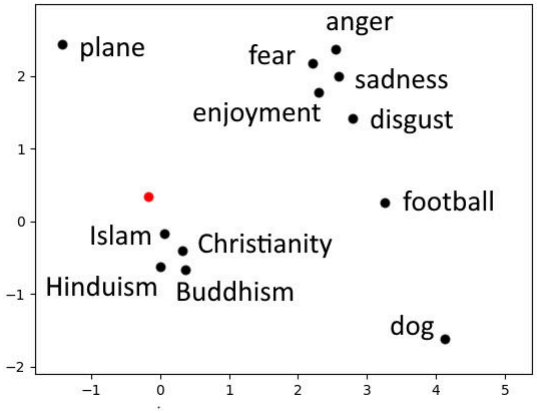


FIGURE 3.3. – Projection en 2D des plongements de mots appartenant à la catégorie décrite. Le point rouge représente la moyenne de l'ensemble {Islam, Christianity, Buddhism, plane} utilisé par l'évaluation OOO.

Topk est une évaluation qui évalue si les plongements de mots sont regroupés en clusters de catégories. Pour cela, cette évaluation compte, parmi les k plus proches voisins d'un mot, le nombre de mots d'une même catégorie. Pour $k = 3$ et pour les mots de la figure 3.3, les 3 plus proches voisins de *Buddhism* (bouddhisme) sont *Christianity* (christianisme), *Hinduism* (hindouisme) et *Islam* (islam), qui appartiennent à la même catégorie que *Buddhism*; le score pour ce mot est donc de 1. Cette opération est

répétée pour chaque mot de la catégorie et obtenons à chaque fois un score de 1. La moyenne du Topk pour chaque mot de la catégorie *Religion* est 1. On peut passer à la catégorie suivante : *Emotion négative*. Les 3 voisins les plus proches de *anger* (colère) sont *fear* (peur), *sadness* (tristesse) et *enjoyment* (plaisir). Le mot *plaisir* n'appartient pas à *Émotion Négative*, le score pour le mot *anger* est de 2/3. Cette opération est répétée pour chaque catégorie, puis le score TopK est la moyenne des scores de toutes les catégories.

OddOneOut (OOO) est une évaluation qui consiste à trouver l'intrus parmi un ensemble de k+1 mots : k mots appartiennent à la catégorie et un mot n'y appartient pas. On calcule alors la moyenne des vecteurs des mots de l'ensemble (incluant l'intrus). Le mot avec le vecteur le plus éloigné de l'ensemble est considéré comme l'intrus. Si l'intrus est bien le mot qui n'appartient pas à la catégorie, le score pour OOO sur cet ensemble est 1, sinon 0.

Par exemple, soit l'ensemble {Islam, Christianity, Buddhism, plane (avion)}. On calcule la moyenne de l'ensemble et on obtient le point rouge visible sur la figure 3.3. Comme *plane* est le mot le plus éloigné de la moyenne, il est considéré comme "l'intrus", à raison, puisqu'il n'appartient pas à la même catégorie que les autres mots.

Topk et OOO ne sont pas des méthodes d'évaluations standards, mais permettent tout de même de pallier le manque de jeu de données d'évaluation pour la plupart de nos langues. De plus, l'hypothèse de base est que les plongements de mots sont regroupés par clusters sémantiques, ce qui n'est pas une certitude. Bien que OOO soit une méthode assez élégante, Topk peut sembler un peu trop stricte, et n'évalue pas nécessairement ce qui nous intéresse. En effet, même si les plongements de mots forment des clusters sémantiques, Topk ne tolère pas qu'un mot d'une autre catégorie soit trop proche du cluster. Nous verrons d'ailleurs dans la section 3.6.2 que les résultats de Topk sont tellement bas qu'il n'est pas vraiment possible de tirer de conclusions sur cette évaluation.

3.5.2. Résultats

Les plongements de mots issus de FastText constituent des représentations fiables, ayant prouvé à plusieurs reprises être une base solide (GRAVE et al. 2018). Nous allons commencer par faire une première évaluation de ces plongements de mots afin d'obtenir un score qui nous servira de référence et nous permettra de comparer l'évolution de la qualité des représentations que nous utilisons après alignement et harmonisation.

Les résultats de l'évaluation par similarité sont disponibles dans le tableau 3.1, et les résultats des évaluations Topk et OOO sont disponibles en annexe, tableaux .4 et .5. Les scores de similarité sont comparés à des résultats d'autres articles évaluant des plongements de mots avec les trois jeux de données que nous utilisons, et nous avons gardé les meilleurs scores pour chaque langue. Ces scores nous permettront de situer la qualité des plongements de mots que nous utilisons par rapport à des plongements de mots de l'état de l'art. Ces scores sont issus de VULIĆ et al. 2020 pour Multisimlex,

Lang.	Score	EDLA	OOV	Taille
Multisimlex				
ar	0.44	-	9	1632
en	0.44	0.54	1	1888
es	0.43	0.54	1	1793
et	0.39	0.45	13	1794
fi	0.57	0.65	6	1803
fr	0.47	0.61	6	1888
he	0.43	0.51	2	1736
pl	0.37	0.47	3	1756
ru	0.36	0.49	2	1832
zh	0.31	0.58	23	1888
moyenne	0.42	0.54	7	1801
Simlex				
de	0.26	0.25	1	999
en	0.30	0.38	0	999
es	0.30	0.29	1	984
fr	0.29	0.20	1	908
it	0.28	0.26	2	999
nl	0.30	0.23	1	989
sv	0.33	0.27	1	959
moyenne	0.30	0.27	1	977

Lang.	Score	EDLA	OOV	Taille
WS-353				
ar	0.57	0.57	2	353
de	0.67	0.78	1	353
en	0.74	0.76	0	353
es	0.59	0.79	1	353
fa	0.67	0.43	2	323
fr	0.57	0.85	1	343
it	0.58	0.76	1	349
nl	0.63	0.85	0	353
pt	0.60	0.81	2	353
ru	0.59	0.76	1	340
sv	0.60	0.78	0	353
zh	0.30	-	20	353
moyenne	0.59	0.74	1	348

TABLEAU 3.1. – Score de similarité avant alignement. OOV est le pourcentage de mots hors vocabulaire parmi nos plongements de mots. Des scores de l'état de l'art sont fournis (EDLA) ainsi que la taille des jeux de données (Taille).

BARZEGAR et al. 2018 pour Simlex et WS-353 et FREITAS et al. 2016 pour WS-353 [8].

Selon le jeu de données utilisé pour l'évaluation, les résultats peuvent varier de façon importante. Avec une moyenne de 0.59, le jeu de données WS-353 obtient les résultats les plus élevés, suivi par Multisimlex avec 0.42, puis Simlex avec 0.30. Les résultats que nous obtenons sont légèrement inférieurs aux meilleurs scores de l'état de l'art (0.15 point d'écart en moyenne pour WS-353 et 0.12 pour Multisimlex). Ils sont cependant meilleurs de 0.03 point pour le jeu de données de Simlex.

3 langues apparaissent dans les 3 jeux de données : l'anglais (en), l'espagnol (es) et le français (fr). Pour Multisimlex, les scores de ces trois langues sont très proches, respectivement 0.44, 0.43 et 0.47. Avec le jeu de données de WS-353, les résultats sont très différents : l'anglais (en) obtient les meilleurs résultats (0.74), devançant le français (fr) (0.57), précédé de près par l'espagnol (es) (0.59).

Avec Simlex les résultats sont bien plus uniformes : 0.30 (en), 0.30 (es) et 0.29 (fr). Si l'on prend en compte toutes les langues, les résultats varient seulement de 0.26 à 0.33, soit un écart de 0.07 point.

Les résultats varient donc beaucoup en fonction du jeu de données, ce qui fait ressortir les inégalités entre les évaluations des langues. Le faible nombre de langues rend difficile l'obtention d'une p-value significative lors du calcul de corrélation entre

8. Des langues ont été ajoutées à certains jeux de données après la publication des articles nous servant de références, expliquant l'absence de scores de référence dans ces cas-là.

les scores des jeux de données. Les scores des corrélations de Pearson entre les 3 paires de jeu de données possibles ne permettent pas de mettre en évidence une corrélation entre les scores de ces jeux de données.

Les résultats des évaluations de Topk et OOO obtiennent également des scores très différents. Alors que Topk obtient des résultats très bas (0.06 en moyenne pour toutes les langues et toutes les catégories), OOO obtient un score de 0.49 en moyenne. Les scores de Topk étant si bas, il est assez difficile de tirer des conclusions sur la qualité des plongements de mots avec cette méthode d'évaluation. Pour OOO, les scores varient beaucoup d'une langue à l'autre : de 0.03 pour le vieux slave (cu) à 0.70 pour le bulgare (bg).

Une corrélation a pu être établie entre l'évaluation OOO et WS-353 avec un score de corrélation de Pearson de 0.80. On peut alors supposer que l'évaluation OOO devrait permettre une bonne approximation de l'évaluation des plongements de mots des langues peu dotées par rapport au jeu de données WS-353.

Les méthodes d'évaluation ne sont pas parfaites et ne permettent pas de donner un score interprétable sur la qualité des plongements de mots. La variabilité des scores selon le jeu de données, et la méthode d'évaluation mettent en évidence la difficulté de cette tâche. C'est pourquoi ces scores seront surtout utilisés pour comparer les étapes de transformations des plongements de mots.

Après alignement L'alignement des espaces vectoriels des plongements de mots que nous utilisons dans un espace commun a-t-il eu un impact sur la qualité de ces représentations ? La question peut se poser, mais MUSE réalisant une simple transformation linéaire, il ne devrait en théorie n'y avoir aucun impact, la structure initiale des espaces étant conservée.

Les scores de similarité, Topk et OOO ont donc été recalculés après alignement, et aucune modification des résultats n'a été constatée, confirmant que l'alignement n'a eu aucun impact sur la qualité des plongements de mots d'un point de vue monolingue.

Après harmonisation L'hypothèse de départ sur l'harmonisation des plongements de mots par le moyennage était que deux mots de formes identiques entre deux langues doivent avoir un sens proche, et donc un vecteur similaire. 90% d'homographes de nos corpus d'entraînement du français (fr) et de l'anglais (en) ont effectivement au moins un sens identique, allant dans le sens de cette hypothèse. Cependant, on constate pourtant que la qualité de nos plongements de mots semble fortement dégradée après cette étape d'harmonisation. Si l'on observe le résumé des scores dans le tableau 3.3 (les résultats détaillés sont dans le tableau 3.2), on constate une baisse très importante des scores de similarité : -0.19 pour Multisimlex, -0.34 pour WS-353, et des résultats presque égaux à zéro pour Simlex!

L'harmonisation a eu un fort impact sur la qualité de nos plongements de mots multilingues. Toutes les langues sont impactées, certaines plus que d'autres : avec Multisimlex, le français (fr) voit ses résultats chuter de 0.33 points, alors que l'hébreu

Lang.	Alignés	Moyennés	EDLA.	Lang.	Alignés	Moyennés	EDLA
Multisimlex				WS-353			
ar	0.44	0.36	-	ar	0.57	0.36	0.57
en	0.44	0.17	0.54	de	0.67	0.16	0.78
es	0.43	0.13	0.54	en	0.74	0.44	0.76
et	0.39	0.22	0.45	es	0.59	0.24	0.79
fi	0.57	0.31	0.65	fa	0.67	0.44	0.43
fr	0.47	0.15	0.61	fr	0.57	0.23	0.85
he	0.43	0.38	0.51	it	0.58	0.19	0.76
pl	0.37	0.19	0.47	nl	0.63	0.29	0.85
ru	0.36	0.26	0.49	pt	0.60	0.17	0.81
zh	0.31	0.14	0.58	ru	0.59	0.24	0.76
moyenne	0.42	0.23	0.54	sv	0.60	0.24	0.78
Simlex				zh	0.30	0.02	-
de	0.26	-0.04	0.25	moyenne	0.59	0.25	0.74
en	0.30	0.10	0.38				
es	0.30	0.04	0.29				
fr	0.29	0.06	0.20				
it	0.28	-0.09	0.26				
nl	0.30	-0.04	0.23				
sv	0.33	0.06	0.27				
moyenne	0.30	0.01	0.27				

TABLEAU 3.2. – Score de similarité après alignement et après harmonisation. Des scores de l'état de l'art sont fournis.

Expé.	Multisimlex	Simlex	WS-353	Topk	OOO
Avant align.	0.42	0.30	0.59	0.06	0.49
Moyennés	0.23	0.01	0.25	0.12	0.47
EDLA	0.54	0.25	0.53	-	-

TABLEAU 3.3. – Résumé des scores de similarité (en moyenne sur toutes les langues disponibles pour chaque jeu de données) et des scores OOO et Topk. Les résultats avant alignement et après sont identiques.

(he) n'en perd que 0.05. Avec WS-353, l'arabe (ar) ne perd "que" 0.21 points, contre 0.52 pour l'allemand (de). Les résultats de Simlex étant désormais proches de 0, nous ne nous y attarderons pas.

Pour les scores OOO, la baisse est beaucoup moins importante : seulement 0.02 point. OOO étant une tâche visant à chercher l'intrus dans un groupe de mots, la faible baisse des résultats suggère que l'harmonisation, bien qu'éloignant le vecteur d'un mot de sa place de départ, il reste suffisamment proche de son emplacement d'origine pour rester proche des mots issus d'une même catégorie, et ne sera pas considéré comme un intrus. Ce résultat est rassurant, suggérant, que les vecteurs des mots ne se sont probablement pas éloignés de manière excessive de leur emplacement d'origine. Le détail des résultats de OOO après harmonisation est disponible en annexe, tableau .6.

Étonnamment, les résultats de Topk augmentent avec l'harmonisation, mais vu les faibles scores sur cette tâche, il n'est pas possible de conclure quoi que ce soit de ce résultat. Il est seulement possible d'émettre des suppositions. Par exemple, il est possible que les catégories *éléments chimiques* ou *capitales* contiennent beaucoup de mots équivalents qui sont homographes entre les langues, favorisant l'évaluation Topk. Le détail des résultats après harmonisation est disponible en annexe, tableau .7.

L'impact de l'harmonisation sur la suite de nos expériences sera exploré plus en profondeur dans la section 3.7.

3.6. Évaluations multilingues

Dans la section précédente, nous nous sommes interessés à l'évaluation monolingue des plongements de mots, c'est-à-dire, la vérification que deux mots ayant un sens proche dans une langue ont bien un vecteur proche. Nous allons à présent évaluer la qualité de l'alignement. Nous dirons que l'alignement a *convergé* si les vecteurs des mots *poulpe* en français (fr) et *octopus* en anglais (en) sont proches dans l'espace vectoriel. Nous vérifierons dans cette section que l'alignement a bien convergé. En particulier, les langues ayant toutes été alignées sur l'espace vectoriel de l'anglais (en), nous vérifierons si deux langues autres que l'anglais (en), par exemple l'allemand (de) et le français (fr), ont également convergé entre elles.

3.6.1. Méthodes

La méthode utilisée pour l'évaluation multilingue sera une évaluation de type induction de lexique bilingue – ou *Bilingual Lexicon Induction* – (BLI), à la manière de GLAVAŠ et al. 2019. BLI est une méthode d'évaluation pour estimer à quel point les mots sont proches de leur traduction en calculant la proportion de voisins les plus proches qui sont des traductions étant donné un dictionnaire bilingue. Un dictionnaire de traduction de mots entre deux langues est nécessaire pour évaluer l'alignement de deux langues avec ce type de tâche.

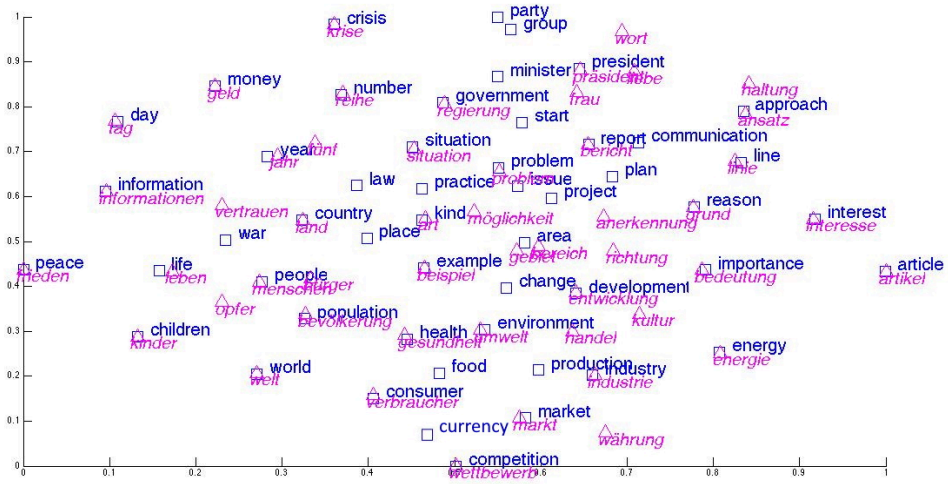


FIGURE 3.4. – Figure issue de RUDER et al. 2019 montrant les plongements de mots bilingues allemand (de) et anglais (en). La plupart des mots en anglais sont correctement alignés avec les mots allemands. Par exemple, *Kinder* est le mot le plus proche de *children* mais *Währung* n'est pas le mot le plus proche de *currency*.

Pour évaluer la proximité d'une paire de mots (w_1, w_2), traduction l'un de l'autre, appartenant respectivement à la langue $L1$ et $L2$, on identifie les plus proches voisins du mot w_1 qui appartiennent à la langue $L2$. Si le plus proche voisin de w_1 est w_2, l'alignement a bien convergé.

Pour trouver ces plus proches voisins, la distance standard utilisée pour BLI est la similarité cosinus. Nous devons ensuite calculer à quel point w_2 est proche de w_1. Pour cela, on calcule le rang de w_2 : on prend la liste des plus proches voisins de w_1 parmi les mots de la langue L_2. Le rang de w_2 est sa position dans cette liste. Par exemple, si w_2 est le voisin le plus proche de w_1, le rang est 1 ; si w_2 est le deuxième voisin le plus proche de w_1, le rang est 2.

Le score de la paire (w_1, w_2) sera le rang réciproque (*RR*), qui est l'inverse du rang. Le calcul de l'inverse du rang permet d'obtenir un score pour chaque paire qui soit compris entre 0 et 1. Ainsi, $RR = \frac{1}{Rang}$.

MRR La première métrique possible est le **rang réciproque moyen – ou *Mean Reciprocal Rank* – (*MRR*)**. Le *MRR* correspond simplement à la moyenne des scores de chaque paire du dictionnaire. Un exemple de calcul du *MRR* est donné dans le tableau 3.4.

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{Rang_{w_i}}$$

où Q est le nombre de paires de mots du dictionnaire.

w_{i_1}	w_{i_2}	Rang	RR
market	Markt	2	0.5
children	Kinder	1	1
year	Jahr	4	0.25
crisis	Krise	1	1

TABLEAU 3.4. – Exemples de paires de traductions du français vers l'allemand. Le *MRR* pour ce très petit dictionnaire est de $\frac{0.5+1+0.25+1}{4} = 0.69$.

MAP La deuxième métrique possible est la **moyenne des précisions moyennes – ou *Mean Average Precision* – (*MAP*)**. L'idée de la *MAP* est de pouvoir évaluer des traductions multiples pour un mot donné. Par exemple, si les voisins les plus proches du mot *big* sont *grand*, *gros* puis *important* : toutes ces traductions sont correctes mais le *MRR* serait de $\frac{1/1+1/2+1/3}{3} = 0.61$, pénalisant les traductions multiples. La *MAP* est une méthode d'évaluation obtenant un score de 1 dans cette situation, en évaluant correctement les mots ayant plusieurs traductions.

Faisons l'hypothèse que dans le dictionnaire, nous ayons le mot *eat* pouvant être traduit par *manger*, *mangé* et *mangerait*. Imaginons maintenant que *manger* est le plus proche voisin de *eat*, *mangé* est le $4^{ème}$ plus proche voisin et *mangerait* est le $5^{ème}$ plus proche voisin. On peut alors calculer la *précision* de chacune des traductions possibles en divisant par le rang de la traduction courante le nombre de traductions correctes déjà rencontrées.

Le rang de *manger* est 1, et une bonne traduction a été rencontrée (*manger* lui-même). La précision de *manger* est donc $\frac{1}{1}$. Le rang de *mangé* est 4, et 2 bonnes traductions ont été rencontrées (lui même, et *manger* qui était le plus proche voisin). La précision de *mangé* est donc $\frac{2}{4}$. On peut alors calculer le score du mot *eat* qui sera la moyenne des précisions – ou *Average Precision* – (*AP*) de chaque traduction :

$$AP(eat) = \frac{1/1 + 2/4 + 3/5}{3} = 0.7$$

Si on reprend le mot *big* dont les trois plus proches voisins sont des voisins corrects (*grand*, *gros* et *important*), alors $AP(big) = \frac{1/1+2/2+3/3}{3} = 1$.

La *MAP* est la moyenne des *AP* pour chacun des w_1 du dictionnaire. Il est préférable d'utiliser la *MAP* plutôt que le *MRR* lorsqu'il existe plusieurs traductions d'un même mot. Dans cette situation, le *MRR* diminuait le score alors que les prédictions sont potentiellement correctes.

Cependant, il est important de noter que l'évaluation d'un dictionnaire fr-en et de ce même dictionnaire inversé, en-fr donc, ne donnera pas les mêmes résultats. Dans la figure 3.5, le plus proche voisin anglais (en) du mot *poulpe* en français (fr) est bien le mot *octopus*. Cependant, le plus proche voisin d'*octopus* dans l'espace du français (fr) est le mot *pieuvre*. Si la seule paire présente dans le dictionnaire est *octopus-poulpe*, alors les résultats seront meilleurs dans le dictionnaire fr-en que

pour le dictionnaire en-fr. Cette situation peut se produire dans le cas d'une langue morphologiquement riche, où un même mot peut être décliné de nombreuses façons. Cette richesse morphologique peut créer un déséquilibre dans la densité des espaces de plongements de mots des langues : une langue morphologiquement riche peut avoir bien plus de mots qu'une autre langue.

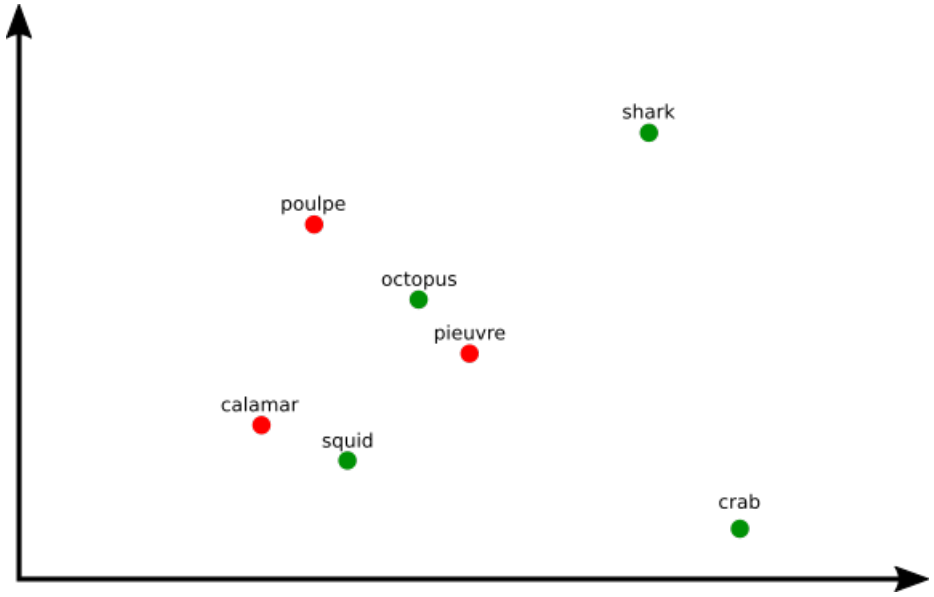


FIGURE 3.5. – Exemple de différence du plus proche voisin qui diffère entre deux langues. Les mots du français (fr) sont en rouge et ceux de l'anglais (en) sont en verts.

Nous évaluerons donc à chaque fois dans les deux sens, et garderons le meilleur résultat des deux pour chaque dictionnaire. Ce choix de garder le meilleur est motivé par l'idée qu'on cherche à estimer si nos plongements de mots sont proches dans l'espace, mais ils n'ont pas nécessairement besoin d'être les plus proches. Pour reprendre l'exemple de la figure 3.5, même si le mot *pieuvre* est plus proche de *octopus* que *poulpe*, la paire *octopus-poulpe* reste proche. Or, cette proximité nous intéresse, même si elle n'est capturée que dans un seul sens d'évaluation. C'est pourquoi nous souhaitons garder le sens qui nous permet de capter le plus de proximités, autrement dit, le meilleur sens d'évaluation pour chaque dictionnaire.

Nous utiliserons deux jeux de données d'évaluation : les dictionnaires de test de MUSE/Panlex de taille 1 500, qui ont été décrits dans la section 3.3 et les dictionnaires de GLAVAŠ et al. 2019, de taille 2 000. Ces derniers sont générés avec Google Translation et ont l'avantage d'être basés sur une même liste de mots traduite en plusieurs langues, permettant d'avoir des dictionnaire en-fr, en-de, mais aussi fr-de. Cette liste de mots étant basée sur les mêmes mots, les résultats d'une langue à l'autre sont plus facilement comparables.

3.6.2. Résultats

On cherche à présent à répondre à la question la plus importante : l'alignement de MUSE a-t-il permis d'aligner les espaces de plongements de mots correctement ? Autrement dit, le vecteur du mot *poulpe* (fr) est-il proche des vecteurs des mots *octopus* (en), *たこ* (ja), *Tintenfisch* (de) et *pulpo* (es) ? Et ces autres vecteurs sont-ils proches entre eux ?

Nous nous comparerons aux résultats de GLAVAŠ et al. 2019, les mêmes dictionnaires seront donc utilisés. Les auteurs ont utilisé des plongements de mots issus de FastText alignés en utilisant la méthode Procrustes, le même algorithme que celui utilisé par MUSE supervisé que nous avons utilisé. Il existe deux principales différences entre les travaux de GLAVAŠ et al. 2019 et le nôtre. La première est que les auteurs n'ont évalué les dictionnaires que dans un sens (en-fr mais pas fr-en), alors que nous avons exécuté les deux et gardé le résultat maximum de chaque langue, ce qui est plus réaliste, mais qui peut biaiser légèrement les résultats en notre faveur. La deuxième est que tous nos alignements ont la même cible (tous les plongements sont alignés vers l'espace latent anglais (en)) alors qu'ils appliquent un alignement pour chaque dictionnaire. Par exemple, ils ont aligné le français (fr) sur l'allemand (de) pour l'évaluation de-fr tandis que nous avons aligné le français (fr) et l'allemand (de) sur l'anglais (en), puis évaluons de-fr sur la tâche d'induction du lexique bilingue. Nous pourrons ainsi vérifier, en nous comparant à GLAVAŠ et al. 2019, si la convergence d'une langue $L1$ et d'une langue $L2$ est possible si elles ont été alignées sur une langue $L3$. En théorie, les résultats devraient être négativement impactés par l'utilisation d'une langue pivot pour l'alignement.

Avant harmonisation Les résultats sont disponibles dans le tableau 3.5. En moyenne, nos plongements alignés obtiennent un score *MAP* de 0.33 points, contre 0.40 pour GLAVAŠ et al. 2019, soit un écart de 0.08 points. Un problème se pose avec ce résultat, celui de l'interprétabilité de la *MAP*. Il n'est pas possible de dire dans l'absolu si cet écart de 0.08 point représente une baisse importante ou non de la qualité de la convergence des langues. C'est pourquoi la suite des résultats et leur interprétation est à prendre avec prudence.

Les résultats les plus proches de l'état de l'art sont naturellement ceux dont une des deux langues du dictionnaire est l'anglais (en), ce qui semble cohérent puisque les langues ont été alignées sur l'espace de l'anglais (en). On remarque que les dictionnaires utilisant du turc (tr), du russe (ru) ou du finlandais (fi) sont les dictionnaires avec l'écart le plus important par rapport à notre référence.

La première paire n'utilisant pas l'anglais (en) est de-fr, suggérant que ces deux langues font partie des langues qui ont le mieux convergé. Avec seulement 0.03 points d'écart de la référence, qui ne passait pas par une langue intermédiaire pour faire l'alignement, le résultat est plutôt élevé. Cette paire est même plus proche de la référence que d'autres paires utilisant l'anglais (en) comme en-fr, en-it ou en-fi.

Dic	Alignés (A)	EDLA (E)	Δ(E - A)	Moyennés (M)	Δ(A - M)
en-de	0.54	0.54	0.01	0.16	0.38
en-ru	0.45	0.46	0.01	0.06	0.39
en-hr	0.31	0.34	0.03	0.10	0.21
en-tr	0.31	0.34	0.03	0.09	0.22
de-fr	0.48	0.51	0.03	0.16	0.32
en-fr	0.62	0.65	0.04	0.23	0.39
ru-fr	0.41	0.47	0.06	0.09	0.33
it-fr	0.61	0.67	0.06	0.16	0.45
en-it	0.57	0.63	0.06	0.12	0.45
en-fi	0.33	0.40	0.06	0.10	0.23
tr-fr	0.27	0.34	0.06	0.10	0.17
hr-ru	0.30	0.37	0.08	0.18	0.11
de-it	0.43	0.51	0.08	0.15	0.28
de-hr	0.25	0.33	0.08	0.17	0.08
de-tr	0.20	0.28	0.08	0.12	0.08
hr-fr	0.29	0.37	0.08	0.12	0.17
fi-it	0.27	0.36	0.08	0.16	0.11
ru-it	0.39	0.47	0.09	0.14	0.25
hr-it	0.27	0.36	0.09	0.17	0.10
fi-fr	0.27	0.36	0.09	0.11	0.15
tr-hr	0.16	0.26	0.10	0.18	-0.02
fi-hr	0.19	0.29	0.10	0.19	0.00
tr-it	0.23	0.34	0.10	0.13	0.11
tr-fi	0.16	0.27	0.11	0.15	0.02
tr-ru	0.18	0.29	0.11	0.09	0.09
de-ru	0.30	0.43	0.13	0.13	0.17
fi-ru	0.21	0.34	0.14	0.10	0.10
de-fi	0.22	0.36	0.14	0.16	0.06
moy.	0.33	0.40	0.08	0.14	0.19

TABLEAU 3.5. – Scores *MAP* des plongements de mots alignés et moyennés, des dictionnaires issus de GLAVAŠ et al. 2019, avec les scores de l'état de l'art issus de l'article. La différence entre les plongements alignés et moyennés est explicitée, ainsi que la différence entre les plongements alignés et les plongements de référence.

Après harmonisation Les résultats de l'évaluation monolingue ont montré que l'harmonisation des plongements de mots par le moyennage avaient un fort impact sur leur qualité. Après harmonisation, les scores *MAP* ont baissé de 0.19 points par rapport aux plongements alignés (voir tableau 3.5), passant de 0.33 points en moyenne à 0.14 après harmonisation. Dans le cadre de l'évaluation multilingue, les baisses de résultats selon les dictionnaires peuvent varier considérablement : le dictionnaire it-fr perd 0.45 points, alors que le dictionnaire fi-hr a une évaluation identique avant et après harmonisation. Le dictionnaire tr-hr gagne même 0.02 point avec l'harmonisation.

Nous avons voulu estimer à quel point l'harmonisation éloignait les mots de leur position d'origine. En théorie, plus le déplacement des vecteurs est important, plus les résultats de l'évaluation seront impactés. Pour chaque mot w de chaque dictionnaire, la distance euclidienne entre le vecteur de w avant l'étape d'harmonisation et son vecteur après harmonisation est mesurée. La moyenne pour tous les mots du dictionnaire est calculée, permettant d'obtenir un score par dictionnaire. Le score de la corrélation de Pearson entre ce score et Δ(Alignés - Moyennés) [9] est de 0.79. Ce résultat va dans le sens de l'hypothèse que plus l'harmonisation a fait se déplacer les vecteurs, plus les scores *MAP* seront dégradés.

Les scores *MAP* des dictionnaires issus de PanLex définis dans la section 3.3 ont également été calculés, avant et après alignement. Malheureusement, aucun score de référence n'est disponible, ces résultats permettent donc surtout d'évaluer l'impact de l'étape d'harmonisation. Les résultats sont disponibles dans le tableau 3.6.

On constate là encore une chute des scores après harmonisation (-0.06 point), même si cette différence est bien moins importante que dans les évaluations précédentes. 10 de nos 42 langues obtiennent même de meilleurs résultats après harmonisation(le vieux slave (cu), l'estonien (et), le basque (eu), le japonais (ja), le kurmanji (kmr), le coréen (ko), le letton (lv), le norvégien nynorsk (nno), le vietnamien (vi) et le chinois (zh).

Cette augmentation n'est cependant pas très significative dans le cas du vieux slave (cu) par exemple, où les plongements de mots de seulement 22 paires [10] de mots ont été identifiées et évaluées. Certaines langues comme le vietnamien (vi) voient leur score augmenter de manière importante (+0.39), tandis que pour d'autres, les résultats sont très dégradés (-0.46 pour le portugais (pt)).

Cependant, les dictionnaires issus de MUSE et PanLex contiennent parfois beaucoup d'homographes, ce qui biaise positivement les résultats. C'est le cas du vietnamien (vi), avec 71.27% d'homographes.

L'harmonisation impacte négativement les résultats des évaluations intrinsèque de la majorité des langues (31/42). Cette chute ne signifie cependant pas que l'harmonisation empêchera un partage de connaissances entre langues sur des tâches

9. Cette différence correspond à "la baisse liée à l'harmonisation"

10. Bien que nous parlions de "paires" de mots, nous traitons en fait un mot source et sa ou ses traductions dans la langue cible. La "paire" de traduction correspondant au mot *octopus* en anglais (en) aurait deux traductions, *poulpe* et *pieuvre* en français (fr). Si le plongement du mot *pieuvre* n'existe pas dans notre ensemble de plongements de mots du français, alors l'évaluation se fera simplement sur la paire *octopus,poulpe*.

Lang.	Align.	Moy.	Δ	Paires	HG	Lang.	Align.	Moy.	Δ	Paires	HG
nno	0.13	0.96	0.83	416	64.00	hi	0.41	0.34	-0.07	744	3.73
ja	0.02	0.67	0.65	920	5.67	hr	0.43	0.33	-0.10	662	13.87
kmr	0.08	0.59	0.51	253	21.27	tr	0.46	0.32	-0.14	677	14.67
vi	0.41	0.80	0.39	1205	71.27	fa	0.39	0.25	-0.15	725	3.13
eu	0.09	0.44	0.35	1067	5.47	ca	0.66	0.51	-0.15	1045	27.47
ko	0.24	0.44	0.20	822	3.67	cs	0.49	0.32	-0.17	571	10.67
et	0.31	0.47	0.16	800	14.60	da	0.52	0.34	-0.19	584	16.07
cu	0.01	0.12	0.11	22	0.00	sv	0.51	0.31	-0.20	565	14.87
zh	0.34	0.44	0.09	727	8.00	id	0.66	0.45	-0.21	702	27.27
lv	0.32	0.37	0.05	612	5.53	ro	0.59	0.38	-0.21	591	16.93
el	0.44	0.44	0.00	591	6.13	nl	0.61	0.38	-0.23	705	21.13
hu	0.44	0.43	-0.02	674	13.33	pl	0.48	0.23	-0.24	497	9.60
nob	0.09	0.07	-0.02	25	0.00	he	0.43	0.17	-0.25	674	1.13
sk	0.37	0.33	-0.04	491	8.40	ar	0.37	0.09	-0.28	543	1.13
fi	0.34	0.29	-0.05	539	8.33	de	0.53	0.20	-0.33	436	11.33
ur	0.06	0.01	-0.05	1252	0.00	ru	0.43	0.09	-0.34	575	2.67
sl	0.36	0.31	-0.05	578	11.07	fr	0.68	0.32	-0.36	573	18.87
ga	0.06	0.01	-0.05	202	0.00	it	0.68	0.28	-0.40	664	18.33
gl	0.21	0.15	-0.06	340	0.00	bg	0.51	0.10	-0.41	583	1.00
kk	0.19	0.12	-0.07	65	0.00	es	0.66	0.22	-0.44	511	13.87
uk	0.38	0.31	-0.07	607	4.93	pt	0.71	0.26	-0.46	553	16.13
moy	0.38	0.33	-0.06	604	12.27						

TABLEAU 3.6. – Scores *MAP* des plongements de mots alignés et moyennés, des dictionnaires de test issus de PanLex et MUSE, décrits dans la section 3.3. Plus le Δ est haut, plus l'impact de l'harmonisation a été négatif. On donne également le nombre de paires de mots pour lesquels les plongements de mots ont été trouvés, et le pourcentage d'homographes du dictionnaire.

extrinsèques. GLAVAŠ et al. 2019 ont montré que la corrélation entre les performances extrinsèques et intrinsèques dépend de la tâche. C'est pourquoi nous allons mener une tâche d'évaluation extrinsèque dans la section 3.7 afin de voir si nos plongements de mots alignés et moyennés peuvent être utiles dans un modèle d'étiquetage de POS.

3.7. Le problème de l'harmonisation

La chute importante des résultats de l'évaluation après harmonisation des plongements de mots était très inattendue. Afin de vérifier l'impact de ce choix sur la suite de nos travaux, nous avons mis en places quelques expériences d'analyse.

La première expérience permet de vérifier dans quelle mesure l'hypothèse de départ est vraie. Cette hypothèse supposait que deux homographes de deux langues devaient avoir un sens proche. Par conséquent, si l'alignement a bien convergé, alors les deux vecteurs des deux homographes devraient être proches.

Nous allons comparer les distances cosinus entre les homographes d'une langue L avec l'anglais (en), afin de les comparer avec les distances cosinus de paires de mots aléatoires. Si notre hypothèse est vraie, alors les homographes devraient avoir une distance cosinus plus faible que des mots aléatoires.

Nous avons donc extrait des homographes entre l'anglais (en) et les 37 autres langues, pour lesquels il existe un plongement de mot dans les deux langues. Ces homographes ont été extraits des corpus d'entraînement monolingue de chaque langue. Il est important de noter que parmi les homographes, on retrouve la ponctuation, les chiffres et les noms propres, qui sont des tokens ayant en effet un sens très proche entre les différentes langues. Certaines langues partagent beaucoup d'homographes avec l'anglais (en), comme le français (fr) avec 868 tokens, tandis que d'autres, comme l'hindi (hi) en partagent très peu (seulement 7, correspondant à de la ponctuation et des chiffres). La taille du lexique d'homographes par langue est disponible en annexe dans le tableau .8.

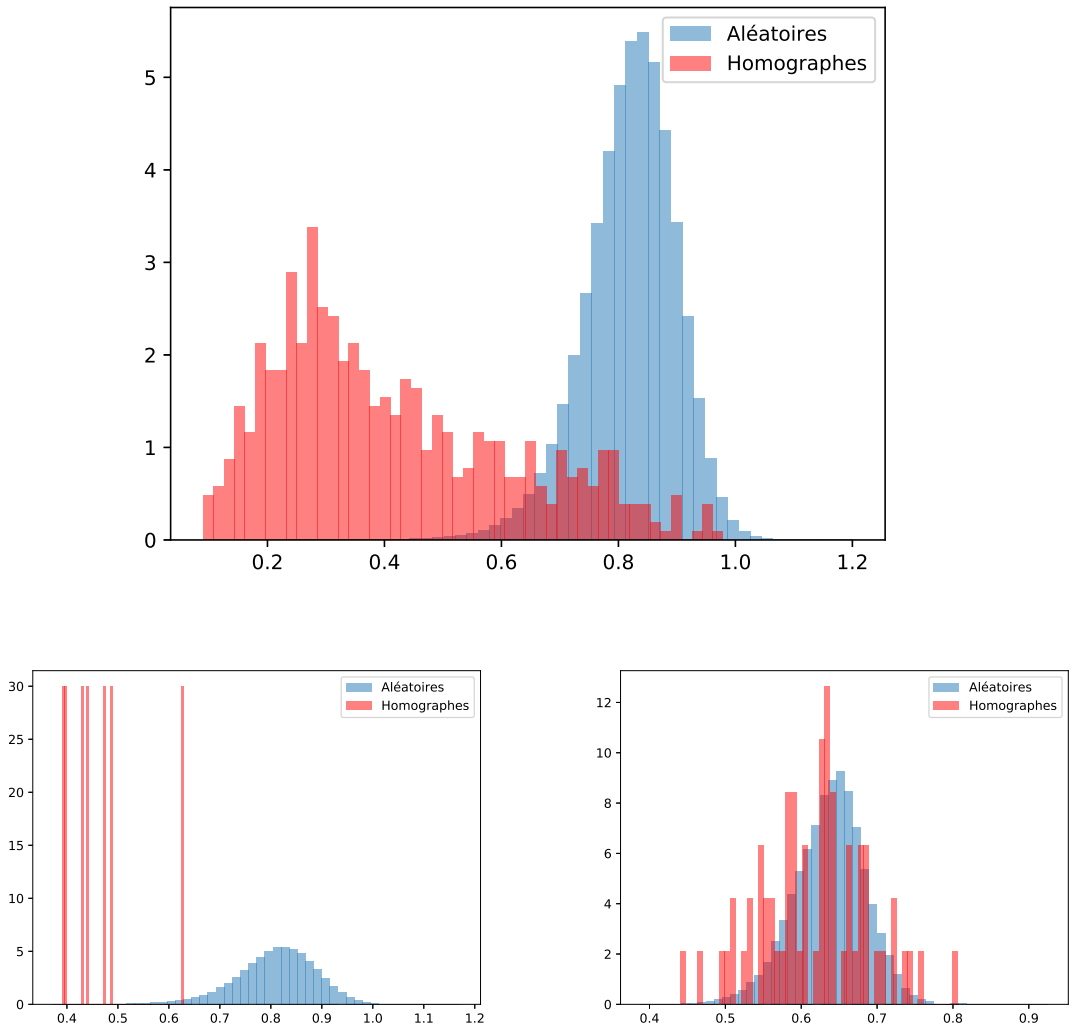


FIGURE 3.6. – Distributions de probabilités de la distance cosinus des homographes et de mots aléatoires entre l'anglais (en) et : l'allemand (de) (en haut), l'hindi (hi) (en bas à gauche) et le japonais (ja) (en bas à droite).

On a calculé la distribution de probabilité des homographes et des paires aléatoires en fonction de la distance cosinus entre les vecteurs des paires de mots. Pour la majorité des langues, une distribution de probabilité similaire des distances cosinus des homographes est observée. Cette distribution "classique" est visible dans la figure 3.6, pour l'exemple de l'allemand (de) (figure du haut).

Les homographes sont bien plus proches entre eux que des mots choisis aléatoirement, confirmant que l'hypothèse de départ était assez réaliste. Cette distribution n'est pas la même pour toutes les langues, comme on peut le voir pour l'hindi (hi) qui n'a que très peu d'homographes, ou pour le japonais (ja), pour qui les homographes ne semblent pas spécialement plus proches que des mots aléatoires.

Paire	Aligné		Moyenné	
	en-fr	fr-en	en-fr	fr-en
pain (fr) - pain (en)	18 697	8 820	1	1
pain (fr) - bread (en)	1	1	1 076	37 934
douleur (fr) - pain (en)	1	1	1 146	19 300
weekend (fr) - weekend (en)	1	1	1	1

TABLEAU 3.7. – Rangs de paires de mots pour les plongements de mots alignés et moyennés pour les deux sens d'évaluation en-fr et fr-en.

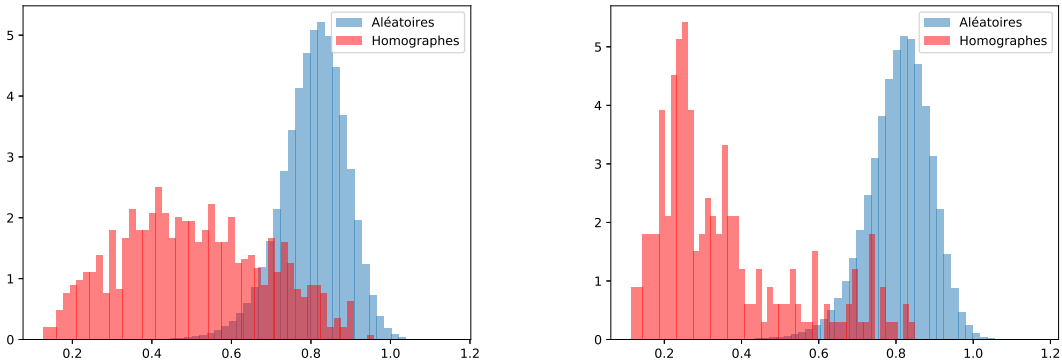


FIGURE 3.7. – Distributions de probabilités de la distance cosinus des homographes et de mots aléatoires entre l'anglais (en) et le français, à gauche avec tous les homographes, et à droite avec les homographes triés, où l'on a retiré les chiffres, la ponctuation et les noms propres.

Cette distribution a également été calculée pour les homographes "épurés" du français, desquels on a retiré les chiffres, les noms propres et la ponctuation qui biaisent positivement les résultats, puisqu'il est très probable que les plongements de mots de ces tokens soient proches. Les mots anglais (en) apparaissant dans le corpus du français (fr) et inversement ont également été retirés.

On constate que les deux distributions sont assez différentes (voir figure 3.7) : les homographes épurés sont encore plus proches que les homographes complets. Cela sous-entend que les mots qui ont été éliminés sont ceux qui sont les moins bien alignés.

Une deuxième expérience consiste à mesurer le rang de paires de mots, à la manière de l'évaluation *MRR* et *MAP*, afin d'estimer la distance de la paire. Pour cela, nous avons pris deux homographes du français (fr) et de l'anglais (en), les mots *pain* et *weekend*. Le mot *pain* n'a pas de sens commun entre les deux langues, alors que *weekend* oui. Nous allons vérifier les rangs de ces paires de mots en fonction des plongements de mots utilisés (alignés ou moyennés). Les résultats sont visibles dans le tableau 3.7.

Lang.	pdm moy.	pdm align.	Diff	Lang.	pdm moy.	pdm align.	Diff
ar	92.94	93.40	0.46	id	90.64	91.78	1.14
bg	95.73	96.81	1.08	it	94.30	96.05	1.75
ca	95.84	96.75	0.91	ja	92.41	89.67	-2.74
cs	95.12	96.10	0.98	ko	93.38	93.66	0.28
da	92.41	94.33	1.92	lv	89.44	90.97	1.53
de	89.16	91.97	2.81	nl	87.78	89.52	1.74
el	95.47	96.39	0.92	nno	90.43	93.17	2.74
en	89.54	92.06	2.52	nob	91.92	94.51	2.59
es	93.22	94.52	1.30	pl	95.19	96.41	1.22
et	90.02	91.61	1.59	pt	93.25	94.83	1.58
eu	90.62	92.77	2.15	ro	94.11	95.56	1.45
fa	94.83	95.09	0.26	ru	94.91	95.97	1.06
fi	84.90	86.78	1.88	sl	92.50	93.71	1.21
fr	93.45	94.95	1.50	sv	92.03	94.33	2.30
ga	89.11	90.70	1.59	tr	90.44	92.81	2.37
he	94.45	95.06	0.61	uk	91.93	92.91	0.98
hi	93.62	94.17	0.55	ur	90.56	91.33	0.77
hr	94.30	95.61	1.31	vi	86.05	87.50	1.45
hu	92.90	94.07	1.17	zh	87.98	90.14	2.16
moy.	92.02	93.37	1.34				

TABLEAU 3.8. – Exactitude des modèles utilisant les plongements de mots moyennés comparé à ceux utilisant les plongements de mots alignés mais non-moyennés.

Avant harmonisation, à l'étape où les plongements sont seulement alignés, on constate que les mots qui sont bien la traduction l'un de l'autre sont les plus proches voisins de leur traduction, confirmant la convergence de l'alignement. Les vecteurs du mot *pain* en français (fr) et en anglais (en) sont éloignés l'un de l'autre, ce qui est cohérent puisque les deux mots n'ont pas de sens commun. Après harmonisation, les résultats sont plus problématiques. Les mots *weekend* et *pain* ont désormais le même vecteur, ce qui est une bonne chose pour *weekend*, mais qui est plus gênant pour *pain*. La moyenne des vecteurs ayant été réalisée pour ce mot, *pain* n'est désormais ni proche de *bread* ni de *douleur*.

La troisième expérience cherche à mesurer l'impact de l'harmonisation sur les plongements de mots de l'anglais (en), notre langue pivot, dont l'espace vectoriel a été utilisé comme référence pour aligner toutes les autres langues. Pour estimer la distance avant et après harmonisation des plongements de mots de l'anglais (en), nous avons calculé la *MAP* entre les plongements de mots alignés et ceux moyennés. Autrement dit, nous avons créé un dictionnaire en_align-en_moy. Si l'harmonisation n'a pas eu d'impact, le score devrait être de 1. Or, avec un score de 0.10, il semble en effet que l'harmonisation des plongements de mots par le moyennage éloigne la majorité des mots de leur position d'origine, impactant leur qualité.

La dernière expérience consiste à comparer sur une tâche d'étiquetage en POS la différence des résultats entre un apprentissage utilisant les plongements de mots moyennés, et une autre avec les plongements de mots alignés avec MUSE, mais

non-moyennés. Les résultats sont disponibles dans le tableau 3.8. Les résultats sont systématiquement meilleurs en utilisant les plongements de mots non-moyennés (à l'exception du japonais (ja)), la différence moyenne n'étant que de 1.34 point comparés aux modèles utilisant les plongements de mots moyennés.

Bien que les résultats soient légèrement inférieurs, la chute est moins importante que ce que laissait prédire les évaluations intrinsèques des plongements de mots. Les très mauvais résultats de nos évaluations pour les plongements de mots moyennés sont donc à relativiser, les résultats en évaluation extrinsèque étant bien moins dramatiques. Cette même conclusion avait par ailleurs déjà été soulignée par GLAVAŠ et al. 2019.

Plus de détails sur la tâche d'étiquetage en POS sont disponibles dans le chapitre 4. Les expériences des chapitres qui vont suivre mériteraient d'être relancées sans passer par l'étape d'harmonisation des plongements de mots. L'évaluation des plongements de mots étant une des dernières étapes de ce travail, il n'a pas été possible de corriger cette erreur par manque de temps, mais les résultats de la tâche d'étiquetage semble montrer une diminution assez uniforme des résultats qui ne devrait pas invalider les conclusions tirées dans les chapitres utilisant ces plongements de mots.

3.8. Lexicalisation de l'analyseur syntaxique

Dans le chapitre 2, nous avions utilisé un analyseur syntaxique délexicalisé, faute de lexique universel à notre disposition. Maintenant que nous disposons des plongements de mots multilingues, nous allons pouvoir les utiliser afin de mesurer l'impact de la lexicalisation sur les résultats. L'architecture de l'analyseur ayant changé, nous vérifierons l'impact de ce changement d'architecture dans la section B.1 en annexe.

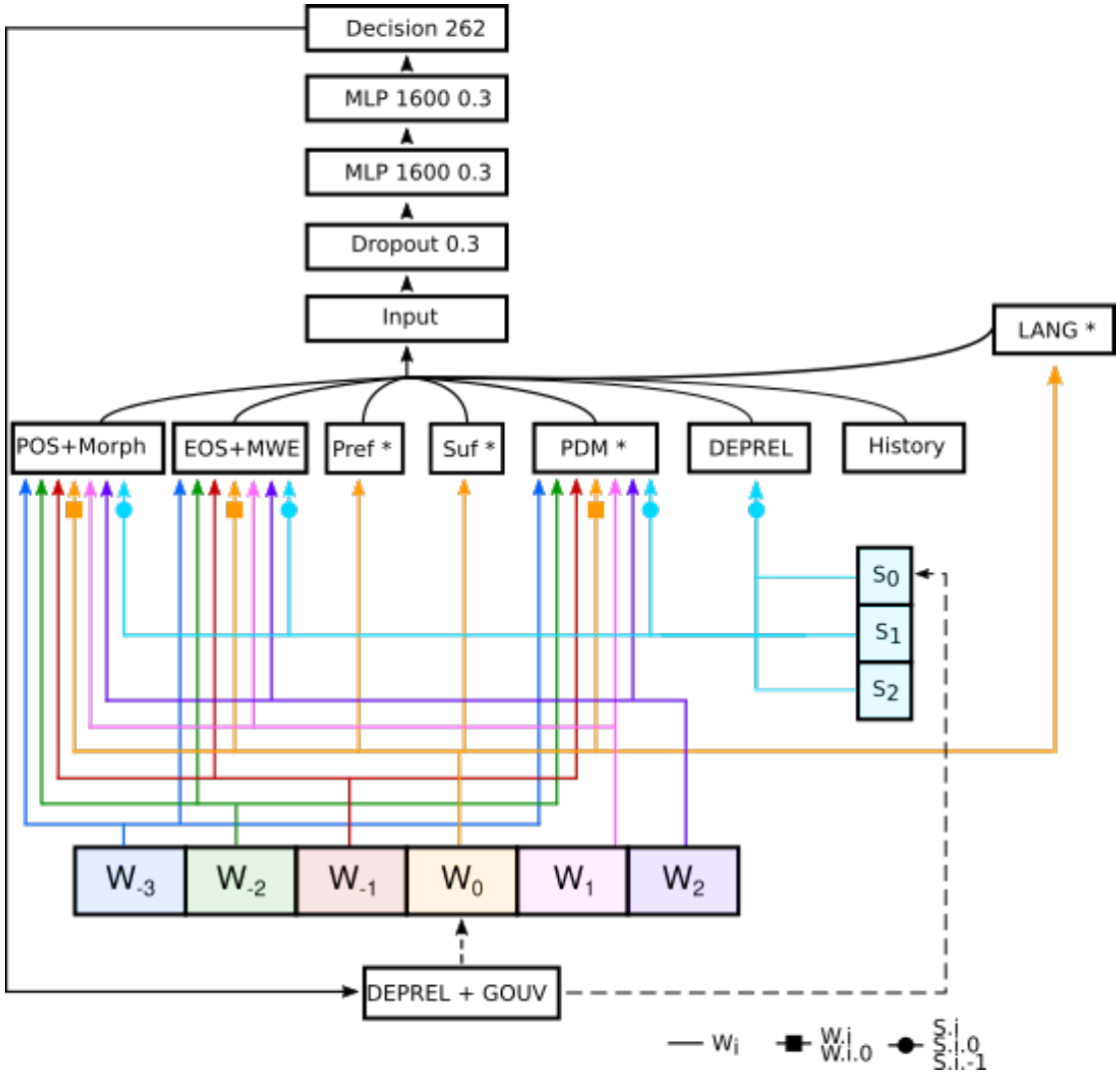


FIGURE 3.8. – Architecture de l'analyseur syntaxique. Le texte est lu en entrée du buffer. Le mot courant est w_0, le mot précédent est w_{-1} et le mot suivant est w_2. Une pile est également à disposition, s_0 étant le sommet de pile. Les 7 bi-LSTM sont représentés par les rectangles POS+Morph, EOS+MWE, ... et prennent en entrée le trait correspondant à leur nom. Les bi-LSTM avec une '*' sont optionnels et sont utilisés selon la configuration (*Multi* $+c$, *Multi* $-c$ W_{22}, ...). LANG n'est pas un bi-LSTM, c'est un vecteur optionnel concaténé directement à la couche d'entrée. La prédiction du mot courant est donnée en entrée pour le mot suivant. w_i correspond aux plongements de POS+Morph, ... et $w_{i.0}$ est le plongement de POS+Morph du fils gauche du nœud dans l'arbre. Idem avec s_i, $s_{i.-1}$ étant le fils droit. Les flèches en pointillées signifient qu'à chaque prédiction, le mouvement prédit est donné en entrée soit au mot courant du buffer, soit au sommet de la pile.

Analyseurs syntaxiques, nouvelles versions L'analyseur utilisé est Macaon [11] (DARY et NASR 2021), un analyseur en transition reposant sur un MLP. Seul le classifieur est expliqué ici. Les entrées de l'analyseur correspondent à une fenêtre de taille 6, centrée sur le mot courant. Des informations sont obtenues à partir de ces mots, telles que leur représentation vectorielle, leurs préfixes et suffixes, les POS des mots précédents, etc. Ces traits sont fournis en entrée de bi-LSTMs qui permettront de contextualiser ces informations, pour ensuite les donner en entrée du MLP. La couche d'entrée de ce MLP consiste alors en la concaténation de la sortie de 4 à 7 bi-LSTM, décrits ci-dessous :

— EOS+MWE : un bi-LSTM prenant en entrée l'information de la présence d'une fin de phrase et la présence d'un token multi-mots parmi le mot courant w_0, les trois mots précédents w_1, w_{-2}, w_3 et les deux mots suivants w_1, w_2, w_3 du buffer, ainsi que le plongement de mot du fils gauche de w_0. La sortie de ce bi-LSTM est de taille 64.

— DEPREL : un bi-LSTM prenant en entrée l'information du type de relation de dépendance entre le mot et sa tête (deprel) pour les trois mots précédents. La sortie de ce bi-LSTM est de taille 64. Lors des 4 premières itérations, les deprels de références des mots précédents sont donnés. À la fin de la quatrième itération, le système fait un décodage de tout le corpus d'entraînement. À partir de la cinquième itération, le système utilise ces deprels prédites à la place de celles de références. Lors de la phase de test, ce sont les prédictions des mots précédents qui sont utilisées.

— POS+Morph : un bi-LSTM prenant en entrée la concaténation du plongement de POS et du plongement des traits morphologiques, pour le mot courant w_0, les trois mots précédents w_1, w_{-2}, w_3 et les deux mots suivants w_1, w_2, w_3 du buffer, ainsi que le plongement de mot du fils gauche de w_0. La sortie de ce bi-LSTM est de taille 64.

— History : un bi-LSTM prenant en entrée la concaténation des plongements des 10 précédentes actions prédites et dont la sortie est de taille 32

— PDM : un bi-LSTM optionnel prenant en entrée les plongements de mots multilingues de taille 300 définis au chapitre 3 et les contextualisant à partir d'une fenêtre de taille 21 centrée sur le mot courant. Ces plongements de mots contextuels obtenus sont de taille 128. Sont alors passés en entrée de ce bi-LSTM :

  — les 8 plongements contextuels de taille 128 correspondant au mot courant w_0, aux trois mots précédents w_1, w_{-2}, w_3 et aux deux mots suivants w_1, w_2, w_3 du buffer, ainsi que le plongement de mot du fils gauche du mot courant w_0

  — les 3 plongements des mots de la pile en haut de la pile, ainsi que leurs fils gauches et droits

  La sortie de PDM est de taille 128 * (8 + 3 * 3) = 2176

11. https://gitlab.lis-lab.fr/franck.dary/macaon

— Pref : un bi-LSTM optionnel prenant le préfixe du mot courant (3 caractères) dont la sortie est un plongement de taille 64.

— Suf : un bi-LSTM optionnel prenant le suffixe du mot courant (3 caractères) dont la sortie est un plongement de taille 64.

— LANG : ce n'est pas un bi-LSTM, mais un simple vecteur optionnel concaténé directement à la couche d'entrée, existant uniquement dans les expériences utilisant W_{22}. Ce vecteur est de taille 80, correspondant aux 80 valeurs du vecteur W_{22} où les 22 traits sont représentés sous forme de one-hot puis concaténés. Alternativement, LANG peut servir de représentation pour l'identifiant de la langue et est alors un vecteur de la taille du nombre de langues dans le corpus d'entraînement.

On obtient ainsi une couche d'entrée de taille 2176 + 64 + 64 + 64 + 32 (+64) (+64) (+80) = 2 400 à 2 608 selon les configurations. Les deux couches cachées sont de taille 1600, avec un *dropout* durant l'entraînement de 0.3.

La couche de sortie de ce réseau de neurones donne une distribution de probabilités, une pour chaque action possible pour le mot courant. Le système étant glouton, il choisira l'action avec la plus haute probabilité. Le nombre d'itérations est de 40, la fonction d'activation est la fonction ReLu, la fonction objectif est un softmax de vraisemblance négative, et l'algorithme d'apprentissage est Adagrad.

Toutes les expériences ont été lancées deux fois [12] à l'exception des expériences en condition zero-shot, en raison de problème de temps d'entraînement. Les résultats des expériences en conditions monolingues et multilingues correspondent à la moyenne des scores des deux modèles. Toutes les entrées (POS, traits morphologiques, ...) sont représentées par des plongements, et sont initialisées aléatoirement, à l'exception des formes de surface qui utilisent les plongements de mots pré-entraînés et alignés (voir chapitre 3). Un schéma de l'architecture de ce système est disponible dans la figure 3.8.

Configurations d'entraînement Nous reprendrons la base des définitions données au chapitre 2 (voir sections 2.5). Nos différentes configurations vont correspondre à une paire ⟨données d'entraînement, architecture⟩. Les données d'entraînement possibles sont *Mono*, *Multi* et *ZS*.

Mono correspond à une unique langue au moment de l'entraînement, et l'analyse syntaxique de cette même langue pour le test.

Multi représente un entraînement basé sur l'ensemble de toutes les langues disponibles, et l'analyse syntaxique de chaque langue par ce modèle multilingue unique.

ZS est identique à *Multi*, auquel on retire une langue L. L'évaluation de *ZS* se fait sur la langue L uniquement.

Les différentes architectures correspondent à la façon de paramétrer notre analyseur. Le modèle peut utiliser les caractères des mots ($+c$) ou non ($-c$) et utiliser

12. afin de limiter l'impact de l'aléatoire, faisant varier les résultats des expériences, parfois de beaucoup.

une représentation de la langue (ID ou W_{22}) ou non. Nos différentes configurations d'entraînement sont donc :

Mono : Corpus monolingue.
Le corpus d'entraînement est L, et le corpus de test et également celui de L. 38 analyseurs sont entraînés, un pour chaque langue. C'est le cas classique d'apprentissage, où l'entraînement se fait sur une langue L, et le décodage se fait sur cette même langue L.

Multi : Corpus multilingue.
Un analyseur est entraîné sur nos 38 langues, sans indication de la langue. Un seul modèle est entraîné, et chaque corpus de test est décodé avec ce modèle. Ce modèle est très bruité, puisqu'il est possible de retrouver des contradictions entre les langues (*a*, en français (fr) est un verbe conjugué (du verbe *avoir*), alors qu'en anglais (en), *a* est un déterminant).

Multi W_{22} : Corpus + WALS.
Idem à *Multi* en utilisant le vecteur LANG qui permet d'associer à chaque mot un vecteur $W(L)$ issu du WALS, qui correspond à la langue du mot. Contrairement à l'utilisation d'un identifiant de la langue, le WALS est porteur d'informations sur la syntaxe, la morphologie des mots, etc. pouvant aider les prédictions.

Multi ID : Corpus multilingue + identifiant de la langue.
Idem à *Multi* en utilisant en plus le vecteur LANG qui permet de représenter l'identifiant de la langue de chaque mot. Cette expérience permet d'expliciter l'information de la langue en cours de traitement pour limiter le bruit apporté par l'apprentissage multilingue, mais sans expliciter les points communs à plusieurs langues.

ZS **:** *Zero-shot*.
38 analyseurs sont entraînés, sans représentation de la langue. L'analyseur de la langue L sera entraîné sur toutes les langues, sauf L, soit un ensemble de 37 langues. Cette configuration représente la situation où un analyseur est entraîné pour une langue pour laquelle aucune donnée d'entraînement n'est disponible. Les plongements de mots sont tout de même disponibles pour les prédictions, ainsi que les caractères des mots.

ZS W_{22} **:** *Zero-shot* + WALS
Ajoute à *ZS* l'indication de la langue de chaque mot à travers son vecteur du WALS, à l'identique de la différence entre *Multi* W_{22} et *Multi*.

Chacune de ces expériences est déclinée en trois versions, selon qu'elle utilise le modèle de caractères, les plongements de mots, un seul des deux ou les deux :

— *Mono* +*c*, *Multi* +*c*, *Multi* +*c* ID, *Multi* +*c* W_{22}, *ZS* +*c* et *ZS* +*c* W_{22} utilisent

les bi-LSTM correspondant au modèle de caractères (Pref et Suf) et celui correspondant au plongements de mot (PDM)

— *Mono −c*, *Multi −c*, *Multi −c* ID, *Multi −c* W_{22}, *ZS −c* et *ZS −c* W_{22} utilisant le bi-LSTM correspondant aux plongements de mots, mais n'utilisant pas ceux correspondant aux caractères des mots (Pref et Suf).

— *Mono_delex*, *Multi_delex*, *Multi_delex* ID, *Multi_delex* W_{22}, *ZS_delex* et *ZS_delex* W_{22} n'utilisent aucun des bi-LSTM liés au lexique (Pref, Suf et PDM).

Cela nous permettra de tester l'impact de l'utilisation des caractères et celui des plongements de mots dans nos expériences. Chaque expérience a été lancée deux fois, les résultats correspondent à la moyenne des scores des deux modèles.

3.8.1. Lexicalisation

L'utilisation des plongements de mots multilingues permet de ré-entraîner les modèles du chapitre 2 de manière lexicalisée. Afin d'avoir des résultats comparables pour les modèles lexicalisés et délexicalisés, les modèles du chapitre 2 ont été ré-entraînés avec l'architecture courante de notre analyseur (voir section B.1 en annexe).

Nous allons chercher à quantifier l'apport de l'utilisation des plongements de mots. Pour cela, on définit pour chaque langue "l'apport de l'utilisation des plongements de mots" :

$$app(pdm, Mono) = (Mono - c) \; - \; (Mono\_delex)$$

On définit également les équivalents multilingues et zero-shot de la même manière.

L'utilisation des plongements de mots devrait fournir au modèle des informations utiles, lui permettant d'obtenir de meilleurs résultats. La lexicalisation de nos modèles consiste en l'ajout d'un modèle de caractères et l'utilisation de plongement de mots multilingues. Afin de mesurer l'impact des plongements de mots uniquement, on compare les expériences *Mono −c* et *Mono_delex*. L'expérience *Mono −c* n'utilise pas le modèle de caractères, mais a accès aux plongements de mots, tandis que l'expérience *Mono_delex* n'utilise ni le modèle de caractères, ni les plongements de mots. La seule différence entre ces deux expériences est donc l'utilisation des plongements de mots. Les comparaisons seront réitérées en contexte multilingue puis en contexte de zero-shot. Un résumé des résultats est disponible dans le tableau 3.9.

Lang.	*Mono* $-c$	*Mono_delex*	Δ	*Multi* $-c$	*Multi_delex*	Δ	*ZS* $-c$	*ZS_delex*	Δ
ar	69.61	65.12	4.49	72.16	66.81	5.35	33.34	43.85	-10.51
cs	68.38	65.84	2.54	73.52	66.87	6.65	63.70	49.51	14.19
es	69.62	70.06	-0.44	73.06	71.36	1.70	74.24	72.78	1.46
et	68.78	69.62	-0.84	75.53	71.98	3.55	59.76	58.58	1.18
fa	76.38	70.44	5.94	77.39	70.67	6.72	42.52	30.70	11.82
hu	63.09	60.13	2.96	69.20	58.81	10.39	47.76	47.04	0.72
id	74.77	72.27	2.50	77.61	67.50	10.11	43.17	46.15	-2.98
ja	87.10	76.68	10.42	88.32	74.28	14.04	9.16	10.02	-0.86
ko	66.47	48.33	18.14	66.83	47.55	19.28	22.98	18.41	4.57
nno	72.79	72.89	-0.10	75.41	75.81	-0.40	73.97	71.90	2.07
nob	76.49	76.71	-0.22	80.13	79.24	0.89	78.63	76.21	2.42
ru	70.44	66.49	3.95	59.02	61.81	-2.79	50.88	50.02	0.86
ur	76.50	73.21	3.29	78.34	72.77	5.57	56.83	67.66	-10.83
vi	58.16	52.22	5.94	63.10	51.31	11.79	38.03	30.61	7.42
zh	64.70	60.91	3.79	68.83	55.73	13.10	11.45	19.21	-7.76
moy.	71.08	68.38	2.70	74.15	68.20	5.95	55.57	54.55	1.02

TABLEAU 3.9. – LAS des expériences avec et sans plongements de mots, en contexte monolingue, multilingue et en contexte de zero-shot. Les colonnes Δ correspondent à $app(pdm, X)$ avec X correspondant à *Mono*, *Multi* ou *ZS* selon le cas.

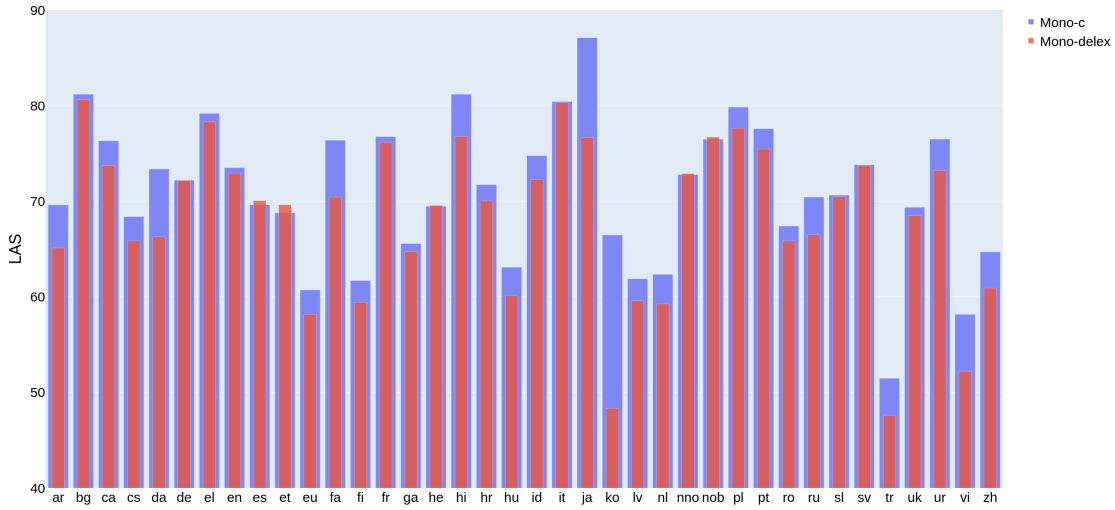


FIGURE 3.9. – LAS pour les expériences en conditions monolingues, sans modèles de caractères, et entièrement délexicalisées, afin de visualiser l'apport des plongements de mots.

En moyenne, les résultats augmentent de 2.70 points de LAS en utilisant les plongements de mots. Certaines langues comme le japonais (ja) ou le coréen (ko) bénéficient énormément de l'utilisation des plongements de mots (respectivement +10.42 et +18.14 points de LAS). Pour la majorité des langues (33/38), l'utilisation des plonge-

ments de mots permet bien une amélioration des résultats dans un contexte monolingue. L'espagnol (es), l'estonien (et), l'hébreu (he), et les deux norvégiens (nno, nob) sont les seules exceptions (voir figure 3.9). La baisse des résultats de ces langues est cependant assez faible (0.84 pour l'estonien (et), la langue pour laquelle la baisse est la plus importante). Il n'est pas impossible d'exclure que cette baisse soit simplement liée à la variabilité aléatoire des entraînements des modèles.

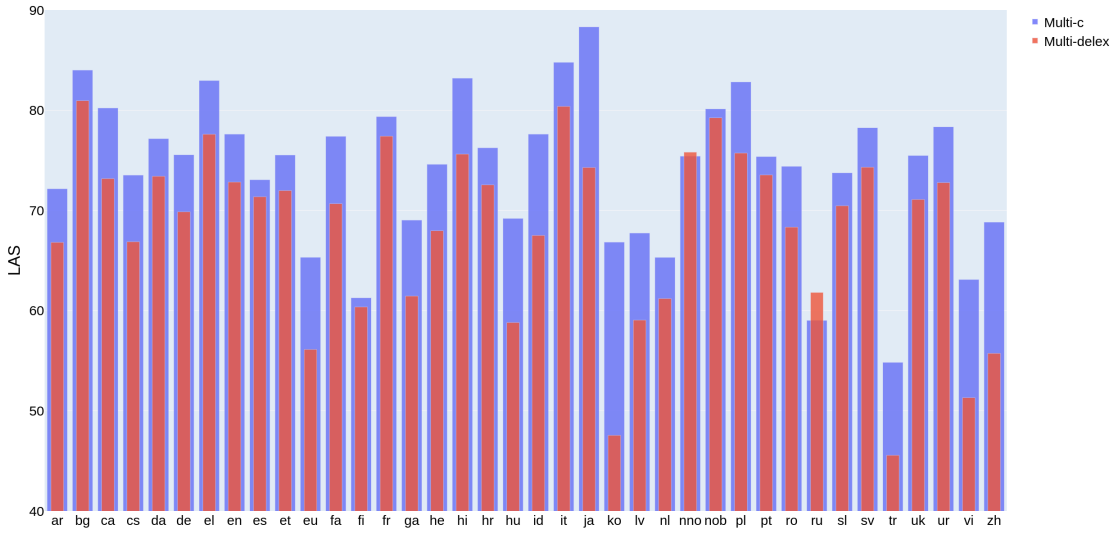


FIGURE 3.10. – LAS pour les expériences en conditions multilingues, sans modèles de caractères, et entièrement délexicalisées, afin de visualiser l'apport des plongements de mots.

En multilingue, les résultats sont similaires à ceux en conditions monolingues (voir figure 3.10), avec un score de corrélation de Pearson de 0.68 entre $app(pdm, Mono)$ et $app(pdm, Multi)$. Cette corrélation semble indiquer que l'apport des plongements de mots est assez similaire en conditions monolingues et multilingues, ce sont globalement les mêmes langues qui en profitent.

Les résultats augmentent en moyenne de 5.95 points, seules 2 langues ne bénéficient pas des plongements de mots (le russe (ru) et le norvégien nynorsk (nno), avec respectivement une baisse de 2.79 et de 0.40 point). 6 langues voient leurs LAS augmenter de plus 10 points : le hongrois (hu), l'indonésien (id), le japonais (ja), le coréen (ko), le vietnamien (vi) et le chinois (zh). Les langues semblant le plus profiter de l'utilisation des plongements de mots sont celles utilisant des alphabets uniques à la langue. Ce résultat peut sembler étonnant, puisqu'aucune de ces expériences n'utilise de modèle de caractères. Cependant, ces langues sont aussi des langues assez isolées de l'ensemble. Les plongements de mots permettent peut-être alors un meilleur partage de connaissances pour les langues isolées.

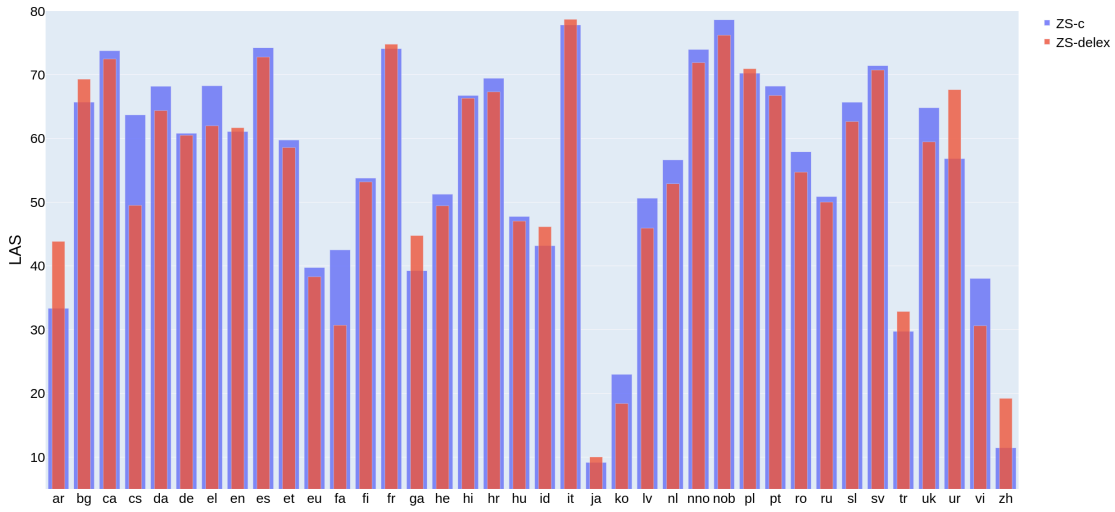


FIGURE 3.11. – LAS pour les expériences en conditions zero-shot, sans modèles de caractères, et entièrement délexicalisées, afin de visualiser l'apport des plongements de mots.

Les résultats de zero-shot sont cependant assez différents. Aucune corrélation n'a d'ailleurs pu être mise en évidence entre $app(pdm, Mono)$ et $app(pdm, ZS)$. Autrement dit, les langues bénéficiant des plongements de mots en monolingue ne sont pas les mêmes qu'en condition zero-shot. Bien qu'en moyenne, l'utilisation de plongements de mots fasse augmenter les résultats de 1.02 point, on constate d'importantes différences selon les langues.

12 langues sur 38 ne bénéficient pas de l'utilisation de plongements de mots : l'arabe (ar), le bulgare (bg), l'anglais (en), le français (fr), l'irlandais (ga), l'indonésien (id), l'italien (it), le japonais (ja), le polonais (pl), le turc (tr), l'ourdou (ur) et le chinois (zh).

Certaines langues comme l'arabe (ar) et l'ourdou (ur) perdent plus de 10 points lors de l'utilisation des plongements de mots multilingues, alors qu'ils bénéficiaient largement de leur utilisation en contexte monolingue et multilingue. D'autres langues comme le tchèque (cs) ou le persan (fa) gagnent au contraire plus de 10 points de LAS avec l'utilisation des plongements de mots.

Comme pour les expériences du chapitre 2, le contexte d'apprentissage zero-shot donne des résultats variant beaucoup d'une langue à l'autre, et il n'est pas évident de tirer une conclusion claire, mais seulement d'en tirer des intuitions sur des explications possibles. Il est par exemple possible que la qualité des plongements de mots pour ces langues soit de moins bonne qualité à cause de l'harmonisation par le moyennage des vecteurs des mots homographes pour les langues avec un même jeu de caractères comme l'arabe (ar) et l'ourdou (ur).

Il est également important de noter que les expériences de type $ZS + c$ sont très longues à entraîner, puisqu'un modèle par langue doit être appris. C'est pourquoi il n'a pas été possible d'apprendre deux modèles pour chaque expérience pour les expériences $ZS + c$ de ce chapitre. L'information de la variabilité des résultats n'est pas

disponible, et il est fort possible que ces expériences varient plus que les expériences *Mono +c* ou *Multi +c*, rendant les différences moins significatives.

La conclusion générale reste tout de même que l'utilisation de plongements de mots est bénéfique, en monolingue, multilingue, et majoritairement en zero-shot bien que ce ne soit pas systématique.

3.9. Conclusions et perspectives

L'objectif de ce chapitre était la création d'un lexique se voulant *universel*. Pour cela, des plongements de mots issus de FastText ont été récupérés et alignés grâce à MUSE. Ces plongements de mots obtiennent des résultats inférieurs à l'état de l'art, mais en restent assez proches, que ce soit pour l'évaluation monolingue ou celle en multilingue. Cette dernière évaluation a pu confirmer que la convergence des langues a eu lieu, en particulier, des langues comme l'allemand (de) et le français (fr) ont convergé, alors qu'elles n'avaient pas été alignées entre elles. En effet, toutes les langues que nous utilisons ayant été alignées sur l'espace vectoriel de l'anglais (en), la convergence de paires de langues ne contenant pas la langue pivot nous conforte dans l'idée qu'il est possible de créer un espace vectoriel universel pour toutes les langues.

Cependant, nous avons également pu conclure que bien que l'hypothèse de la proximité des homographes semble vérifiée pour une majorité de langue, l'harmonisation a un impact très négatif sur les évaluations intrinsèques. Néanmoins, des évaluations extrinsèques sur une tâche d'étiquetage en POS nous ont montré que cette chute des évaluations après harmonisation serait moins importante en pratique pour la suite des expériences.

Les évaluations extrinsèques sur l'analyse syntaxique ont montré que l'utilisation de plongements de mots permet d'améliorer les résultats, en particulier en monolingue et multilingue. En zero-shot, bien que l'utilité des plongements de mots ne soit pas systématique pour toutes les langues, le lexique semble tout de même être majoritairement utile.

De nombreuses améliorations pourraient être proposées pour rendre les plongements de mots plus performants. Par exemple, les mots des dictionnaires issus de PanLex étaient rangés par fréquence lorsqu'un corpus d'apprentissage était disponible pour apprendre ces fréquences. Mais pour les langues n'en disposant pas, aucun traitement n'était effectué, alors qu'il aurait été possible de les ranger par ordre de fréquence de leur traduction en anglais (en). Nous aurions pu éviter l'étape d'harmonisation, ou nous aurions pu moyenner en pondérant avec la fréquence d'apparition du mot dans chaque langue, ce qui semble être une meilleure solution, puisque les corpus d'apprentissage de plongement de mots sont connus pour nos expériences. L'alignement des espaces de plongements de mots aurait pu se faire en utilisant MUSE en mode non supervisé, afin de voir si les résultats perdent en qualité ou non. Et évidemment, il serait très intéressant de tester l'utilisation de plongements de mots contextuels de type BERT multilingue, afin de voir si ceux-ci obtiennent de meilleurs

résultats que nos plongements de mots actuels.

Un lexique *universel* étant maintenant à notre disposition, nous allons pouvoir utiliser les mots comme entrées de nouvelles tâches, comme cela a été fait dans ce chapitre en lexicalisant l'analyseur. Nous verrons ainsi dans le chapitre suivant comment le lexique peut être utilisé pour une tâche de prédiction des POS.

Chapitre 4.

Étiquetage en parties du discours

Sommaire

4.1 Introduction 89
4.2 État de l'art 91
4.3 Étiqueteur 92
4.4 Résultats et analyses 96
4.4.1 Monolingue 97
4.4.2 Multilingue 98
4.4.3 Zero-shot 103
4.4.4 Variabilité en zero-shot 109
4.4.5 Familles de langues 116
4.5 Conclusions 121

4.1. Introduction

Avec les traits morphologiques, les POS sont les seules informations qui étaient données en entrée de l'analyseur syntaxique du modèle délexicalisé présenté au chapitre 2. Les entrées de ce modèle étaient les POS et les traits morphologiques de références. Il n'est cependant pas réaliste de considérer que ces annotations de références soient toujours à notre disposition. Toujours dans l'optique de s'émanciper le plus possible du besoin de données annotées, la prédiction des POS semble être une condition nécessaire afin de tendre vers des modèles multilingues plus performants. Leur prédiction est donc incontournable.

Dans le chapitre 2 sur l'analyse syntaxique délexicalisée, nous nous reposions sur les traits morphologiques et les POS de références qui étaient annotés en suivant le guide d'annotation de UD, et qui sont donc universels entre les langues. Or, pour prédire les POS, nous n'avons plus beaucoup de données disponibles pour l'entraînement : les mots, et les caractères qui les composent. Or, les lexiques et les alphabets ne sont pas universels à travers les langues.

Pour pallier ce problème, nous avons construit nos plongements de mots décrits dans le chapitre 3 afin de disposer d'une représentation lexicale universelle. Tous les plongements de mots appartiennent donc à un espace vectoriel commun à toutes les langues.

L'expression de la morphologie est souvent portée par les préfixes et les suffixes des mots. C'est pourquoi l'utilisation des caractères des mots semble essentielle. Nous allons utiliser un *modèle de caractères* dans notre système nous permettant d'accéder aux préfixes et suffixes de chaque mot afin d'aider à identifier les marqueurs morphologiques qui aideront pour la prédiction des POS.

Déjà d'un point de vue monolingue, la morphologie du mot est une vraie aide pour la prédiction des POS. Par exemple, le suffixe *ont* en français (fr) indique une forte probabilité que le mot soit un verbe.

De plus, certaines langues ont des marqueurs morphologiques communs, comme le 's' en fin de mot qui peut être un marqueur du pluriel en français (fr), en anglais (en), en espagnol (es) et bien d'autres.

— Chat ⇒ Chat**s** (français)

— Cat ⇒ Cat**s** (anglais)

— Gato ⇒ Gato**s** (espagnol)

Ces similarités peuvent être une caractéristique intéressante pour identifier des régularités utilisables dans un cadre d'apprentissage zero-shot, où il faut tirer parti de connaissances apprises sur des langues et les appliquer à une langue jamais vue à l'entraînement.

Les caractères sont la plupart du temps contenus dans un alphabet assez limité, et sont souvent partagés entre plusieurs langues. Les caractères peuvent être un point commun entre les langues partageant un même alphabet pouvant potentiellement aider à un partage d'informations entre les langues.

Nous verrons cependant que l'utilisation des caractères est une arme à double tranchant, pouvant se montrer très utile pour deux langues proches (espagnol (es) et catalan (ca)) ou au contraire poser de vrais problèmes en rapprochant des langues utilisant le même alphabet sans être proches (arabe (ar) et ourdou (ur)), ou en éloignant des langues très proches ayant des alphabets différents (hindi (hi) et ourdou (ur)).

Bien que n'ayant pas montré une amélioration nette des résultats au chapitre 2, le vecteur de traits typologiques W_{22} [1] issu du WALS pourrait potentiellement se montrer utile dans le cadre d'une tâche d'étiquetage en POS, et son utilité sera testée dans ces expériences où W_{22} servira de représentation de nos langues.

Nous allons explorer dans ce chapitre l'impact des différents éléments pouvant aider le partage de connaissances entre les langues pour une tâche d'étiquetage en POS : *les plongements de mots multilingues* permettant de capturer le "sens" des mots, *le modèle de caractères* qui permettra quant à lui de capturer les caractéristiques morphologiques, et enfin *un vecteur issu du WALS* qui permettra d'expliciter au modèle les points communs et différences entre les langues afin de l'inciter à rapprocher et éloigner certaines langues.

Les impacts du modèle de caractères et du WALS seront explorés dans des configurations multilingues (voir section 4.4.2) et de zero-shot (voir section 4.4.3), où l'on

1. W_{22} a été créé à partir du WALS dans le cadre de nos travaux du chapitre 2. Il permet d'expliciter des informations syntaxiques comme l'ordre du Sujet et du Verbe pour chaque langue.

mettra en évidence leur importance pour rapprocher ou éloigner certaines langues entre elles. On explorera également la question de l'importante variabilité des résultats en zero-shot dans la section 4.4.4, où contrairement aux intuitions de départ, nous verrons que la variabilité des résultats en zero-shot n'est que peu corrélée au taux de couverture lexicale. Nous verrons qu'en revanche, la présence d'une langue proche est déterminante pour les prédictions de zero-shot. Nous finirons en apprenant de nouveaux modèles entraînés sur de plus petits ensembles de langues dans la section 4.4.5, où nous constaterons qu'il n'est pas nécessairement plus intéressant d'entraîner les modèles sur un sous-ensemble de langues proches.

4.2. État de l'art

L'étiquetage en POS est une tâche ancienne, permettant de faciliter d'autres tâches telle que l'analyse syntaxique d'une phrase. Le premier corpus majeur annoté en POS, le *Brown Corpus*, date des années 60 (FRANCIS et KUCERA 1979). L'utilisation de modèles de Markov cachés dans les années 80 (MARSHALL 1983) a permis d'obtenir une exactitude des résultats comprise entre 93 et 95%. Depuis, de nombreuses méthodes ont été essayées, allant des étiqueteurs non supervisés (GOLDWATER et GRIFFITHS 2007) aux étiqueteurs multilingues de type *zero-shot* (LAUSCHER et al. 2020) similaires à BERT (DEVLIN et al. 2019) tels que des *transformers* massivement multilingues.

Les résultats pour la tâche d'étiquetage en POS dépassent aujourd'hui les 95% d'exactitude sur les langues bien dotées comme l'anglais (en) ou le français (fr). Cependant, la question du traitement des langues ayant peu de ressources annotées se pose, les modèles nécessitant des corpus annotés pour apprendre. Des solutions basées sur de l'apprentissage non-supervisé (BERG-KIRKPATRICK et al. 2010; CARDENAS et al. 2019; GOLDWATER et GRIFFITHS 2007; STRATOS et al. 2016) permettent de se passer d'annotations tout en obtenant des résultats exploitables (75.5% d'exactitude pour BERG-KIRKPATRICK et al. 2010).

D'autres solutions explorées sont celles exploitant les données annotées d'autres langues en créant des systèmes multilingues (LAUSCHER et al. 2020; PLANK et al. 2016; YASUNAGA et al. 2018). De nombreuses manières d'apprendre des systèmes multilingues existent. La première, extrêmement simple, consiste à prendre le modèle d'une langue $L1$ et de décoder un texte d'une langue $L2$ avec. En prenant deux langues proches, ce genre de système peut obtenir de bons résultats (ESKANDER et al. 2020 ont obtenu 88.7% d'exactitude lors du décodage d'un corpus de portugais avec un modèle de l'espagnol). J.-K. KIM et al. 2017 ont montré que leur système, appris sur une famille de langues (Germanique, Slaves, Romanes,...) permet d'obtenir de meilleurs résultats que des modèles monolingues.

Contrairement à la tâche de prédiction de l'analyse syntaxique, l'utilisation de ressources issues du WALS n'a que peu été explorée dans le cadre de prédictions des POS. ZHANG, REICHART et al. 2012 proposent d'utiliser des traits typologiques issus du WALS afin de convertir des jeux d'étiquettes de POS propres à une langue vers un jeu d'étiquettes de POS universel, et proposent dans ZHANG, GADDY et al. 2016 d'utiliser

des traits du WALS pour évaluer la qualité de plongements de mots multilingues. BJERVA et AUGENSTEIN 2021 ont montré qu'en empêchant l'accès à l'information de traits issus du WALS à l'aide de méthodes inspirées des réseaux antagonistes, les performances de modèle multilingues sur différentes tâches, incluant la prédiction des POS, été impactées négativement. Ce résultat implique que les traits du WALS sont potentiellement appris de manière spontanée. Fournir le WALS en entrée des systèmes pourrait alors être redondant. Les auteurs ont également montré que les traits issus de la catégorie de l'ordre des mots et ceux issus de la catégorie morphologie étaient utiles pour la tâche de prédictions des POS, contrairement aux traits issus de la catégorie phonologie et les traits liés à la généalogie des langues.

DEVLIN et al. 2019 a récemment permis d'obtenir des plongements de mots contextuels multilingues de très grande qualité, permettant de battre l'état de l'art sur de nombreuses tâches différentes, notamment sur des tâches d'apprentissage zero-shot (WU et DREDZE 2019) .

4.3. Étiqueteur

L'étiqueteur utilisé est celui du logiciel Macaon [2] (DARY et NASR 2021). Cet étiqueteur traite les mots de façon séquentielle dans le sens de la lecture. Pour chaque mot, un vecteur de traits représentant le contexte du mot à étiqueter ainsi que certaines caractéristiques de ce dernier sont extraits d'une fenêtre glissante centrée sur le mot. Ce contexte est donné en entrée d'un perceptron multicouche (MLP) jouant le rôle de classifieur. Ce MLP consiste en une couche d'entrée, deux couches cachées et une couche de sortie. Un schéma de ce réseau de neurones est disponible dans la figure 4.1. Les activations au niveau de la couche de sortie représentent une distribution de probabilités des parties de discours possibles pour le mot courant. Le système est glouton, il choisit la POS avec la plus haute probabilité pour le mot courant et passe au mot suivant.

2. https://gitlab.lis-lab.fr/franck.dary/macaon

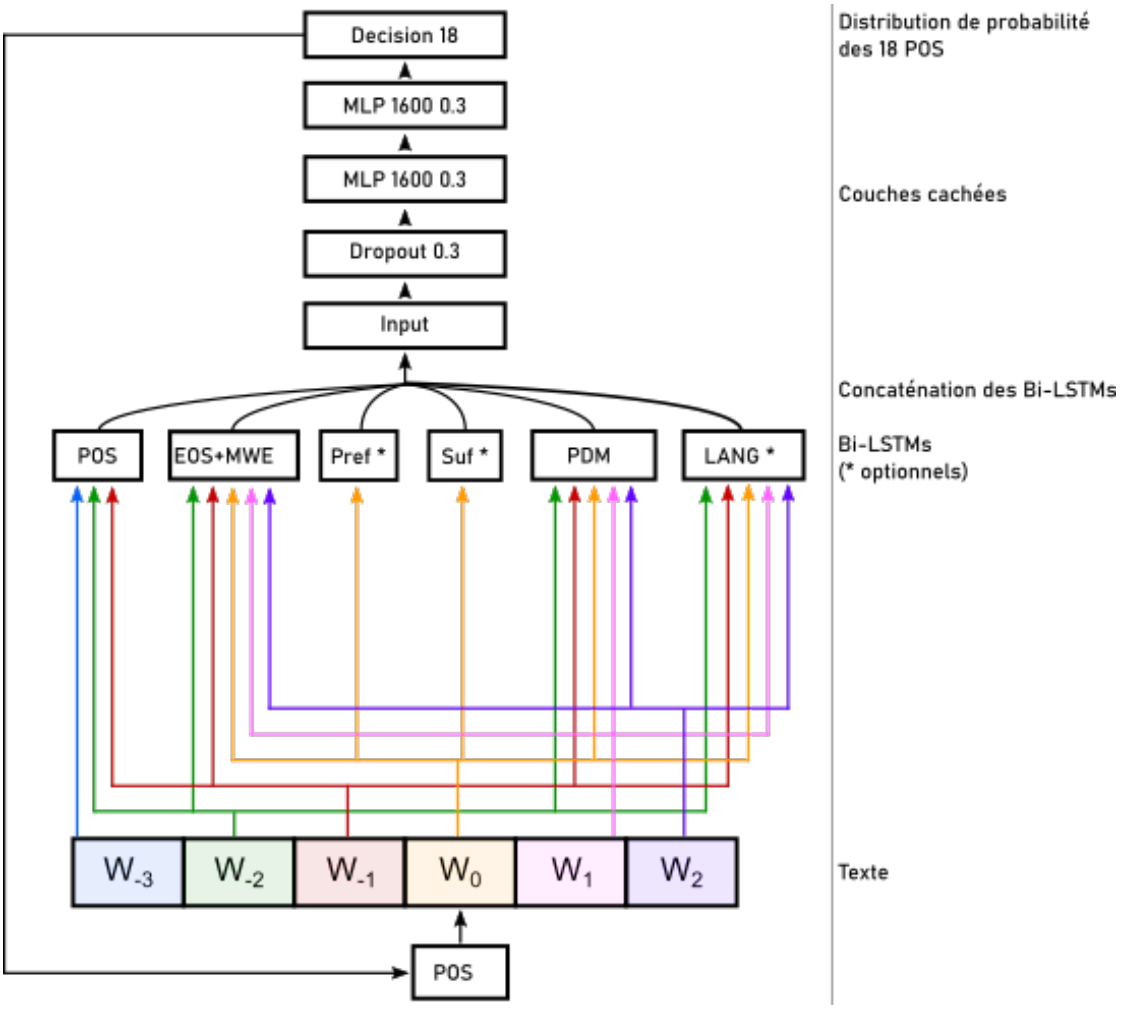


FIGURE 4.1. – Architecture de l'étiqueteur. Le texte est lu en entrée. Le mot courant est w_0, le mot précédent est w_{-1} et le mot suivant est w_1. 6 bi-LSTM sont représentés par les rectangles POS, EOS+MWE, ... et prennent en entrée le trait correspondant à leur nom. Les bi-LSTM avec une '*' sont optionnels et sont utilisés ou non selon la configuration de l'étiqueteur (voir section 4.3). La prédiction du mot courant est donnée en entrée pour le mot suivant.

Les entrées de l'étiqueteur correspondent à une fenêtre de taille 6, centrée sur le mot courant. Des informations sont obtenues à partir de ces mots, telles que leur représentation vectorielle, leurs préfixes et suffixes, les POS des mots précédents, etc... Ces traits sont fournis en entrée de réseaux récurrents bi-LSTMs qui permettront de contextualiser ces informations, pour ensuite les donner en entrée du MLP. La couche d'entrée de ce MLP consiste alors en la concaténation de la sortie de 3 à 6 bi-LSTM, décrits ci-dessous :

— PDM : un bi-LSTM prenant en entrée les plongements de mots multilingues de taille 300 définis au chapitre 3 et les contextualisant à partir d'une fenêtre

de taille 21 centrée sur le mot courant. Ces plongements de mots contextuels obtenus sont de taille 128. Les 5 plongements contextuels correspondant au mot courant, aux deux mots précédents et aux deux mots suivants sont alors fournis à la couche d'entrée. La sortie de PDM est donc de taille 128 * 5 = 640

— EOS+MWE : un bi-LSTM prenant en entrée l'information de la présence d'une fin de phrase (EOS) et la présence d'un token multi-mots (MWE, pour la présence d'expressions idiomatiques, *MultiWord Expressions*) parmi le mot courant, les deux mots précédents et les deux mots suivants. La sortie de ce bi-LSTM est de taille 64.

— POS : un bi-LSTM prenant en entrée les POS des trois mots précédents. La sortie de ce bi-LSTM est de taille 64. Lors des 4 premières itérations, les POS de références des mots précédents sont donnés. À la fin de la quatrième itération, le système fait un décodage de tout le corpus d'entraînement. À partir de la cinquième itération, le système utilise ces POS prédites à la place de celles de références. Lors de la phase de test, ce sont les prédictions des mots précédents qui sont utilisées.

— Pref : un bi-LSTM optionnel existant uniquement pour les expériences utilisant les caractères du mot. Il utilise le préfixe du mot courant (3 caractères) et sa sortie est un plongement de taille 64.

— Suf : un bi-LSTM optionnel existant uniquement pour les expériences utilisant les caractères du mot. Il utilise le suffixe du mot courant (3 caractères) et sa sortie est un plongement de taille 64.

— LANG : un bi-LSTM optionnel, existant uniquement dans les expériences utilisant W_{22}, dont l'entrée est de taille 80, correspondant aux 80 valeurs du vecteur W_{22} où les 22 traits sont représentés sous forme de one-hot puis concaténés. En sortie, on obtient un plongement de taille 64. Alternativement, ce bi-LSTM peut servir de représentation pour l'identifiant de la langue. [3]

On obtient ainsi une couche d'entrée de taille 640 + 64 + 64 (+64) (+64) (+64) = 832 à 1024 selon les configurations. Les deux couches cachées sont de taille 1600, avec un *dropout* durant l'entraînement de 0.3. Le nombre d'itération est de 40, la fonction d'activation est la fonction ReLu, la fonction d'erreur est un softmax de vraisemblance négative, et l'algorithme d'apprentissage est Adagrad.

Configurations d'entraînement Nos différentes configurations correspondent à une paire ⟨données d'entraînement, architecture⟩. Les données d'entraînement possibles sont *Mono*, *Multi* et *ZS*.

En tout, 234 étiqueteurs ont été entraînés sans compter les expériences de la section 4.4.5. Nous allons décrire quels sont les différentes configurations d'entraînement correspondant à ces étiqueteurs.

3. Pour des raisons de simplicité au moment de l'implémentation, l'information de la langue se base sur les 5 mots du contexte courant au lieu de se contenter du mot courant uniquement. L'utilisation du mot courant uniquement a été implémentée pour les expériences des chapitres 5 et 3, sans utiliser de bi-LSTM, mais un simple vecteur concaténé à la couche d'entrée

Nous reprendrons la base des définitions données au chapitre 2 (voir section 2.5).

Mono correspond à une unique langue au moment de l'entraînement, et l'étiquetage de cette même langue pour le test.

Multi représente un entraînement basé sur l'ensemble de toutes les langues disponibles, et l'étiquetage de chaque langue par ce modèle multilingue unique.

ZS est identique à *Multi*, auquel on retire une langue L. L'évaluation de *ZS* se fait sur l'étiquetage de la langue L uniquement.

Les différentes architectures correspondent à la façon de paramétrer notre étiqueteur. Le modèle peut utiliser les caractères des mots ($+c$) ou non ($-c$) et utiliser une représentation de la langue (ID ou W_{22}) ou non.

Nos différentes configurations d'entraînement sont donc :

Mono : Corpus monolingue.
Le corpus d'entraînement est L, et le corpus de test et également celui de L. 38 étiqueteurs sont entraînés, un pour chaque langue. C'est le cas classique d'apprentissage, où l'entraînement se fait sur une langue L, et le décodage se fait sur cette même langue L.

Multi : Corpus multilingue.
Un étiqueteur est entraîné sur nos 38 langues, sans indication de la langue. Un seul modèle est entraîné, et chaque corpus de test est décodé avec ce modèle. Ce modèle est très bruité, puisqu'il est possible de retrouver des contradictions entre les langues (*a*, en français (fr) est un verbe conjugué (du verbe *avoir*), alors qu'en anglais (en), *a* est un déterminant).

Multi W_{22} : Corpus + WALS.
Idem à *Multi* en utilisant le bi-LSTM LANG qui permet d'associer à chaque mot un vecteur $W(L)$ issu du WALS, qui correspond à la langue du mot. Contrairement à l'utilisation d'un identifiant de la langue, le WALS est porteur d'informations sur la syntaxe, la morphologie des mots, etc... pouvant aider les prédictions.

Multi ID : Corpus multilingue + identifiant de la langue.
Idem à *Multi* en utilisant en plus le bi-LSTM LANG qui permet de représenter l'identifiant de la langue de chaque mot. Cette expérience permet d'expliciter l'information de la langue en cours de traitement pour limiter le bruit apporté par l'apprentissage multilingue, mais sans expliciter les points communs à plusieurs langues.

ZS **:** *Zero-shot.*
38 étiqueteurs sont entraînés, sans représentation de la langue. L'étiqueteur de la langue L sera entraîné sur toutes les langues, sauf L, soit un ensemble de 37 langues. Cette configuration représente la situation où un étiqueteur est entraîné pour une langue pour laquelle aucune donnée d'entraînement n'est disponible. Les plongements de mots sont tout de même disponibles pour les prédictions, ainsi que les caractères des mots.

ZS W_{22} **:** *Zero-shot* + WALS

Ajoute à *ZS* l'indication de la langue de chaque mot à travers son vecteur du WALS, à l'identique de la différence entre *Multi* W_{22} et *Multi*.

Chacune de ces configurations est déclinée en deux versions, selon qu'elle utilise ou non un modèle de caractères : *Mono* $+c$, *Multi* $+c$, *Multi* $+c$ ID, *Multi* $+c$ W_{22}, *ZS* $+c$ et *ZS* $+c$ W_{22} utilisent les bi-LSTM correspondant au modèle de caractères (Pref et Suf), alors que *Mono* $-c$, *Multi* $-c$, *Multi* $-c$ ID, *Multi* $-c$ W_{22}, *ZS* $-c$ et *ZS* $-c$ W_{22} ne les utilisent pas. Cela nous permettra de tester l'impact de l'utilisation des caractères dans nos expériences.

Chaque expérience est réalisée deux fois, afin de limiter l'impact de la variabilité des expériences. Les résultats correspondent à la moyenne des scores des deux modèles.

Corpus Les corpus utilisés dans ces expériences sont les mêmes que ceux des expériences précédentes, décrits dans la section 2.5 à une langue près, le norvégien (no). Dans nos expériences sur l'analyse délexicalisée, nous avions fait le choix de considérer le norvégien nynorsk (nno) et le bokmål (nob) comme une seule et même langue, puisque dans le WALS, une seule langue correspondant au norvégien était disponible. Cependant, il s'avère que les deux langues sont en réalité assez différentes [4]. Ces langues n'étant finalement pas si proches, nous avons fait le choix de les séparer pour les nouvelles expériences.Afin de ne pas modifier les corpus pour pouvoir se comparer aux précédentes expériences, nous avons conservé nos corpus tels qu'ils étaient. Par conséquent, le norvégien nynorsk (nno) et le bokmål (nob) n'ont que 10 000 (resp. 1 000) tokens dans leur corpus d'entraînement (resp. de développement), contrairement au 20 000 (resp. 2 000) des autres langues.

Métriques La métrique utilisée pour l'évaluation de l'étiquetage de POS est l'exactitude.

Comme dans le chapitre 2, le script d'évaluation de la campagne d'évaluation CoNLL 2017 a été utilisé. il permet de calculer l'exactitude par langue, ainsi que la moyenne des exactitudes, qui est la macro-moyenne de l'exactitude de toutes les langues qui ont un corpus d'entraînement. Cette macro-moyenne permet d'être indépendant de la taille des corpus de test, évitant de favoriser les langues ayant des corpus de test de taille importante.

4.4. Résultats et analyses

Analysons à présent les résultats des différentes configurations qui ont été réalisées sur la tâche d'étiquetage en POS. Les résultats des configurations en condition *Mono* sont étudiés, afin d'établir une base de résultats de référence. L'impact du WALS et du modèle de caractères sur les résultats dans les configurations *Multi* et *ZS* sont également étudiés, et nous verrons comment ces entrées aident à rapprocher et éloigner

4. le norvégien du WALS correspondant au bokmål

certaines langues entre elles. Une forte variabilité des résultats est constatée en condition de zero-shot. L'origine de cette variabilité est très fortement liée à la présence de langue proche de la langue cible au moment de l'entraînement. Nous finirons nos analyses avec des expériences réalisées sur des sous-ensembles de langues, des *familles* de langues, pour voir s'il est réellement nécessaire de s'entraîner sur notre ensemble complet de langues.

Seule une partie des résultats sera présentée, mais les résultats complets sont disponibles en annexes (voir tableau .12).

4.4.1. Monolingue

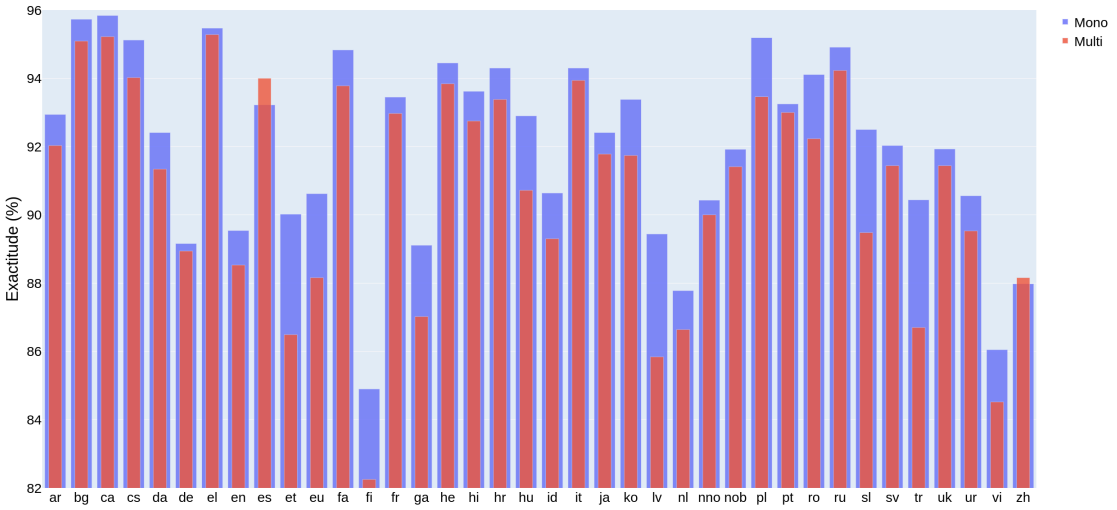


FIGURE 4.2. – Exactitude de la prédiction des POS pour les expériences en conditions monolingues et multilingues

L'objectif de cette première expérience est d'établir une base de référence constituée de nos résultats monolingues, afin de pouvoir comparer l'impact du passage à la configuration multilingue sur les résultats.

Les résultats sont présentés dans le tableau 4.1 et la figure 4.2. Une importante variabilité des résultats est constatée : pour l'expérience *Mono +c*, la moyenne est de 92.02 points, mais les résultats varient de 84.90 pour le finnois (fi) à 95.47 pour le grec (el), soit un écart de 10.57 point d'exactitude.

Cette variabilité se retrouvait également dans le cadre de l'analyse syntaxique délexicalisée, avec des écarts cependant plus marqués : le LAS variait de 46.78 pour le turc (tr) à 81.44 points l'italien (it), soit 34.66 points d'écart. Les résultats de ces deux tâches sont assez corrélés, avec un score de corrélation de Spearman de 0.61. Cette corrélation pourrait mettre en évidence l'existence de difficultés intrinsèques aux

Lang.	*Mono +c*	*Multi +c*	Δ
es	93.22	94.00	-0.78
zh	87.98	88.16	-0.18
el	95.47	95.28	0.19
he	94.45	93.84	0.61
vi	86.05	84.52	1.53
fi	84.90	82.25	2.65
lv	89.44	85.84	3.60
tr	90.44	86.70	3.74
Moy.	92.02	90.81	1.21

TABLEAU 4.1. – Exactitude de la prédiction des POS pour les expériences en conditions monolingues et multilingues de certaines langues.

langues ou aux corpus utilisés. Si les langues ayant les résultats les plus bas et celles ayant les résultats les plus hauts sont globalement les mêmes, quelle que soit la tâche, cela va dans le sens de l'hypothèse qu'une langue, ou un corpus, présente des difficultés intrinsèques (une morphologie riche, beaucoup d'ambiguïté, des dépendances syntaxiques lointaines, ...).

4.4.2. Multilingue

De *Mono* +*c* à *Multi* +*c* Lors du passage au contexte multilingue avec modèle de caractères, les résultats diminuent de seulement 1.21 point en moyenne. Cette diminution de seulement 1.21 en moyenne semble assez surprenante, car vraiment très faible comparée à la chute de 5.68 points constatée lors des expériences menées sur l'analyse syntaxique[5]. Une différence majeure entre ces expériences est qu'à présent, nous utilisons les mots, sous la forme de plongements de mots multilingues et d'un modèle de caractères.

Comme pour l'analyse syntaxique, certaines les langues souffrent de ce passage en contexte multilingue, comme le turc (tr) ou le letton (lv), perdant respectivement 3.74 et 3.60 points, qui sont les deux langues dont les résultats diminuent le plus lors du passage au multilingue. On notera deux exceptions, l'espagnol (es) et le chinois (zh) qui gagnent respectivement 0.78 et 0.18 point lors du passage en multilingue. Dans les expériences sur l'analyse syntaxique délexicalisée, l'espagnol (es) était l'unique langue profitant du passage en multilingue, ce qui reste cohérent avec les résultats pour l'étiquetage en POS. Cependant, le chinois (zh) perdait 14.37 points, contrairement à l'augmentation de 0.18 point actuelle. C'était la langue dont les résultats chutaient le plus lors du passage en multilingue. Une explication possible de ce résultat serait que le chinois (zh) peut tirer parti d'autres langues présentes dans l'ensemble d'entraînement pour la tâche d'étiquetage, mais que les langues proches pour cette tâche ne le

5. Les résultats sont comparés à ceux du chapitre 2 et pas à ceux du chapitre 3 utilisant la nouvelle architecture, pour des raisons chronologiques.

Lang.	*Multi* $+c$	*Multi* $-c$	$app(mdc, Multi)$
ur	89.53	87.84	1.69
hi	92.75	90.39	2.36
ar	92.03	89.09	2.94
fa	93.78	90.44	3.34
fi	82.25	75.53	6.72
vi	84.52	76.84	7.68
hu	90.72	82.92	7.80
et	86.49	78.15	8.34
tr	86.70	77.80	8.90
ko	91.74	80.60	11.14
Moy.	90.81	85.36	5.44

TABLEAU 4.2. – Exactitude de la prédiction des POS pour les expériences en conditions multilingues, avec et sans modèle de caractères, de certaines langues.

sont pas nécessairement pour la tâche d'analyse syntaxique.

Modèle de caractères Une différence majeure avec les résultats de l'analyse syntaxique délexicalisée est l'utilisation d'un modèle de caractères. Afin d'estimer l'apport du modèle de caractères ($app(mdc, Multi)$), les résultats des expériences *Multi* $+c$ et *Multi* $-c$ sont comparés. Les résultats sont disponibles dans le tableau 4.2.

En moyenne, l'$app(mdc, Multi)$ est de 5.44 points d'exactitude. Toutes les langues bénéficient du modèle de caractères, allant d'une augmentation de 1.69 point pour l'ourdou (ur) à 11.14 points pour le coréen (ko). L'utilité d'un modèle de caractères en condition multilingue est donc évidente.

En théorie, les langues morphologiquement riches comme le turc (tr) devraient bénéficier du modèle de caractères. Nous avons vu précédemment que le modèle de caractères devrait permettre de mieux capturer les marqueurs morphologiques des langues, ce qui devrait améliorer les résultats.

Les langues très riches d'un point de vue morphologique devraient également bénéficier du modèle de caractères, les préfixes et suffixes des mots étant souvent porteurs des marqueurs morphologiques.

Avec des augmentations de +6.72 pour le finnois (fi), +7.80 pour le hongrois (hu), +8.34 pour l'estonien (et) ou encore +8.90 pour le turc (tr), l'hypothèse de l'importance du modèle de caractères pour la prédiction des POS pour les langues morphologiquement riches semble vérifiée. On notera tout de même une exception pour l'arabe (ar), avec une augmentation de seulement +2.94. L'alphabet de l'arabe (ar) étant le même que pour le persan (fa) et l'ourdou (ur), il est possible que le modèle de caractères ait introduit du bruit lié à ces langues, correspondant par exemple à des homographes ayant des POS complètement différentes, ce qui expliquerait la faible augmentation des résultats comparé à ce qui était attendu.

Le modèle de caractères est indubitablement utile, puisque bénéfique pour toutes

les langues. Ces résultats sont cohérents avec les conclusions du chapitre 3 sur l'étiquetage des POS (voir section B.2 en annexe). $app(mdc, ZS)$ de l'analyseur syntaxique et $app(mdc, ZS)$ de l'étiqueteur en POS du chapitre 4 sont d'ailleurs corrélés à 0.46 avec la mesure de Pearson.

Nous avons également cherché à tester l'impact des plongements de mots, comme nous l'avons fait pour le modèle de caractères. Une expérience envisagée serait donc d'entraîner un modèle avec uniquement les caractères en entrée. L'apport des plongements de mots dans le cadre d'une tâche d'analyse syntaxique a cependant été abordée dans la section 3.8.1, où le système pouvait se reposer sur d'autres entrées, telles que les POS et les traits morphologiques.

WALS Les résultats du chapitre 2 ont montré que l'ajout d'un vecteur issu du WALS aidait à partager des informations entre les langues. Nous cherchons à vérifier si ce phénomène a toujours lieu dans le cadre d'une tâche d'étiquetage en POS. Les informations typologiques tels que l'ordre du verbe et du sujet semblent être pertinentes pour la prédiction des POS, et nous testerons la capacité du système à tirer profit de ces informations. Les résultats sont présents dans le tableau 4.3 et la figure 4.3.

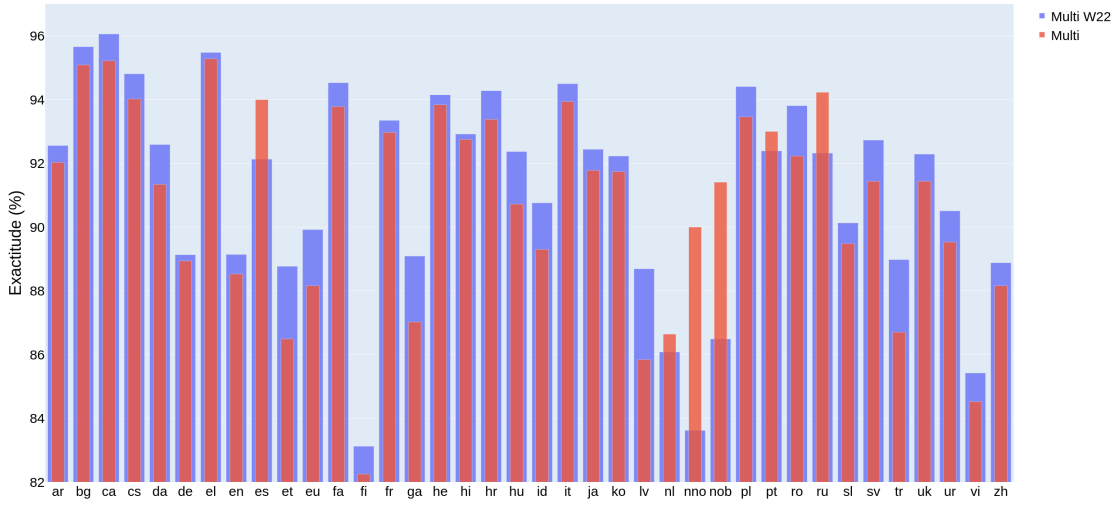


FIGURE 4.3. – Exactitude de la prédiction des POS pour chaque langue pour les expériences multilingues, avec et sans vecteur de traits typologiques issu du WALS

L'ajout du vecteur W_{22} fait gagner, en moyenne, 0.42 point, l'ajout du vecteur a donc permis de limiter un peu la baisse des résultats liée au passage en multilingue, sans pouvoir la compenser entièrement, le passage de *Mono +c* à *Multi +c* ayant fait baisser les résultats de 1.21 point.

Lang.	*Multi +c*	*Multi +c* W_{22}	Δ	*Multi −c*	*Multi −c* W_{22}	Δ
nno	90.00	83.62	-6.38	84.32	78.33	-5.99
nob	91.41	86.49	-4.92	85.96	81.54	-4.42
ru	94.23	92.32	-1.91	88.47	86.73	-1.74
es	94.00	92.13	-1.87	90.83	89.88	-0.95
pt	93.00	92.39	-0.61	88.87	88.75	-0.12
nl	86.64	86.08	-0.56	80.64	81.16	0.52
hi	92.75	92.92	0.17	90.39	91.19	0.80
de	88.94	89.13	0.19	82.06	82.46	0.40
ur	89.53	90.51	0.98	87.84	88.70	0.86
lv	85.84	88.69	2.85	78.53	82.69	4.16
moy.	90.81	91.23	0.42	85.36	86.46	1.10

TABLEAU 4.3. – Exactitude de la prédiction des POS pour chaque langue pour les expériences multilingues, avec et sans vecteur de traits typologiques issu du WALS. Les colonnes *Multi −c* et *Multi −c* W_{22} ont été ajoutées pour faciliter la lecture.

Bien que majoritairement utile (augmentation des résultats pour $\frac{32}{38}$ langues) les résultats varient d'une langue à l'autre. Le letton (lv) est la langue qui bénéficie le plus de l'ajout de W_{22}, avec un gain de 2.85 points. L'ajout du vecteur W_{22} semble bénéficier plus aux langues qui avaient le plus souffert du passage en multilingue. Pour vérifier cela, nous avons calculé la corrélation de Spearman entre Δ(*Multi +c* W_{22}, *Multi +c*), "l'apport du WALS", et Δ(*Mono +c*, *Multi +c*), "la chute liée au multilinguisme", et avons obtenu un score de corrélation de 0.70, indiquant l'existence d'une corrélation entre ces résultats. L'augmentation des résultats lors de l'ajout de W_{22} est donc très corrélée à la baisse des résultats liée au passage en contexte multilingue. Autrement dit, le WALS permet d'aider les langues dont les résultats baissaient le plus lors du passage au multilingue. L'espagnol (es), le portugais (pt) et le russe (ru), au contraire, ne tirent aucun bénéfice de ce vecteur, faisant baisser leurs résultats. L'aide que le WALS apporte aux langues souffrant le plus du passage au multilingue pourrait être liée à une possible isolation des langues par rapport à l'ensemble, que le WALS permet de compenser en explicitant des points communs entre la langue cible et les différentes langues de l'entraînement.

Le cas des deux norvégiens (nynorsk nno et bokmål nob) est particulièrement intéressant, l'ajout du vecteur faisant chuter les résultats de respectivement 6.38 et 4.92 points. Ces deux langues ayant un vecteur W_{22} identique, il est possible que le système *Multi +c* W_{22} ait perdu sa capacité à les distinguer, menant à une baisse des résultats.

Pour comparer le comportement des systèmes lors de l'ajout du vecteur W_{22} et lors de l'ajout de l'identifiant ID, nous avons calculé la corrélation entre Δ(*Multi +c* W_{22}, *Multi +c*), "l'apport du WALS", et Δ(*Multi +c* ID, *Multi +c*), "l'apport d'un identifiant de la langue". On obtient un résultat de 0.95 pour Spearman, et 0.98 pour Pearson. Ce

résultat met en évidence que les langues bénéficiant du WALS et celles qui n'en tirent pas parti sont les mêmes que lors de l'utilisation d'un simple identifiant de la langue, appuyant fortement l'hypothèse que W_{22} serait utilisé comme un simple identifiant de la langue, le comportement des systèmes étant le même, qu'on utilise le WALS ou un identifiant de la langue.

Deux autres paires de langues possèdent des vecteurs identiques : le néerlandais (nl) et l'allemand (de) ainsi que l'hindi (hi) et l'ourdou (ur). Si la baisse des résultats pour les norvégiens est causée par l'incapacité du système à les distinguer à cause de leurs vecteurs identiques, alors les scores des paires nl-de et hi-ur devraient également baisser. C'est le cas du néerlandais (nl), qui est une des six langues pour lesquelles les résultats diminuent. Ce n'est cependant pas le cas de l'allemand (de), même si l'augmentation est faible (+0.19). Ces résultats vont tout de même légèrement dans le sens de l'hypothèse selon laquelle le système n'est pas capable de distinguer des langues ayant un vecteur identique, menant à une légère baisse des résultats plutôt qu'à une augmentation.

Le cas de l'hindi (hi) et l'ourdou (ur) est particulier. Ces deux langues sont à l'origine une seule et même langue (l'hindoustani). Elles sont cependant basées sur deux systèmes d'alphabets différents (le devanagari pour l'hindi (hi), et l'alphabet perso-arabe pour l'ourdou (ur)). L'utilisation de deux alphabets est un indice fort pour le système, qui doit être capable de distinguer ces deux langues, même si leur vecteur issu du WALS est identique.

Afin d'empêcher le système d'utiliser les caractères pour différencier l'hindi (hi) et l'ourdou (ur), on observera la différence entre *Multi −c* et *Multi −c* W_{22}, où le modèle de caractères n'est pas utilisé. La paire hi-ur ne devrait plus être distinguable, à l'instar de la paire nno-nob, et les résultats devraient baisser en utilisant le vecteur du WALS. Or ce n'est pas le cas, les résultats de l'hindi (hi) et de l'ourdou (ur) augmentant avec l'utilisation de W_{22}, même sans modèle de caractères (respectivement +0.80 et +0.86). Il n'est donc toujours pas possible d'affirmer que la baisse des résultats entre les deux norvégiens provienne de leur vecteur identique du WALS.

Une autre hypothèse est que dans le WALS, il n'existe qu'un norvégien, qui semble correspondre au bokmål (nob). S'il existe des différences entre les deux langues, le vecteur du nynorsk (nno) ne décrit pas correctement les caractéristiques de la langue. Mais le norvégien bokmål (nob) ne devrait alors pas souffrir du passage à *Multi +c* W_{22}.

Il est également possible que les deux norvégiens, ayant moins de tokens à l'entraînement que les autres langues, aient pu être lésés lors de l'apprentissage. Mais si c'était le cas, les résultats pour *Multi +c* devraient être impactés également.

Il semble donc qu'aucune des hypothèses que nous avons envisagées ne soit entièrement satisfaisante. Une étude plus poussée devra être mise en place pour apporter une réponse à cette question.

En conclusion, le passage au multilingue ne semble pas faire baisser les résultats aussi drastiquement que pour la tâche d'analyse syntaxique, avec une baisse moyenne de seulement 1.21 point. Le modèle de caractères a montré être systématiquement

utile, et sera considéré comme faisant partie de l'expérience de base dans la suite de nos travaux. L'utilisation du WALS semble quant à lui être bénéfique pour la majorité des langues, mais il n'est pas possible d'affirmer que ce vecteur n'est pas utilisé comme un simple identifiant de la langue.

4.4.3. Zero-shot

Nous allons à présent nous intéresser aux résultats dans un cadre d'apprentissage en zero-shot. Le contexte de zero-shot correspond à la situation où aucune donnée d'apprentissage annotée n'est disponible. Nous commencerons par observer les différences avec la configuration multilingue vue précédemment, puis on explorera l'impact du modèle de caractères et du WALS dans cette configuration.

De *Multi* à ZS

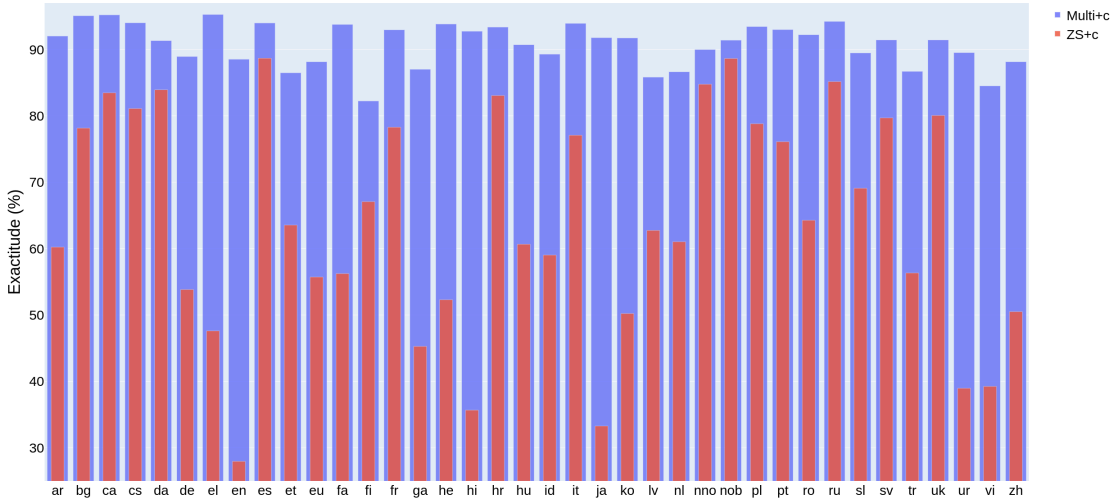


FIGURE 4.4. – Exactitude de la prédiction des POS pour chaque langue pour les expériences multilingues *Multi* $+c$ et celles de zero-shot ZS $+c$.

En moyenne, le passage du système *Multi* $+c$ au système ZS $+c$ fait chuter l'exactitude de 26.65 points (voir tableau 4.4 et figure 4.4). Cette chute était attendue, les conditions d'entraînement en zero-shot étant particulièrement difficiles. Certaines langues comme l'anglais (en) subissent une baisse très importante de leur résultat (-60.57 points). La plupart des langues ayant une telle baisse (japonais (ja), hindi (hi), ourdou (ur), ...) sont des langues très isolées ou ont un alphabet unique, qui n'aura par conséquent jamais été vu à l'entraînement. Le cas de l'anglais est particulier, car cette chute ne semble pas cohérente. Il sera étudié plus en profondeur à la fin de cette section.

Lang.	*Multi +c*	*ZS +c*	Δ
en	88.53	27.96	60.57
ja	91.78	33.28	58.50
hi	92.75	35.67	57.08
ur	89.53	38.97	50.56
es	94.00	88.69	5.31
nno	90.00	84.77	5.23
nob	91.41	88.65	2.76
moy	90.81	64.16	26.65

TABLEAU 4.4. – Exactitude de la prédiction des POS pour chaque langue pour les expériences multilingues *Multi +c* et celles de zero-shot *ZS +c*.

Certaines langues, au contraire, ne subissent qu'une légère baisse, comme les norvégiens (nno et nob) ou l'espagnol (es). Il s'agit de langues pour lesquelles il existe des langues proches dans l'ensemble d'apprentissage qui sont similaires, par exemple le catalan (ca) pour l'espagnol (es), cette dernière subit une baisse de seulement 5.31 points, une diminution bien inférieure à la diminution moyenne (-26.65).

Modèle de caractères Le modèle de caractères a prouvé son utilité dans le cadre de l'apprentissage multilingue. Mais se pourrait-il qu'au contraire, il puisse être néfaste dans un cadre d'apprentissage zero-shot ?

En effet, pour des langues comme l'hindi (hi), qui ne partage son alphabet avec aucune autre langue de l'ensemble d'entraînement, l'utilité d'un modèle de caractères peut être remise en question. Le modèle, lors de l'apprentissage, utilisera l'information des caractères pour améliorer ses prédictions. Lors du décodage sur une nouvelle langue dont il n'a jamais vu l'alphabet, le système ne pourra plus utiliser cette information sur laquelle il s'appuyait lors de l'apprentissage.

Deux langues peuvent également partager un alphabet mais n'avoir que très peu de racines lexicales ou de caractéristiques morphologiques communes. C'est le cas par exemple de l'ourdou (ur) et de l'arabe (ar). L'utilisation du modèle de caractères de l'une pour l'autre semble alors plus néfaste qu'utile. Nous allons vérifier ces hypothèses.

Lang.	$ZS+c$	$ZS-c$	Δ
hi	35.67	60.55	-24.88
ur	38.97	56.66	-17.69
fa	56.24	68.72	-12.48
ja	33.28	45.72	-12.44
el	47.61	60.02	-12.41
ar	60.23	66.88	-6.65
es	88.69	85.81	2.88
sl	69.10	65.31	3.79
ca	83.48	77.39	6.09
Moy	64.16	65.15	-0.99

TABLEAU 4.5. – Exactitude pour chaque langue pour les expériences zero-shot, avec et sans modèle de caractères, et différence entre les deux.

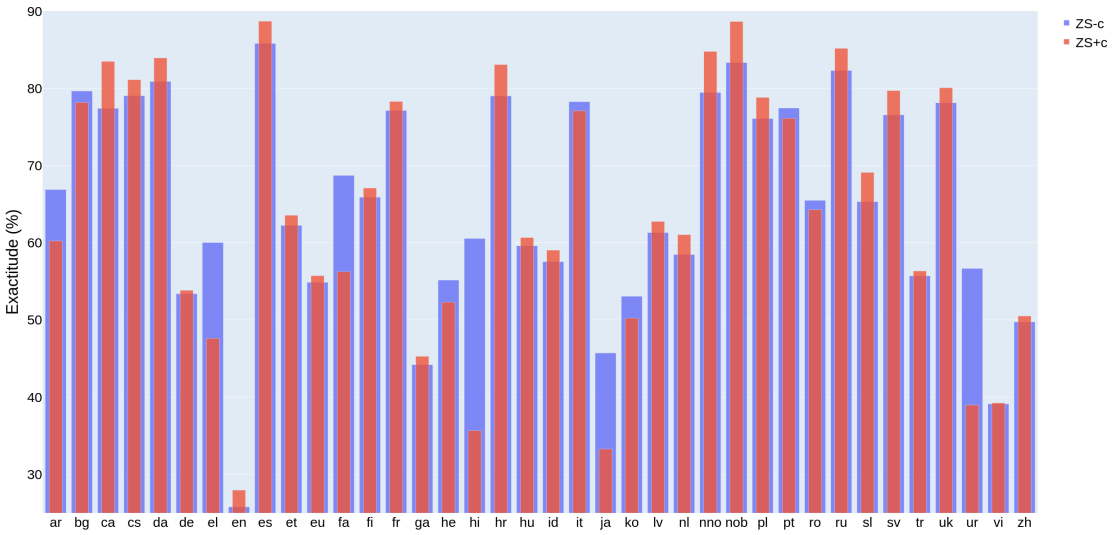


FIGURE 4.5. – Exactitude pour chaque langue pour les expériences zero-shot, avec et sans modèle de caractères

Nous comparons l'apprentissage des systèmes ZS avec et sans modèle de caractères (voir table 4.5 et figure 4.5).

En condition monolingue (*Mono*), le retrait du modèle de caractères a provoqué une chute moyenne de 5.31 points. En condition de zero-shot (ZS), lors du retrait du modèle de caractères, les résultats augmentent en moyenne de 0.99 point. Mais comme pour les expériences précédentes, ces résultats ne sont pas très intéressants en moyenne, les résultats variant énormément d'une langue à l'autre.

Six langues bénéficient fortement du retrait du modèle de caractères : l'arabe (ar), le grec (el), le persan (fa), l'hindi (hi), le japonais (ja) et l'ourdou (ur). Cette augmentation va de +6.65 pour l'arabe (ar) jusqu'à +24.88 pour l'hindi (hi).

Ces langues correspondent effectivement à des langues ayant un alphabet unique comme l'hindi (hi) et des binômes de langues comme l'ourdou (ur) et l'arabe (ar) qui sont basées sur le même alphabet alors que ce sont deux langues avec peu de points communs. Ces résultats vont ainsi dans le sens de notre hypothèse. Les langues qui tirent profit du modèle de caractères sont des langues comme l'espagnol (es) et le catalan (ca), qui partagent beaucoup de leurs racines lexicales et de leurs marqueurs morphologiques avec au moins une autre langue.

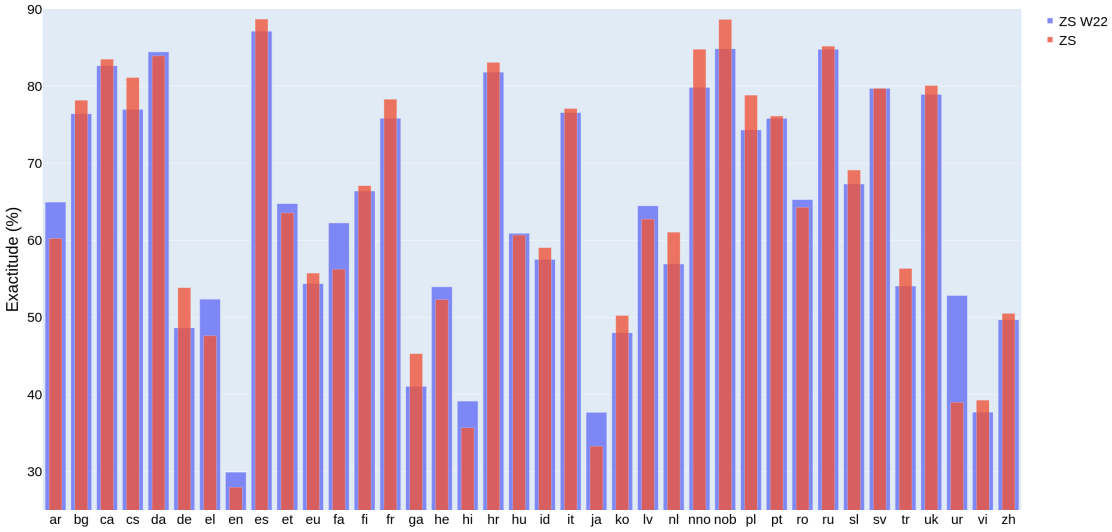


FIGURE 4.6. – Exactitude pour les expériences en conditions de zero-shot, avec et sans vecteurs de traits typologique issus du WALS.

WALS Le WALS permet de faire le rapprochement entre certaines langues, ou au contraire d'éloigner des langues partageant peu de caractéristiques communes, en explicitant des points communs entre les langues. En forçant ainsi le modèle à apprendre ces informations, on espère réussir à inciter les modèles à partager plus de connaissances entre les langues.

On observe que l'ajout de W_{22} diminue en moyenne les résultats de 0.22 point. Ce premier résultat, bien qu'un peu décevant, met en évidence que contrairement à la configuration *Multi +c*, l'utilité du vecteur est moins évidente. Il est bien plus intéressant d'observer les résultats langue par langue afin de tirer des conclusions.

L'allemand (de) et le néerlandais (nl) font partie des langues qui subissent une importante chute des résultats lors de l'ajout du vecteur W_{22}. Leurs vecteurs étant

Lang.	$ZS+c$	$ZS+c\ W_{22}$	Δ
de	53.84	48.63	-5.21
nl	61.04	56.91	-4.13
en	27.96	29.90	1.94
hi	35.67	39.12	3.45
ar	60.23	64.94	4.71
fa	56.24	62.24	6.00
ur	38.97	52.82	13.85
Macro moyenne	64.16	63.94	-0.22

TABLEAU 4.6. – Exactitude pour chaque langue pour les expériences zero-shot, avec et sans vecteurs de traits typologiques issus du WALS.

identiques, et la langue de test n'ayant pas été vue lors de l'apprentissage, le système ne peut à priori pas les distinguer. Ces langues étant probablement trop différentes en réalité, le vecteur W_{22} ne permet peut être pas de capturer correctement les différences entre ces langues.

Cependant, on remarque le phénomène inverse pour d'autres langues. L'ourdou (ur) gagne 13.85 points avec l'ajout de W_{22}. L'ourdou (ur) partage son vecteur avec celui de l'hindi (hi), mais pas son alphabet. On peut supposer que le vecteur du WALS a alors pu rapprocher ces deux langues, qui n'étaient pas proches à l'origine dans le système $ZS+c$. L'hindi (hi) profite aussi de l'ajout du vecteur W_{22} (+3.45), même si ce gain est moins prononcé que celui de l'ourdou (ur).

L'arabe (ar), le persan (fa) et l'ourdou (ur) partagent le même alphabet alors que ces langues sont assez éloignées les unes des autres. L'arabe (ar) et l'ourdou (ur) ont des valeurs différentes pour 12 des 22 traits de W_{22}, le persan (fa) et l'ourdou (ur) pour 12 traits également, et l'arabe (ar) et le persan (fa) pour 11 des 22 traits. L'ajout du vecteur W_{22} au système $ZS+c$ semble avoir éloigné ces langues malgré le partage de leur alphabet, permettant au système d'obtenir de meilleurs résultats.

L'ajout d'un vecteur de traits typologiques issu du WALS semble être bénéfique pour les langues présentant des similarités avec des langues de l'entraînement mais qu'il est nécessaire d'éloigner en réalité. Il est également utile pour les langues ayant une langue proche dans l'ensemble d'entraînement (comme pour l'hindi (hi) avec l'ourdou (ur)) qui sont en apparence assez éloignées pour le système (à cause de l'utilisation d'un alphabet différent) et qu'il faut tenter de rapprocher pour permettre un meilleur partage de connaissances entre les deux langues.

Les langues bénéficiant du vecteur W_{22} semblent être les mêmes que celles bénéficiant de l'absence du modèle de caractères. On a mesuré la corrélation de Pearson entre $\Delta(ZS-c,\ ZS+c)$, "ce que fait perdre le modèle de caractères" et $\Delta(ZS+c\ W_{22},\ ZS+c)$, "ce que fait gagner le WALS" pour vérifier cette intuition. On trouve un score de 0.74, indiquant que ces résultats sont bel et bien corrélés. Ce résultat indique que la présence du WALS ou l'absence du modèle de caractères permet d'aider les mêmes langues. On peut supposer que chacun à leur manière, le WALS et le retrait du modèle

de caractères permettent de faire des rapprochement entre langues, et d'en éloigner d'autres.

Le cas de l'anglais (en) Plusieurs hypothèses ont été envisagées pour comprendre l'incohérence des résultats de cette langue. En effet, alors que les résultats de l'anglais (en) était corrects en condition monolingue et multilingue, les résultats s'effondrent en zero-shot, bien plus que pour les autres langues, avec une chute de plus de 60 points lors du passage de $Multi+c$ à $ZS+c$, alors que cette langue ne semble pas spécialement isolée, contrairement au japonais (ja) ou à l'hindi (hi).

La première hypothèse était une possible différence de domaine entre le corpus d'apprentissage et de test. Après vérification, le corpus de test de l'anglais (en) contient beaucoup de texte oral, mais ce n'est pas la seule langue pour laquelle c'est le cas (danois (da), français (fr), grec (el), ...). Ce corpus de test ne semble pas particulièrement inhabituel, cette hypothèse est donc mise de côté.

Une autre piste consiste à vérifier le taux de couverture des mots du corpus de test par le corpus d'entraînement (voir tableau .11 en annexe). En effet, des mots de l'anglais (en) apparaissent dans presque toutes les langues, et donc certains mots ont pu être vus lors de l'entraînement. Est-ce que cela a pu avoir un impact sur les performances ? Nous avons calculé le taux de couverture des mots du corpus de test de l'anglais (en) par le corpus d'entraînement, donné par la formule suivante :

$$couvTestByTrain(L) = 100 * \frac{nb\ mots\ trouvés}{nb\ mots}$$

où $nb\ mots$ est le nombre de tokens du corpus de test et $nb\ mots\ trouvés$ est le nombre de tokens du corpus de test dont il existe au moins une occurrence dans le corpus d'entraînement de la langue L.

Nous avons également calculé le pourcentage de mots pour lesquels la POS a une valeur identique dans le train et le test, à l'aide d'une méthode similaire à la distance de Wasserstein.

On calcule $distribTrain(w,p)$ et $distribTest(w,p)$ le pourcentage de fois où w apparaît avec le POS p dans le train et dans le test.

Puis, pour chaque w, on calcule $match(w)$, correspondant au pourcentage de fois où w apparaît avec le même POS dans le train et le test.

$$match(w) = \sum min(distribTrain(w,p), distribTest(w,p))$$

POS	$distribTrain(w,p)$	$distribTest(w,p)$
p1	40	3
p2	60	92
p3	0	5
$match(w) = 3 + 60 + 0 = 63$		

Enfin, on calcule $matchTotal(l)$, qui correspond à la moyenne des $match(w)$ de tous les mots w du corpus de test, pondérée par le taux d'apparition de chaque mot. Les résultats sont disponibles en annexe, tableau .11.

Le score de corrélation de Pearson entre les résultats des expériences $ZS+c$ et les scores de $matchTotal(l)$ est de 0.81, et est de 0.62 entre $ZS+c$ et $couvTestByTrain(l)$, ce qui indique que la cohérence des POS des mots entre le train et le test est très importante pour les résultats des expériences $ZS+c$, et que le taux de couverture des mots du test par le train a un impact sur les résultats des expériences $ZS+c$. Pourtant, le score de l'anglais (en) pour ces deux mesures ne semble pas différent des autres langues. $couvTestByTrain(en)$ vaut 67.09, $couvTestByTrain(es)$ est assez similaire (67.84), pourtant, le résultat de l'anglais (en) pour l'expérience $ZS+c$ est de 27.96 et celui de l'espagnol (es) est de 88.69. $matchTotal(en)$ = 30.49, est un score similaire à celui du tchèque (cs), soit 29.67. Pourtant, le tchèque obtient un score de 81.11 pour l'expérience $ZS+c$.

L'anglais (en) fait figure d'outsider là encore. Ce ne sont donc ni la couverture des POS ni des mots entre l'entraînement et le test qui permettent d'expliquer les mauvais résultats de l'anglais (en).

Aucune des hypothèses explorées ne permet d'expliquer les résultats de l'anglais (en). Plus de recherches mériteraient d'être effectuées pour expliquer ce phénomène.

4.4.4. Variabilité en zero-shot

Nous avons observé que dans un cadre de zero-shot, les résultats varient énormément d'une langue à l'autre, avec un écart type de 17.06 entre les langues. Cette variation est bien plus importante que pour les expériences en monolingue (2.72 d'écart type) ou multilingue (3.23 d'écart type).

Deux hypothèses vont être explorées pour tenter d'expliquer cette variabilité. La première concerne la question du taux de couverture des mots de nos corpus par nos plongements de mots. Autrement dit, y a-t-il beaucoup de mots inconnus dans les textes du test? La deuxième hypothèse est celle de l'existence d'une langue "proche" dans l'ensemble d'entraînement, qui permettrait un meilleur partage de connaissances avec la langue cible. Deux manières de définir une langue proche seront définies :

— une langue proche au sens empirique, et une langue proche (en fonction des résultats d'un modèle monolingue d'une langue $L1$ appliqué à une langue $L2$)

— une langue proche définie en fonction du WALS, en comparant les vecteurs des langues

Mots inconnus Les plongements de mots sont bien souvent la donnée la plus fiable qu'aura l'étiqueteur à sa disposition dans un cadre d'apprentissage zero-shot. Nous avons vu en effet que dans certains cas, le modèle de caractères peut poser bien des problèmes, dans le cas où la langue cible est la seule utilisant son alphabet (comme pour l'hindi (hi), le coréen (ko), ...), comme vu dans la section 4.4.3. De plus, si le taux de couverture du corpus de test par les plongements de mots est faible, le modèle n'a

Lang.	couvEmbed(Lang.)	Lang.	couvEmbed(Lang.)	Lang.	couvEmbed(Lang.)
es	97.92	sv	95.22	he	92.16
fa	97.81	bg	95.06	hu	91.92
pl	97.65	sl	94.90	tr	91.90
pt	96.92	it	94.86	lv	91.34
el	96.60	ur	94.58	et	91.16
hr	96.36	da	94.53	ja	91.15
hi	96.32	nno	94.40	fi	90.75
id	96.21	uk	94.07	zh	90.57
ar	96.06	ro	94.02	kmr	90.08
ru	95.94	nl	93.80	ko	82.52
nob	95.93	ca	93.37	vi	78.03
de	95.80	eu	93.21	bxr	70.39
cs	95.78	ga	92.27	sme	69.86
en	95.46	fr	92.20		
moy.	92.42				

TABLEAU 4.7. – Taux de couverture des corpus de test de chaque par nos plongements de mots.

alors que vraiment peu d'informations pour ses prédictions (l'ordre des mots dans la phrase). Si les résultats sont corrélés à ce taux de couverture, cela pourrait être un début d'explication de la variabilité des résultats d'une langue à une autre. On définit le taux de couverture lexical de la façon suivante :

$$couvEmbed(l) = 100 * \frac{nb\ mots\ trouvés}{nb\ mots}$$

où $nb\ mots$ est le nombre de tokens du corpus, et $nb\ mots\ trouvés$ est le nombre de tokens pour lesquels un plongement de mot existe .

Le calcul des taux de couverture pour chaque corpus de train et de test de chaque langue est présenté dans le tableau .10 disponible en annexe, et un résumé des résultats est disponible dans le tableau 4.7.

La plupart des langues ont un bon $couvEmbed(l)$ avec des scores supérieurs à 90%, même s'il existe quelques langues pour lesquelles ce n'est pas le cas : le coréen (ko) avec 82.52%, le vietnamien (vi) avec 78.03%, le bouriate (bxr) avec 70.39% et le sami du nord (sme) avec 69.86%.

Pour estimer l'influence du taux de couverture sur les résultats de zero-shot, on calcule la corrélation de Pearson entre $couvEmbed(l)$ et les résultats des expériences $ZS+c$. La corrélation obtenue est de seulement 0.37. Ces mesures sont donc faiblement corrélées, la couverture des mots a un faible impact sur les résultats des expériences $ZS+c$. Le taux de couverture ne semble pas suffisant pour expliquer la variabilité des résultats en zero-shot, et il est nécessaire d'explorer d'autres pistes d'explications.

Existence d'une langue proche de façon empirique En zero-shot, on entraîne les modèles sur les 38 langues moins une, la langue cible. Parmi les 37 langues restantes dans l'ensemble d'entraînement, la présence d'une langue proche de la langue cible est-elle importante ? La variabilité des résultats en condition zero-shot pourrait-elle venir de là ?

L'existence d'une langue proche dans l'ensemble d'entraînement pourrait aider à avoir un meilleur partage de connaissances avec la langue cible. Pour calculer l'existence d'une langue proche de façon empirique, on s'intéresse aux résultats d'un étiqueteur monolingue de la langue $L1$ appliqué à la langue $L2$. On obtient une matrice 38x41 [6] permettant d'estimer de manière empirique la proximité des langues entre elles (voir tableaux .13 et .14 en annexe). La diagonale correspond à la situation classique apprentissage sur une langue $L1$, suivi du test sur cette même langue $L1$ (ce sont les résultats de l'expérience *Mono* $+c$).

On définit une nouvelle mesure, la langue la plus proche (LPP) de la langue L comme étant la langue dont le modèle donnera le meilleur score lors de l'étiquetage de la langue L. Le résultat du score de l'étiquetage de la langue L par la LPP nous donne une mesure empirique d'isolation de la langue L (voir tableau 4.8), où plus le résultat est bas, plus la langue est isolée.

Pour estimer si le score de la LPP joue un rôle dans les résultats de l'expérience $ZS+c$, on a mesuré la corrélation de Pearson entre les scores de ces deux mesures. On trouve alors une corrélation de 0.95. La présence d'une langue proche semble donc être déterminante pour expliquer les résultats qui seront obtenus dans un contexte de zero-shot, et les résultats des expériences de zero-shot dépendent du score de la LPP, et il semble possible de prévoir les résultats des expériences $ZS+c$ avec le score de la LPP. On peut voir figure 4.7 que l'augmentation des résultats $ZS+c$ est en effet très liée à celle des scores des LPP.

6. 38 langues disposant de données d'entraînement et donc d'un modèle, et 41 langues ayant des données de test (les trois langues de différence sont bxr, kmr, et smr).

Lang.	LPP	Score LPP	Lang.	LPP	Score LPP
ar	fa	53.00	it	es	66.46
bg	ru	71.63	ja	pl	39.99
bxr	cs	45.98	kmr	tr	38.12
ca	es	79.20	ko	cs	52.91
cs	sl	73.63	lv	sl	57.85
da	nob	80.12	nl	nob	54.61
de	nl	47.42	nno	nob	83.02
el	cs	44.74	nob	nno	84.07
en	fr	47.18	pl	cs	71.72
es	ca	82.20	pt	es	68.98
et	fi	62.67	ro	fr	55.91
eu	fi	52.73	ru	bg	76.50
fa	sl	54.94	sl	hr	66.54
fi	et	61.36	sme	fi	39.15
fr	ca	67.80	sv	da	73.60
ga	pl	38.87	tr	eu	53.71
he	cs	45.82	uk	ru	75.12
hi	sl	40.94	ur	sl	44.36
hr	sl	76.33	vi	ko	46.18
hu	pt	51.92	zh	ja	44.16
id	fi	56.12			

TABLEAU 4.8. – LPP pour pour les 41 langues de test. Les langues suivies en jaune appartiennent à la famille des langues Slaves, en rouge à la famille des langues Germaniques et en bleu à la famille des langues Romanes

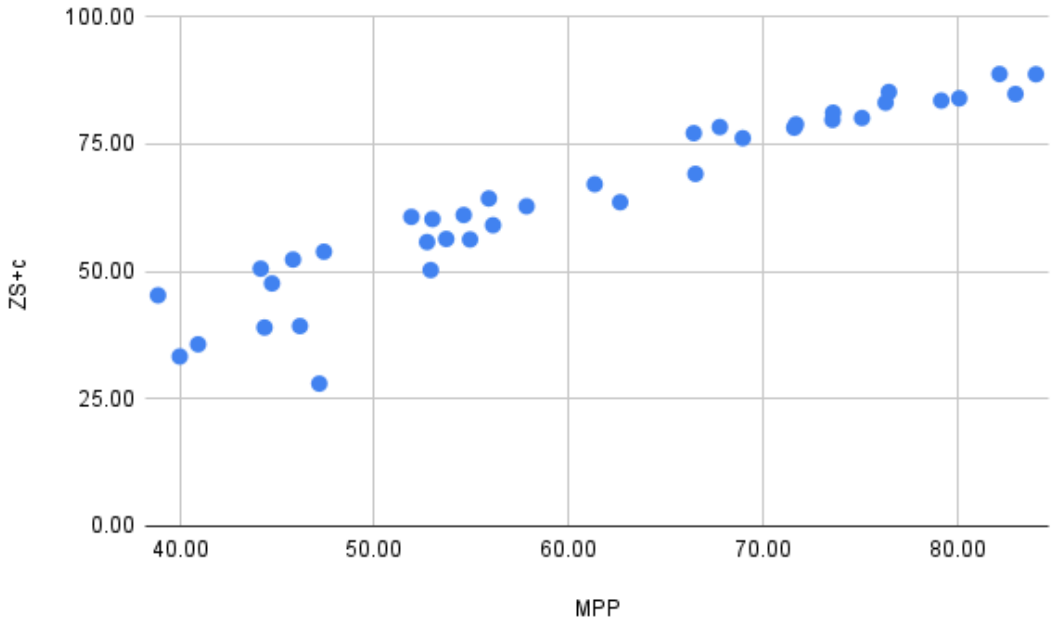


FIGURE 4.7. – Scores de $ZS+c$ par rapport aux scores des LPP.

Nous avons vu précédemment que le modèle de caractères pouvait avoir un fort impact sur les résultats en zero-shot. Cela ne change cependant pas l'importance du score de la LPP dans l'explication des résultats de zero-shot, la corrélation de Pearson entre $ZS-c$ et LPP-c étant de 0.92. [7]

Existence d'une langue proche au sens du WALS Le calcul de la LPP, bien qu'extrêmement utile, coûte aussi très cher en termes de temps et de calcul. Mais surtout, dans des conditions de zero-shot extrêmes, où il n'existe absolument aucune donnée annotée, pas même pour l'évaluation, il n'est même pas possible de trouver la LPP, puisqu'il ne sera pas possible de décoder puis d'évaluer un corpus de test. Les scores de la LPP étant très corrélés aux résultats de zero-shot, elle nous permettait de donner à l'avance une estimation de l'exactitude des prédictions. Nous aimerions trouver, à l'aide du WALS, une mesure d'isolation de la langue qui serait corrélée aux résultats $ZS+c$, à la façon de la LPP.

Pour quantifier l'isolement d'une langue, nous définissons l'indice de connectivité – ou *Connectedness Index* – (CI) d'une langue comme le nombre moyen de valeurs de traits qu'il partage avec les autres langues. Le CI permet de donner une mesure d'isolation d'une langue par rapport à un ensemble d'après le WALS. Cette mesure consiste à comparer deux à deux le vecteur W_{22} de la langue L avec toutes les autres langues d'un ensemble, et faire la moyenne du nombre de traits partagés avec d'autre langue :

7. Nous avons défini la même matrice 38x41 pour les modèles *Mono* $-c$ (voir tableaux .15 et .16 en annexe).

Lang.	CI	Lang.	CI
ar	57.00	id	58.23
bg	63.64	it	61.06
ca	53.93	ja	30.71
cs	54.55	ko	41.65
da	59.83	lv	54.05
de	47.17	nl	47.17
el	62.16	nno	59.09
en	65.11	nob	59.09
es	61.79	pl	64.25
et	62.04	pt	66.34
eu	39.07	ro	59.95
fa	43.24	ru	67.32
fi	57.13	sl	66.34
fr	58.85	sv	59.83
ga	53.69	tr	31.57
he	61.67	uk	68.30
hi	48.03	ur	48.03
hr	64.37	vi	57.49
hu	53.19	zh	53.69

TABLEAU 4.9. – CI pour W_{22} pour les 38 langues d'entraînement

$$CI(L) = \frac{100}{k} \sum_{f=1}^{k} \frac{1}{N-1} \sum_{L' \neq L} \delta(W(L', f), W(L, f))$$

où k est la dimension du vecteur du WALS, N[8] est le nombre de langues, $W(l, f)$ est la valeur du trait f pour la langue L, et δ est le Delta de Kronecker[9]. $CI(L)$ indique à quel point le vecteur du WALS pour la langue L partage ses valeurs avec les autres langues. $CI(L) = 0$ signifie que L est un cas particulier où la valeur des traits de la langue L ne se retrouve dans aucune autre langue. Dans le cas où $CI(L) = 100$, L partage toutes les valeurs de tous ses traits avec toutes les langues de l'ensemble. Cette situation n'arrive que si toutes les langues ont un vecteur identique. Le CI pour W_{22} pour les 38 langues d'entraînement est disponible dans le tableau 4.9.

8. ici, $N = 38$, pour les 38 langues de l'ensemble d'entraînement
9. $\delta(x, y) = 1$ si $x = y$ et 0 sinon

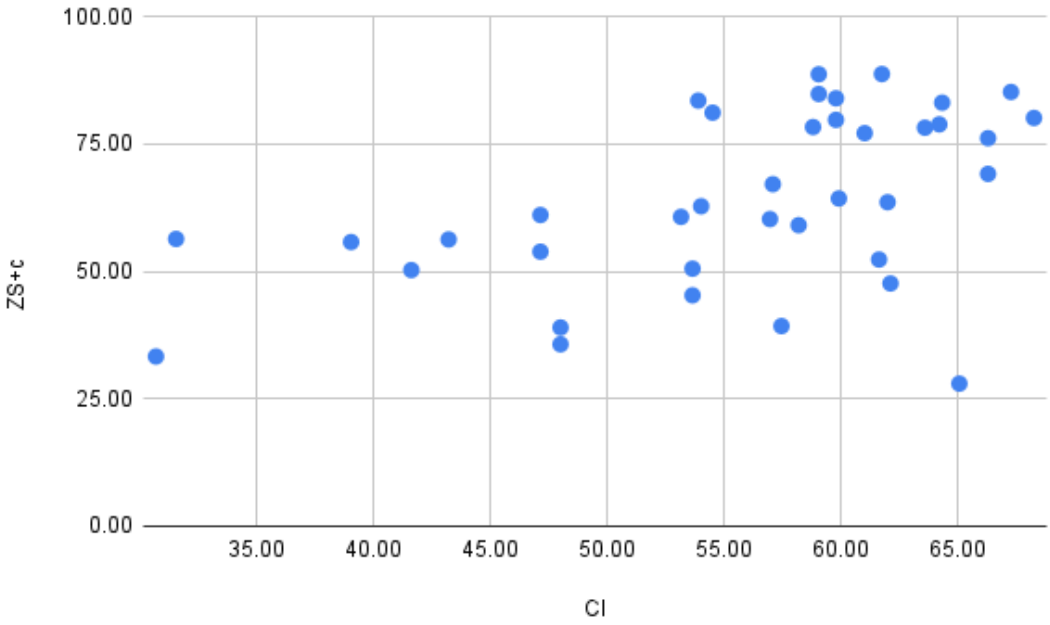


FIGURE 4.8. – Scores de $ZS+c$ par rapport aux CI de chaque langue.

Lorsqu'on mesure la corrélation de Pearson entre le CI et $ZS+c$, on obtient un score de 0.50. Ce score est malgré tout bien inférieur à la corrélation de 0.95 obtenue entre le score de la LPP et les scores de l'expérience $ZS+c$. Le WALS ne permet donc pas d'estimer aussi bien le score d'une langue en condition de zero-shot. Figure 4.8, on peut constater que les scores de $ZS+c$ sont bien moins dépendants du CI que ce qu'ils étaient de la LPP. Ils restent cependant suffisamment corrélés pour donner un début d'indication sur la variabilité d'une langue à l'autre en zero-shot.

En conclusion, la variabilité des résultats en zero-shot peut donc s'expliquer en partie par le taux de couverture lexicale, mais c'est principalement la présence d'une langue proche dans le corpus d'entraînement qui aura un impact sur les résultats pour les expériences de type $ZS+c$. Plus une langue est isolée, d'après le WALS mais surtout au sens empirique du terme, plus les résultats en contexte $ZS+c$ seront bas. Nous avons en effet pu constater une corrélation très importante (0.95 avec la mesure de Pearson) entre les résultats de zero-shot et le scores des LPP. La corrélation avec les scores des CI, bien que moins importante (0.50 avec Pearson) reste une source d'explication de la variabilité qui a l'avantage de nécessiter uniquement le WALS.

L'utilisation uniquement de la LPP ou de l'ensemble complet de toutes nos langues semble tout de même être très arbitraire. Nous pourrions envisager l'utilisation de x langues, où x pourrait être compris entre 0 et $N-1$. L'entraînement sur un ensemble réduit de langues permettrait-il d'égaler les performances d'un entraînement sur l'ensemble complet? Nous allons tenter de répondre à cette question dans la section 4.4.5.

4.4.5. Familles de langues

Nous avons observé l'importance de la présence d'une langue proche lors de l'entraînement, qui semble être un critère déterminant pour obtenir de bons résultats pour des expériences de zero-shot. On peut alors se poser la question de la nécessité d'un apprentissage sur toutes les langues. Un apprentissage sur un sous-ensemble de langues d'une même famille pourrait-il obtenir de meilleurs résultats ? Les résultats de J.-K. KIM et al. 2017 ont montré que leur modèle bénéficiait d'un apprentissage par famille de langue.

Nous allons nous servir de familles de langues pour entraîner des modèles *Multi* et *ZS*. La restriction à une famille de langues plutôt que l'ensemble de toutes les langues semble pouvoir, en théorie, limiter le bruit et donner de meilleurs résultats en limitant l'hétérogénéité des langues rencontrées.

Création L'idéal aurait été de se baser sur la LPP pour créer des familles de langues. Pour une langue L, il est possible d'ordonner les langues en fonction de leur proximité avec L grâce à leur score lors de l'étiquetage de la langue L par les autres langues. Ainsi, on pourrait ajouter le rang au LPP. LPP(L,n) serait la $n^{ème}$ langue la plus proche de L. LPP(L,1) correspond donc à la langue la plus proche.

On pourrait alors créer 37 familles, dont les tailles augmenteraient petit à petit. Un premier entraînement aurait lieu en utilisant LPP(L,1), puis l'entraînement suivant sur LPP(L,1) et LPP(L,2), etc... On chercherait alors quelle est la taille de famille donnant les résultats optimaux pour L. L'intuition de cette méthode est donnée dans la figure 4.9.

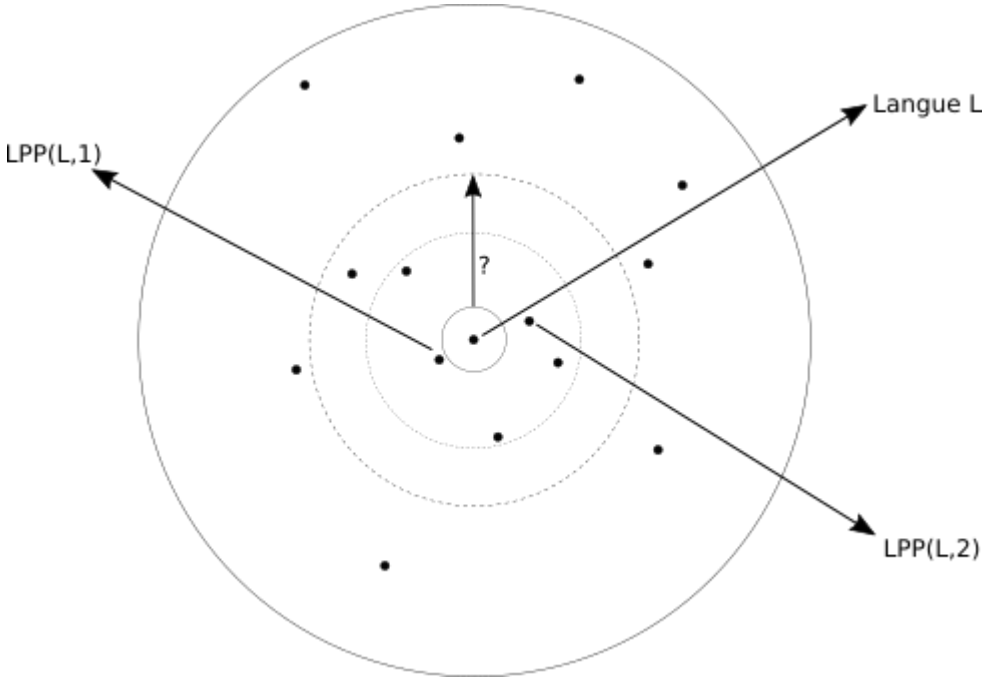


FIGURE 4.9. – Choix des langues d'entraînement pour une langue L. On augmente petit à petit le nombre de langues à inclure dans la famille jusqu'à trouver la taille optimale.

Cependant, cette méthode est extrêmement coûteuse et n'est pas raisonnablement testable, puisqu'elle nécessite l'entraînement de 37 familles pour les 38 langues ayant un corpus d'entraînement.

Une alternative consiste à se baser sur le WALS pour identifier des familles de langues ayant des caractéristiques communes. L'avantage de se baser sur un vecteur du WALS est que cette mesure ne dépend pas de l'existence de corpus annotés.

Deux types de familles ont été définis : les familles dites *naturelles*, issues de la phylogénie des langues, et des familles *artificielles*.

Trois familles naturelles existent dans notre ensemble de langues : les langues Romanes, Germaniques et Slaves.

Les familles artificielles ont été créées en utilisant le WALS en regroupant des ensembles de 6 langues [10].

Il existe donc $\binom{6}{37}$ combinaisons de 6 langues possibles, soit 2 324 784 sous-ensembles, dont nous allons calculer le MID [11] de façon exhaustive.

10. des familles de 6 langues ont été créées, pour rester proches du nombre de langues dans les familles naturelles (6 langues pour les langues romanes, 6 pour les germaniques et 7 pour les slaves).

11. défini chapitre 2 section 2.3 : c'est la distance interne moyenne d'un ensemble de langues $L = \{l_1, \ldots, l_n\}$, la moyenne des distances de chaque paire dans un ensemble de langues L :

$$MID(L) = \frac{1}{n^2 - n} \sum_{\substack{(l_i, l_j) \in L \times L \\ i \neq j}} d(l_i, l_j)$$

Nous conserverons l'ensemble ayant le plus petit MID, le plus grand, et l'ensemble le plus proche de l'ensemble moyen.

Les 6 familles de langues suivantes sont définies :

— Les langues Romanes : Français (fr), Catalan (ca), Espagnol (es), Italien (it), Portugais (pt), Roumain (ro) **(MID = 4.13)** qu'on appellera aussi **Romance**

— Les langues Slaves : Bulgare (bg), Tchèque (cs), Croate (hr), Polonais (pl), Russe (ru), Ukrainien (uk), Slovène (sl) **(MID = 4.19)** qu'on appellera aussi **Slavic**

— Les langues Germaniques : Allemand (de), Danois (da), Anglais (en), Néerlandais (nl), Norvégien (no), Suédois (sv) **(MID = 4.47)** qu'on appellera aussi **German**

— Les langues proches (artificielle) : Bulgare (bg), Anglais (en), Croate (hr), Russe (ru), Slovène (sl), Ukrainien (uk) **(MID = 2.93)** qu'on appellera aussi **Close**

— Les langues moyennement éloignées (artificielle) : Arabe (ar), Bulgare (bg), Catalan (ca), Tchèque (cs), Grec (el), Hindi (hi) **(MID = 9.73**, le MID moyen étant 9.76) qu'on appellera aussi **Mean**

— Les langues distantes (artificielle) : Catalan (ca), Tchèque (cs), Allemand (de), Basque (eu), Irlandais (ga), Japonais (ja) **(MID = 14.13)** qu'on appellera aussi **Distant**

W_{22}, bien qu'étant une description extrêmement partielle de la langue, permet tout de même d'estimer assez correctement si des langues d'un ensemble sont proches d'un point de vue phylogénique, l'ensemble de langues proches défini à partir de W_{22} étant un sous-ensemble des langues slaves, à l'exception de l'anglais (en). Ce résultat tend à montrer la pertinence du MID pour estimer la proximité de langues entre elles. [12]

À présent, nous allons comparer les apprentissages par familles de langues et les apprentissages sur l'ensemble de toutes les langues. Nous comparerons ces résultats dans un contexte multilingue avec un apprentissage sur 6 langues, puis en zero-shot avec un apprentissage sur 5 langues.

Multilingue Une augmentation quasi systématique des résultats par rapport au *Multi +c* classique a lieu lorsqu'on apprend le système sur une famille de langues (voir tableau 4.10), et ce quelque soit la famille. En particulier, dans le cas de l'ensemble des langues distantes, qui est censé être le pire des cas, les résultats augmentent en moyenne de 1.02 point. C'est au final la famille bénéficiant le plus du passage à l'apprentissage par famille. Ce résultat étonnant laisse penser que c'est surtout la réduction du nombre de langues dans les données d'entraînement qui améliore les résultats. En effet, $\frac{1}{6}^{ème}$ du corpus d'entraînement est la langue cible lors de l'entraînement par famille, contre $\frac{1}{37}^{ème}$ pour l'entraînement classique de *Multi +c*.

12. Rappelons cependant que, les MID de l'ensemble des langues Slaves et l'ensemble des langues proches ne sont pas tout à fait comparables puisque le nombre de langues qui les compose est différent (6 pour *Close* et 7 pour *Slavic*).

Lang.	Famille	*Multi +c* (fam)	*Multi +c*	Δ	Lang.	Famille	*Multi +c* (fam)	*Multi +c*	Δ
da	German	92.38	91.34	1.04	ar	Mean	92.80	92.03	0.77
de	German	89.02	88.94	0.08	bg	Mean	95.98	95.09	0.89
en	German	89.06	88.53	0.53	ca	Mean	95.73	95.22	0.51
nl	German	87.45	86.64	0.81	cs	Mean	94.84	94.02	0.82
nno	German	84.81	90.00	-5.19	el	Mean	95.54	95.28	0.26
nob	German	91.19	91.41	-0.22	hi	Mean	93.36	92.75	0.61
sv	German	91.95	91.44	0.51	moy.	Mean	94.71	94.07	0.64
moy.	German	89.41	89.76	-0.35					
					bg	Close	95.59	95.09	0.50
bg	Slavic	95.62	95.09	0.53	en	Close	89.46	88.53	0.93
cs	Slavic	94.93	94.02	0.91	hr	Close	93.87	93.38	0.49
hr	Slavic	94.12	93.38	0.74	ru	Close	94.81	94.23	0.58
pl	Slavic	94.53	93.46	1.07	sl	Close	91.62	89.48	2.14
ru	Slavic	94.84	94.23	0.61	uk	Close	92.07	91.44	0.63
sl	Slavic	91.79	89.48	2.31	moy.	Close	92.90	92.03	0.87
uk	Slavic	92.21	91.44	0.77					
moy.	Slavic	94.01	93.01	1.00	ca	Distant	95.55	95.22	0.33
					cs	Distant	94.59	94.02	0.57
ca	Romance	95.83	95.22	0.61	de	Distant	89.25	88.94	0.31
es	Romance	94.06	94.00	0.06	eu	Distant	90.12	88.16	1.96
fr	Romance	93.48	92.97	0.51	ga	Distant	88.70	87.02	1.68
it	Romance	94.29	93.94	0.35	ja	Distant	93.09	91.78	1.31
pt	Romance	93.60	93.00	0.60	moy.	Distant	91.88	90.86	1.02
ro	Romance	93.52	92.23	1.29					
moy.	Romance	94.13	93.56	0.57					

TABLEAU 4.10. – Détails des résultats des entraînements par famille (fam), comparés à ceux sur le *Multi +c* classique, entraîné sur 38 langues.

Lang.	Famille	*ZS +c* (fam)	*ZS +c*	Δ	Lang.	Famille	*ZS +c* (fam)	*ZS +c*	Δ
da	German	84.01	83.93	0.08	ar	Mean	38.16	60.23	-22.07
de	German	49.05	53.84	-4.79	bg	Mean	55.92	78.16	-22.24
en	German	29.36	27.96	1.40	ca	Mean	49.46	83.48	-34.02
nl	German	63.21	61.04	2.17	cs	Mean	58.59	81.11	-22.52
nno	German	85.06	84.77	0.29	el	Mean	44.41	47.61	-3.20
nob	German	89.14	88.65	0.49	hi	Mean	37.27	35.67	1.60
sv	German	78.87	79.70	-0.83	moy.	Mean	47.30	64.38	-17.08
moy.	German	68.39	68.56	-0.17					
					bg	Close	77.10	78.16	-1.06
bg	Slavic	77.59	78.16	-0.57	en	Close	19.00	27.96	-8.96
cs	Slavic	82.96	81.11	1.85	hr	Close	80.19	83.07	-2.88
hr	Slavic	82.18	83.07	-0.89	ru	Close	83.98	85.17	-1.19
pl	Slavic	75.74	78.82	-3.08	sl	Close	67.95	69.10	-1.15
ru	Slavic	84.52	85.17	-0.65	uk	Close	79.41	80.07	-0.66
sl	Slavic	72.61	69.10	3.51	moy.	Close	67.94	70.59	-2.65
uk	Slavic	79.56	80.07	-0.51					
moy.	Slavic	79.31	79.36	-0.05	ca	Distant	51.43	83.48	-32.05
					cs	Distant	56.61	81.11	-24.50
ca	Romance	83.69	83.48	0.21	de	Distant	34.97	53.84	-18.87
es	Romance	89.16	88.69	0.47	eu	Distant	51.42	55.73	-4.31
fr	Romance	73.50	78.29	-4.79	ga	Distant	39.68	45.28	-5.60
it	Romance	74.12	77.09	-2.97	ja	Distant	18.87	33.28	-14.41
pt	Romance	76.58	76.11	0.47	moy.	Distant	42.16	58.79	-16.63
ro	Romance	62.86	64.28	-1.42					
moy.	Romance	76.65	77.99	-1.34					

TABLEAU 4.11. – Détails des résultats des entraînements par famille, comparés à ceux sur les *ZS +c* classiques, chacun entraîné sur 38-1 langues.

Ce résultat un peu décevant nous pousse à nous intéresser plus aux résultats en condition de zero-shot, pour lesquels l'entraînement par famille de langues ne sera pas biaisé de la sorte.

Zero-shot La diminution du nombre de langues lors de l'entraînement zero-shot devrait réduire l'hétérogénéité de phénomènes que le système doit apprendre. L'hypothèse est qu'en se concentrant uniquement sur des langues proches de la langue cible, l'exactitude des résultats devrait augmenter.

Contrairement aux résultats des expériences *Multi +c*, dans le contexte *ZS +c*, les résultats des entraînements sur les familles de langues sont souvent moins bons que ceux des entraînements classiques (voir tableau 4.11), en particulier pour *Distant* et *Mean*. Cela peut s'expliquer par l'absence de langue proche à l'entraînement, ce qui va dans le sens de notre hypothèse.

Cependant, on ne constate pas d'amélioration des résultats pour les familles qui devraient être proches (les familles naturelles et *Close*). Mais même si le score diminue en moyenne, cette baisse n'est pas systématique langue par langue. Comme souvent, les résultats fluctuent beaucoup entre les langues, ne permettant pas de tirer de conclusions fermes. On peut tout de même remarquer que pour la famille des langues

slaves, même si les résultats n'augmentent pas en moyenne, ils ne diminuent pas non plus. Il est donc possible d'obtenir d'aussi bons résultats avec seulement $\frac{7}{37}^{ème}$ des données.

Il est intéressant d'observer le comportement d'une même langue se trouvant dans plusieurs familles différentes, afin de voir l'impact que peut avoir la proximité de la famille pour les prédictions dans un contexte de zero-shot. La présence de la LPP ou non devrait impacter les résultats de façon conséquente.

Le bulgare (bg) apparaît dans plusieurs familles de langues (*Slavic*, *Mean* et *Close*). La LPP du bulgare (bg) est le russe (ru), qui est présent dans la famille *Slavic*, dans *Close*, et bien sur dans le modèle $ZS + c$ classique entraîné sur toutes les langues. Il n'est cependant pas présent dans *Mean*.

Dans les configurations où le russe (ru) apparaît dans l'ensemble d'entraînement, les scores pour le bulgare (bg) sont élevés (avec un score supérieur à 77). Cependant, pour l'ensemble *Mean*, où le russe (ru) n'est pas présent [13], on observe alors une chute du score pour le bulgare (bg) dans cet ensemble : 55.92 d'exactitude, soit une baisse de 22.24 points comparé au modèle $ZS + c$ classique. Cette baisse est également observée pour les autres langues apparaissant dans plusieurs ensembles de langues comme le catalan (ca) ou le tchèque (cs). Ces résultats vont dans le sens des conclusions de la section 4.4.4, la LPP reste l'élément déterminant dans la prédiction des résultats.

Il semble que l'entraînement classique sur toutes les langues disponibles reste supérieur à l'entraînement par famille de langue. Ce dernier reste cependant meilleur que de simplement prendre la LPP.

Nous n'avons pas pu trouver une façon raisonnable de choisir les N meilleurs langues pour entraîner un modèle de zero-shot sur une langue L. Il semble que la meilleure solution soit d'entraîner le modèle sur toutes les langues à disposition, afin que le système puisse voir le plus de langues proches potentielles de L à l'apprentissage.

Conclusion Il semble que l'apprentissage par famille de langues soit le reflet assez proche d'un apprentissage classique sur toutes les langues, l'élément déterminant lors des apprentissages étant la présence ou non d'un modèle proche de la langue de test. Se réduire à un plus petit ensemble de langues peut donc être utile pour réduire le temps d'apprentissage des modèles tout en obtenant de meilleurs résultats que de simplement prendre la LPP de la langue cible. Le gain de temps, la nécessité de moins de données annotées pour des résultats relativement bons permet de se questionner sur l'utilité de créer des modèles sur l'ensemble complet de toutes les langues.

4.5. Conclusions

Dans ce chapitre, nous avons testé l'apprentissage de modèles multilingues et de modèles zero-shot pour de nombreuses langues sur une tâche d'étiquetage de

13. la LPP du bulgare (bg) est le grec (el) dans la famille *Mean*, le modèle *Mono* $+c$ du grec donne un score de 51.63 pour le décodage du bulgare (bg)

POS. Nous avons pu conclure que, comme pour l'analyse syntaxique délexicalisée, une représentation de la langue basée sur des traits typologiques issus du WALS permet d'obtenir de meilleurs résultats dans un cadre multilingue. Cependant, dans un contexte de zero-shot, son utilité n'est pas systématique pour toutes les langues, mais joue un rôle déterminant pour rapprocher ou éloigner des langues d'une autre trop éloignée ou au contraire trop proche.

Nous avons surtout pu constater l'importance de la présence d'une langue proche dans le corpus d'entraînement dans un contexte de zero-shot, ainsi que la nécessité de retirer le modèle de caractères lorsque plusieurs langues avec le même alphabet ne sont pas proches.

L'entraînement sur un sous-ensemble de langues a également été exploré, et a permis de montrer que les résultats sur des familles de langues proches (au sens du WALS) permettait d'obtenir de bons résultats tout en réduisant les temps d'apprentissage, même si l'apprentissage sur l'ensemble complet de nos langues obtient de meilleurs résultats.

Les tâches de prédiction de l'analyse syntaxique et celle de prédiction des POS ont été réalisées depuis le début de cette thèse. Les apports du WALS, d'une langue proche ou encore d'un modèle de caractères ont été étudiés afin d'identifier quelles sont les ressources pouvant favoriser le partage de connaissances entre langues. Les POS et les traits morphologiques de références étaient fournis en entrée de l'analyseur. Nous avons vu dans ce chapitre qu'il est possible d'obtenir des résultats exploitables lors de la prédiction des POS, et la question de la prédiction à la fois des POS et de l'analyse syntaxique se pose et sera le sujet du chapitre suivant.

Chapitre 5.

Système de prédictions complètes

Sommaire

5.1 Introduction 123
5.2 État de l'art 124
5.3 Tagparser 126
5.4 Résultats et analyses 131
5.4.1 Morphologie 132
5.4.2 POS et traits morphologiques de références VS prédits 135
5.4.3 WALS 137
5.5 Conclusions et perspectives 140

5.1. Introduction

La tâche de prédiction de l'analyse syntaxique de la phrase a été explorée dans le chapitre 2 et la tâche d'étiquetage des POS a été étudiée dans le chapitre 4. Nous cherchons à présent à étudier l'impact sur les prédictions de la chaîne de traitement : étiquetage des POS, des traits morphologiques et analyse syntaxique de la phrase. Nous nommerons cette chaîne de traitement une tâche de *tagparsing*, et l'architecture de ce système un *tagparser*. La morphologie est également prédite, car les entrées utilisées dans le chapitre 2 étaient les POS et les traits morphologiques. Ces derniers ont donc également été prédits par soucis de cohérence avec le chapitre 2.

Des travaux présentés dans cette thèse, la tâche de tagparsing est celle qui correspond le plus aux besoins réels du TAL, puisqu'elle nécessite très peu de données annotées de référence pour ses prédictions (tokenisation et segmentation en phrases uniquement), et va jusqu'à la prédiction de l'analyse syntaxique. La prédiction de la tokenisation et de la segmentation en phrases n'a pas été réalisée car cette tâche est assez éloignée d'une tâche d'étiquetage ou d'analyse syntaxique. De plus, la question de l'apport du WALS est une question centrale de nos travaux, mais l'utilisation de traits typologiques issus de cette ressource ne semble pas très pertinente pour une tâche de tokenisation, puisque les traits du WALS que nous avons choisi sont très orientés autour de la syntaxe. L'ajout d'une prédiction supplémentaire risque d'introduire plus de bruit dans les prédictions, rendant l'analyse des résultats plus compliquée encore à interpréter. Les expériences de tagparsing nous permettront de comparer l'impact

des ressources et modèles utilisés à travers les différentes tâches de prédiction, et de vérifier si les conclusions tirées dans les chapitres précédents restent vraies dans un contexte de chaîne de prédiction. Le tagparser sera présenté dans la section 5.3.

Dans ce chapitre, nous aborderons 3 questions de recherche principales :

1. les résultats de la prédiction de la morphologie sont-ils cohérents avec les résultats de la prédiction des POS et de l'analyse syntaxique des chapitres précédents ? (voir section 5.4.1). La morphologie n'ayant pas encore été prédite dans nos travaux, nous analyserons l'impact du WALS et du modèle de caractères dans les différentes configurations (monolingue, multilingue et zero-shot). Le modèle de caractères permettant de capturer la plupart des marqueurs morphologiques des mots, son utilisation devrait être très bénéfique pour la tâche de prédiction des traits morphologiques. Les traits du WALS étant très orientés autour de la syntaxe, il est possible que son utilisation soit moins efficace pour aider au partage de connaissances entre langues que sur la tâche d'analyse syntaxique ou d'étiquetage des POS.
2. quel impact a la prédiction des POS et des traits morphologiques, comparée à l'utilisation des POS et des traits morphologiques de référence ? (voir section 5.4.2). L'utilisation des étiquettes prédites à la place de celles de références devrait mener à une chute des résultats de l'analyse syntaxique. Nous tenterons de quantifier cette baisse, et de mesurer qui des prédictions des POS ou des traits morphologiques a le plus d'impact sur les prédictions de la syntaxe.
3. l'utilisation du WALS permet-il un rapprochement ou éloignement des langues entre elles en condition multilingue et de zero-shot ? Dans la continuité des chapitres précédents, nous observerons dans la section 5.4.3 quelle est l'utilité d'un vecteur de traits typologiques issus du WALS sur ces expériences, dans un cadre multilingue et un cadre d'apprentissage zero-shot. Nous constaterons si WALS permet de faciliter le partage d'informations entre langues dans une tâche de tagparsing.

5.2. État de l'art

Les prédictions réalisées à partir de textes bruts sont des questions qui ont été largement abordées grâce à des campagnes d'évaluation qui ont incité de nombreux chercheurs à s'intéresser à ces problématiques. BUCHHOLZ et MARSI 2006 ; NIVRE, HALL et al. 2007 ; ZEMAN, HAJIČ et al. 2018 ; ZEMAN, POPEL et al. 2017 présentent les résultats de 4 campagnes d'évaluation mise en place par CoNLL (*the Conference on Computational Natural Language Learning*) qui ont largement permis de faire avancer la recherche sur l'analyse syntaxique en dépendance, incitant leurs participants à s'intéresser aux problématiques du multilinguisme.

La première campagne d'évaluation, (BUCHHOLZ et MARSI 2006), était une tâche de prédiction de l'arbre syntaxique de la phrase pour 13 langues, en utilisant les mots en entrées, ainsi que les POS, les traits morphologiques,... De grandes variations entre les

langues avaient été constatées : de 67.5% pour le meilleur système pour le turc (tr) à 91.7% pour le japonais (ja). Les auteurs expliquaient que le corpus du turc (tr) était plus petit que celui d'autres langues (bien que comparable au slovène (sl) qui obtenait 73.4% de LAS) tout en étant plus riche d'un point de vue morphologique. Déterminer ce qui rend un corpus plus simple qu'un autre à analyser était la principale perspective de ces travaux.

La deuxième campagne (NIVRE, HALL et al. 2007) ajoutait à la première la problématique de l'adaptation de domaine. Autrement dit, une tâche supplémentaire était demandée, consistant à entraîner un modèle sur des données issues du Wall Street Journal, et à décoder du texte provenant de résumés biomédicaux, chimiques, et de dialogues parent-enfant. Cette campagne n'était cependant disponible que pour de l'anglais (en). L'apprentissage zero-shot est une forme de problème d'adaptation de domaine. De nombreuses techniques ont été mises en place pour cette tâche, certaines se rapprochant de nos travaux, notamment une approche basée sur l'utilisation uniquement de traits devant bien se transférer d'un domaine à l'autre (DREDZE et al. 2007). Les auteurs de ces travaux ont tenté de retirer les traits devant mal se transférer en théorie (comme certains traits liés aux dépendances reliant des POS `nom-nom`) menant à une légère augmentation de leurs résultats sur la prédiction de l'analyse syntaxique. L'ajout de traits devant bien se transférer (des traits liés à la position lexicale des mots étiquetés `nom` par exemple) ne mène cependant pas à une augmentation des résultats.

10 ans plus tard, CoNLL met de nouveau en place une campagne d'évaluations sur le sujet de l'analyse syntaxique en dépendances multilingue (ZEMAN, POPEL et al. 2017), essayant cette fois-ci de se mettre dans des conditions plus "réalistes". Le but était de partir de texte brut, et de réaliser l'analyse syntaxique de la phrase sans avoir la tokenisation, les POS ni les traits morphologiques de référence. Il fallait donc s'en passer, ou bien les prédire. L'utilisation de systèmes multilingues a pu être grandement encouragée grâce à l'utilisation des UD, permettant l'utilisation d'un schéma d'annotation commun aux différentes langues. 45 langues disposaient de corpus d'entraînement, et 4 langues surprises sans données d'entraînement ou très peu (une vingtaine de phrases) ont été utilisées. Pour les 4 langues surprises, qui n'ont été connues qu'une semaine avant la phase d'évaluation, et pour certaines langues avec très peu de données d'entraînement comme l'ouïghour (ug) ou le kazakh (kk), les participants ont généralement utilisé des approches multilingues avec une ou plusieurs langues proches, le plus souvent de manière délexicalisée. Les données que nous utilisons dans nos travaux sont issues de cette campagne d'évaluation.

Le succès de la campagne d'évaluation de 2017 a incité les organisateurs à réitérer l'opération l'année suivante. Ainsi, la campagne d'évaluations CoNLL de 2018 (ZEMAN, HAJIČ et al. 2018) s'intéresse aux mêmes problématiques qu'en 2017, mettant cette fois-ci l'accent sur la question de l'évaluation et des métriques utilisées, en proposant deux nouvelles :

— le MLAS (*Morphology-Aware Labeled Attachment Score*), une extension du CLAS (NIVRE, MARNEFFE et al. 2016), permettant dévaluer à la fois les POS, la morphologie, et le LAS des mots pleins, aussi appelés *content words*, qui correspondent

aux mots porteurs de sens dans la phrase (ce n'est par exemple pas le cas des déterminants ou des prépositions).

— le BLEX (*Bilexical Dependency Score*), similaire au MLAS, mais évaluant la lemmatisation plutôt que la morphologie.

Ces deux nouvelles mesures ont été créées afin d'avoir une meilleure comparabilité des résultats entre les langues, pour que les langues ayant beaucoup de mots vides (l'inverse des mots pleins) ne soient pas biaisées positivement, les prédictions pour ces mots étant plus simples à réaliser. Nos premiers résultats ont été évalués à l'aide du script d'évaluation de la campagne d'évaluation de 2017, lorsque ces mesures n'étaient pas encore standards. Afin d'avoir des résultats comparables tout au long de la thèse, nous avons continué une évaluation classique basée sur le l'UAS, le LAS et l'exactitude.

Ces deux dernières campagnes d'évaluation ont permis d'établir une base solide sur les connaissances que nous avons sur la chaîne de prédiction complète à partir du texte brut. Dans ce chapitre, nous nous rapprochons de ce type de tâche, puisque nous allons faire la chaîne de prédiction complète à partir de texte segmenté en mots, sur des tâches multilingues et d'apprentissage zero-shot. Nous nous intéressons en particulier à l'apport que le WALS et un modèle de caractères peuvent avoir sur ces tâches dans des contextes multilingues et de zero-shot.

5.3. Tagparser

Architecture La prédiction complète des POS, des traits morphologiques et de l'analyse syntaxique a été réalisée de façon jointe. Alors que classiquement, tous les traits morphologiques sont prédits, puis donnés en entrée d'un étiqueteur en POS, dont les résultats sont utilisés en entrées d'un analyseur, nous avons fait le choix de réaliser toutes les prédictions pour un mot, puis de traiter le mot suivant. Pour réaliser ces prédictions, nous avons utilisé ce que nous appelons un *tagparser*, que nous allons présenter dans cette section.

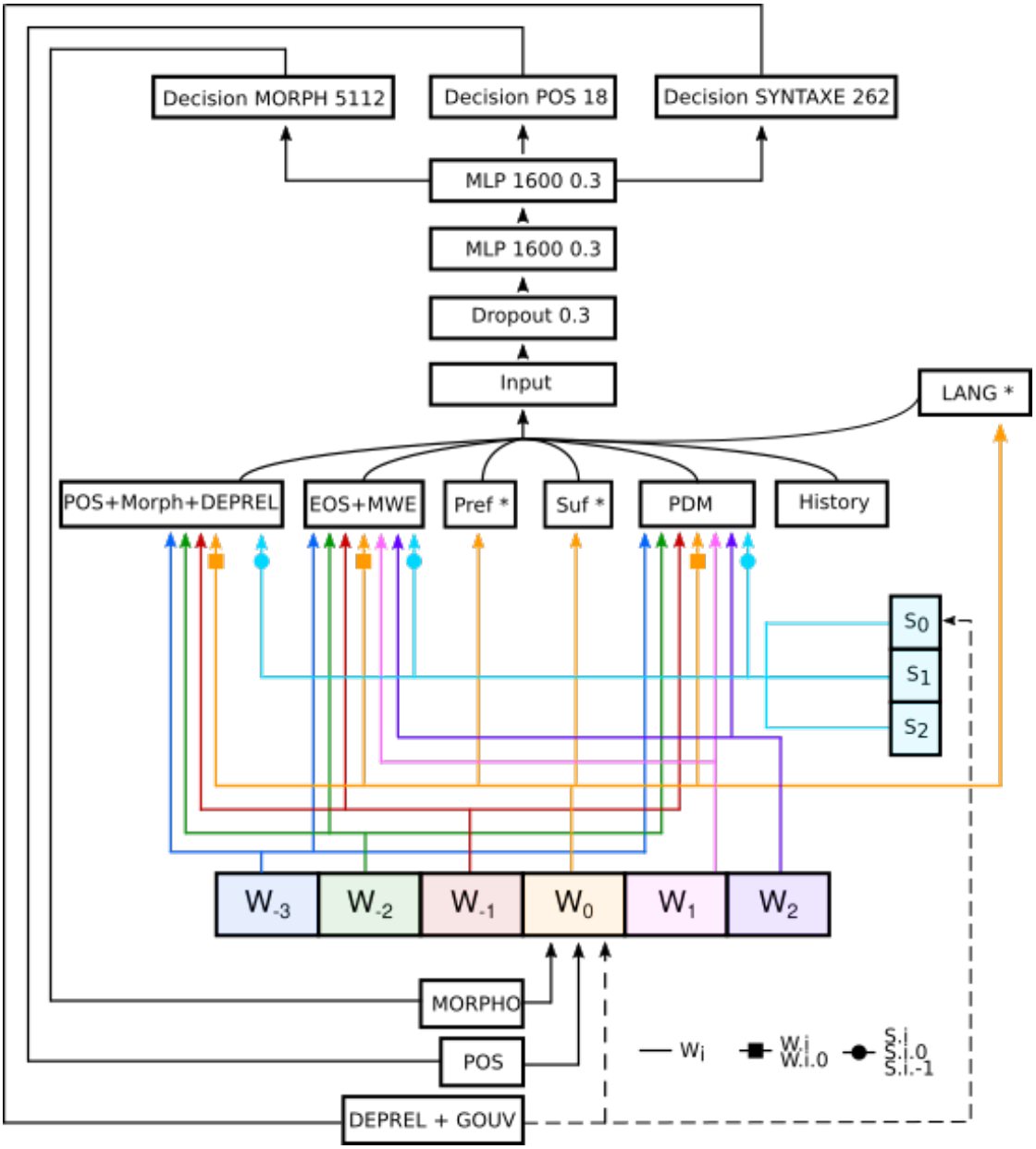


FIGURE 5.1. – Architecture du tagparser. Le texte est lu en entrée du buffer. Le mot courant est w_0, le mot précédent est w_{-1} et le mot suivant est w_1. Une pile est également à disposition, s_0 étant le sommet de pile. Les 6 bi-LSTM sont représentés par les rectangles POS+Morph, EOS+MWE, ... et prennent en entrée le trait correspondant à leur nom. Les bi-LSTM avec une '*' sont optionnels et sont utilisés selon la configuration (*Multi +c*, *Multi −c* W_{22}, ...). La prédiction du mot courant est donnée en entrée pour le mot suivant. w_i correspond aux plongements de POS+Morph, ... et $w_{i.0}$ est le plongement de POS+Morph du fils gauche du nœud dans l'arbre. Idem avec s_i, $s_{i.-1}$ étant le fils droit. Les flèches en pointillée signifient qu'à chaque prédiction, le mouvement prédit est donné en entrée soit au mot courant du buffer, soit au sommet de la pile.

Les entrées du tagparser correspondent à une fenêtre de taille 6, centrée sur le mot courant. Des informations sont obtenues à partir de ces mots, telles que leur représentation vectorielle, leurs préfixes et suffixes, les POS des mots précédents, etc. Ces traits sont fournis en entrée de bi-LSTMs qui permettront de contextualiser ces informations, pour ensuite les donner en entrée d'un perceptron multicouches (MLP). La couche d'entrée de ce MLP consiste alors en la concaténation de la sortie de 4 à 6 bi-LSTM, décrits ci-dessous :

— PDM : un bi-LSTM prenant en entrée les plongements de mots multilingues de taille 300 définis au chapitre 3 et les contextualisant à partir d'une fenêtre de taille 21 centrée sur le mot courant. Ces plongements de mots contextuels obtenus sont de taille 128. Sont alors passés en entrée de ce bi-LSTM :

  — les 8 plongements contextuels de taille 128 correspondant au mot courant w_0, aux trois mots précédents w_1, w_{-2}, w_3 et aux deux mots suivants w_1, w_2, w_3 du buffer, ainsi que le plongement de mot du fils gauche du mot courant w_0

  — les 3 plongements des mots en haut de la pile, ainsi que leurs fils gauches et droits

  La sortie de PDM est de taille 128 * (8 + 3 * 3) = 2176

— EOS+MWE : idem précédemment, un bi-LSTM prenant en entrée l'information de la présence d'une fin de phrase et la présence d'un token multi-mots parmi le mot courant w_0, les trois mots précédents w_1, w_{-2}, w_3 et les deux mots suivants w_1, w_2, w_3 du buffer, ainsi que le plongement de mot du fils gauche de w_0. La sortie de ce bi-LSTM est de taille 64.

— POS+Morph+DEPREL : un bi-LSTM prenant en entrée la POS, les traits morphologiques et l'information du type de relation de dépendance entre le mot et sa tête (deprel) pour les trois mots précédents. La sortie de ce bi-LSTM est de taille 64. Lors des 4 premières itérations, les étiquettes de référence des mots précédents sont donnés. À la fin de la quatrième itération, le système fait un décodage de tout le corpus d'entraînement. À partir de la cinquième itération, le système utilise ces étiquettes prédites à la place de celles de références.

— History : idem précédemment, un bi-LSTM prenant en entrée la concaténation des plongements des 10 précédentes actions prédites et dont la sortie est de taille 32

— Pref : idem précédemment, un bi-LSTM optionnel prenant le préfixe du mot courant (3 caractères) dont la sortie est un plongement de taille 64.

— Suf : idem précédemment, un bi-LSTM optionnel prenant le suffixe du mot courant (3 caractères) dont la sortie est un plongement de taille 64.

— LANG : ce n'est pas un bi-LSTM, mais un simple vecteur optionnel concaténé directement à la couche d'entrée, existant uniquement dans les expériences utilisant W_{22}. Ce vecteur est de taille 80, correspondant aux 80 valeurs du vecteur W_{22} où les 22 traits sont représentés sous forme de one-hot puis concaténés.

Alternativement, LANG peut servir de représentation pour l'identifiant de la langue et est alors un vecteur de la taille du nombre de langues dans le corpus d'entraînement

On obtient ainsi une couche d'entrée de taille 2176 + 64 + 64 + 32 (+64) (+64) (+80) = 2 327 à 2 535 selon les configurations. Les deux couches cachées sont de taille 1600, avec un *dropout* durant l'entraînement de 0.3. La couche de sortie de ce réseau de neurones donne une distribution de probabilités, une pour chaque action possible pour le mot courant. Le système étant glouton, il choisira l'action avec la plus haute probabilité. Le nombre d'itérations est de 40, la fonction d'activation est la fonction ReLu, la fonction objectif est un softmax de vraisemblance négative, et l'algorithme d'apprentissage est Adagrad. Toutes les expériences ont été réalisées deux fois, sauf les expériences en condition zero-shot, en raison de problème de temps d'entraînement. Les résultats des expériences en conditions monolingues et multilingues correspondent à la moyenne des deux modèles. Toutes les entrées (POS, traits morphologiques, ...) sont représentées par des plongements, et sont initialisées aléatoirement, à l'exception des formes de surface qui utilisent les plongements de mots pré-entraînés et alignés. Un schéma de l'architecture de ce système est disponible dans la figure 5.1.

Configurations d'entraînement Nous reprenons la base des définitions données dans les chapitres précédents. Nos différentes configurations correspondent à une paire ⟨données d'entraînement, architecture⟩. Les données d'entraînement possibles sont *Mono*, *Multi* et *ZS*.

Mono correspond à une unique langue au moment de l'entraînement, et le tagparsing de cette même langue pour le test.

Multi représente un entraînement basé sur l'ensemble de toutes les langues disponibles, et le tagparsing de chaque langue par ce modèle multilingue unique.

ZS est identique à *Multi*, auquel on retire une langue L. L'évaluation de *ZS* se fait sur la langue L uniquement.

Les différentes architectures correspondent à la façon de paramétrer notre tagparser. Le modèle peut utiliser les caractères des mots ($+c$) ou non ($-c$) et utiliser une représentation de la langue (ID ou W_{22}) ou non. Nos différentes configurations d'entraînement sont donc :

Mono : Corpus monolingue.
Le corpus d'entraînement est L, et le corpus de test et également celui de L. 38 tagparsers sont entraînés, un pour chaque langue. C'est le cas classique d'apprentissage, où l'entraînement se fait sur une langue L, et le décodage se fait sur cette même langue L.

Multi : Corpus multilingue.
Un tagparser est entraîné sur nos 38 langues, sans indication de la langue. Un seul modèle est entraîné, et chaque corpus de test est décodé avec ce modèle. Ce modèle est très bruité, puisqu'il est possible de retrouver des contradictions entre les langues (*a*, en français (fr) est un verbe conjugué (du verbe *avoir*), alors qu'en anglais (en), *a*

est un déterminant).

Multi W_{22} : Corpus + WALS.
Idem à *Multi* en utilisant le vecteur LANG qui permet d'associer à chaque mot un vecteur $W(L)$ issu du WALS, qui correspond à la langue du mot. Contrairement à l'utilisation d'un identifiant de la langue, le WALS est porteur d'informations sur la syntaxe, la morphologie des mots, etc... pouvant aider les prédictions.

Multi ID : Corpus multilingue + identifiant de la langue.
Idem à *Multi* en utilisant en plus le vecteur LANG qui permet de représenter l'identifiant de la langue de chaque mot. Cette expérience permet d'expliciter l'information de la langue en cours de traitement pour limiter le bruit apporté par l'apprentissage multilingue, mais sans expliciter les points communs à plusieurs langues.

ZS **:** *Zero-shot.*
38 tagparsers sont entraînés, sans représentation de la langue. Le tagparser de la langue L sera entraîné sur toutes les langues, sauf L, soit un ensemble de 37 langues. Cette configuration représente la situation où un tagparser est entraîné pour une langue pour laquelle aucune donnée annotée pour l'entraînement n'est disponible.

ZS W_{22} **:** *Zero-shot* + WALS
Ajoute à *ZS* l'indication de la langue de chaque mot à travers son vecteur du WALS, à l'identique de la différence entre *Multi* W_{22} et *Multi*.

Chacune de ces expériences est déclinée en trois versions, selon qu'elle utilise le modèle de caractères, les plongements de mot, un seul des deux ou les deux. Cela nous permet de tester l'impact de l'utilisation des caractères et celui des plongements de mots dans nos expériences.

— *Mono* $+c$, *Multi* $+c$, *Multi* $+c$ ID, *Multi* $+c$ W_{22}, *ZS* $+c$ et *ZS* $+c$ W_{22} utilisent les bi-LSTM correspondant au modèle de caractères (Pref et Suf) et celui correspondant au plongement de mot (PDM)

— *Mono* $-c$, *Multi* $-c$, *Multi* $-c$ ID, *Multi* $-c$ W_{22}, *ZS* $-c$ et *ZS* $-c$ W_{22} utilisent le bi-LSTM correspondant aux plongements de mots, mais n'utilisent pas ceux correspondants aux caractères des mots (Pref et Suf).

— *Mono_delex*, *Multi_delex*, *Multi_delex* ID, *Multi_delex* W_{22}, *ZS_delex* et *ZS_delex* W_{22} n'utilisent aucun des bi-LSTM liés au lexique (Pref, Suf et PDM).

Métriques Pour l'analyse syntaxique, on s'intéresse ici aux scores obtenus pour le LAS, les résultats entre les UAS et le LAS étant très similaires (plus de 0.95 de corrélation de Pearson entre les résultats de UAS et du LAS pour toutes les expériences). Pour les POS et les traits morphologiques, nous continuons d'utiliser l'exactitude.

Traits morphologiques Les systèmes faisant les prédictions jointes sont les premiers de nos travaux à prédire également les traits morphologiques. Pour rester cohérents avec les traits utilisés dans le chapitre 2, nous n'avons gardé que les 16 traits les plus fréquents (voir section 2.4).

Contrairement aux expériences du chapitre 2, nous avons voulu vérifier si la prédiction des traits morphologiques de manière indépendante, ou au contraire lorsque les traits sont concaténés, changeait quelque chose. Par exemple, si un mot a les traits `Gender=Masc|Number=Sing`, on peut soit prédire le genre, puis le nombre. Ou bien, on peut faire la prédiction comme si `Gender=Masc|Number=Sing` était une étiquette à part entière. Nous avons donc entraîné plusieurs tagparsers, et aucune des deux méthodes ne semblait obtenir des résultats significativement supérieurs à l'autre. Nous avons donc opté pour la méthode permettant de minimiser le temps d'entraînement. C'est la raison pour laquelle les systèmes ont été entraînés avec la méthode concaténant les traits morphologiques.

Dans les UD, il existe pour certaines langues des traits marqués plus d'une fois pour un même mot. Ce sont les *layered features*[1]. Par exemple, il est possible de voir l'étiquette de morphologie suivante : `Gender=Masc|Gender[psor]=Fem`. Cette étiquette indique que le mot est masculin, mais que le mot porte également le marqueur du féminin associé à un possessif. Cette étiquette peut être utile pour des langues comme le tchèque (cz), l'hébreu (he) ou l'arabe (ar). Nous avons fait le choix de supprimer ces traits afin de simplifier les systèmes.

5.4. Résultats et analyses

Dans cette section, nous allons répondre à 3 questions :

1. Les prédictions des traits morphologiques sont-elles cohérentes avec les résultats des autres tâches ?
2. Quel impact a eu la prédiction des POS et des traits morphologiques comparés à l'utilisation des références ?
3. Quel utilité le WALS a-t-il dans une tâche de tagparsing ?

De très nombreux résultats vont être présentés dans ce chapitre. Seulement une partie des résultats sera donnée pour simplifier la lecture, mais les résultats complets sont disponibles en annexe (voir tableaux .18 : résultats de la prédiction des traits morphologiques. .19 : résultats de la prédiction des POS. .20 : prédictions de l'analyse syntaxique, métrique UAS. .21 : prédictions de l'analyse syntaxique, métrique LAS). Une comparaison avec un système de l'état de l'art (UDPipe) est également disponible en annexe dans la section D.1.

1. https://universaldependencies.org/u/overview/feat-layers.html

Modèle	Exactitude
Mono $+c$	81.14
Mono $-c$	72.33
Multi $+c$	78.37
Multi $-c$	69.29
Multi $+c$ ID	77.68
Multi $-c$ ID	68.09
Multi $+c$ W_{22}	75.18
Multi $-c$ W_{22}	67.88
ZS $+c$	42.24
ZS $-c$	42.51
ZS $+c$ W_{22}	43.67
ZS $-c$ W_{22}	44.62

TABLEAU 5.1. – Moyenne sur toutes les langues des prédictions de la morphologie pour les différentes expériences.

5.4.1. Morphologie

Les expériences de ce chapitre correspondent à la seule situation où les traits morphologiques sont prédits, c'est-à-dire dans le cadre d'une tâche jointe avec la prédiction des POS et de l'analyse syntaxique. Les traits morphologiques de référence étaient donnés en entrée de l'analyseur du chapitre 2, et afin de pouvoir continuer de les donner en entrée, il est désormais nécessaire de les prédire. Nous allons explorer dans cette section si l'utilisation du WALS et d'un modèle de caractères a des effets positifs ou négatifs sur la prédiction de la morphologie. Nous allons également vérifier si les traits morphologiques prédits ont un impact sur les prédictions des autres tâches. Les résultats de la prédiction des traits morphologiques sont disponibles en annexe, tableau .18. Un résumé des résultats en moyenne sur toutes les langues est disponible dans le tableau 5.1.

Le premier résultat étonnant est qu'il n'a pas été possible d'établir de corrélation entre les résultats de la prédiction des traits morphologiques et les résultats de la prédiction des POS, et ce peu importe l'expérience. Ce résultat suggère que ces deux tâches sont indépendantes et n'ont que peu d'influence l'une sur l'autre, même dans le cadre d'une tâche jointe. Nous allons donc tenter d'analyser les résultats indépendamment des résultats des POS.

Les résultats, comme bien souvent, varient beaucoup d'une langue à l'autre. La plupart des langues asiatiques (indonésien (id), japonais (ja), coréen (ko), vietnamien (vi) et chinois (zh)) obtiennent un score de presque 100%, ces langues étant morphologiquement très pauvres. La proportion de mots n'ayant aucun trait morphologique dépasse les 85% pour toutes ces langues, alors qu'en moyenne sur toutes les langues, seulement 38.5 % des mots n'ont pas de traits morphologiques (voir tableau .22 en annexe). Même en condition zero-shot, ces langues obtiennent des résultats supérieurs à 90%, exceptée le coréen (ko), qui obtient un score de 64.61 d'exactitude dans le meilleur des cas zero-shot, avec le modèle ZS $-c$ W_{22}.

Les langues slaves, au contraire, ne s'en sortent pas très bien pour la plupart : pour

l'expérience *Mono +c*, 88.53 pour le bulgare (bg), 61.30 pour le tchèque (cs), 72.92 pour le croate (hr), 59.22 pour le polonais, 56.77 pour le russe (ru), 74.35 pour le slovène (sl) et 54.55 pour l'ukrainien (uk), ce dernier étant le score le plus bas. Le score moyen sur toutes les langues est de 81.14, alors que le score moyen des langues slaves et de seulement 66.80. Les langues slaves sont morphologiquement assez riches : aucune ne dépasse les 35% de mots n'ayant pas de traits morphologiques. Les résultats de prédiction de la morphologie semblent assez liés aux pourcentages de mots n'ayant pas de traits morphologiques dans les corpus de test. Nous avons donc mesuré la corrélation de Pearson pour vérifier cette intuition : avec un score de 0.58, ces résultats semblent donc en effet un peu corrélés. Cette corrélation est d'ailleurs plus importante avec l'expérience *Multi +c*, avec un score de 0.62, et encore plus en *ZS +c*, avec un score de 0.67.

WALS Les traits du WALS sélectionnés pour le vecteur W_{22} ont été choisis pour aider la tâche d'analyse syntaxique. Il peut toujours permettre d'éventuellement rapprocher des langues ayant des vecteurs similaires, mais son utilité dans le cadre d'une tâche de prédiction de la morphologie n'est pas garantie.

En configuration *Multi +c*, les résultats en moyenne restent assez proches, bien que légèrement inférieurs (-2.77 points) des résultats de *Mono +c*. L'utilisation du WALS avec l'expérience *Multi +c* W_{22} ne semble cependant pas utile, en raison d'une diminution de 3.19 points en moyenne. Cependant, en observant le détail langue par langue, on observe d'importantes différences entre les langues. Pour commencer, l'utilisation du WALS fait diminuer les résultats de 39.12 et 35.35 points pour les deux norvégiens (respectivement nynorsk (nno) et bokmål (nob)). Ces deux résultats extrêmes peuvent en partie s'expliquer par le caractère identique de leur deux vecteurs issu du WALS. 8 langues sur 38 perdent plus de 10 points avec l'ajout du vecteur du WALS. Trois langues voient leurs résultats augmenter de plus de 10 points : l'arabe (ar), l'hindi (hi) et le turc (tr).

En condition de zero-shot, les meilleurs résultats sont obtenus avec l'expérience *ZS −c* W_{22}, qui n'utilise pas de modèle de caractères mais utilise le WALS. Les résultats de *ZS −c* W_{22} sont de 2.38 points supérieurs à *ZS +c*.

Modèle de caractères La question de l'utilité du modèle de caractères pour prédire les traits morphologiques est assez intéressante. Une hypothèse de départ ayant été que les marqueurs morphologiques étaient portés dans les préfixes et suffixes des mots, expliquant l'importance de l'utilisation du modèle de caractères. L'existence de marqueurs morphologiques communs d'une langue à l'autre a également été évoquée, comme avec le 's' en fin de mot qui peut être un marqueur du pluriel en français ou en anglais (*chat**s***, *cat**s***).

L'utilisation d'un modèle de caractères pour la prédiction des traits morphologiques semble essentielle dans les configurations *Mono +c* et *Multi +c*, les marqueurs morphologiques étant bien souvent contenus dans les préfixes et suffixes des mots. Cette importance se ressent dans les prédictions. Dans les configurations *Mono +c*

et *Multi +c*, le modèle de caractères aide systématiquement, avec une augmentation d'un peu moins de 10 points en moyenne du $app(mdc, X)$ où X peut être n'importe quelle expérience monolingue ou multilingue, avec une augmentation des scores pour presque toutes les langues. Il n'est cependant pas possible d'affirmer qu'il existe un partage de connaissances au niveau des expériences multilingues, les caractères de la langue étant contenus dans ces expériences. Nous allons vérifier alors vérifier si le modèle de caractères permet toujours d'améliorer les prédictions dans un cadre de zero-shot.

En condition de zero-shot, la situation s'inverse : le $app(mdc, ZS)$ diminue de 0.27 point en moyenne et $app(mdc, ZS\ W_{22})$ diminue de 0.95 point. 18/38 langues ont leur $app(mdc, ZS)$ qui augmente, et 20/38 pour $app(mdc, ZS\ W_{22})$. Certains phénomènes surprenants ont cependant lieu : alors que le catalan (ca) bénéficie du modèle de caractères même en condition zero-shot, avec un gain de 7.51 points, l'espagnol (es), qui devrait bénéficier de la présence d'une langue comme le catalan (ca) dans l'ensemble d'apprentissage, perd 7.08 points. Plus d'investigations seraient nécessaires pour comprendre ce type de phénomène. Il serait intéressant par exemple de récupérer tous les préfixes et les suffixes du catalan et de l'espagnol associé à leurs traits morphologiques, et trouver les points communs et différences pouvant éventuellement expliquer ce phénomène. On remarquera cependant que des langues comme le finlandais (fi), le hongrois (hu), l'estonien (et), le turc (tr) ou l'arabe (ar), qui sont des langues morphologiquement très riches, bénéficient toutes de l'absence du modèle de caractères en condition de zero-shot. Ces résultats suggèrent que les préfixes et suffixes de ces langues, porteurs de marqueurs morphologiques, ne sont pas partagés avec d'autres langues, ne permettant pas de partage de connaissances.

Conclusion Le vecteur W_{22} issu du WALS étant très syntaxique, il peut sembler cohérent que son utilité soit moins visible sur une tâche de prédiction de la morphologie. Alors que dans les chapitres précédents le WALS était utile en multilingue, mais obtenait des résultats moins clairs en zero-shot, il semble ici pouvoir aider la prédiction des traits morphologiques en zero-shot, mais pas en multilingue. Cependant, les langues ayant une morphologie riche semble tout de même pouvoir tirer parti du WALS en multilingue. Il est alors possible que le vecteur W_{22} ne permettent pas d'aider les prédictions de la morphologie, mais qu'il pourrait aider à rapprocher certaines langues entre elles.

Le modèle de caractères permet d'améliorer presque systématiquement les résultats en monolingue et multilingue. Cependant, il semble plutôt gêner les prédictions en zero-shot, en particulier pour les langues à morphologie riche.

Il a aussi été possible de mettre en évidence que les prédictions des POS et des traits morphologiques ne sont pas corrélés, suggérant que ces deux tâches n'ont pas d'influence l'une sur l'autre, même dans le cadre d'une tâche jointe. Il pourrait être intéressant de tester l'utilisation des POS de référence pour la prédiction de la morphologie et inversement, afin de vérifier si ces tâches n'ont réellement pas d'influence l'une sur l'autre.

X	Moy. des diff.	σ	min	max	Corr. POS	Corr. Morpho
Mono $+c$	-9.30	-4.50	-0.20 ro	-18.09 nob	0.43*	-0.17
Mono $-c$	-15.79	-5.19	-5.69 he	-26.12 et	0.60*	0.04
Multi $+c$	-14.68	-5.68	-4.57 cs	-26.63 hi	0.70*	-0.23
Multi $-c$	-24.45	-10.24	-7.41 es	-56.95 ja	0.82*	-0.09
Multi $-c$ ID	-23.12	-8.19	-9.29 es	-43.09 ro	0.82*	0.17
ZS $+c$	-25.71	-11.07	-5.57 ko	-51.21 el	0.33*	0.40*
ZS $-c$	-26.29	-11.01	-2.68 ja	-43.87 el	0.20	0.41*

TABLEAU 5.2. – Différences des scores entre l'expérience X avec les étiquettes prédites et cette même expérience X avec les étiquettes de référence : moyenne des différences, écart type (σ), score et langue dont les résultats diminuent le moins (min) et le plus (max), corrélation de Pearson avec les résultats de la prédiction des POS et avec ceux de la morphologie. Les corrélations significatives (p-value < 0.05) sont suivies d'une étoile.

5.4.2. POS et traits morphologiques de références VS prédits

Les expériences du chapitre 2 utilisaient les POS et la morphologie de référence pour prédire l'arbre syntaxique. Lors des nouvelles expériences, les POS et la morphologie utilisées en entrée ne sont plus celles de référence, mais sont prédites par le modèle. TIEDEMANN 2015 a montré le fort impact de la qualité des POS dans la tâche d'analyse syntaxique de la phrase, et qu'une baisse importante des résultats est à prévoir lors de l'utilisation de POS prédites, en particulier dans un cadre multilingue. L'objectif de cette section est d'étudier l'impact de la prédiction des POS et des traits morphologiques prédits sur la prédiction de l'analyse syntaxique, comparé à l'utilisation des étiquettes de référence.

Monolingue Dans un cadre monolingue, on commence par observer la différence des résultats entre le LAS des expériences avec POS et morphologie de référence, et ceux des expériences avec les POS et la morphologie prédits (voir tableau 5.2).

La prédiction des POS et des traits morphologiques entraîne une baisse moyenne de 9.30 points en condition *Mono* $+c$. Ces résultats varient cependant de manière importante, l'écart type des différences étant de 4.51 points. Les deux langues perdant le plus sont les deux norvégiens, avec -17.45 points pour le nynorsk (nno) et -18.09 pour le bokmål (nob). Ces deux langues disposent de moitié moins de données d'entraînement que les autres et il est très probable que la très faible quantité de données (10 000 tokens seulement à l'apprentissage) impacte les résultats de façon marquée.

Les deux langues suivantes perdant le plus sont le vietnamien (vi) et le chinois (zh), avec respectivement une baisse de 15.69 et de 16.49 points de LAS en utilisant les étiquettes prédites. Ces langues morphologiquement très pauvres devaient utiliser principalement les POS pour faire l'analyse de la phrase. Or leurs résultats pour la prédiction de leur POS sont assez bas : 82.54 pour le vietnamien (vi) et 85.40 pour le chinois (zh), là où les résultats moyens de la prédiction des POS est de 89.40. Ce sont les deux langues avec les scores les plus bas pour la prédiction des POS.

Les langues qui, au contraire, ne souffrent pas trop (une perte de moins de 2 points) de l'absence de POS et de morphologie de référence sont le roumain (ro), avec une baisse de seulement 0.20 point, et le turc (tr) avec une baisse de 1.46 point. Le turc (tr) étant la langue avec les résultats les plus bas (51.22 de LAS avec les étiquettes de référence), il est possible que la qualité des annotations de POS et de morphologie de références du turc (tr) soit assez basse.

L'écart type des résultats des deux modèles *Mono +c* utilisant les POS et traits morphologiques de référence du roumain (ro) est assez important (3.10) et il se peut que ce modèle soit assez instable [2]. Par conséquent, il n'est pas possible de tirer de conclusion sur la faible baisse des résultats en configuration délexicalisée pour le roumain (ro).

La différence de scores entre les expériences avec les étiquettes prédites et les étiquettes de référence donne une mesure de l'impact de la prédiction des étiquettes pour chaque langue. La corrélation de Pearson est calculée entre cette mesure d'impact, et le score de prédiction des étiquettes de POS. Un score de 0.43 est obtenu, indiquant l'existence d'une corrélation entre ces résultats. Plus le score de prédiction des POS est bas, plus la chute des résultats est prononcée avec les étiquettes prédites. Cependant, lorsqu'on calcule la corrélation avec la prédiction des étiquettes de traits morphologiques aucune corrélation n'a pu être mise en évidence. La qualité des prédictions des POS semble être plus importante que celle des traits morphologiques pour la prédiction de l'arbre syntaxique.

Pour les modèles n'utilisant pas les caractères, on constate les mêmes tendances, mais de manière plus prononcée encore : en moyenne, les résultats baissent plus (-15.79 points) et la corrélation avec la prédiction des POS est de 0.60. Sans les caractères, les prédictions des POS sont d'autant plus importantes et nécessitent une meilleure qualité de prédiction encore.

Multilingue En contexte multilingue, on observe des comportements similaires aux expériences monolingues. La moyenne des différences est plus élevée, atteignant -14.68 points de LAS de différence entre les expériences avec et sans prédiction des étiquettes. Sans les caractères, cette différence atteint -24.45 points. L'écart type de 5.68 est également plus élevé. La langue pâtissant le plus de la prédiction des étiquettes est l'hindi (hi) avec une baisse de 26.63 points de LAS. Celle qui, au contraire, s'en sort le mieux est le tchèque (cs), avec une chute de "seulement" 4.57 points de LAS. Le score d'exactitude de la prédiction des POS pour cette langue était de 92.96, le deuxième meilleur score après celui du bulgare (bg), de 94.03. Le score de corrélation de Pearson entre la prédiction des POS et la différence de score entre les modèles avec et sans prédiction des étiquettes est calculée, et de nouveau, ces deux scores sont corrélés, avec une corrélation de 0.70, plus importante encore que pour les résultats monolingues. Cette forte corrélation semble indiquer qu'en multilingue, le système s'appuie plus encore sur les POS qu'en monolingue. Aucune corrélation n'est cependant établie avec la prédiction de la morphologie.

2. L'écart type entre les deux expériences est de 0.41 en moyenne.

Zero-shot Comme vu dans les différentes expériences de cette thèse, les conditions des expériences zero-shot sont très difficiles. En utilisant les étiquettes de POS et de morphologie de référence, le score de prédiction de l'analyse syntaxique est de 55.26 points de LAS en moyenne sur toutes les langues. En prédisant également les POS et la morphologie, le score pour l'analyse syntaxique chute à 29.54 points de LAS, soit une baisse de 25.71 points. L'écart type des différences est de 11.07 points, indiquant une très forte variation de l'importance de l'utilisation des étiquettes de référence selon les langues.

Le coréen (ko) est la langue qui, en apparence, s'en sort le mieux avec une baisse de seulement 5.57 points. Ce résultat est cependant à nuancer, le score du LAS du coréen (ko) étant de 23.40 en utilisant les étiquettes de référence. Ce score baisse à 17.83 points avec la prédiction des étiquettes de POS et de morphologie, soit une baisse de 5.57 points. La baisse n'est certes pas très importante, mais cela s'explique par les très faibles résultats de la prédiction de la syntaxe, même en utilisant les étiquettes de référence. La langue subissant la baisse la plus importante est le grec (el) avec une baisse de 51.21 points de LAS. Avec un score d'exactitude de 46.02 pour la prédiction des POS, il est possible que, là encore, la prédiction des POS joue un rôle dans les faibles résultats du modèle n'utilisant pas les étiquettes de référence.

On remarque que pour les expériences $ZS+c$, les résultats de la prédiction des POS sont moins corrélés avec la différence entre les expériences avec et sans étiquettes de référence. Avec un score de Pearson de 0.33, ces scores restent tout de même légèrement corrélés. Contrairement aux expériences *Mono* $+c$ et *Multi* $+c$, on observe cette fois une corrélation significative avec les scores de la prédiction de la morphologie, avec un score de Pearson de 0.40. On peut supposer qu'avec le peu d'informations disponibles en contexte de zero-shot, le modèle se retrouve à se reposer plus sur les traits morphologiques que pour les expériences précédentes.

Globalement Il peut être intéressant de noter que les langues bénéficiant et pâtissant le plus de la prédiction des POS changent énormément selon les expériences, et il ne semble pas y avoir une langue qui revienne de façon systématique. Le roumain (ro) apparaît même à la fois dans le min et dans le max, respectivement pour les expériences *Mono* $+c$ et pour les expériences *Multi* $-c$ ID. La corrélation avec les résultats de la prédiction des POS est presque systématique, exceptée pour l'expérience $ZS-c$.

5.4.3. WALS

Le WALS a prouvé à travers les différents chapitres de ce travail qu'il pouvait se montrer utile dans des cadres de prédictions multilingues. Son utilité semble plus ambivalente dans un cadre de zero-shot : il semble pouvoir aider à un meilleur partage de connaissances pour certaines langues proches (par exemple l'hindi (hi) et l'ourdou (ur), deux langues presque identiques, mais utilisant un alphabet différent). Au contraire, il semble également aider à l'éloignement de langues se ressemblant, mais étant assez éloignées en réalité (par exemple, l'arabe (ar) et l'ourdou (ur), qui partagent leur alphabet, mais sont très différentes).

X	*Multi +c*			*Multi −c*			*ZS +c*			*ZS −c*		
Lang.	POS	Morpho	LAS	POS	Morpho	LAS	POS	Morpho	LAS	POS	Morpho	LAS
ar	12.12	12.15	16.10	-5.15	-4.97	-6.46	5.83	-0.29	-2.91	4.49	-2.04	-3.37
bg	1.48	0.77	2.88	-4.09	-4.79	-9.14	-2.82	-1.80	3.14	-1.21	0.11	5.16
ca	-11.68	-13.31	-13.58	-8.54	-8.86	-12.97	1.89	2.96	3.81	0.65	-0.44	0.09
cs	-3.49	-4.11	-8.18	5.03	5.82	10.17	3.21	11.92	12.67	4.03	11.87	12.62
da	-1.08	-4.96	-8.10	2.86	-0.06	-0.17	-0.01	0.49	1.05	-2.01	-8.53	-11.17
de	-0.19	-0.10	-0.28	0.52	0.36	1.30	-9.72	-5.15	-6.72	-12.60	-4.33	-4.21
el	7.62	6.97	14.15	-7.03	-7.22	-13.47	0.15	5.96	8.81	1.30	2.43	4.32
en	-2.50	-3.21	-6.96	-1.45	-2.09	-3.97	0.55	-6.07	-2.85	0.16	-4.09	-6.77
es	-1.04	-0.61	-0.53	-2.63	-1.90	-3.63	2.25	0.07	0.14	1.10	-0.96	-0.63
et	1.85	1.65	2.91	5.30	5.85	9.00	0.06	-1.98	-1.89	-0.74	1.87	2.08
eu	1.90	1.48	3.58	-0.87	-1.20	-1.00	9.21	1.47	-2.00	7.76	-3.98	-3.86
fa	-0.26	0.00	-0.32	1.02	1.12	4.64	5.43	-2.42	-4.15	6.83	2.93	3.95
fi	0.39	0.28	-0.39	0.79	0.68	0.22	-1.89	-8.01	-7.14	1.12	-4.24	-5.43
fr	-13.17	-14.77	-20.03	-0.27	0.46	1.81	0.22	-3.81	-4.25	-0.53	-1.65	-3.58
ga	-0.56	-0.55	-1.41	11.73	13.17	15.33	-0.29	8.09	7.60	0.06	0.14	3.93
he	-0.46	-0.05	0.69	4.59	3.91	6.50	1.00	2.97	1.04	-0.74	-3.34	-1.97
hi	10.81	10.17	17.15	13.78	14.24	22.42	-4.40	0.17	3.67	3.76	5.39	7.46
hr	4.70	3.63	6.07	11.76	9.56	16.27	5.61	10.40	15.02	7.16	11.06	13.27
hu	-19.08	-17.56	-23.39	4.90	4.27	7.67	6.01	-8.57	-4.96	9.31	1.65	1.15
id	1.34	8.21	14.89	0.39	1.01	2.19	12.10	6.32	8.88	29.78	6.46	16.04
it	-0.43	-0.67	-0.30	3.31	3.86	4.44	4.00	9.99	16.13	-7.29	6.40	7.12
ja	0.19	1.20	3.60	2.25	24.76	33.20	-0.46	4.54	1.31	1.45	-0.69	-0.71
ko	0.45	6.42	12.51	0.50	1.74	4.05	-17.21	-2.07	-3.18	2.81	-1.90	0.86
lv	-1.72	-2.96	-6.37	5.04	6.94	10.44	-0.29	-0.98	-2.06	-0.23	6.04	8.29
nl	-4.60	-0.18	-0.04	-4.24	-0.96	-0.97	2.22	4.93	8.39	-0.40	0.02	2.47
nno	-39.12	-9.68	-11.91	-26.78	-3.94	-1.31	-26.21	-8.58	-9.50	-16.31	-4.07	-2.37
nob	-35.35	-6.23	-9.90	-24.99	-3.08	-2.89	-17.87	-2.93	-6.81	-5.91	10.66	13.99
pl	-4.32	-6.32	-12.29	1.14	1.21	1.04	2.66	-4.30	-7.14	4.94	-0.41	2.03
pt	-19.02	-0.40	-0.77	-19.52	-3.11	-5.71	1.14	2.61	4.36	1.71	0.00	-0.02
ro	7.80	6.25	11.15	2.17	1.34	2.60	0.49	4.15	5.66	1.36	8.22	9.28
ru	-12.75	1.56	-9.21	-8.72	5.81	-1.65	-1.80	8.73	10.03	-1.72	7.83	5.49
sl	-10.85	-12.79	-15.65	2.46	-0.09	1.57	0.29	1.25	0.63	2.39	1.34	2.39
sv	-5.76	-7.73	-12.72	-8.54	-14.97	-18.55	7.21	1.26	0.89	7.09	1.24	0.29
tr	15.24	16.37	19.06	0.63	0.88	1.29	5.62	1.99	3.36	6.76	5.54	4.12
uk	0.83	0.16	0.41	-6.95	-13.70	-19.10	3.41	-2.84	-1.05	4.17	-0.64	3.59
ur	-0.68	0.01	0.15	-5.20	-5.57	-9.43	23.86	14.05	11.98	16.76	2.54	11.44
vi	0.07	0.03	0.64	0.81	5.12	4.80	32.84	-0.61	1.25	4.40	-8.43	1.19
zh	0.01	0.13	0.70	0.17	-0.49	1.00	0.30	2.69	2.27	-1.56	-3.31	1.59
moy.	-3.19	-0.76	-0.94	-1.42	0.92	1.36	1.44	1.23	1.72	2.11	1.07	2.63

TABLEAU 5.3. – $app(wals, X)$ pour la prédiction des POS, des traits morphologiques, et la prédiction de l'analyse syntaxique pour les modèles X.

L'utilité du WALS est de nouveau estimée dans le cadre des prédictions jointes. Les résultats complets des modèles joints sont disponibles dans les tableaux .18, .19, .20, .21, en annexe.

On mesure "l'apport de l'utilisation du WALS" en calculant pour chaque langue :

$$app(wals, Multi+c) = (Multi+c\ W_{22}) - (Multi+c)$$

On définit l'équivalent zero-shot de la même manière, avec et sans modèle de caractères. La valeur des $app(wals, X)$ pour chaque modèle, pour les différentes prédictions est disponible dans le tableau 5.3.

Contrairement aux expériences de prédictions de POS du chapitre 4, le WALS dans le cadre des prédictions jointes ne semble plus utile pour l'expérience $Multi+c$, que ce soit au niveau de la prédiction des POS, des traits morphologiques ou de l'analyse syntaxique (respectivement -3.19, -0.76 et -0.94 en moyenne). Contrairement aux expériences d'analyse syntaxique ou d'étiquetage seules, les langues du modèle joint sont nombreuses à ne pas profiter du WALS. Environ la moitié en bénéficie, avec certaines langues comme l'arabe (ar), l'hindi (hi) ou le turc (tr) qui en bénéficient beaucoup (+de 10 points pour les trois tâches de prédictions). Parallèlement à ça, certaines langues en pâtissent au contraire beaucoup : le catalan (ca), le français (fr), le hongrois (hu), et le slovène (sl) perdent tous plus de 10 points dans les trois tâches de prédiction. Les deux norvégiens ont également des chutes de résultats importantes, avec une baisse de plus de 35 points pour la prédiction des traits morphologiques avec l'utilisation du WALS. Un fait surprenant à propos du catalan (ca) et du français (fr), dont la présence du WALS fait fortement baisser les résultats, est que ces deux langues sont les seules à avoir pour valeur "OptDoubleNeg" pour deux traits (134A et 144A, voir tableau .1 en annexe). Ces traits sont liés à la position des marqueurs de la négation dans ces deux langues. Il serait intéressant de mener plus d'expériences afin d'étudier plus en profondeur les partages de valeurs de traits entre les langues afin de voir s'il existe une corrélation avec les résultats.

En contexte zero-shot, les résultats augmentent en moyenne, quel que soit le score de prédiction qui nous intéresse. De très grosses améliorations des résultats pour certaines langues comme le vietnamien (vi) (jusqu'à +32.84 points pour la prédiction des POS) sont à mettre en parallèle avec de grosses dégradations des résultats comme pour le coréen (ko) (-17.21 points pour les POS). Ces écarts importants sont plus marqués sur les scores des POS. Il ne semble d'ailleurs toujours pas y avoir de lien entre l'apport du WALS pour prédire les POS et l'apport du WALS pour la prédiction de la morphologie et de l'arbre syntaxique.

La corrélation de Pearson entre l'apport du WALS pour des paires de scores a alors été calculée. Autrement dit, nous avons cherché à vérifier si l'apport du WALS pour la prédiction des POS était corrélé à l'apport du WALS pour la prédiction des traits morphologiques par exemple. Les résultats de ces corrélations sont disponibles dans le tableau 5.4.

Toutes les mesures sont corrélées de façon significative (p-value < 0.05) à l'exception de la paire POS-morphologie pour l'expérience $ZS-c$. Tandis que les corrélations

Modèle X-Y	POS-Morpho	POS-LAS	Morpho-LAS
Multi $+c$	0.72*	0.72*	0.96*
Multi $-c$	0.59*	0.60*	0.97*
$ZS+c$	0.39*	0.40*	0.93*
$ZS-c$	0.23	0.40*	0.87*

TABLEAU 5.4. – Pour chaque modèle M, calcul de la corrélation entre $app(wals, M)$ de la prédiction X et celui de la prédiction Y

sont légères en zero-shot pour POS-Morpho et POS-LAS, elles sont plus importantes dans un contexte multilingue. La corrélation entre l'apport du WALS pour la prédiction des traits morphologiques et la prédiction de l'analyse syntaxique est, quant à elle, extrêmement élevée, allant de 0.87 pour $ZS-c$ jusqu'à 0.97 pour *Multi* $-c$, alors qu'aucune corrélation n'a pu être mise en évidence entre les scores de la prédiction des traits morphologiques et les scores de la prédiction de l'analyse syntaxique. On peut en déduire que le WALS est donc utilisé plus ou moins de la même manière pour les prédictions des traits morphologiques et l'analyse syntaxique, mais est utilisé différemment pour la prédiction des POS. Les traits du WALS sont plutôt de nature syntaxique, ce qui explique que l'apport du WALS est souvent plus important sur les tâches de prédiction de l'analyse syntaxique. Il serait intéressant de créer d'autres vecteurs de représentation de la langue (issu du WALS ou autre) qui soient plus lexicaux que syntaxique, afin de tester si l'impact sur la prédiction des POS pourrait être plus bénéfique.

5.5. Conclusions et perspectives

Le tagparsing a été réalisé et évalué dans ce chapitre. Ce système prédit les POS, les traits morphologiques et l'analyse syntaxique de façon jointe grâce à un *tagparser*. C'est dans ce contexte que les traits morphologiques ont été prédits pour la première fois.

Nous avons pu constater que dans le cas de la morphologie, les langues morphologiquement riches obtiennent, de façon assez cohérente, des résultats plus bas que les autres langues. Le WALS n'est pas utilisé de la même manière que dans les chapitres précédents : peu utile en condition multilingue, le WALS permet cependant d'aider en zero-shot, contrairement aux conclusions du chapitre 2 et du chapitre 4. L'utilisation d'un modèle de caractères est extrêmement utile pour la prédiction de traits morphologiques en monolingue et multilingue, presque toutes les langues bénéficient de cet ajout. Cette utilité n'est cependant plus certaine en condition de zero-shot, où le modèle de caractères fait baisser les résultats en moyenne. Aucun apport n'est d'ailleurs constaté pour les langues à morphologie riche.

Les traits morphologiques et les POS utilisés pour l'analyse syntaxique sont désormais prédits. Nous avons mesuré l'impact de cette prédiction comparé à l'utilisation des étiquettes de référence. L'écart est très important, la prédiction des étiquettes

faisant perdre jusqu'à 26.29 points de LAS en moyenne pour l'expérience $ZS-c$. Les résultats de l'analyse syntaxique semblent dépendre des POS bien plus que des traits morphologiques en monolingue et multilingue. En zero-shot, les résultats de l'analyse syntaxique sont plus corrélés à la morphologie. Il est possible qu'avec le peu d'informations disponibles en zero-shot, le système se repose un peu plus sur les traits morphologiques que pour les expériences monolingues et multilingues.

L'utilité du WALS a alors été étudiée pour vérifier si son impact est similaire aux expériences des chapitres précédents sur les résultats de tagparsing. Contrairement aux chapitres précédents, le WALS ne semble plus être utile en multilingue, constat également établi pour les prédictions des traits morphologiques. Il est cependant utile en moyenne en zero-shot, même si, comme habituellement, cela varie énormément langue par langue. Le WALS semble être utilisé de façon très similaire pour la prédiction des traits morphologiques et pour l'analyse syntaxique. Plus d'analyses seraient nécessaires pour comprendre ce phénomène, en particulier sur la nature très syntaxique des vecteurs que nous avons extraits du WALS.

Le système multilingue de prédictions complètes présenté dans ce chapitre constitue une première étape importante. Il permet notamment de prédire les POS, les traits morphologiques et l'arbre syntaxique d'un texte (segmenté en mots) dans une langue pour laquelle aucune donnée annotée n'est disponible, en dehors d'un corpus brut pour apprendre les plongements de mots, et un petit lexique bilingue vers une langue pivot. Cependant, étant donné la complexité de la tâche et l'ambition des objectifs, de nombreuses autres pistes restent à explorer. Il serait intéressant de mettre en place une expérience masquant l'information de chaque trait du WALS afin de voir quels sont ceux qui ont le plus d'impact dans les prédictions, afin de permettre d'identifier si ce sont des traits plutôt en lien avec la syntaxe, la morphologie, ou autre. Cependant, la possibilité que les traits du WALS soient dépendants les uns des autres est importante, et il serait alors nécessaire de masquer des ensembles de traits pour vérifier l'impact de chaque trait sur les expériences, ce qui nécessiterait de faire une très grande quantité d'expériences supplémentaires. L'ajout de la prédiction de la segmentation en phrases et en mots pourrait également permettre d'obtenir un vrai système de bout en bout en partant du texte brut. Certaines analyses réalisées dans les chapitres précédents auraient leur place ici aussi, entre autre l'analyse des activations des neurones du réseau (section 2.7) ou les tests sur des familles de langues (section 4.4.5).

Conclusions et perspectives

Le but que nous souhaitions atteindre dans cette thèse était de réaliser un système capable d'annoter un texte de n'importe quelle langue grâce à cet unique système. Nous avons tenté de réaliser un tel système à l'aide de trois ressources et outils : les *Universal Dependencies* (UD), le *World Atlas of Language Structures* (WALS) et des plongements de mots multilingues. Tout au long de ces travaux, nous avons testé quels étaient les éléments pouvant permettre de faciliter le partage de connaissances dans des systèmes multilingues et de zero-shot sur une tâche d'analyse syntaxique délexicalisée, une tâche de prédiction des POS des mots, et sur une tâche jointe de prédiction des POS, des traits morphologiques et d'analyse syntaxique.

Les deux grandes questions explorées dans le chapitre 2 étaient : comment représenter les données de manière universelle, et comment représenter l'information de l'origine de la langue ? Dans le cadre d'une tâche de prédiction de l'analyse syntaxique de la phrase, nous avons créé des corpus afin de profiter au mieux des annotations universelles mises à disposition par UD. Ces corpus permettent une représentation équilibrée des différentes langues à notre disposition. Nous avons ensuite testé l'utilité de diverses représentations de la langue. Ces représentations sont basées sur des statistiques sur les annotations syntaxiques d'une langue (W_d et W_{df}), sur le WALS (W_N, W_{22}) ou sur un simple identifiant de la langue, et chacune a ses avantages et ses défauts. Bien que les meilleurs résultats en multilingue soient obtenus via l'utilisation des représentations basées sur les statistiques des langues, ces vecteurs ont l'inconvénient de ne pas pouvoir être appris en l'absence de données d'entraînement, ce qui empêche leur utilisation dans un contexte de zero-shot. Nous avons pu mettre en évidence que l'utilisation de représentations de la langue permet d'augmenter nettement les résultats en conditions multilingues, et que les vecteurs W_d et W_{df} permettent d'obtenir de meilleurs résultats qu'un identifiant de la langue, suggérant un partage de connaissances entre langues supplémentaire comparé à l'utilisation d'un simple identifiant de la langue. De plus, nous avons montré, à l'aide d'analyses sur les activations des neurones, que l'utilisation du vecteur W_{22} permet de rapprocher des langues proches, et d'éloigner des langues distantes lors d'un apprentissage en condition zero-shot.

Les initiatives créant des représentations universelles sont de plus en plus mises sur le devant de la scène. Les *Universal Dependencies* (UD), en particulier, proposent un schéma d'annotation commun pour toutes les langues pour les POS, les traits morphologiques et les dépendances syntaxiques. Cependant, l'utilisation des UD ne permet pas de représenter le lexique d'une manière universelle. L'utilisation de plongements de mots multilingues est une solution élégante pour représenter le lexique d'une manière universelle pour toutes les langues. La création de plongements

de mots multilingues et leur évaluation ont été étudiées dans le chapitre 3. Nous avons montré qu'en partant de plongements de mots pré-entraînés de FastText (BOJANOWSKI et al. 2017) puis en les alignant à l'aide de MUSE (CONNEAU et al. 2017), il était possible d'obtenir des plongements de mots multilingues dans un même espace commun, qui obtiennent de bons résultats sur des évaluations extrinsèques (voir section 4.4) malgré une étape d'harmonisation problématique.

Nous nous sommes ensuite intéressés à l'impact de l'utilisation de W_{22} pour la prédiction d'étiquettes de POS dans le chapitre 4. Nous avons pu conclure, comme dans le chapitre 2, que le WALS est presque toujours utile en condition multilingue, et peut aider à rapprocher et éloigner certaines langues proches (hindi et ourdou)) ou distantes (arabe et ourdou) en condition d'apprentissage zero-shot. Il est surtout utile pour les langues qui sont assez isolées par rapport à l'ensemble d'entraînement. Nous avons également pu constater que l'utilisation d'un modèle de caractères, bien que systématiquement utile en monolingue et multilingue, est bien plus ambivalent en zero-shot. Si une langue L est un peu isolée (ou bien qu'une langue avec un même alphabet mais avec très peu de points communs avec la langue L était présente à l'entraînement) alors il est préférable de retirer le modèle de caractère. La présence d'une langue proche dans l'ensemble d'entraînement s'est montrée être un facteur déterminant lors des expériences en zero-shot. Les résultats de zero-shot étaient pourtant souvent meilleurs que l'utilisation du simple modèle de la langue la plus proche (au sens empirique). Nous avons testé un apprentissage intermédiaire, consistant à prendre une famille de 6 langues proches. Bien que là encore, supérieur à la simple utilisation du modèle de la langue la plus proche, les systèmes s'entraînant sur l'ensemble de toutes nos langues restaient supérieurs. La diversité des langues vues à l'entraînement semble donc être un atout en zero-shot, car la simple augmentation de la quantité de données n'est probablement pas la seule responsable de l'augmentation des scores.

Dans le cadre d'une tâche jointe de prédiction des POS, des traits morphologiques et de l'analyse syntaxique réalisée dans le chapitre 5, nous avons prédit les traits morphologiques pour la première fois. Nous avons pu constater que dans le cas de la prédiction des traits morphologiques, le WALS est en moyenne utile en zero-shot, alors que ce n'est pas le cas en multilingue. Au contraire, l'utilisation des caractères, bien qu'extrêmement utile en monolingue et multilingue, ne semble pas permettre de partager des connaissances en zero-shot. Sur un système de prédictions complètes, nous avons pu constater que l'apport du WALS est très similaire sur la tâche de prédiction des traits morphologiques et sur celle de prédiction de l'analyse syntaxique.

Est-il possible d'aider le partage de connaissances entre langues dans des systèmes d'apprentissages multilingues et zero-shot? Cette question, qui est au cœur de cette thèse, n'a pas de réponse définitive. Cependant, les analyses mises en place tout au long de ces travaux ont permis de mettre en évidence que des ressources comme le WALS pouvait permettre un meilleur traitement des langues très isolées. Néanmoins, l'élément déterminant pour traiter au mieux une langue en zero-shot est la présence d'une langue proche dans l'ensemble d'entraînement. Cette langue proche peut cependant ne pas toujours être une évidence, comme dans le cas de l'hindi et de

l'ourdou n'utilisant pas le même système d'écriture, alors que les deux langues sont très similaires. L'utilisation des caractères des mots peut porter préjudice à ces langues, même si les caractères sont très utiles dans un cadre multilingue et en zero-shot pour des langues proches comme l'espagnol et le catalan. L'utilisation d'un lexique *universel* a également été testée, et bien qu'utile en moyenne, il n'a pas été possible de conclure sur une utilité définitive des plongements de mots multilingues dans un contexte de zero-shot. Même si les scores de performance des analyseurs et étiqueteurs utilisés dans cette thèse restent bas comparés à l'état de l'art, ce travail a permis de proposer des approches pouvant permettre d'inclure des langues disposant de peu de ressources dans les systèmes de TAL. Bien que nous ayons pu mettre en évidence certaines clefs pouvant aider au partage de connaissances entre langues, l'hypothèse de l'existence d'une grammaire universelle évoquée par CHOMSKY et LASNIK 2008 n'a pas pu être démontrée à l'aide des travaux de cette thèse.

Perspectives De nombreuses pistes restent encore à explorer. Pour commencer, les dernières étapes de prédictions pour réaliser la chaîne complète de traitement pourraient être mise en place : la segmentation en phrases et en tokens. La prédiction des lemmes également, qui comme les POS, les traits morphologiques et les arbres syntaxiques est une brique de base utilisée pour de nombreuses autres tâches de TAL.

Quelques questions restent encore en suspend et auraient méritées une recherche plus approfondie :

— quelles dépendances syntaxiques, POS, traits morphologiques, ... sont impactées par quel trait du WALS ? Des systèmes pourrait être ré-entraînés en retirant un trait du WALS, et il serait ainsi possible de mesurer l'impact de ce trait sur les prédictions.

— quelle est l'origine des variations de résultats selon les langues en monolingue ? Comment définir la difficulté intrinsèque d'une langue ?

Certaines étapes mériteraient d'être retravaillées avec le recul actuel que nous avons. Par exemple, il serait intéressant de revoir le travail effectué sur l'harmonisation par le moyennage des plongements de mots, par exemple à l'aide d'une pondération par la fréquence d'apparition des mots dans chaque langue. L'utilisation de plongements de mots contextuels issus de BERT multilingue par exemple (DEVLIN et al. 2019) pourrait également permettre l'obtention d'un lexique *universel* de meilleur de qualité.

Le travail sur l'obtention d'un vecteur du WALS pourrait également être revu, en particulier la façon de remplacer les valeurs manquantes, qui pourrait passer par une prédiction plus élaborée, ainsi qu'une évaluation de la qualité des méthodes de remplacement. De nombreux traits du WALS n'ont pas été utilisés, et certains d'entre eux qui ont été mis de côté pourraient éventuellement permettre d'aider le partage de connaissances entre langue. Plusieurs méthodes de prise en compte des représentations de la langue par le système mériteraient d'être testées, ainsi que d'autres ressources en lien avec la typologie des langues comme Glottolog (HAMMARSTRÖM et al. 2021) à la place du WALS.

Certaines pistes explorées puis mises de côté n'ont pas permis de tirer de conclusions. Pourtant, un certain nombre d'idées auraient pu mener à des résultats intéressants si étudiés plus en profondeur. En particulier, un travail autour de Morfessor (VIRPIOJA et al. 2013) avait été mené dans le but de créer des *morphèmes universels*. L'idée étant de segmenter les mots en morphèmes, puis d'identifier quels étaient les marqueurs morphologiques de chaque langue pour les aligner dans un espace commun, à la manière de l'alignement des plongements de mots réalisés dans le chapitre 3. Une autre piste que nous souhaitions explorer était de s'abstraire du système d'écriture des langues en ramenant l'ensemble des données à l'alphabet phonétique international. Une idée pouvant être explorée plus à long terme aurait été de traiter, à l'aide de méthodes multilingues, des langues orales ne disposant pas de systèmes d'écriture. Nous avons vu que des milliers de langues sont dans cette situation, et des ressources comme PHOIBLE (MORAN et MCCLOY 2019) sont sources de connaissances sur la phonologie des langues, et pourrait être utilisés, à la façon du WALS, sur des corpus oraux.

Bibliographie

AGIĆ, Željko, Anders JOHANNSEN, Barbara PLANK, Héctor Martínez ALONSO, Natalie SCHLUTER et Anders SØGAARD (juill. 2016). « Multilingual Projection for Parsing Truly Low-Resource Languages ». en. In : *Transactions of the Association for Computational Linguistics* 4.0, p. 301-312. ISSN : 2307-387X (cf. p. 12).

AMMAR, Waleed, George MULCAIRE, Miguel BALLESTEROS, Chris DYER et Noah A. SMITH (fév. 2016). « Many Languages, One Parser ». In : *arXiv :1602.01595 [cs]*. arXiv : 1602.01595 (cf. p. 15, 16, 34, 36, 37, 46).

AMMAR, Waleed, George MULCAIRE, Yulia TSVETKOV, Guillaume LAMPLE, Chris DYER et Noah A. SMITH (fév. 2016). « Massively Multilingual Word Embeddings ». In : *arXiv :1602.01925 [cs]*. arXiv : 1602.01925 (cf. p. 34, 53, 54).

BARZEGAR, Siamak, Brian DAVIS, Siegfried HANDSCHUH et Andre FREITAS (jan. 2018). « Multilingual Semantic Relatedness Using Lightweight Machine Translation ». In : *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, p. 108-114. DOI : 10.1109/ICSC.2018.00024 (cf. p. 63).

BERG-KIRKPATRICK, Taylor, Alexandre BOUCHARD-CÔTÉ, John DENERO et Dan KLEIN (juin 2010). « Painless Unsupervised Learning with Features ». In : *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California : Association for Computational Linguistics, p. 582-590 (cf. p. 91).

BJERVA, Johannes et Isabelle AUGENSTEIN (avr. 2021). « Does Typological Blinding Impede Cross-Lingual Sharing? » In : *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*. Online : Association for Computational Linguistics, p. 480-486 (cf. p. 92).

BOJANOWSKI, Piotr, Edouard GRAVE, Armand JOULIN et Tomas MIKOLOV (déc. 2017). « Enriching Word Vectors with Subword Information ». en. In : *Transactions of the Association for Computational Linguistics* 5, p. 135-146. ISSN : 2307-387X. DOI : 10.1162/tacl_a_00051 (cf. p. 53, 143).

BUCHHOLZ, Sabine et Erwin MARSI (2006). « CoNLL-X shared task on multilingual dependency parsing ». In : *Proceedings of the Tenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, p. 149-164 (cf. p. 124).

CAMACHO-COLLADOS, Jose, Mohammad Taher PILEHVAR, Nigel COLLIER et Roberto NAVIGLI (août 2017). « SemEval-2017 Task 2 : Multilingual and Cross-lingual Semantic Word Similarity ». In : *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada : Association for Computational Linguistics, p. 15-26. DOI : 10.18653/v1/S17-2002 (cf. p. 60).

CARDENAS, Ronald, Ying LIN, Heng JI et Jonathan MAY (avr. 2019). « A Grounded Unsupervised Universal Part-of-Speech Tagger for Low-Resource Languages ». In : *arXiv :1904.05426 [cs]*. arXiv : 1904.05426 (cf. p. 91).

CHEN, Danqi et Christopher MANNING (oct. 2014). « A Fast and Accurate Dependency Parser using Neural Networks ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics, p. 740-750 (cf. p. 39).

CHOMSKY, Noam et Howard LASNIK (2008). « The theory of principles and parameters ». In : *Syntax*. De Gruyter Mouton, p. 506-569 (cf. p. 13, 144).

CONNEAU, Alexis, Guillaume LAMPLE, Marc'Aurelio RANZATO, Ludovic DENOYER et Hervé JÉGOU (oct. 2017). « Word Translation Without Parallel Data ». In : *arXiv :1710.04087 [cs]*. arXiv : 1710.04087 (cf. p. 53-55, 143).

DARY, Franck et Alexis NASR (2021). « The Reading Machine : a Versatile Framework for Studying Incremental Parsing Strategies ». In : *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies* (cf. p. 80, 92).

DEVLIN, Jacob, Ming-Wei CHANG, Kenton LEE et Kristina TOUTANOVA (mai 2019). « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *arXiv :1810.04805 [cs]*. arXiv : 1810.04805 (cf. p. 17, 27, 52, 53, 91, 92, 144).

DREDZE, Mark, John BLITZER, Partha Pratim TALUKDAR, Kuzman GANCHEV, João GRAÇA et Fernando PEREIRA (juin 2007). « Frustratingly Hard Domain Adaptation for Dependency Parsing ». In : *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic : Association for Computational Linguistics, p. 1051-1055 (cf. p. 125).

DRYER, Matthew S. et Martin HASPELMATH, éd. (2013). *WALS Online*. Leipzig : Max Planck Institute for Evolutionary Anthropology (cf. p. 14, 23, 34, 35, 37).

DUONG, Long, Trevor COHN, Steven BIRD et Paul COOK (juill. 2015). « Low Resource Dependency Parsing : Cross-lingual Parameter Sharing in a Neural Network Parser ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*. Beijing, China : Association for Computational Linguistics, p. 845-850 (cf. p. 13).

ESKANDER, Ramy, Smaranda MURESAN et Michael COLLINS (nov. 2020). « Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios ». In : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online : Association for Computational Linguistics, p. 4820-4831. DOI : 10.18653/v1/2020.emnlp-main.391 (cf. p. 91).

FARUQUI, Manaal, Yulia TSVETKOV, Pushpendre RASTOGI et Chris DYER (août 2016). « Problems With Evaluation of Word Embeddings Using Word Similarity Tasks ». In : *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany : Association for Computational Linguistics, p. 30-35. DOI : 10.18653/v1/W16-2506 (cf. p. 54).

FINKELSTEIN, Lev, Evgeniy GABRILOVICH, Yossi MATIAS, Ehud RIVLIN, Zach SOLAN, Gadi WOLFMAN et Eytan RUPPIN (avr. 2001). « Placing search in context : the concept revisited ». In : *Proceedings of the 10th international conference on World Wide Web*. WWW '01. New York, NY, USA : Association for Computing Machinery, p. 406-414. ISBN : 978-1-58113-348-6. DOI : 10.1145/371920.372094 (cf. p. 60).

FIRTH, John R (1957). « A synopsis of linguistic theory, 1930-1955 ». In : *Studies in linguistic analysis* (cf. p. 26).

FISCH, Adam, Jiang GUO et Regina BARZILAY (nov. 2019). « Working Hard or Hardly Working : Challenges of Integrating Typology into Neural Dependency Parsers ». In : *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, p. 5714-5720. DOI : 10.18653/v1/D19-1574 (cf. p. 36).

FRANCIS, W Nelson et Henry KUCERA (1979). « Brown corpus manual ». In : *Letters to the Editor* 5.2, p. 7 (cf. p. 91).

FREITAS, André, Siamak BARZEGAR, Juliano Efson SALES, Siegfried HANDSCHUH et Brian DAVIS (2016). « Semantic Relatedness for All (Languages) : A Comparative Analysis of Multilingual Semantic Relatedness Using Machine Translation ». en. In : *Knowledge Engineering and Knowledge Management*. Sous la dir. d'Eva BLOMQVIST, Paolo CIANCARINI, Francesco POGGI et Fabio VITALI. Lecture Notes in Computer Science. Cham : Springer International Publishing, p. 212-222. ISBN : 978-3-319-49004-5. DOI : 10.1007/978-3-319-49004-5_14 (cf. p. 63).

GANIN, Yaroslav, Evgeniya USTINOVA, Hana AJAKAN, Pascal GERMAIN, Hugo LAROCHELLE, François LAVIOLETTE, Mario MARCHAND et Victor LEMPITSKY (2016). « Domain-adversarial training of neural networks ». In : *The Journal of Machine Learning Research* 17.1, p. 2096-2030 (cf. p. 50).

GERZ, Daniela, Ivan VULIĆ, Felix HILL, Roi REICHART et Anna KORHONEN (nov. 2016). « SimVerb-3500 : A Large-Scale Evaluation Set of Verb Similarity ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas : Association for Computational Linguistics, p. 2173-2182. DOI : 10.18653/v1/D16-1235 (cf. p. 60).

GLAVAŠ, Goran, Robert LITSCHKO, Sebastian RUDER et Ivan VULIĆ (juill. 2019). « How to (Properly) Evaluate Cross-Lingual Word Embeddings : On Strong Baselines, Comparative Analyses, and Some Misconceptions ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy : Association for Computational Linguistics, p. 710-721. DOI : 10.18653/v1/P19-1070 (cf. p. 54, 66, 69-71, 74, 78).

GOLDBERG, Yoav et Joakim NIVRE (2012). « A dynamic oracle for arc-eager dependency parsing ». In : *Proceedings of COLING 2012*, p. 959-976 (cf. p. 39).

GOLDWATER, Sharon et Tom GRIFFITHS (2007). « A fully Bayesian approach to unsupervised part-of-speech tagging ». In : *Proceedings of the 45th annual meeting of the association of computational linguistics*, p. 744-751 (cf. p. 91).

GRAVE, Edouard, Piotr BOJANOWSKI, Prakhar GUPTA, Armand JOULIN et Tomas MIKOLOV (mars 2018). « Learning Word Vectors for 157 Languages ». In : *arXiv :1802.06893 [cs]*. arXiv : 1802.06893 (cf. p. 62).

GUO, Jiang, Wanxiang CHE, David YAROWSKY, Haifeng WANG et Ting LIU (2015). « Cross-lingual Dependency Parsing Based on Distributed Representations. » In : *ACL (1)*, p. 1234-1244 (cf. p. 34).

HAMMARSTRÖM, Harald, Robert FORKEL, Martin HASPELMATH et Sebastian BANK (2021). *Glottolog 4.4*. Published : Max Planck Institute for Evolutionary Anthropology. Leipzig. DOI : 10.5281/zenodo.4761960 (cf. p. 14, 23, 35, 144).

HERMANN, Karl Moritz et Phil BLUNSOM (juin 2014). « Multilingual Models for Compositional Distributed Semantics ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Baltimore, Maryland : Association for Computational Linguistics, p. 58-68. DOI : 10.3115/v1/P14-1006 (cf. p. 53).

HILL, Felix, Roi REICHART et Anna KORHONEN (déc. 2015). « SimLex-999 : Evaluating Semantic Models With (Genuine) Similarity Estimation ». In : *Computational Linguistics* 41.4, p. 665-695. DOI : 10.1162/COLI_a_00237 (cf. p. 60).

HOCHREITER, Sepp et Jürgen SCHMIDHUBER (nov. 1997). « Long Short-Term Memory ». In : *Neural Computation* 9.8, p. 1735-1780. ISSN : 0899-7667. DOI : 10.1162/neco.1997.9.8.1735 (cf. p. 25).

JASTRZEBSKI, Stanisław, Damian LEŚNIAK et Wojciech Marian CZARNECKI (fév. 2017). « How to evaluate word embeddings ? On importance of data efficiency and simple supervised tasks ». In : *arXiv :1702.02170 [cs]*. arXiv : 1702.02170 (cf. p. 54).

JOSHI, Pratik, Sebastin SANTY, Amar BUDHIRAJA, Kalika BALI et Monojit CHOUDHURY (juill. 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, p. 6282-6293. DOI : 10.18653/v1/2020.acl-main.560 (cf. p. 12).

KAMHOLZ, David, Jonathan POOL et S. COLOWICK (2014). « PanLex : Building a Resource for Panlingual Lexical Translation ». In : *LREC* (cf. p. 54, 55).

KIM, Joo-Kyung, Young-Bum KIM, Ruhi SARIKAYA et Eric FOSLER-LUSSIER (sept. 2017). « Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources ». In : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark : Association for Computational Linguistics, p. 2832-2838. DOI : 10.18653/v1/D17-1302 (cf. p. 91, 116).

KIM, Yunsu, Petre PETROV, Pavel PETRUSHKOV, Shahram KHADIVI et Hermann NEY (nov. 2019). « Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages ». In : *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, p. 866-876. DOI : 10.18653/v1/D19-1080 (cf. p. 55).

LAUSCHER, Anne, Vinit RAVISHANKAR, Ivan VULIĆ et Goran GLAVAŠ (nov. 2020). « From Zero to Hero : On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers ». In : *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP). Online : Association for Computational Linguistics, p. 4483-4499. DOI : 10.18653/v1/2020.emnlp-main.363 (cf. p. 91).

LIU, Haitao (juin 2010). « Dependency direction as a means of word-order typology : A method based on dependency treebanks ». en. In : *Lingua* 120.6, p. 1567-1578. ISSN : 00243841. DOI : 10.1016/j.lingua.2009.10.001 (cf. p. 36).

LYNN, Teresa, Jennifer FOSTER, Mark DRAS et Lamia TOUNSI (août 2014). « Cross-lingual Transfer Parsing for Low-Resourced Languages : An Irish Case Study ». In : *Proceedings of the First Celtic Language Technology Workshop*. Dublin, Ireland : Association for Computational Linguistics et Dublin City University, p. 41-49 (cf. p. 35).

MARSHALL, Ian (sept. 1983). « Choice of grammatical word-class without global syntactic analysis : Tagging words in the lob corpus ». en. In : *Computers and the Humanities* 17.3, p. 139-150. ISSN : 1572-8412. DOI : 10.1007/BF02259886 (cf. p. 91).

MCDONALD, Ryan, Slav PETROV et Keith HALL (2011). « Multi-source transfer of delexicalized dependency parsers ». In : *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, p. 62-72 (cf. p. 34, 35).

MIKOLOV, Tomas, Quoc V LE et Ilya SUTSKEVER (2013). « Exploiting similarities among languages for machine translation ». In : *arXiv preprint arXiv :1309.4168* (cf. p. 14, 53).

MORAN, Steven et Daniel MCCLOY, éd. (2019). *PHOIBLE 2.0*. Jena : Max Planck Institute for the Science of Human History (cf. p. 145).

MOSELEY, Christopher (2010). *Atlas of the World's Languages in Danger*. Unesco (cf. p. 12).

NASEEM, Tahira, Regina BARZILAY et Amir GLOBERSON (2012). « Selective sharing for multilingual dependency parsing ». In : *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*. Association for Computational Linguistics, p. 629-637 (cf. p. 16, 35-38).

NELSON, Douglas L., Cathy L. MCEVOY et Thomas A. SCHREIBER (août 2004). « The University of South Florida free association, rhyme, and word fragment norms ». en. In : *Behavior Research Methods, Instruments, & Computers* 36.3, p. 402-407. ISSN : 0743-3808, 1532-5970. DOI : 10.3758/BF03195588 (cf. p. 60).

NEUBIG, Graham, Chris DYER, Yoav GOLDBERG, Austin MATTHEWS, Waleed AMMAR, Antonios ANASTASOPOULOS, Miguel BALLESTEROS, David CHIANG, Daniel CLOTHIAUX, Trevor COHN, Kevin DUH, Manaal FARUQUI, Cynthia GAN, Dan GARRETTE, Yangfeng JI, Lingpeng KONG, Adhiguna KUNCORO, Gaurav KUMAR, Chaitanya MALAVIYA, Paul MICHEL, Yusuke ODA, Matthew RICHARDSON, Naomi SAPHRA, Swabha SWAYAMDIPTA et Pengcheng YIN (2017). « DyNet : The Dynamic Neural Network Toolkit ». In : *arXiv preprint arXiv :1701.03980* (cf. p. 40).

NIVRE, Joakim (juill. 2004). « Incrementality in Deterministic Dependency Parsing ». In : *Proceedings of the Workshop on Incremental Parsing : Bringing Engineering and Cognition Together*. Barcelona, Spain : Association for Computational Linguistics, p. 50-57 (cf. p. 28).

– (2008). « Algorithms for deterministic incremental dependency parsing ». In : *Computational Linguistics* 34.4, p. 513-553 (cf. p. 39).

NIVRE, Joakim, Johan HALL, Sandra KÜBLER, Ryan MCDONALD, Jens NILSSON, Sebastian RIEDEL et Deniz YURET (2007). « The CoNLL 2007 shared task on dependency parsing ». In : *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (cf. p. 124, 125).

NIVRE, Joakim, Marie-Catherine de MARNEFFE, Filip GINTER, Yoav GOLDBERG, Jan HAJIC, Christopher D. MANNING, Ryan T. MCDONALD, Slav PETROV, Sampo PYYSALO et Natalia SILVEIRA (2016). « Universal Dependencies v1 : A Multilingual Treebank Collection. » In : *LREC* (cf. p. 14, 22, 33, 35, 125).

NIVRE, Joakim et Jens NILSSON (2005). « Pseudo-projective dependency parsing ». In : *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 99-106 (cf. p. 40).

PILEHVAR, Mohammad Taher, Dimitri KARTSAKLIS, Victor PROKHOROV et Nigel COLLIER (oct. 2018). « Card-660 : Cambridge Rare Word Dataset - a Reliable Benchmark for Infrequent Word Representation Models ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium : Association for Computational Linguistics, p. 1391-1401. DOI : 10.18653/v1/D18-1169 (cf. p. 60).

PLANK, Barbara, Anders SØGAARD et Yoav GOLDBERG (juill. 2016). « Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss ». In : *arXiv :1604.05529 [cs]*. arXiv : 1604.05529 (cf. p. 91).

PONTI, Edoardo Maria, Roi REICHART, Anna KORHONEN et Ivan VULIĆ (juill. 2018). « Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP ». In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Melbourne, Australia : Association for Computational Linguistics, p. 1531-1542 (cf. p. 36).

RUDER, Sebastian, Ivan VULIĆ et Anders SØGAARD (août 2019). « A Survey of Cross-lingual Word Embedding Models ». en. In : *Journal of Artificial Intelligence Research* 65, p. 569-631. ISSN : 1076-9757. DOI : 10.1613/jair.1.11640 (cf. p. 54, 67).

SHI, Tianze, Felix G. WU, Xilun CHEN et Yao CHENG (août 2017). « Combining Global Models for Parsing Universal Dependencies ». In : *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada : Association for Computational Linguistics, p. 31-39. DOI : 10.18653/v1/K17-3003 (cf. p. 36).

SMITH, Aaron, Bernd BOHNET, Miryam de LHONEUX, Joakim NIVRE, Yan SHAO et Sara STYMNE (oct. 2018). « 82 Treebanks, 34 Models : Universal Dependency Parsing with Multi-Treebank Models ». In : *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium : Association for Computational Linguistics, p. 113-123. DOI : 10.18653/v1/K18-2011 (cf. p. 29).

STRAKA, Milan et Jana STRAKOVÁ (août 2017). « Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe ». In : *Proceedings of the CoNLL 2017 Shared Task :*

Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada : Association for Computational Linguistics, p. 88-99. DOI : 10.18653/v1/K17-3009 (cf. p. 179).

STRATOS, Karl, Michael COLLINS et Daniel HSU (2016). « Unsupervised Part-Of-Speech Tagging with Anchor Hidden Markov Models ». In : *Transactions of the Association for Computational Linguistics* 4, p. 245-257. DOI : 10.1162/tacl_a_00096 (cf. p. 91).

STRINGHAM, Nathan et Mike IZBICKI (nov. 2020). « Evaluating Word Embeddings on Low-Resource Languages ». In : *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Online : Association for Computational Linguistics, p. 176-186. DOI : 10.18653/v1/2020.eval4nlp-1.17 (cf. p. 54, 60, 61).

TÄCKSTRÖM, Oscar, Ryan MCDONALD et Joakim NIVRE (juin 2013). « Target Language Adaptation of Discriminative Transfer Parsers ». In : *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Atlanta, Georgia : Association for Computational Linguistics, p. 1061-1071 (cf. p. 16, 36, 38).

TIEDEMANN, Jörg (2015). « Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels. » In : *DepLing*, p. 340-349 (cf. p. 135).

TSARFATY, Reut, Shoval SADDE, Stav KLEIN et Amit SEKER (nov. 2019). « What's Wrong with Hebrew NLP? And How to Make it Right ». In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) : System Demonstrations*. Hong Kong, China : Association for Computational Linguistics, p. 259-264. DOI : 10.18653/v1/D19-3044 (cf. p. 180).

VILARES, David, Carlos GÓMEZ-RODRÍGUEZ et Miguel A. ALONSO (juill. 2015). « One model, two languages : training bilingual parsers with harmonized treebanks ». In : *arXiv :1507.08449 [cs]*. arXiv : 1507.08449 (cf. p. 35).

VIRPIOJA, Sami, Peter SMIT, Stig-Arne GRÖNROOS, Mikko KURIMO et al. (2013). « Morfessor 2.0 : Python implementation and extensions for Morfessor Baseline ». In : Publisher : Aalto University (cf. p. 145).

VULIĆ, Ivan, Simon BAKER, Edoardo Maria PONTI, Ulla PETTI, Ira LEVIANT, Kelly WING, Olga MAJEWSKA, Eden BAR, Matt MALONE, Thierry POIBEAU, Roi REICHART et Anna KORHONEN (fév. 2020). « Multi-SimLex : A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity ». In : *Computational Linguistics* 46.4, p. 847-897. ISSN : 0891-2017. DOI : 10.1162/coli_a_00391 (cf. p. 60, 62).

WANG, Dingquan et Jason EISNER (déc. 2017). « Fine-Grained Prediction of Syntactic Typology : Discovering Latent Structure with *Supervised* Learning ». en. In : *Transactions of the Association for Computational Linguistics* 5, p. 147-161. ISSN : 2307-387X. DOI : 10.1162/tacl_a_00052 (cf. p. 51).

– (déc. 2018). « Surface Statistics of an Unknown Language Indicate How to Parse It ». en. In : *Transactions of the Association for Computational Linguistics* 6, p. 667-685. ISSN : 2307-387X. DOI : 10.1162/tacl_a_00248 (cf. p. 36, 37).

WIKIPEDIA (2021). *Hausa language — Wikipedia, The Free Encyclopedia* (cf. p. 33).

WU, Shijie et Mark DREDZE (nov. 2019). « Beto, Bentz, Becas : The Surprising Cross-Lingual Effectiveness of BERT ». In : *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, p. 833-844. DOI : 10.18653/v1/D19-1077 (cf. p. 92).

YASUNAGA, Michihiro, Jungo KASAI et Dragomir RADEV (avr. 2018). « Robust Multilingual Part-of-Speech Tagging via Adversarial Training ». In : *arXiv :1711.04903 [cs]*. arXiv : 1711.04903 (cf. p. 91).

ZEMAN, Daniel, Jan HAJIČ, Martin POPEL, Martin POTTHAST, Milan STRAKA, Filip GINTER, Joakim NIVRE et Slav PETROV (oct. 2018). « CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies ». In : *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium : Association for Computational Linguistics, p. 1-21. DOI : 10.18653/v1/K18-2001 (cf. p. 15, 35, 124, 125).

ZEMAN, Daniel, Martin POPEL, Milan STRAKA, Jan HAJIČ, Joakim NIVRE, Filip GINTER, Juhani LUOTOLAHTI, Sampo PYYSALO, Slav PETROV, Martin POTTHAST, Francis TYERS, Elena BADMAEVA, Memduh GOKIRMAK, Anna NEDOLUZHKO, Silvie CINKOVÁ, Jan HAJIČ JR., Jaroslava HLAVÁČOVÁ, Václava KETTNEROVÁ, Zdeňka URešOVÁ, Jenna KANERVA, Stina OJALA, Anna MISSILÄ, Christopher D. MANNING, Sebastian SCHUSTER, Siva REDDY, Dima TAJI, Nizar HABASH, Herman LEUNG, Marie-Catherine de MARNEFFE, Manuela SANGUINETTI, Maria SIMI, Hiroshi KANAYAMA, Valeria de PAIVA, Kira DROGANOVA, Héctor MARTÍNEZ ALONSO, Çağrı ÇÖLTEKIN, Umut SULUBACAK, Hans USZKOREIT, Vivien MACKETANZ, Aljoscha BURCHARDT, Kim HARRIS, Katrin MARHEINECKE, Georg REHM, Tolga KAYADELEN, Mohammed ATTIA, Ali ELKAHKY, Zhuoran YU, Emily PITLER, Saran LERTPRADIT, Michael MANDL, Jesse KIRCHNER, Hector Fernandez ALCALDE, Jana STRNADOVÁ, Esha BANERJEE, Ruli MANURUNG, Antonio STELLA, Atsuko SHIMADA, Sookyoung KWAK, Gustavo MENDONÇA, Tatiana LANDO, Rattima NITISAROJ et Josie LI (août 2017). « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies ». In : *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada : Association for Computational Linguistics, p. 1-19. DOI : 10.18653/v1/K17-3001 (cf. p. 15, 35, 41, 124, 125, 179).

ZEMAN, Daniel et Philip RESNIK (2008). « Cross-Language Parser Adaptation between Related Languages. » In : *IJCNLP*, p. 35-42 (cf. p. 34, 35).

ZHANG, Yuan et Regina BARZILAY (sept. 2015). « Hierarchical Low-Rank Tensors for Multilingual Transfer Parsing ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal : Association for Computational Linguistics, p. 1857-1867 (cf. p. 16, 35, 38).

ZHANG, Yuan, David GADDY, Regina BARZILAY et Tommi JAAKKOLA (juin 2016). « Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings ». In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. San Diego, California : Association for Computational Linguistics, p. 1307-1317. DOI : 10.18653/v1/N16-1156 (cf. p. 91).

ZHANG, Yuan, Roi REICHART, Regina BARZILAY et Amir GLOBERSON (juill. 2012). « Learning to Map into a Universal POS Tagset ». In : *Proceedings of the 2012 Joint Confe-*

rence on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea : Association for Computational Linguistics, p. 1368-1378 (cf. p. 91).

ANNEXES

A. Annexes chapitre 2

Traits	Description	Valeurs possibles	Présence dans W
81A	Ordre du Sujet, de l'Objet et du Verbe	SOV SVO VSO VOS OVS OSV Pas d'ordre dominant	W_N W_{22}
82A	Ordre du Sujet et du Verbe	SV VS Pas d'ordre dominant	W_{22}
83A	Ordre de l'Objet et du Verbe	OV VO Pas d'ordre dominant	W_{22}
85A	Ordre de l'Adposition et de la Phrase nominale	Postpositions Prépositions Inpositions Pas d'ordre dominant Pas d'adpositions	W_N W_{22}
86A	Ordre du génitif et du Nom	Génitif-Nom Nom-Génitif Pas d'ordre dominant	W_N W_{22}
87A	Ordre de l'Adjectif et du Nom	Adjectif-Nom Nom-Adjectif Pas d'ordre dominant Uniquement les clauses relatives à tête interne	W_N W_{22}
88A	Ordre du démonstratif et du nom	Démonstratif-Nom Nom-Démonstratif Préfixe Démonstratif Suffixe Démonstratif Démonstratifs avant et après le Nom Mixés	W_N W_{22}

89A	Ordre du Nombre et du Nom	Nombre-Nom Nom-Nombre Pas d'ordre dominant Les nombres modifient uniquement les verbes	W_N W_{22}
90A	Ordre de la Clause Relative et du Nom	Nom-Clause Relative Clause Relative-Nom Internally headed Corrélatif Adjoint Doubly headed Mixé	W_{22}
92A	Position des particules de question polaire	Initiale Finale Seconde position Autre position Dans l'une des deux position Aucune particule de question	W_{22}
94A	Ordre de la subordonné adverbial et de la clause	Mot du subordonné initial Mot du subordonné final Mot subordonné interne Suffixe subordonné Mixé	W_{22}
95A	Relation entre l'ordre de l'objet et du verbe et l'ordre d'adposition et la phrase nominale	OV & Postpositions OV & Prépositions VO & Postpositions VO & Prépositions Autre	W_{22}
96A	Relation entre l'ordre de l'objet et du verbe et l'ordre de la clause relative et du nom	OV & RelN OV & NRel VO & RelN VO & NRel Autres	W_{22}
97A	Relation entre l'ordre de l'objet et du verbe et l'ordre de l'adjectif et du nom	OV & AdjN OV & NAdj VO & AdjN VO & NAdj Autre	W_{22}

101A	Expression des sujets pronominaux	Pronoms obligatoires en position sujet Affixes du sujet sur le verbe Clitiques du sujet sur hôte variable Pronoms sujets dans des positions différentes Pronoms optionnels en position de sujet Mixés	W_{22}
112A	Morphèmes négatifs	Affixe négatif Particule négative Verbe auxiliaire négatif Mot négatif, incertain si verbe ou particule Variation entre mot négatif et affixe Double négation	W_{22}
116A	Questions polaires	Particule de question Morphologie sur les verbes interrogatifs Mélange des deux types précédents Ordre des mots interrogatifs Absence de morphèmes déclaratifs Intonation interrogative uniquement Pas de distinction interrogative-déclarative	W_{22}
143A	Ordre du morphème négatif et du verbe	NegV VNeg [Neg-V] [V-Neg] Ton négatif Type 1 / Type 2 Type 1 / Type 3 Type 1 / Type 4 Type 2 / Type 3 Type 2 / Type 4 Type 3 / Type 4 Type 3 / Negative Infix OptSingleNeg ObligDoubleNeg OptDoubleNeg OptTripleNeg&ObligDoubleNeg OptTripleNeg&OptDoubleNeg	W_{22}

143E	Morphèmes négatifs préverbaux	NegV [Neg-V] NegV&[Neg-V] Aucun	W_{22}
143F	Morphèmes négatifs post-verbaux	VNeg [V-Neg] VNeg&[V-Neg] Aucun	W_{22}
143G	Moyens morphologiques mineurs pour signaliser la négation	NegTone NegInfix NegStemChange Aucun	W_{22}
144A	Position du mot négatif par rapport au sujet, à l'objet et au verbe	NegSVO SNegVO SVNegO SVONeg NegSOV SNegOV SONegV SOVNeg NegVSO NegVSO VSNegO VSONeg NegVOS ONegVS OVNegS OSVNeg Plus d'une position OptSingleNeg ObligDoubleNeg OptDoubleNeg MorphNeg Autre	W_{22}

TABLEAU .1. – Description des traits du WALS. Les mots en rouge sont des valeurs apparaissant dans W_{22}, et celles en violet apparaissent à la fois dans W_{22} et dans W_N. Toutes les valeurs de W_N apparaissent dans W_{22}.

B. Annexes chapitre 3

B.1. Ré-entraînement des analyseurs syntaxiques

L'architecture de Macaon ayant changée depuis les expériences du chapitre 2, les expériences ont été relancées afin d'avoir des résultats comparables entre les différentes expériences de ce chapitre. Les résultats complets de ces deux séries d'expériences sont disponibles dans le tableau .3.

On définit pour chaque langue "l'apport de la nouvelle architecture" :

$$app(arch, Mono) = (Mono_{chap2}) - (Mono_{chap5})$$

On définit de la même manière les équivalents multilingues, zero-shot, et avec et sans méthode de représentation de la langue. L'apport de l'architecture de chaque modèle pour chaque langue est disponible dans le tableau .2.

En condition monolingue, les nouvelles expériences obtiennent de légèrement moins bonnes performances que celles du chapitre 2 (-1.12 points de LAS en moyenne). L'indonésien (id) est la langue bénéficiant le plus de la nouvelle architecture (+3.63 points) et le russe (ru) est celle en pâtissant le plus (-7.57 points). Pour toutes les autres expériences, les résultats augmentent en moyenne, jusqu'à une augmentation de +4.30 points pour les expériences *Multi +c*. Certaines langues tirent profit de cette nouvelle architecture de façon systématique et importante à travers les 6 différentes expériences. C'est le cas de l'indonésien (id), qui gagne jusqu'à 13.45 points pour l'expérience $ZS + c\ W_{22}$. Au contraire, le russe (ru) ne profite jamais de cette nouvelle architecture, perdant jusqu'à 23.10 points pour l'expérience *Multi +c* ID.

Cette expérience nous a permis d'établir notre nouvelle base de référence afin d'avoir des résultats utilisant une même architecture et étant ainsi comparables.

B.2. Modèle de caractères

Nous avons également cherché à quantifier l'apport de l'utilisation du modèle de caractères. Pour cela, on définit pour chaque langue "l'apport de l'utilisation du modèle de caractères" :

$$app(mdc, Mono) = (Mono + c) - (Mono - c)$$

On définit de la même manière les équivalents multilingues et zero-shot. La lexicalisation du modèle a également introduit l'utilisation d'un modèle de caractères. Pour estimer l'apport du modèle de caractères, on compare les expériences *Mono +c* et *Mono −c* ainsi que les versions multilingues et zero-shot de ces expériences. *Mono +c* utilise les plongements de mots et le modèle de caractères alors que *Mono −c* n'utilise que les plongements de mots.

Lang. \ X	*Mono +c*	*Multi +c*	*Multi +c* ID	*Multi +c* W_{22}	*ZS +c*	*ZS +c* W_{22}
ar	-0.77	6.22	4.30	3.50	16.35	2.86
bg	2.03	6.63	4.05	5.00	2.26	10.70
ca	-3.47	0.41	1.68	-1.25	2.00	2.44
cs	-3.08	-1.14	-0.01	-1.30	-12.57	-4.57
da	-7.34	6.03	5.15	4.65	2.82	3.76
de	1.11	6.10	4.73	5.23	0.85	2.24
el	1.19	6.34	4.60	4.64	-1.75	2.65
en	2.89	6.81	4.79	5.12	0.82	2.62
es	-1.41	-0.62	-4.53	-1.60	1.38	1.00
et	2.64	8.22	6.42	6.35	0.06	5.96
eu	-5.09	0.36	0.42	1.13	4.80	5.90
fa	-2.41	4.65	2.35	2.22	-0.55	-0.47
fi	-1.55	4.07	-1.07	-0.78	2.50	4.97
fr	0.41	3.16	2.10	2.46	2.11	2.91
ga	-1.82	1.04	0.17	-1.22	1.26	0.59
he	-0.65	4.94	1.65	1.75	-1.92	0.66
hi	-2.13	1.76	2.94	2.24	12.09	10.85
hr	-1.00	5.06	3.46	3.82	4.44	2.57
hu	-6.95	-3.74	-1.24	-3.85	-1.20	-0.96
id	3.63	9.12	9.95	9.38	0.58	13.45
it	-1.11	3.93	1.72	1.51	2.88	1.48
ja	-1.58	6.06	0.24	1.28	2.38	4.85
ko	0.65	10.38	10.80	10.84	0.42	5.72
lv	-0.30	4.94	5.33	4.71	1.52	1.76
nl	-3.29	3.99	4.49	2.39	1.78	2.68
no	0.21	4.34	-4.13	-6.46	20.97	12.58
pl	-3.60	1.48	0.00	-0.43	1.80	0.99
pt	3.47	7.80	2.39	1.42	3.90	2.28
ro	-5.15	0.61	-0.82	-0.79	-1.12	-2.85
ru	-7.57	0.46	-23.10	-21.65	-5.07	-6.56
sl	3.38	6.34	4.59	4.64	2.14	1.80
sv	1.68	4.34	4.04	3.89	5.94	3.80
tr	0.84	4.55	7.30	7.83	3.39	1.17
uk	-3.10	1.69	-1.62	-1.62	-7.61	3.90
ur	-1.14	3.62	3.66	2.86	8.73	10.15
vi	-2.18	8.89	3.09	6.16	4.75	1.28
zh	1.08	10.27	4.37	8.17	-3.03	-1.67
moy.	-1.12	4.30	2.01	1.95	2.16	3.07

TABLEAU .2. – $app(arch, X)$ de chaque langue pour chaque modèle X, entre les expériences réalisées avec l'architecture du chapitre 2 et l'architecture du chapitre 3. Un résultat négatif indique que le modèle du chapitre 2 obtenait un meilleur résultat.

Lang.	*Mono +c*		*Multi +c*		*Multi +c* ID		*Multi +c* W_{22}		*ZS +c*		*ZS +c* W_{22}	
	chap2	chap5	chap2	chap5	chap2	chap5	chap2	chap5	chap2	chap5	chap2	chap5
ar	65.89	65.12	60.59	66.81	64.04	68.34	64.38	67.88	27.50	43.85	34.34	37.20
bg	78.59	80.62	74.32	80.95	79.25	83.30	77.47	82.47	67.05	69.31	63.73	74.43
ca	77.18	73.71	72.76	73.17	76.63	78.31	76.27	75.02	70.48	72.48	68.88	71.32
cs	68.92	65.84	68.01	66.87	69.41	69.40	69.61	68.31	62.08	49.51	59.47	54.90
da	73.62	66.28	67.38	73.41	70.26	75.41	70.25	74.90	61.56	64.38	63.87	67.63
de	71.07	72.18	63.76	69.86	70.36	75.09	69.22	74.45	59.65	60.50	60.51	62.75
el	77.11	78.30	71.26	77.60	76.16	80.76	75.84	80.48	63.74	61.99	65.96	68.61
en	70.05	72.94	66.02	72.83	71.17	75.96	70.19	75.31	60.87	61.69	62.11	64.73
es	71.47	70.06	71.98	71.36	73.83	69.30	73.22	71.62	71.40	72.78	70.76	71.76
et	66.98	69.62	63.76	71.98	68.41	74.83	67.79	74.14	58.52	58.58	58.77	64.73
eu	63.26	58.17	55.76	56.12	60.11	60.53	59.39	60.52	33.50	38.30	31.68	37.58
fa	72.85	70.44	66.02	70.67	69.50	71.85	70.00	72.22	31.25	30.70	34.27	33.80
fi	60.97	59.42	56.29	60.36	59.65	58.58	59.28	58.50	50.69	53.19	48.72	53.69
fr	75.74	76.15	74.25	77.41	76.16	78.26	75.82	78.28	72.68	74.79	71.97	74.88
ga	66.55	64.73	60.41	61.45	66.86	67.03	65.96	64.74	43.51	44.77	42.75	43.34
he	70.21	69.56	63.03	67.97	67.97	69.62	67.45	69.20	51.36	49.44	52.98	53.64
hi	78.91	76.78	73.86	75.62	74.86	77.80	74.45	76.69	54.23	66.32	57.33	68.18
hr	71.03	70.03	67.49	72.55	71.37	74.83	70.40	74.22	62.88	67.32	65.71	68.28
hu	67.08	60.13	62.55	58.81	66.84	65.60	67.51	63.66	48.24	47.04	49.62	48.66
id	68.64	72.27	58.38	67.50	63.98	73.93	64.57	73.95	45.57	46.15	43.03	56.48
it	81.44	80.33	76.45	80.38	80.56	82.28	79.83	81.34	75.82	78.70	77.04	78.52
ja	78.26	76.68	68.22	74.28	75.81	76.05	75.56	76.84	7.64	10.02	28.53	33.38
ko	47.68	48.33	37.17	47.55	38.61	49.41	39.66	50.50	17.99	18.41	23.18	28.90
lv	59.89	59.59	54.11	59.05	59.98	65.31	60.17	64.88	44.41	45.93	47.38	49.14
nl	62.56	59.27	57.21	61.20	59.28	63.77	58.59	60.98	51.11	52.89	48.98	51.66
no	74.59	74.80	73.19	77.53	76.95	72.82	75.93	69.48	53.09	74.06	55.49	68.07
nno		72.89		75.81		71.24		68.39		71.90		67.27
nob		76.71		79.24		74.40		70.56		76.21		68.87
pl	81.24	77.64	74.24	75.72	80.52	80.52	79.02	78.59	69.15	70.95	70.72	71.71
pt	72.00	75.47	65.74	73.54	69.83	72.22	69.86	71.28	62.85	66.75	65.86	68.14
ro	70.99	65.84	67.72	68.33	71.47	70.65	70.61	69.82	55.84	54.72	61.01	58.16
ru	74.06	66.49	61.35	61.81	74.72	51.62	74.45	52.80	55.09	50.02	55.50	48.94
sl	67.10	70.48	64.12	70.46	66.44	71.03	66.36	71.00	60.52	62.66	63.26	65.06
sv	72.05	73.73	69.97	74.31	72.55	76.59	71.88	75.77	64.80	70.74	67.05	70.85
tr	46.78	47.62	41.01	45.56	43.16	50.46	41.62	49.45	29.46	32.85	30.33	31.50
uk	71.60	68.50	69.40	71.09	75.30	73.68	72.81	71.19	67.08	59.47	64.92	68.82
ur	74.35	73.21	69.15	72.77	70.93	74.59	70.76	73.62	58.93	67.66	59.41	69.56
vi	54.40	52.22	42.42	51.31	53.75	56.84	51.72	57.88	25.86	30.61	41.07	42.35
zh	59.83	60.91	45.46	55.73	58.48	62.85	54.87	63.04	22.24	19.21	24.48	22.81
moy.	69.32	68.20	63.64	67.94	68.25	70.25	67.64	69.60	51.86	54.02	53.80	56.87

TABLEAU .3. – Comparaison du LAS pour chaque langue entre les expériences du chapitre 2 pour l'analyse syntaxique délexicalisée et celles du chapitre 5, pour l'analyse syntaxique lexicalisée.

$app(pdm, Mono)$ n'est pas corrélé à $app(mdc, Mono)$ dans un contexte d'apprentissage monolingue. Ce score de corrélation est également calculé en condition d'apprentissage zero-shot, et aucune corrélation n'est détectée non plus. Cette absence de corrélation pourrait indiquer que les plongements de mots et le modèle de caractère sont complémentaires dans ces expériences, puisqu'ils n'aident pas les mêmes langues de la même manière. Cependant, dans un contexte d'apprentissage multilingue, on trouve une corrélation de Pearson significative de -0.5. Cette corrélation est négative et est donc inversement proportionnelle : plus le gain lié à l'utilisation de plongements de mots est important, moins l'utilisation d'un modèle de caractères est utile.

En moyenne, l'utilisation d'un modèle de caractères permet d'augmenter les résultats de 0.70 point de LAS en condition monolingue. Mais ces résultats varient selon les langues, et l'utilité n'est pas systématique, puisque 8/38 langues ne bénéficient pas du modèle de caractères. Les augmentations et diminutions ne sont cependant pas très importantes, la langue dont les résultats diminuent le plus est le roumain (ro) avec une baisse de 1.44 point, et celle augmentant le plus est l'hébreu (he) avec une augmentation de 2.62 points.

En multilingue, le modèle de caractères semble encore moins utile : seulement 10/38 langues bénéficient de l'utilisation d'un modèle de caractères, pour une moyenne de -0.11 point. Cette différence va de -2.92 pour le tchèque (cs) à +5.04 pour le portugais (pt) en passant par le cas particulier du russe (ru) qui voit son score augmenter de 13.39 points avec l'utilisation du modèle de caractères.

Enfin, en contexte d'apprentissage zero-shot, les résultats sont plus difficilement interprétables. 19 langues semblent bénéficier de l'ajout du modèle de caractères. Les résultats varient cependant de généralement moins d'un point entre $ZS+c$ et $ZS-c$. Quelques langues voient leurs résultats augmenter ou diminuer de manière importante lors de l'ajout du modèle de caractères, comme l'arabe (ar) avec +9.43, le persan (fa) avec -4.33, l'hindi (hi) avec -5.06 et l'ourdou avec -10.98.

Lang.	Capitales	Villes anciennes	Religions	Pays	Élément Chimique	Moyenne
ar	0.64	0.36	0.30	0.98	0.83	0.62
bg	0.76	0.52	0.61	0.94	0.67	0.70
bxr	0.11	0.00	0.00	0.51	0.71	0.44
ca	0.53	0.15	0.30	0.89	0.41	0.46
cs	0.58	0.37	0.45	0.92	0.73	0.61
cu	0.02	0.00	0.00	0.09	0.00	0.03
da	0.47	0.24	0.31	0.85	0.88	0.55
de	0.63	0.52	0.52	0.90	0.49	0.61
el	0.58	0.17	0.25	0.94	0.66	0.52
en	0.64	0.21	0.50	0.95	0.51	0.56
es	0.64	0.29	0.51	0.92	0.55	0.58
et	0.47	0.23	0.63	0.78	0.77	0.58
eu	0.43	0.22	0.06	0.88	0.15	0.35
fa	0.55	0.22	0.52	0.91	0.76	0.59
fi	0.60	0.10	0.40	0.80	0.69	0.52
fr	0.52	0.36	0.42	0.71	0.48	0.50
ga	0.17	0.01	0.01	0.54	0.57	0.26
gl	0.61	0.36	0.45	0.88	0.77	0.61
got	0.00	0.00	0.00	0.00	0.00	0.00
he	0.54	0.27	0.46	0.80	0.80	0.57
hi	0.20	0.06	0.23	0.83	0.50	0.36
hr	0.46	0.41	0.56	0.92	0.58	0.59
hu	0.74	0.35	0.53	0.66	0.59	0.57
id	0.55	0.27	0.27	0.88	0.59	0.51
it	0.59	0.46	0.42	0.93	0.52	0.59
ja	0.17	0.03	0.00	0.24	0.53	0.20
kk	0.35	0.32	0.40	0.69	0.93	0.54
kmr	-	-	-	-	-	
ko	0.83	0.33	0.36	0.97	0.49	0.59
lv	0.40	0.08	0.52	0.92	0.57	0.50
nl	0.54	0.17	0.51	0.74	0.41	0.47
nno	-	-	-	-	-	
nob	-	-	-	-	-	
pl	0.66	0.50	0.42	0.94	0.32	0.57
pt	0.57	0.42	0.61	0.90	0.65	0.63
ro	0.49	0.26	0.48	0.90	0.61	0.55
ru	0.53	0.28	0.47	0.82	0.43	0.51
sk	0.54	0.44	0.59	0.89	0.51	0.59
sl	0.37	0.17	0.54	0.89	0.84	0.56
sme	-	-	-	-	-	
sv	0.49	0.32	0.49	0.82	0.82	0.59
tr	0.55	0.39	0.55	0.84	0.78	0.62
ug	0.00	0.00	0.00	0.22	0.00	0.05
uk	0.49	0.36	0.44	0.75	0.89	0.58
ur	0.50	0.05	0.18	0.85	0.13	0.34
vi	0.57	0.38	0.00	0.86	0.81	0.52
zh	0.01	0.00	0.00	0.22	0.78	0.20
moy	0.47	0.25	0.35	0.76	0.57	0.4859

TABLEAU .4. – Score OddOneOut par catégories et par langues avant alignement.

Lang.	Capitales	Villes anciennes	Religions	Pays	Élément Chimique	Moyenne
ar	0.00	0.00	0.00	0.01	0.12	0.03
bg	0.05	0.02	0.04	0.15	0.12	0.08
bxr	0.07	0.00	0.00	0.36	0.17	0.15
ca	0.03	0.02	0.04	0.19	0.10	0.07
cs	0.01	0.01	0.02	0.06	0.16	0.05
cu	0.00	0.00	0.00	0.08	0.00	0.02
da	0.02	0.00	0.00	0.02	0.23	0.06
de	0.02	0.03	0.03	0.09	0.13	0.06
el	0.04	0.02	0.04	0.12	0.24	0.09
en	0.05	0.02	0.02	0.13	0.25	0.09
es	0.07	0.04	0.03	0.19	0.17	0.10
et	0.01	0.01	0.09	0.01	0.17	0.06
eu	0.00	0.01	0.00	0.01	0.01	0.01
fa	0.01	0.01	0.01	0.08	0.15	0.05
fi	0.01	0.00	0.00	0.02	0.14	0.03
fr	0.04	0.02	0.01	0.12	0.21	0.08
ga	0.03	0.00	0.00	0.04	0.40	0.10
gl	0.03	0.02	0.01	0.10	0.17	0.07
got	0.00	0.00	0.00	0.02	0.00	0.01
he	0.03	0.02	0.00	0.02	0.20	0.05
hi	0.01	0.01	0.00	0.06	0.20	0.06
hr	0.01	0.00	0.00	0.01	0.07	0.02
hu	0.01	0.01	0.01	0.09	0.20	0.07
id	0.03	0.01	0.01	0.09	0.20	0.07
it	0.04	0.03	0.02	0.15	0.14	0.07
ja	0.06	0.01	0.01	0.30	0.43	0.16
kk	0.02	0.00	0.02	0.08	0.26	0.07
kmr	-	-	-	-	-	
ko	0.02	0.00	0.01	0.02	0.32	0.08
lv	0.02	0.01	0.04	0.04	0.20	0.06
nl	0.03	0.01	0.02	0.13	0.15	0.07
nno	-	-	-	-	-	
nob	-	-	-	-	-	
pl	0.02	0.00	0.02	0.13	0.02	0.04
pt	0.03	0.02	0.04	0.17	0.24	0.10
ro	0.01	0.00	0.02	0.04	0.06	0.03
ru	0.02	0.01	0.00	0.13	0.07	0.05
sk	0.02	0.01	0.02	0.05	0.22	0.06
sl	0.02	0.00	0.00	0.03	0.12	0.03
sme	-	-	-	-	-	
sv	0.01	0.01	0.01	0.03	0.26	0.06
tr	0.02	0.03	0.03	0.11	0.37	0.11
ug	0.00	0.00	0.00	0.13	0.01	0.03
uk	0.01	0.00	0.02	0.10	0.10	0.05
ur	0.01	0.01	0.00	0.04	0.08	0.03
vi	0.02	0.01	0.00	0.16	0.11	0.06
zh	0.04	0.00	0.01	0.31	0.41	0.15
moy	0.02	0.01	0.02	0.10	0.17	0.06

TABLEAU .5. – Score Topk par catégories et par langues avant alignement.

Lang.	Capitales	Villes anciennes	Religions	Pays	Élément Chimique	Moyenne
ar	0.60	0.35	0.30	0.93	0.79	0.59
bg	0.73	0.47	0.55	0.92	0.66	0.66
bxr	0.10	0.00	0.00	0.43	0.48	0.33
ca	0.59	0.13	0.31	0.68	0.32	0.40
cs	0.56	0.37	0.43	0.83	0.70	0.58
cu	0.01	0.00	0.00	0.08	0.00	0.02
da	0.51	0.26	0.29	0.77	0.87	0.54
de	0.71	0.55	0.51	0.86	0.48	0.62
el	0.52	0.15	0.25	0.87	0.64	0.49
en	0.78	0.21	0.46	0.97	0.50	0.58
es	0.70	0.27	0.48	0.89	0.50	0.57
et	0.53	0.24	0.60	0.68	0.72	0.55
eu	0.53	0.23	0.05	0.77	0.14	0.34
fa	0.51	0.24	0.44	0.85	0.72	0.55
fi	0.65	0.11	0.42	0.76	0.67	0.52
fr	0.61	0.34	0.41	0.69	0.50	0.51
ga	0.15	0.01	0.01	0.47	0.55	0.24
gl	0.56	0.33	0.44	0.76	0.63	0.54
got	0.00	0.00	0.00	0.00	0.00	0.00
he	0.51	0.26	0.45	0.77	0.80	0.56
hi	0.16	0.05	0.21	0.77	0.53	0.34
hr	0.53	0.42	0.56	0.83	0.56	0.58
hu	0.72	0.33	0.45	0.66	0.55	0.54
id	0.56	0.24	0.25	0.79	0.53	0.47
it	0.63	0.43	0.45	0.89	0.48	0.58
ja	0.22	0.04	0.00	0.27	0.57	0.22
kk	0.37	0.31	0.38	0.63	0.91	0.52
kmr	-	-	-	-	-	
ko	0.82	0.32	0.35	0.96	0.49	0.59
lv	0.35	0.11	0.52	0.86	0.55	0.48
nl	0.62	0.20	0.49	0.65	0.37	0.46
nno	-	-	-	-	-	
nob	-	-	-	-	-	
pl	0.63	0.52	0.40	0.83	0.27	0.53
pt	0.52	0.39	0.62	0.69	0.55	0.55
ro	0.53	0.27	0.46	0.79	0.51	0.51
ru	0.52	0.27	0.45	0.82	0.43	0.50
sk	0.56	0.46	0.61	0.85	0.44	0.58
sl	0.41	0.20	0.52	0.80	0.76	0.54
sme	-	-	-	-	-	
sv	0.76	0.47	0.57	0.82	0.80	0.68
tr	0.53	0.39	0.50	0.70	0.69	0.56
ug	0.00	0.00	0.00	0.21	0.00	0.05
uk	0.47	0.33	0.43	0.75	0.82	0.56
ur	0.41	0.05	0.14	0.74	0.13	0.29
vi	0.61	0.45	0.00	0.85	0.49	0.48
zh	0.01	0.00	0.00	0.14	0.66	0.16
moy	0.48	0.25	0.34	0.70	0.53	0.4652

TABLEAU .6. – Score OddOneOut par catégories et par langues après moyennage.

Lang.	Capitales	Villes anciennes	Religions	Pays	Élément Chimique	Moyenne
ar	0.01	0.00	0.00	0.03	0.12	0.03
bg	0.09	0.03	0.04	0.26	0.36	0.15
bxr	0.13	0.00	0.00	0.37	0.17	0.17
ca	0.11	0.02	0.06	0.19	0.07	0.09
cs	0.11	0.04	0.03	0.16	0.42	0.15
cu	0.01	0.00	0.00	0.07	0.00	0.02
da	0.16	0.03	0.02	0.29	0.53	0.20
de	0.11	0.08	0.08	0.18	0.38	0.17
el	0.02	0.02	0.04	0.06	0.22	0.07
en	0.13	0.05	0.06	0.34	0.53	0.22
es	0.09	0.05	0.03	0.17	0.19	0.11
et	0.15	0.00	0.11	0.19	0.30	0.15
eu	0.12	0.02	0.01	0.29	0.10	0.11
fa	0.01	0.01	0.01	0.06	0.13	0.04
fi	0.14	0.02	0.00	0.24	0.43	0.17
fr	0.12	0.04	0.04	0.16	0.34	0.14
ga	0.14	0.02	0.00	0.05	0.40	0.12
gl	0.13	0.04	0.08	0.21	0.27	0.14
got	0.00	0.00	0.00	0.02	0.00	0.01
he	0.02	0.02	0.00	0.01	0.20	0.05
hi	0.00	0.01	0.00	0.04	0.19	0.05
hr	0.15	0.02	0.04	0.11	0.16	0.09
hu	0.13	0.01	0.04	0.18	0.29	0.13
id	0.16	0.03	0.02	0.28	0.33	0.16
it	0.11	0.04	0.04	0.26	0.16	0.12
ja	0.05	0.01	0.00	0.30	0.43	0.16
kk	0.11	0.00	0.20	0.26	0.53	0.22
kmr	-	-	-	-	-	
ko	0.02	0.00	0.01	0.02	0.32	0.08
lv	0.07	0.01	0.04	0.09	0.22	0.09
nl	0.12	0.03	0.04	0.21	0.41	0.16
nno	-	-	-	-	-	
nob	-	-	-	-	-	
pl	0.08	0.04	0.05	0.20	0.02	0.08
pt	0.08	0.04	0.06	0.13	0.23	0.11
ro	0.14	0.02	0.05	0.29	0.07	0.12
ru	0.05	0.01	0.04	0.30	0.32	0.14
sk	0.13	0.02	0.05	0.24	0.36	0.16
sl	0.13	0.03	0.00	0.13	0.18	0.10
sme	-	-	-	-	-	
sv	0.14	0.04	0.09	0.25	0.51	0.21
tr	0.11	0.06	0.02	0.19	0.36	0.15
ug	0.00	0.00	0.00	0.13	0.01	0.03
uk	0.03	0.01	0.04	0.14	0.11	0.07
ur	0.02	0.02	0.00	0.07	0.07	0.04
vi	0.18	0.06	0.00	0.36	0.06	0.13
zh	0.04	0.00	0.00	0.25	0.38	0.13
moy	0.09	0.02	0.03	0.18	0.25	0.12

TABLEAU .7. – Score Topk par catégories et par langues après moyennage.

Lang.	nb hom.	Lang.	nb hom.
ar	14	id	624
bg	21	it	352
ca	497	ja	65
cs	586	ko	42
da	367	lv	109
de	583	nl	459
el	41	nno	295
en	-	nob	333
es	572	pl	117
et	127	pt	560
eu	205	ro	365
fa	7	ru	155
fi	259	sl	229
fr	868	sv	281
ga	170	tr	122
he	9	uk	16
hi	7	ur	9
hr	281	vi	61
hu	124	zh	96

TABLEAU .8. – Nombre d'homographes avec l'anglais pour chaque langue.

Lang	*Mono +c*	*Mono −c*	*Mono_delex*	*Multi +c*	*Multi −c*	*Multi_delex*	*Multi +c* ID	*Multi −c* ID	*Multi_delex* ID	*Multi +c* W_{22}	*Multi −c* W_{22}	*Multi_delex* W_{22}	*ZS +c*	*ZS −c*	*ZS_delex*	*ZS +c* W_{22}	*ZS −c* W_{22}	*ZS_delex* W_{22}
ar	70.31	69.61	65.12	71.20	72.16	66.81	70.22	71.37	68.34	71.47	71.11	67.88	42.77	33.34	43.85	39.05	38.44	37.20
bg	81.45	81.19	80.62	83.15	84.00	80.95	82.50	84.11	83.30	83.44	83.69	82.47	64.21	65.70	69.31	72.24	72.89	74.43
ca	76.95	76.32	73.71	79.46	80.22	73.17	77.92	79.43	78.31	79.06	79.69	75.02	73.09	73.77	72.48	70.83	71.57	71.32
cs	68.72	68.38	65.84	70.60	73.52	66.87	70.45	71.07	69.40	71.18	71.89	68.31	64.05	63.70	49.51	62.20	66.55	54.90
da	74.70	73.37	66.28	76.29	77.16	73.41	76.03	76.66	75.41	76.41	76.45	74.90	67.50	68.19	64.38	69.25	69.20	67.63
de	72.43	72.20	72.18	74.97	75.55	69.86	74.30	74.33	75.09	74.57	74.66	74.45	62.15	60.81	60.50	61.97	63.06	62.75
el	79.56	79.19	78.30	83.04	82.96	77.60	81.61	82.33	80.76	82.32	81.84	80.48	66.61	68.27	61.99	66.68	68.60	68.61
en	73.89	73.51	72.94	77.29	77.61	72.83	76.97	77.23	75.96	77.06	77.22	75.31	64.05	61.08	61.69	64.72	64.40	64.73
es	71.52	69.62	70.06	76.97	73.06	71.36	75.97	72.55	69.30	75.16	74.88	71.62	73.30	74.24	72.78	74.00	73.99	71.76
et	70.66	68.78	69.62	73.20	75.53	71.98	73.01	74.77	74.83	74.41	73.84	74.14	60.56	59.76	58.58	63.77	64.08	64.73
eu	62.86	60.71	58.17	62.80	65.32	56.12	61.46	62.53	60.53	64.20	63.66	60.52	38.65	39.74	38.30	36.82	39.45	37.58
fa	75.48	76.38	70.44	76.91	77.39	70.67	76.06	76.68	71.85	76.44	76.20	72.22	38.19	42.52	30.70	32.83	39.50	33.80
fi	62.36	61.70	59.42	64.53	61.28	60.36	63.66	59.88	58.58	60.91	60.48	58.50	53.24	53.78	53.19	54.81	54.35	53.69
fr	76.33	76.75	76.15	79.53	79.36	77.41	79.04	78.87	78.26	79.02	78.75	78.28	73.24	74.09	74.79	74.19	73.66	74.88
ga	67.09	65.57	64.73	66.38	69.03	61.45	65.00	67.08	67.03	67.52	66.99	64.74	40.70	39.25	44.77	45.03	45.96	43.34
he	72.09	69.47	69.56	73.44	74.60	67.97	71.21	72.49	69.62	73.92	71.92	69.20	51.63	51.25	49.44	53.02	54.71	53.64
hi	81.38	81.19	76.78	83.15	83.20	75.62	82.50	82.66	77.80	82.59	83.09	76.69	61.69	66.75	66.32	55.15	65.98	68.18
hr	70.98	71.73	70.03	74.94	76.25	72.55	74.63	75.92	74.83	75.12	75.12	74.22	68.96	69.45	67.32	68.95	68.88	68.28
hu	64.92	63.09	60.13	66.34	69.20	58.81	64.72	65.91	65.60	68.09	68.34	63.66	48.63	47.76	47.04	52.88	49.15	48.66
id	75.34	74.77	72.27	76.25	77.61	67.50	74.91	76.32	73.93	77.02	75.75	73.95	43.39	43.17	46.15	59.13	58.60	56.48
it	81.53	80.43	80.33	84.13	84.78	80.38	83.97	84.12	82.28	83.91	83.59	81.34	77.48	77.81	78.70	78.75	79.38	78.52
ja	87.59	87.10	76.68	88.36	88.32	74.28	87.35	87.28	76.05	88.37	87.41	76.84	9.37	9.16	10.02	15.63	19.22	33.38
ko	67.44	66.47	48.33	66.79	66.83	47.55	64.51	64.96	49.41	66.55	65.05	50.50	23.40	22.98	18.41	29.32	29.00	28.90
lv	61.67	61.89	59.59	65.53	67.74	59.05	63.92	66.48	65.31	66.09	66.47	64.88	50.10	50.62	45.93	49.45	50.23	49.14
nl	61.73	62.34	59.27	66.17	65.31	61.20	64.79	64.12	63.77	64.07	64.08	60.98	55.85	56.64	52.89	55.01	56.07	51.66
nno	73.78	72.79	72.89	77.92	75.41	75.81	77.18	75.07	71.24	74.07	73.88	68.39	73.64	73.97	71.90	71.89	71.17	67.27
nob	77.21	76.49	76.71	81.74	80.13	79.24	81.16	79.12	74.40	78.08	78.18	70.56	77.85	78.63	76.21	76.83	76.58	68.87
pl	80.62	79.85	77.64	80.31	82.82	75.72	79.86	81.27	80.52	81.12	81.00	78.59	71.07	70.24	70.95	72.44	71.75	71.71
pt	78.91	77.59	75.47	80.41	75.37	73.54	79.33	74.62	72.22	75.62	75.58	71.28	68.71	68.21	66.75	70.36	70.35	68.14
ro	65.96	67.40	65.84	72.07	74.40	68.33	72.47	73.00	70.65	72.67	72.06	69.82	58.07	57.91	54.72	60.36	60.56	58.16
ru	71.88	70.44	66.49	72.41	59.02	61.81	71.41	55.36	51.62	55.89	55.39	52.80	52.10	50.88	50.02	53.29	52.74	48.94
sl	71.89	70.64	70.48	73.32	73.75	70.46	73.24	72.59	71.03	72.69	72.47	71.00	66.01	65.69	62.66	66.35	67.00	65.06
sv	74.12	73.81	73.73	77.37	78.25	74.31	76.74	77.74	76.59	77.53	77.16	75.77	71.91	71.43	70.74	71.82	70.62	70.85
tr	51.22	51.47	47.62	53.64	54.83	45.56	50.64	52.37	50.46	53.46	53.26	49.45	26.65	29.73	32.85	35.60	35.50	31.50
uk	70.72	69.36	68.50	73.19	75.48	71.09	72.81	73.02	73.68	74.38	74.94	71.19	65.55	64.82	59.47	68.58	67.94	68.82
ur	75.89	76.50	73.21	77.98	78.34	72.77	78.00	77.86	74.59	78.02	77.97	73.62	45.85	56.83	67.66	55.44	63.37	69.56
vi	60.48	58.16	52.22	63.08	63.10	51.31	59.89	59.96	56.84	62.59	60.63	57.88	38.87	38.03	30.61	40.29	40.79	42.35
zh	65.91	64.70	60.91	68.56	68.83	55.73	66.95	67.27	62.85	69.17	66.36	63.04	10.72	11.45	19.21	14.24	13.80	22.81
moy	71.78	71.08	68.38	74.04	74.15	68.20	73.06	72.85	70.32	73.27	72.92	69.59	55.26	55.57	54.55	56.93	57.87	57.16

TABLEAU .9. – LAS pour chaque langue dans une configuration délexicalisée et lexicalisée pour de l'analyse syntaxique utilisant les POS et la morphologie de référence.

C. Annexes chapitre 4

l.	Corpus	Couv(l)	CouvUniq(l)	Nb mots	Nb mots uniq	Nb mots trouvés	Nb mots uniq. trouvés
ar	test	96.06	93.17	32128	8754	30861	8156
ar	train	96.26	94.16	22675	7248	21826	6825
bg	test	95.06	89.94	15724	5656	14948	5087
bg	train	95.00	89.71	20028	6315	19027	5665
bxr	test	70.39	44.21	10032	4053	7062	1792
ca	test	93.37	91.76	58017	9245	54172	8483
ca	train	93.61	93.12	20070	4896	18788	4559
cs	test	95.78	90.52	195879	37128	187620	33608
cs	train	96.45	94.44	20051	8149	19340	7696
da	test	94.53	86.47	10023	3259	9475	2818
da	train	94.73	87.13	20009	5291	18954	4610
de	test	95.80	94.47	16537	5173	15842	4887
de	train	96.24	93.27	20339	6780	19575	6324
el	test	96.60	93.17	10422	3208	10068	2989
el	train	96.99	92.60	20007	4715	19404	4366
en	test	95.46	89.90	36640	6405	34975	5758
en	train	95.96	92.87	20013	4602	19205	4274
es	test	97.92	94.38	65068	11865	63714	11198
es	train	97.70	94.36	20220	5836	19754	5507
et	test	91.16	80.34	12410	4787	11313	3846
et	train	90.43	79.51	20002	5876	18088	4672
eu	test	93.21	83.36	24374	8353	22719	6963
eu	train	92.90	83.64	20008	6607	18588	5526
fa	test	97.81	95.59	16122	3945	15769	3771
fa	train	97.73	94.59	20194	4900	19735	4635
fi	test	90.75	80.57	37381	13971	33925	11256
fi	train	90.71	82.86	20020	9270	18160	7681
fr	test	92.20	92.74	31668	6312	29198	5854
fr	train	92.11	92.38	20570	5563	18946	5139
ga	test	92.27	80.64	10138	3037	9354	2449
ga	train	91.90	80.93	20007	3073	18387	2487
he	test	92.16	89.48	15134	5115	13947	4577
he	train	91.78	91.01	24646	8098	22620	7370
hi	test	96.32	86.64	35430	5335	34126	4622
hi	train	96.53	88.61	20010	4381	19316	3882
hr	test	96.36	93.23	13228	5275	12746	4918
hr	train	95.64	91.93	20002	7309	19129	6719
hu	test	91.92	84.36	10448	4436	9604	3742

hu	train	93.82	87.75	20012	5918	18776	5193
id	test	96.21	93.84	11780	4027	11334	3779
id	train	95.71	93.58	20000	5610	19142	5250
it	test	94.86	95.21	11153	3380	10580	3218
it	train	95.20	94.73	21358	5258	20333	4981
ja	test	91.15	79.54	12615	3676	11498	2924
ja	train	90.80	78.92	20007	5170	18167	4080
kmr	test	90.08	78.42	10090	2734	9089	2144
ko	test	82.52	75.34	10926	7059	9016	5318
ko	train	81.84	73.50	20012	11070	16378	8136
lv	test	91.34	79.62	10528	3690	9616	2938
lv	train	91.71	81.53	20019	5696	18359	4644
nl	test	93.80	87.55	21566	5542	20228	4852
nl	train	95.07	89.41	20007	5640	19021	5043
nno	test	94.40	83.90	24773	5614	23386	4710
nno	train	95.07	87.03	10288	3138	9781	2731
nob	test	95.93	87.94	29966	6258	28745	5503
nob	train	96.08	91.01	9715	3202	9334	2914
pl	test	97.65	95.07	10906	4828	10650	4590
pl	train	97.34	94.31	20197	7344	19660	6926
pt	test	96.92	93.12	44580	9100	43205	8474
pt	train	96.82	93.68	21409	5427	20729	5084
ro	test	94.02	91.37	16324	5492	15348	5018
ro	train	94.23	91.43	20024	6428	18869	5877
ru	test	95.94	94.28	129027	30357	123784	28620
ru	train	96.58	94.88	20019	8112	19335	7697
sl	test	94.90	91.37	24077	7546	22848	6895
sl	train	95.31	91.00	20013	7066	19074	6430
sme	test	69.86	35.57	10010	3854	6993	1371
sv	test	95.22	87.11	20377	4810	19402	4190
sv	train	95.79	88.79	20008	5244	19166	4656
tr	test	91.90	85.95	10256	4862	9425	4179
tr	train	91.82	85.13	20598	6972	18914	5935
uk	test	94.07	89.32	12926	5223	12159	4665
uk	train	94.81	90.17	20004	4150	18965	3742
ur	test	94.58	84.13	14806	2949	14003	2481
ur	train	95.30	84.09	20018	3784	19077	3182
vi	test	78.03	44.91	11955	2543	9329	1142
vi	train	77.08	42.53	20017	2779	15430	1182
zh	test	90.57	79.90	12012	4055	10879	3240
zh	train	90.97	78.69	20032	5608	18223	4413

TABLEAU .10. – Taux de couvertures des plongements de mots pour chaque corpus pour chaque langue. *nb mots* est le nombre de tokens apparaissant dans le corpus, *nb mots uniq.* est le nombre de mots différents qui apparaissent dans le corpus, *nb mots trouvés* est le nombre de mots pour lesquels un plongement de mot existe, et *nb mots uniq. trouvés* est le nombre de mots différents pour lesquels un plongement de mot a été trouvé. *Couv.* est la couverture pour le nombre total de mots présents dans le corpus (nb mots trouvés*100/nb mots), et *Couv. uniq.* est la couverture en prenant en compte le nombre de mots différents qui apparaissent (nb mots uniq.*100/nb mots uniq)

Certaines langues, comme l'hébreu (he) ou le français (fr) ont plus de ~20 000 tokens dans le corpus d'entraînement (resp. 24 646 et 20 570 tokens) dans le tableau .10. Cela vient du fait que les mots multi-tokens sont séparés pour expliciter leurs composantes. Par exemple, en français (fr), le mot “au" peut être décomposé en deux mots : “à" et “le". Lors de l'entraînement du modèle, le système verra le plongement de mot du mot multi-tokens mais ne prédira rien pour ce dernier. Le plongement est uniquement utilisé pour les prédictions des mots l'encadrant. C'est pourquoi ils n'étaient pas comptés lors de l'équilibrage du nombre de tokens de chaque corpus. Cependant, nous les comptons pour calculer $couvEmbed(l)$ puisque leurs plongements sont utilisés par l'étiqueteur.

Identifiant	Forme de surface	Lemme	...
1	Il	il	...
2	parle	parler	...
3-4	au	_	...
3	à	à	...
4	le	le	...
5	poulpe	poulpe	...

Dans l'exemple ci-dessus, on compte 5 tokens pour la création de corpus (puisqu'il y aura 5 prédictions réalisées par l'étiqueteur), mais on compte 6 tokens pour le calcul de $couvEmbed(l)$ puisque 6 plongements seront utilisés par l'étiqueteur.

Lang.	couv(l)	matchTotal(l)
ar	24.24	17.32
bg	43.76	36.65
ca	66.00	46.07
cs	40.24	29.67
da	69.33	52.09
de	49.04	26.71
el	11.54	11.15
en	67.09	30.49
es	67.84	46.52
et	38.82	24.35
eu	31.37	19.44
fa	38.72	23.59
fi	30.46	22.31
fr	63.14	37.62
ga	46.97	19.95
he	10.24	10.16
hi	2.62	2.35
hr	47.68	34.01
hu	35.97	19.97
id	34.81	24.12
it	61.91	37.68
ja	17.66	12.96
ko	9.49	9.30
lv	44.19	29.55
nl	52.83	25.07
nno	69.83	54.25
nob	75.24	60.63
pl	38.47	26.42
pt	68.50	40.49
ro	41.73	25.13
ru	41.26	38.46
sl	53.85	32.97
sv	54.36	35.60
tr	30.31	19.24
uk	35.84	33.17
ur	23.07	10.68
vi	23.68	16.05
zh	42.32	27.10

TABLEAU .11. – Pourcentage de mots du test vu à l'entraînement, et pourcentage de mots du test dont les POS sont similaires à ceux de l'entraînement.

Lang.	L		L-c		SIGMA		SIGMA_W80		SIGMA_ID		SIGMA-c		SIGMA-c_W80		SIGMA-c_ID		LBAR		LBAR-c		LBAR_W80		LBAR-c_W80	
	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev	acc	stdev
ar	92.94	0.10	90.80	0.05	92.03	0.23	92.56	0.02	92.57	0.05	89.09	0.02	90.35	0.13	90.38	0.17	60.23	0.58	66.88	1.73	64.94	0.85	62.42	1.41
bg	95.73	0.08	92.25	0.09	95.09	0.04	95.66	0.11	95.77	0.02	90.84	0.05	91.78	0.04	92.00	0.02	78.16	0.38	79.65	0.08	76.40	0.88	77.80	0.11
ca	95.84	0.05	90.87	0.10	95.22	0.04	96.06	0.01	96.11	0.06	89.60	0.05	91.46	0.05	91.44	0.05	83.48	0.06	77.39	0.17	82.63	0.60	76.61	0.08
cs	95.12	0.07	90.71	0.05	94.02	0.01	94.81	0.01	94.89	0.05	88.97	0.08	90.41	0.07	90.64	0.07	81.11	0.09	79.03	0.10	76.97	2.27	76.06	1.48
da	92.41	0.05	87.50	0.13	91.34	0.10	92.59	0.05	93.12	0.08	87.27	0.14	88.86	0.25	89.25	0.16	83.93	0.05	80.88	0.26	84.43	0.16	80.89	0.02
de	89.16	0.04	81.91	0.57	88.94	0.15	89.13	0.05	89.53	0.09	82.06	0.02	82.46	0.22	83.01	0.05	53.84	1.59	53.39	0.54	48.63	0.26	49.37	0.98
el	95.47	0.05	91.56	0.17	95.28	0.04	95.48	0.05	95.31	0.01	90.96	0.08	91.62	0.22	91.38	0.16	47.61	0.45	60.02	0.21	52.34	1.69	60.79	0.02
en	89.54	0.12	85.81	0.16	88.53	0.00	89.14	0.12	89.24	0.14	84.89	0.02	85.89	0.02	86.01	0.15	27.96	0.11	25.78	0.78	29.90	6.20	27.51	1.38
es	93.22	0.02	90.11	0.10	94.00	0.01	92.13	0.07	92.16	0.03	90.83	0.01	89.88	0.07	89.66	0.04	88.69	0.12	85.81	0.44	87.11	0.16	85.62	0.14
et	90.02	0.13	81.50	0.10	86.49	0.16	88.77	0.04	88.78	0.12	78.15	0.24	81.80	0.19	81.64	0.17	63.55	0.08	62.24	0.31	64.74	0.11	62.82	0.35
eu	90.62	0.10	84.73	0.33	88.16	0.11	89.92	0.05	89.75	0.02	81.16	0.11	83.94	0.08	83.94	0.05	55.73	0.11	54.87	0.20	54.36	0.38	50.77	0.61
fa	94.83	0.02	92.47	0.14	93.78	0.00	94.53	0.08	94.47	0.19	90.44	0.03	92.01	0.05	91.99	0.03	56.24	1.00	68.72	0.70	62.24	4.23	62.30	1.30
fi	84.90	0.10	77.45	0.15	82.25	0.02	83.12	0.07	83.25	0.12	75.53	0.17	77.14	0.03	77.12	0.22	67.08	0.03	65.89	0.15	66.38	0.42	64.89	0.45
fr	93.45	0.02	88.67	0.09	92.97	0.02	93.35	0.06	93.59	0.04	88.63	0.15	89.40	0.03	89.49	0.02	78.29	1.07	77.12	0.90	75.81	0.44	69.78	3.72
ga	89.11	0.09	82.84	0.12	87.02	0.03	89.09	0.23	88.78	0.12	80.23	0.14	83.30	0.17	83.16	0.11	45.28	0.32	44.20	0.36	41.03	2.33	39.20	0.48
he	94.45	0.05	92.57	0.03	93.84	0.03	94.15	0.02	94.10	0.05	91.05	0.12	92.25	0.07	92.00	0.12	52.30	0.80	55.16	0.70	53.95	2.64	55.31	0.80
hi	93.62	0.08	92.30	0.01	92.75	0.15	92.92	0.08	93.05	0.00	90.39	0.20	91.19	0.05	91.42	0.02	35.67	2.82	60.55	0.23	39.12	0.00	58.09	4.14
hr	94.30	0.07	89.42	0.19	93.38	0.12	94.28	0.10	94.22	0.07	88.00	0.01	89.66	0.09	89.78	0.25	83.07	0.16	79.01	0.59	81.79	0.81	77.98	0.58
hu	92.90	0.01	84.85	0.18	90.72	0.27	92.37	0.07	92.68	0.15	82.92	0.28	85.69	0.16	85.72	0.16	60.67	1.02	59.59	0.83	60.88	1.45	62.59	2.09
id	90.64	0.04	85.61	0.05	89.30	0.11	90.76	0.18	90.76	0.17	83.00	0.01	86.01	0.08	85.92	0.03	59.04	0.54	57.54	0.30	57.50	0.11	52.88	1.70
it	94.30	0.03	91.30	0.15	93.94	0.02	94.50	0.03	94.79	0.01	90.50	0.13	91.47	0.02	91.67	0.02	77.09	0.45	78.26	0.35	76.56	0.92	77.04	0.29
ja	92.41	0.17	88.05	0.19	91.78	0.15	92.44	0.04	92.18	0.08	87.88	0.02	88.18	0.18	88.12	0.07	33.28	1.64	45.72	2.52	37.66	1.20	40.25	0.42
ko	93.38	0.03	83.64	0.20	91.74	0.12	92.23	0.11	91.94	0.11	80.60	0.07	82.23	0.11	82.53	0.16	50.23	0.37	53.06	0.12	47.99	0.81	50.87	0.33
lv	89.44	0.15	82.25	0.06	85.84	0.12	88.69	0.12	88.84	0.21	78.53	0.05	82.69	0.27	82.70	0.11	62.75	0.10	61.31	0.12	64.46	0.23	60.21	0.09
nl	87.78	0.17	81.78	0.02	86.64	0.05	86.08	0.12	86.34	0.26	80.64	0.12	81.16	0.10	81.78	0.21	61.04	0.10	58.47	0.71	56.91	0.38	58.09	0.51
nno	90.43	0.09	84.11	0.20	90.00	0.02	83.62	0.27	83.78	1.07	84.32	0.09	78.33	0.18	70.25	3.98	84.77	0.03	79.45	0.50	79.82	0.27	75.15	0.05
nob	91.92	0.03	85.72	0.07	91.41	0.03	86.49	0.32	84.84	1.77	85.96	0.05	81.54	0.22	70.87	4.16	88.65	0.20	83.33	0.05	84.84	0.20	79.95	0.23
pl	95.19	0.00	89.12	0.28	93.46	0.11	94.41	0.03	94.55	0.08	87.09	0.10	89.27	0.12	89.16	0.05	78.82	0.16	76.08	0.08	74.30	0.22	72.78	1.76
pt	93.25	0.03	89.19	0.12	93.00	0.12	92.39	0.00	92.39	0.05	88.87	0.07	88.75	0.01	88.70	0.09	76.11	0.31	77.44	0.37	75.80	0.16	76.92	0.94
ro	94.11	0.16	89.60	0.01	92.23	0.06	93.81	0.10	93.69	0.10	86.43	0.04	89.42	0.11	89.27	0.22	64.28	0.77	65.48	0.09	65.26	0.05	65.22	1.17
ru	94.91	0.03	89.69	0.07	94.23	0.02	92.32	0.20	92.08	0.13	88.47	0.02	86.73	0.11	86.16	0.10	85.17	0.12	82.30	0.17	84.76	0.41	82.10	0.46
sl	92.50	0.05	85.06	0.09	89.48	0.01	90.13	0.30	90.79	0.01	81.64	0.20	82.95	0.14	82.88	0.20	69.10	0.29	65.31	0.07	67.29	0.40	64.44	0.62
sv	92.03	0.17	87.83	0.04	91.44	0.01	92.73	0.16	93.37	0.09	87.17	0.06	88.69	0.02	89.35	0.19	79.70	0.49	76.56	0.38	79.69	0.26	76.53	0.24
tr	90.44	0.22	80.93	0.02	86.70	0.16	88.98	0.08	88.81	0.01	77.80	0.18	81.14	0.24	81.37	0.04	56.34	0.35	55.69	0.12	54.05	0.76	49.70	1.72
uk	91.93	0.02	84.09	0.10	91.44	0.00	92.29	0.10	92.27	0.05	85.18	0.03	86.56	0.04	86.32	0.27	80.07	0.08	78.12	0.15	78.91	0.07	76.94	0.24
ur	90.56	0.01	88.78	0.29	89.53	0.10	90.51	0.12	90.44	0.08	87.84	0.14	88.70	0.03	89.02	0.09	38.97	0.21	56.66	0.51	52.82	0.03	64.95	0.33
vi	86.05	0.19	77.43	0.15	84.52	0.04	85.42	0.14	85.43	0.13	76.84	0.16	77.72	0.13	77.75	0.02	39.25	0.50	39.13	2.42	37.68	0.45	30.03	0.03
zh	87.98	0.03	82.51	0.08	88.16	0.29	88.88	0.16	88.75	0.03	83.98	0.19	84.75	0.07	84.72	0.15	50.51	0.49	49.74	0.06	49.66	0.77	52.14	0.36
Moy.	92.02	0.08	86.71	0.13	90.81	0.08	91.23	0.10	91.25	0.15	85.36	0.10	86.46	0.11	86.02	0.32	64.16	0.47	65.15	0.48	63.94	0.88	63.34	0.83

TABLEAU .12. – Exactitude de la prédiction de POS de chaque système pour chaque langue

Models\Tests	ar	bg	bxr	ca	cs	da	de	el	en	es	et	eu	fa	fi	fr	ga	he	hi	hr	hu
ar	92.94	27.19	19.53	18.42	31.23	20.87	17.47	33.36	13.86	18.96	28.31	33.58	45.07	31.49	15.05	19.42	41.66	14.12	28.49	28.88
bg	41.00	95.74	33.09	24.64	45.69	26.02	20.38	39.13	19.81	22.70	35.27	34.53	42.98	36.59	16.84	19.62	39.50	33.68	41.87	35.63
ca	28.18	45.82	25.04	95.84	46.98	40.54	28.44	34.28	32.90	82.20	42.10	45.79	25.24	43.63	67.80	34.65	29.49	17.79	45.13	44.78
cs	45.92	47.84	45.98	48.33	95.12	42.75	34.23	44.74	38.67	47.78	48.38	48.85	46.89	50.34	44.13	37.97	45.82	30.54	71.69	48.34
da	38.81	51.02	26.50	43.86	52.38	92.41	38.43	42.89	26.78	42.67	48.06	47.89	41.30	49.19	37.34	38.00	40.22	28.70	48.73	49.99
de	12.19	22.30	26.12	28.38	33.55	37.54	89.17	17.23	27.16	27.09	34.05	34.64	10.70	32.44	22.43	23.05	20.23	15.98	27.47	33.12
el	48.11	51.63	25.80	11.90	24.07	16.06	13.60	95.47	13.61	13.15	19.28	17.84	52.06	17.70	12.85	13.25	40.71	40.35	15.16	19.42
en	21.45	38.14	26.19	38.91	42.18	41.62	31.18	32.70	89.54	38.54	37.72	37.43	21.64	37.54	37.53	35.62	30.92	25.10	39.20	49.80
es	29.16	36.55	25.85	79.20	44.85	38.87	28.49	31.09	34.71	93.23	40.49	45.12	25.97	42.98	58.03	36.89	28.74	26.11	42.16	44.05
et	42.27	41.76	48.40	39.21	49.67	40.79	26.92	37.69	34.45	39.48	90.02	48.15	47.72	61.36	35.06	37.06	39.43	28.09	47.02	47.91
eu	36.68	43.31	33.26	39.62	49.77	39.48	28.84	37.18	35.43	39.38	46.98	90.62	39.96	49.38	38.18	36.47	40.02	28.41	47.00	48.03
fa	53.00	37.07	44.21	27.81	40.53	31.06	32.92	33.81	29.22	29.45	39.94	40.06	94.84	40.98	30.07	35.71	32.94	29.81	37.60	37.26
fi	33.32	41.41	25.38	39.03	53.35	39.98	24.34	38.59	21.51	40.62	62.67	52.73	35.72	84.90	30.23	34.20	38.51	28.83	50.00	51.24
fr	38.99	42.01	18.63	68.53	46.99	40.95	33.09	33.93	47.18	69.79	41.49	44.94	29.59	42.62	93.46	35.09	45.41	17.68	40.99	40.41
ga	32.21	33.22	29.78	43.25	43.17	39.13	28.77	27.03	37.50	42.60	36.67	41.27	28.92	39.68	37.65	89.11	34.36	23.64	40.28	42.27
he	46.88	43.89	43.27	30.46	45.55	32.36	24.54	39.81	20.08	27.73	40.06	44.19	39.01	43.58	21.45	25.97	94.45	33.55	39.81	40.23
hi	27.17	34.85	25.47	27.67	37.69	26.98	23.51	32.56	19.61	26.12	30.54	36.62	32.64	31.90	17.18	24.20	31.09	93.63	32.16	35.92
hr	42.79	51.77	26.26	42.45	73.57	43.52	28.89	39.23	29.67	41.38	46.04	46.52	45.47	48.03	35.03	37.00	35.77	35.46	94.30	47.14
hu	35.80	39.92	25.96	39.32	50.68	40.94	27.89	29.61	35.40	38.03	47.47	49.45	42.60	48.42	33.20	36.79	37.91	34.50	47.40	92.90
id	18.80	29.42	24.60	30.87	41.10	31.63	22.18	23.27	21.45	28.36	41.07	43.15	17.91	42.62	18.72	31.05	27.09	21.92	39.85	40.09
it	37.61	38.86	25.82	67.22	47.39	44.19	31.39	30.73	37.17	65.93	40.93	45.09	35.34	41.80	51.35	38.53	38.68	28.32	42.57	43.36
ja	19.62	23.26	38.04	18.97	16.79	16.50	17.44	16.22	15.91	16.35	16.41	19.69	20.59	17.15	14.66	12.76	22.91	17.38	14.31	19.12
ko	29.06	34.41	39.09	29.45	41.48	32.35	32.74	31.19	30.08	29.86	42.29	42.48	33.09	43.09	31.55	36.66	31.56	20.91	38.37	38.93
lv	43.69	52.93	40.43	39.05	56.27	41.60	28.18	44.33	20.04	36.85	47.48	47.74	48.53	50.01	28.15	34.32	42.57	39.44	52.60	46.22
nl	34.34	45.14	26.07	47.88	51.02	49.36	47.42	36.23	42.69	45.77	42.00	47.31	33.50	46.52	40.21	34.48	34.89	33.45	48.93	46.16
nno	35.56	41.95	38.39	39.81	48.88	71.61	39.33	29.74	32.23	42.55	41.80	44.36	39.24	47.36	27.87	38.75	35.57	29.80	49.65	45.71
nob	17.87	32.47	25.92	45.77	50.23	80.12	37.26	26.86	38.23	46.47	44.96	46.91	17.85	48.97	36.00	36.81	31.74	24.02	49.92	46.78
pl	46.10	48.14	44.45	44.21	69.58	39.78	28.61	40.98	36.67	43.62	46.69	48.14	46.91	47.81	41.18	38.87	45.51	31.52	66.20	46.34
pt	38.99	44.09	28.01	64.44	46.77	42.62	30.34	40.10	33.84	69.51	42.82	47.49	33.86	45.23	53.20	36.20	40.97	25.62	41.01	51.92
ro	45.34	48.10	37.59	52.18	50.94	40.88	31.75	41.53	34.98	51.58	45.35	48.61	44.74	46.62	50.12	38.21	43.60	30.85	46.84	48.09
ru	43.41	71.63	31.01	28.51	48.67	30.48	22.86	44.12	21.83	27.54	41.59	42.51	46.24	44.99	20.55	29.28	41.70	32.84	44.90	43.52
sl	48.00	53.53	37.19	38.89	73.63	42.26	24.38	41.90	24.82	38.58	44.82	46.18	54.94	47.50	34.21	35.52	40.73	40.94	76.33	46.88
sv	35.61	44.09	27.27	44.64	51.50	74.39	39.77	38.13	39.30	42.13	46.96	47.30	35.06	48.62	34.53	38.22	39.50	31.73	46.87	48.32
tr	36.88	40.97	43.23	36.00	45.59	36.73	31.64	39.25	30.53	37.26	47.12	48.95	45.37	49.24	34.08	36.94	28.77	29.31	42.31	47.47
uk	41.72	64.98	37.74	31.41	49.27	35.76	30.37	39.41	30.95	33.58	43.97	43.97	43.29	46.70	32.35	32.41	38.45	27.69	43.45	39.91
ur	36.91	38.53	25.59	26.20	34.45	24.05	19.67	33.88	17.77	22.55	26.86	32.74	43.20	30.89	16.58	23.97	30.56	37.12	30.88	34.56
vi	25.07	33.80	38.66	28.12	36.43	33.13	36.03	29.77	29.79	30.43	39.57	39.44	22.51	40.27	31.78	28.57	29.38	18.07	32.29	32.85
zh	19.54	25.08	38.29	10.13	15.71	13.23	13.03	21.02	12.17	10.89	15.33	15.39	19.12	15.07	11.13	10.72	26.88	20.64	11.53	14.24
max	53.00	71.63	45.98	79.20	73.63	80.12	47.42	44.74	47.18	82.20	62.67	52.73	54.94	61.36	67.80	38.87	45.82	40.94	76.33	51.92
argmax	fa	ru	cs	es	sl	nob	nl	cs	fr	ca	fi	fi	sl	et	ca	pl	cs	sl	sl	pt

TABLEAU .13. – Matrice des décodages de tous les corpus de tests par toutes les expériences *Mono +c* (partie 1).

Models\Tests	id	it	ja	kmr	ko	lv	nl	nno	nob	pl	pt	ro	ru	sl	sme	sv	tr	uk	ur	vi	zh
ar	22.77	20.87	5.64	16.48	29.45	36.52	20.74	16.87	17.41	30.93	22.19	27.59	29.45	21.64	14.13	19.07	31.57	32.00	28.19	24.49	16.21
bg	40.78	19.17	19.86	17.30	40.10	40.54	25.76	22.47	24.42	43.82	26.73	27.70	76.50	30.00	17.24	24.54	37.26	69.23	32.41	22.97	33.36
ca	51.57	49.36	12.74	25.68	30.39	47.10	39.18	39.71	40.77	46.37	62.42	53.93	48.11	35.50	20.98	39.11	42.35	41.23	20.39	26.63	28.61
cs	54.09	46.97	18.50	30.89	52.91	57.58	40.95	39.38	40.69	71.72	49.68	52.70	54.85	57.84	23.99	40.74	50.03	51.52	34.85	34.58	30.32
da	52.18	40.72	25.52	27.58	50.48	52.82	53.68	75.07	81.20	50.92	44.41	44.34	55.71	37.58	27.55	73.60	46.39	49.74	31.79	36.09	39.53
de	38.79	24.00	4.31	23.50	18.98	38.61	42.93	34.16	34.41	30.68	30.56	27.22	27.57	21.87	21.22	31.49	32.64	24.87	20.06	22.10	20.25
el	17.22	12.91	29.26	14.33	44.32	27.24	16.68	12.62	13.31	22.28	16.06	17.01	54.64	12.76	13.77	12.77	20.80	51.73	42.36	24.33	35.36
en	45.55	36.75	18.02	21.83	25.20	45.97	41.35	35.64	38.64	39.34	46.75	38.58	42.07	27.86	21.86	33.66	35.69	36.59	20.97	27.76	27.34
es	52.55	66.46	22.80	30.83	28.55	46.84	40.49	35.59	37.95	44.48	68.98	55.75	39.09	34.34	19.44	36.00	42.40	38.27	28.59	25.27	25.64
et	51.87	37.68	28.33	36.54	49.77	53.80	40.85	38.05	38.17	49.18	40.56	43.84	48.05	37.60	41.98	38.48	48.71	46.62	33.32	36.65	39.17
eu	49.01	36.84	30.70	34.23	47.45	52.24	40.63	36.24	36.67	48.74	42.40	45.17	48.53	34.74	33.05	37.14	53.71	46.03	31.92	34.73	42.84
fa	35.49	32.41	14.00	37.31	48.66	43.13	29.92	30.54	29.23	41.18	30.43	36.80	41.04	27.83	36.84	33.41	48.01	40.28	34.98	43.05	23.78
fi	56.12	35.52	29.88	32.17	45.03	54.50	38.22	36.05	37.57	49.29	43.78	45.06	42.84	39.17	39.15	38.12	52.25	46.82	33.48	29.88	41.05
fr	54.36	62.56	19.09	27.52	34.50	47.14	43.45	38.93	40.05	46.02	61.07	55.91	46.14	33.85	19.14	37.11	42.73	41.54	21.29	34.85	35.03
ga	46.35	44.66	22.04	42.40	30.33	44.96	38.24	38.30	37.27	43.82	45.00	50.44	37.29	29.97	33.78	37.17	40.04	36.31	22.75	36.67	34.88
he	47.94	29.60	24.55	23.67	44.26	46.34	31.21	27.69	29.19	40.91	37.36	37.25	50.17	29.41	25.72	29.96	42.05	47.60	34.88	27.15	41.98
hi	37.93	23.51	12.78	16.88	29.89	37.56	27.76	23.97	25.21	35.20	31.98	29.69	38.70	24.35	17.17	24.96	36.11	38.24	36.66	23.18	21.74
hr	50.78	38.81	38.43	25.49	28.84	55.32	44.28	41.53	41.94	69.44	43.16	46.75	55.74	66.54	21.83	40.17	47.04	50.05	36.44	31.87	40.87
hu	49.99	30.54	26.41	32.91	42.03	49.60	40.02	37.79	37.75	45.80	45.93	41.00	45.26	36.07	34.31	39.15	50.01	45.69	33.97	31.80	33.84
id	90.64	33.54	6.15	23.62	28.93	45.07	31.39	28.73	29.81	37.31	34.50	35.03	34.20	26.68	18.31	28.94	42.10	32.70	24.01	32.97	22.31
it	54.86	94.30	25.19	25.65	30.08	47.68	42.70	39.98	42.21	51.85	60.90	54.93	45.62	34.92	20.16	39.28	40.88	41.69	28.58	24.14	37.77
ja	30.69	13.68	92.41	15.21	35.22	19.06	18.67	13.21	14.97	16.80	20.21	14.32	22.05	9.46	16.79	9.97	17.09	26.06	18.22	16.02	44.16
ko	38.07	31.64	24.75	36.75	93.38	45.11	30.61	30.77	29.27	41.76	31.49	39.69	40.72	27.87	37.61	34.64	48.07	35.63	19.16	46.18	38.94
lv	47.49	39.08	11.79	36.69	46.33	89.45	36.51	36.09	37.97	57.34	39.09	46.44	56.29	41.84	36.50	41.23	48.05	53.64	42.19	29.48	19.84
nl	52.64	48.70	16.56	29.95	31.00	48.74	87.78	47.20	47.83	48.31	44.68	46.09	48.24	37.33	31.00	47.11	45.22	41.90	32.35	30.56	30.61
nno	45.90	39.68	31.19	37.75	35.80	49.40	41.60	90.43	84.07	46.14	40.60	43.16	46.48	34.32	37.63	67.23	43.36	42.29	30.25	35.39	39.95
nob	54.60	40.07	28.19	28.75	29.01	51.83	54.61	83.02	91.92	47.02	45.77	43.33	36.01	36.35	27.93	71.12	43.64	34.59	22.01	31.81	39.27
pl	50.66	37.61	39.99	36.26	47.54	55.60	40.69	38.30	37.99	95.19	45.60	48.70	56.81	50.83	35.76	39.52	49.68	53.70	39.15	34.48	36.72
pt	52.98	61.96	24.35	28.95	38.91	50.11	40.31	39.07	41.68	47.74	93.25	52.23	49.33	32.29	26.31	39.08	42.77	46.12	24.86	28.19	35.29
ro	50.67	54.03	30.28	36.62	44.43	50.19	41.84	38.43	38.95	51.42	60.54	94.12	53.21	36.55	24.74	36.23	45.86	48.71	32.81	36.21	42.10
ru	46.80	23.82	25.01	19.48	49.81	48.22	29.49	28.18	28.02	47.42	33.48	36.15	94.91	35.27	20.43	31.43	45.15	75.12	34.69	32.17	37.19
sl	52.28	37.46	36.10	26.30	39.33	57.85	38.56	41.01	40.39	67.71	39.97	46.40	56.20	92.50	26.30	40.72	47.37	52.66	44.36	29.62	40.45
sv	50.19	38.12	31.03	30.13	40.08	51.87	52.79	68.28	72.51	50.97	42.14	44.93	46.61	37.44	29.48	92.04	45.80	43.50	32.47	32.95	42.73
tr	45.32	34.12	28.33	38.12	52.60	50.44	36.82	34.75	34.60	45.66	40.46	43.48	49.68	32.14	35.22	37.72	90.44	44.74	29.12	37.71	37.97
uk	44.27	31.29	28.49	27.91	43.99	50.96	36.48	33.91	33.88	50.49	35.75	42.20	75.40	36.16	23.23	37.66	44.91	91.93	26.33	36.61	41.44
ur	35.53	16.73	27.10	17.12	32.10	36.18	26.73	21.31	22.56	30.06	29.56	28.58	37.32	21.70	17.18	25.18	29.26	38.18	90.56	29.12	33.09
vi	36.49	29.56	21.27	34.99	35.11	46.19	29.29	29.48	29.32	38.44	29.67	36.47	40.95	25.91	33.85	30.88	45.93	37.15	14.04	86.05	34.64
zh	15.81	10.04	28.99	14.40	27.32	19.48	15.15	12.16	12.71	15.81	13.31	13.73	29.68	9.41	13.51	10.97	17.06	26.46	22.62	20.53	87.98
max	56.12	66.46	39.99	38.12	52.91	57.85	54.61	83.02	84.07	71.72	68.98	55.91	76.50	66.54	39.15	73.60	53.71	75.12	44.36	46.18	44.16
argmax	fi	es	pl	tr	cs	sl	nob	nob	nno	cs	es	fr	bg	hr	fi	da	eu	ru	sl	ko	ja

TABLEAU .14. – Matrice des décodages de tous les corpus de tests par toutes les expériences *Mono +c* (partie 2).

Models\Tests	ar	bg	bxr	ca	cs	da	de	el	en	es	et	eu	fa	fi	fr	ga	he	hi	hr	hu
ar	90.80	35.64	26.85	26.50	38.81	27.24	17.83	39.40	14.16	27.20	33.15	37.87	53.81	37.27	20.66	28.82	50.03	26.30	35.14	35.70
bg	40.19	92.25	26.13	30.62	56.35	38.30	27.53	49.78	21.72	30.53	45.61	45.12	43.60	49.21	20.79	30.11	41.34	29.25	53.21	46.60
ca	28.62	38.83	25.30	90.87	47.72	44.01	29.87	32.12	27.35	77.41	41.53	43.90	30.96	41.42	64.73	29.31	27.84	21.17	47.29	41.92
cs	41.70	50.57	28.66	47.08	90.71	43.34	34.30	51.11	37.90	47.63	49.39	47.96	44.44	50.12	39.67	38.73	43.67	30.16	70.71	51.00
da	34.89	48.17	25.95	42.06	55.35	87.51	38.35	40.26	23.32	41.81	46.78	46.66	35.67	49.69	40.67	36.14	36.76	27.29	49.49	50.61
de	20.32	33.43	29.12	30.65	40.27	39.65	81.91	27.92	23.18	28.86	36.51	36.90	19.72	37.96	22.72	24.85	26.14	20.50	36.54	37.79
el	35.65	47.36	24.54	25.08	48.72	29.47	17.58	91.57	13.34	22.40	38.59	39.71	33.30	42.57	14.55	25.07	43.78	28.30	44.80	41.91
en	15.42	26.79	24.74	31.84	34.51	39.63	33.16	24.80	85.82	30.07	29.91	32.12	16.23	28.48	37.00	26.98	23.39	15.93	29.72	41.21
es	30.98	41.59	23.54	68.79	45.21	40.44	26.24	38.19	31.40	90.11	39.74	43.48	31.36	40.81	50.08	30.91	33.26	22.68	42.16	42.53
et	40.37	47.39	39.41	35.68	49.04	43.62	35.25	41.05	33.37	36.13	81.51	47.78	41.33	59.77	33.08	33.23	40.96	28.77	45.15	47.28
eu	38.43	43.48	33.57	39.81	49.23	39.53	29.76	40.71	31.89	39.08	45.40	84.74	38.79	48.49	38.24	33.76	40.36	27.50	45.97	46.90
fa	56.20	43.09	23.35	31.10	44.57	33.89	32.38	37.84	25.28	32.98	42.87	41.91	92.47	44.83	27.29	35.51	36.33	33.83	41.45	40.90
fi	39.34	46.69	29.45	35.99	55.20	41.24	33.66	43.48	17.74	36.33	57.98	49.25	41.33	77.45	25.65	31.22	39.78	30.87	48.04	48.36
fr	34.26	36.48	20.89	65.19	43.03	42.75	35.05	34.72	40.21	69.25	39.21	40.75	32.67	39.21	88.68	30.36	39.80	23.14	37.30	38.80
ga	36.88	38.62	27.37	40.73	43.03	40.47	28.99	32.76	34.10	41.91	34.39	38.58	39.32	38.01	40.12	82.85	42.60	27.16	38.60	39.21
he	48.97	44.64	29.27	29.68	45.26	33.90	22.90	38.60	18.31	27.00	41.39	44.91	43.22	46.16	20.92	28.26	92.57	32.37	40.36	42.62
hi	42.05	47.54	24.88	29.58	45.73	33.25	23.55	39.17	19.72	28.01	37.04	42.50	48.12	39.41	21.80	31.78	42.98	92.30	40.43	43.61
hr	43.22	49.20	26.70	44.46	71.60	41.75	27.09	46.01	23.15	43.81	46.94	46.71	44.53	49.26	30.74	38.06	44.26	28.62	89.43	49.57
hu	41.59	46.04	24.57	34.93	53.16	43.27	34.21	42.70	26.62	35.80	46.56	46.87	43.96	50.26	27.73	34.48	41.11	29.78	47.31	84.85
id	35.41	42.41	24.64	34.22	47.53	35.56	23.37	40.29	19.76	32.00	42.13	47.68	38.61	43.67	21.33	30.93	40.29	28.71	45.10	43.46
it	33.65	40.53	25.10	65.15	48.92	44.40	36.76	34.67	42.37	66.87	41.83	44.86	30.52	43.12	38.20	33.34	32.93	22.39	45.65	43.01
ja	11.18	16.11	18.75	13.31	13.25	11.91	12.10	11.05	9.16	10.98	12.79	15.23	11.58	14.42	8.49	8.75	14.16	17.41	12.27	15.75
ko	38.59	40.00	48.90	30.14	43.57	33.52	32.91	38.51	29.31	31.18	43.38	45.74	44.28	44.60	31.45	34.72	39.85	28.13	39.60	40.42
lv	36.95	52.15	25.28	39.72	56.58	44.51	26.43	46.37	17.87	39.85	49.45	48.09	40.67	51.60	32.19	36.87	44.17	29.72	51.75	46.53
nl	31.85	46.22	29.15	48.79	57.39	52.88	45.92	39.20	38.22	47.50	42.25	46.34	32.22	44.97	36.27	31.64	36.94	27.79	49.53	48.13
nno	35.82	42.87	34.51	42.80	53.20	66.80	39.27	37.10	29.47	44.57	42.85	42.85	35.61	44.98	32.52	36.01	35.54	26.95	49.49	45.59
nob	33.49	43.12	27.63	44.12	54.01	76.30	43.65	36.49	35.94	47.37	42.38	44.25	34.83	45.03	38.10	35.52	37.75	26.56	48.02	45.04
pl	38.63	48.74	39.36	43.09	67.28	41.09	28.96	43.88	36.31	44.18	48.27	46.68	46.42	49.99	37.99	39.21	36.80	29.58	64.17	48.74
pt	30.15	40.89	25.27	61.48	48.46	42.31	35.26	38.26	22.62	70.71	40.53	44.86	30.24	43.09	56.70	35.45	34.89	24.09	42.50	50.49
ro	40.89	44.00	26.88	51.63	51.07	43.56	37.78	42.61	31.90	54.16	46.03	46.54	39.55	47.35	55.50	37.11	40.67	29.45	46.54	50.76
ru	42.51	72.72	26.03	34.51	57.55	38.79	27.38	53.11	24.14	33.31	45.59	47.82	44.90	49.99	26.41	32.46	42.49	29.64	51.56	47.58
sl	41.74	50.88	30.21	43.09	70.47	45.68	28.76	45.55	17.52	46.90	48.89	46.03	43.32	49.47	36.15	35.26	40.39	29.55	71.90	50.96
sv	39.36	47.65	28.27	43.39	55.49	68.40	40.99	42.49	38.46	46.49	48.40	46.92	42.60	49.67	41.35	40.10	40.77	27.85	48.57	46.80
tr	33.56	41.39	40.94	32.08	44.66	39.31	35.88	41.37	24.58	34.73	46.12	47.56	41.50	48.62	33.87	31.87	28.51	25.53	40.09	49.00
uk	41.24	68.96	27.29	27.96	56.73	36.10	25.41	47.80	25.86	32.08	47.47	44.02	44.53	49.69	20.65	30.28	41.82	29.98	52.41	44.43
ur	37.31	35.92	25.70	25.35	37.20	25.17	17.02	33.98	16.37	22.53	30.73	37.05	46.10	36.51	15.72	25.16	35.85	42.52	34.70	39.89
vi	27.81	32.92	37.22	28.05	38.62	34.94	34.91	29.30	28.12	32.26	38.41	39.32	25.44	38.77	31.22	28.41	35.03	18.26	33.10	34.53
zh	23.57	27.55	41.57	11.09	23.91	14.32	13.66	26.65	9.77	10.89	23.18	26.91	18.75	25.08	9.10	12.75	29.59	20.49	21.16	23.20
max	56.20	72.72	41.57	68.79	71.60	76.30	45.92	53.11	42.37	77.41	57.98	49.25	53.81	59.77	64.73	40.10	50.03	42.52	71.90	51.00
argmax	fa	ru	zh	es	hr	nob	nl	ru	it	ca	fi	fi	ar	et	ca	sv	ar	ur	sl	cs

TABLEAU .15. – Matrice des décodages de tous les corpus de tests par toutes les expériences *Mono* $-c$ (partie 1).

Models\Tests	id	it	ja	kmr	ko	lv	nl	nno	nob	pl	pt	ro	ru	sl	sme	sv	tr	uk	ur	vi	zh
ar	29.69	31.68	10.31	27.06	43.00	42.01	23.87	25.30	25.26	35.58	32.02	35.37	40.53	26.63	20.66	27.46	37.55	38.25	33.30	25.70	16.92
bg	46.57	30.42	30.72	17.87	47.86	50.88	37.49	36.70	36.48	52.64	33.42	41.47	75.17	40.27	17.64	41.87	46.83	67.22	33.10	32.13	41.86
ca	45.63	54.24	12.37	20.76	37.86	46.73	38.06	40.16	41.70	51.51	56.89	48.52	42.58	36.34	18.04	40.54	39.31	40.45	21.98	23.13	20.01
cs	49.50	45.14	17.68	26.34	45.92	56.33	42.69	40.39	39.60	71.02	44.49	50.17	56.61	57.01	19.93	44.81	50.00	53.70	33.13	31.69	27.76
da	49.83	34.36	11.94	21.22	46.55	51.68	52.81	70.47	76.08	51.13	37.82	43.54	52.01	41.09	19.92	69.65	46.44	47.39	27.13	27.39	19.99
de	40.60	24.05	11.61	17.95	30.65	40.44	39.85	35.85	35.64	37.82	35.34	29.74	39.37	27.80	18.41	37.47	37.34	36.13	23.53	21.61	18.40
el	32.41	23.66	10.66	14.99	42.47	47.89	25.14	28.14	27.68	45.18	26.64	40.34	52.47	35.74	16.09	31.28	42.55	52.62	26.03	23.23	21.96
en	36.85	32.42	13.95	17.49	22.36	36.95	34.69	34.11	36.78	32.90	36.56	28.38	32.23	22.18	17.15	33.29	29.39	28.63	17.84	20.49	20.79
es	51.04	63.68	10.72	22.03	32.74	45.56	36.83	35.54	37.74	43.86	65.32	51.39	46.83	34.31	18.00	36.03	39.73	42.42	26.57	23.77	16.10
et	41.77	36.86	30.99	32.44	54.57	54.29	39.46	36.72	37.69	48.49	35.06	45.46	52.87	38.38	42.40	42.98	49.10	51.37	32.70	33.02	37.13
eu	45.02	38.30	30.30	23.19	52.11	51.71	38.12	36.41	36.75	49.42	41.78	44.72	49.64	34.67	22.52	38.79	51.32	45.61	30.75	33.73	39.72
fa	37.87	35.80	17.52	24.29	51.33	48.18	32.88	32.39	31.53	45.14	34.07	39.45	46.41	32.35	22.81	36.26	49.69	46.55	31.86	33.55	19.96
fi	48.20	36.23	25.89	24.10	51.95	54.72	37.66	36.20	36.43	50.50	38.71	44.40	51.36	40.53	29.82	40.32	50.07	49.91	32.59	30.31	34.41
fr	46.10	62.73	4.69	20.18	32.29	43.12	42.13	39.55	41.29	43.49	65.87	50.30	40.58	31.52	16.29	37.35	38.71	37.71	25.58	28.80	13.29
ga	40.90	37.96	22.96	35.34	39.36	43.37	39.57	41.51	39.78	43.54	46.34	49.98	42.06	31.10	23.01	40.66	40.17	40.26	28.02	34.88	29.64
he	45.66	28.67	18.75	21.76	47.19	46.37	30.58	29.95	31.67	40.73	34.20	38.13	47.70	30.95	21.29	32.37	42.97	46.74	35.77	24.46	30.82
hi	41.81	24.83	20.07	15.95	42.80	44.43	31.98	32.40	32.40	41.14	32.56	34.33	44.77	30.66	16.81	33.91	43.77	46.62	55.80	30.93	24.56
hr	48.74	40.27	32.35	21.50	45.21	53.76	40.19	39.66	39.63	65.41	43.78	47.21	54.36	62.58	18.91	42.62	46.09	51.43	29.72	30.94	34.89
hu	43.89	24.21	27.03	17.45	43.07	51.09	41.14	38.88	38.86	48.83	45.63	39.52	50.09	37.97	20.39	42.78	50.15	48.61	30.81	26.52	36.25
id	85.61	42.90	10.10	19.26	45.57	48.28	29.54	31.98	33.07	42.21	39.93	43.73	47.02	34.87	16.64	33.93	47.87	43.90	32.83	31.55	24.18
it	53.03	91.30	13.77	23.37	34.34	45.51	45.19	40.41	42.92	50.78	63.79	49.73	44.87	36.57	18.18	41.43	43.11	40.50	25.78	22.68	21.70
ja	21.15	8.50	88.05	11.18	12.98	16.57	14.22	9.25	10.30	10.00	13.36	10.50	15.87	7.73	12.26	7.91	14.19	16.84	17.69	13.66	43.69
ko	41.07	31.91	28.80	36.80	83.64	48.20	32.03	32.68	31.33	46.08	33.29	40.11	46.37	31.29	35.97	36.82	50.01	44.53	31.62	43.85	40.51
lv	46.60	37.15	22.83	23.15	52.90	82.25	36.54	38.71	40.99	56.95	45.16	50.34	54.72	43.14	23.74	44.83	48.70	53.31	29.64	25.34	33.95
nl	51.18	45.02	23.91	21.86	41.92	48.91	81.78	49.48	51.24	51.89	43.10	47.78	51.70	39.30	20.39	51.88	45.29	45.03	29.41	27.72	34.38
nno	42.96	40.72	24.94	28.33	39.27	47.88	41.83	84.12	75.91	48.66	41.60	42.79	45.58	38.84	24.86	62.66	42.58	42.38	26.66	36.14	33.42
nob	47.66	41.66	25.05	20.45	37.04	48.81	52.40	77.58	85.73	49.29	42.52	43.02	48.29	39.00	18.51	67.61	42.90	43.74	27.17	33.98	29.17
pl	48.62	41.17	35.61	26.11	49.17	55.25	39.62	39.39	38.80	89.13	44.73	48.15	54.88	48.63	27.96	42.02	51.34	52.11	31.12	33.78	41.07
pt	48.14	64.56	13.12	19.62	34.65	46.01	39.99	38.93	40.53	48.11	89.19	51.84	46.47	35.21	18.74	38.77	41.04	41.84	24.03	23.51	22.29
ro	49.63	48.32	24.27	25.81	43.17	49.35	38.90	39.80	39.21	50.42	60.73	89.60	47.21	39.66	20.13	40.77	46.74	45.04	28.82	31.11	35.22
ru	48.28	29.08	26.65	17.09	50.95	51.83	39.14	35.68	35.93	55.32	37.72	46.71	89.70	41.52	20.07	40.03	49.69	69.74	32.05	31.60	39.07
sl	45.59	39.53	27.46	24.18	44.94	56.37	38.68	41.45	42.34	64.12	42.66	48.88	56.55	85.07	24.18	47.45	47.42	52.72	30.71	28.99	36.64
sv	44.01	50.04	31.20	21.79	48.36	52.75	51.62	62.57	65.22	53.75	39.89	46.18	51.73	40.73	24.84	87.83	47.15	48.84	28.78	31.77	39.07
tr	41.32	34.56	32.56	31.51	53.86	51.94	34.39	36.42	35.88	43.53	38.36	45.01	49.51	32.30	32.48	38.79	80.93	44.14	30.01	30.86	38.40
uk	43.90	27.29	26.51	19.76	46.99	51.72	31.91	32.94	32.30	54.04	32.64	40.88	74.19	42.57	23.10	37.35	45.67	84.09	32.26	31.28	34.94
ur	37.17	17.59	29.09	14.88	40.05	39.53	27.64	21.61	22.68	34.41	27.55	29.80	35.62	24.81	17.33	22.91	35.60	36.03	88.78	27.33	30.60
vi	38.95	37.13	23.94	35.33	38.17	43.55	31.77	31.54	31.54	39.81	31.71	36.78	40.34	27.25	37.55	32.41	41.99	37.65	18.40	77.43	37.57
zh	20.95	12.20	43.56	14.35	35.82	29.96	15.77	12.53	12.62	23.31	13.21	19.41	30.74	15.85	19.22	15.34	26.61	32.40	16.79	20.07	82.51
max	53.03	64.56	43.56	35.34	54.57	56.37	52.81	77.58	76.08	71.02	65.87	51.84	75.17	62.58	37.55	69.65	51.34	69.74	55.80	43.85	43.69
argmax	it	pt	zh	ga	et	sl	da	nob	da	cs	fr	pt	bg	hr	vi	da	pl	ru	hi	ko	ja

TABLEAU .16. – Matrice des décodages de tous les corpus de tests par toutes les expériences *Mono* $-c$ (partie 2).

Lang.	POS			Morpho			UAS			LAS		
	UDPIPE	*Mono +c*	Δ	UDPIPE	*Mono +c*	Δ	UDPIPE	*Mono +c*	Δ	UDPIPE	*Mono +c*	Δ
bg	97.72	94.12	-3.60	95.55	88.53	-7.02	88.82	79.59	-9.23	84.92	72.98	-11.94
ca	98.00	93.83	-4.17	97.20	91.72	-5.48	88.69	75.45	-13.24	85.53	69.87	-15.66
da	95.28	89.77	-5.51	94.37	87.66	-6.71	78.91	66.25	-12.66	75.28	60.23	-15.05
el	95.35	92.94	-2.41	89.89	83.87	-6.02	84.31	76.53	-7.78	80.67	71.84	-8.83
et	87.60	85.66	-1.94	81.14	79.16	-1.98	68.65	62.77	-5.88	60.01	52.73	-7.28
eu	92.33	87.68	-4.65	87.25	75.72	-11.53	75.59	64.59	-11.00	70.45	56.83	-13.62
fa	96.02	92.41	-3.61	96.09	91.84	-4.25	84.18	72.27	-11.91	80.33	66.33	-14.00
ga	88.86	84.75	-4.11	76.27	74.17	-2.10	73.10	71.23	-1.87	62.87	59.58	-3.29
he	80.87	92.22	11.35	77.57	86.98	9.41	62.06	74.50	12.44	57.86	67.39	9.53
hr	95.88	92.28	-3.60	84.34	72.92	-11.42	83.73	71.15	-12.58	77.73	62.92	-14.81
hu	90.80	89.56	-1.24	70.59	84.99	14.40	72.36	67.80	-4.56	66.54	60.41	-6.13
id	93.43	88.78	-4.65	99.52	99.33	-0.19	81.67	73.31	-8.36	75.47	65.34	-10.13
ko	94.22	91.77	-2.45	99.34	99.53	0.19	66.64	69.99	3.35	60.30	62.96	2.66
lv	88.40	86.31	-2.09	82.02	70.66	-11.36	68.38	61.34	-7.04	61.80	52.95	-8.85
nno	96.54	87.10	-9.44	95.02	79.94	-15.08	85.86	63.80	-22.06	82.74	56.33	-26.41
nob	96.83	88.78	-8.05	95.25	79.36	-15.89	86.62	66.22	-20.40	83.89	59.12	-24.77
pl	95.43	92.53	-2.90	83.46	59.22	-24.24	86.31	79.40	-6.91	80.21	71.44	-8.77
ro	96.62	91.39	-5.23	96.05	88.27	-7.78	85.74	73.80	-11.94	80.32	65.76	-14.56
uk	87.33	88.72	1.39	71.00	54.55	-16.45	69.28	69.04	-0.24	61.09	60.59	-0.50
ur	92.13	87.82	-4.31	80.31	76.17	-4.14	83.86	77.58	-6.28	77.09	68.69	-8.40
vi	75.29	82.54	7.25	83.93	99.39	15.46	44.99	54.73	9.74	39.97	44.79	4.82
zh	83.47	85.40	1.93	88.28	98.03	9.75	61.81	57.89	-3.92	57.89	49.42	-8.47
moy.	91.75	89.38	-2.37	87.47	82.82	-4.66	76.43	69.51	-6.92	71.04	61.75	-9.29

TABLEAU .17. – Comparaison de notre système *Mono +c* avec UDPipe.

D. Annexes chapitre 5

D.1. Comparaison avec l'état de l'art

Les scores des résultats sur la tâche de tagparsing que nous avons présentés jusqu'ici peuvent sembler très bas. Dans cette section, nous allons chercher à mesurer la perte subie par les conditions extrêmes de nos expériences (20 000 tokens à l'entraînement seulement) par rapport à un analyseur de l'état de l'art, UDPipe (STRAKA et STRAKOVÁ 2017). Ces résultats nous permettront d'estimer quels résultats nous pourrions espérer obtenir si les conditions d'entraînement étaient plus favorables.

Lors de la campagne d'évaluation CoNLL 2017 (ZEMAN, POPEL et al. 2017), les organisateurs proposaient une baseline. Cette baseline utilise UDPipe (STRAKA et STRAKOVÁ 2017), et est considérée comme un solide système de référence. Cela permettra de se faire une idée d'où se place notre meilleur système, *Mono +c*, faisant la prédiction des POS, des traits morphologiques et l'analyse syntaxique du texte, par rapport à UDPipe. Afin d'obtenir des résultats qui soient le plus comparables possibles, nous avons gardé uniquement les langues n'ayant qu'un corpus arboré de test. En effet, pour nos résultats, nous concaténons l'ensemble des corpus arborés de test disponibles. En conservant uniquement les langues n'ayant qu'un corpus arboré de test, nous obtenons des corpus de test identiques. Les résultats de UDPipe sont ceux issus de la campagne d'évaluation CoNLL 2017, et n'ont pas été ré-entraînés. Les résultats sont visibles en annexe dans le tableau .17.

Il est important de rappeler que même lors d'un apprentissage en condition monolingue, chaque modèle de chaque langue ne dispose que de 20 000 tokens pour

s'entraîner, ce qui est bien inférieur à ce qui est utilisé classiquement, notamment avec UDPipe. Quant à lui, UDPipe prédit également la segmentation en phrases et en tokens, ce qui lui donne un désavantage que nous n'avons pas, puisque nous utilisons les segmentations de référence. Les résultats ne sont donc pas tout à fait comparables, et il est important de garder ces différences à l'esprit lors de l'analyse des résultats, leur comparaison étant limitée.

Les résultats de notre système sont inférieurs de 2.37 points et 4.66 d'exactitude en moyenne pour la prédiction des POS et des traits morphologiques respectivement par rapport à UDPipe. La tâche de prédiction de l'arbre syntaxique de la phrase est une tâche plus difficile, et les résultats sont plus inférieurs encore sur la tâche de prédiction de l'analyse syntaxique : -6.92 et -9.29 points de UAS et de LAS respectivement.

Sur les 22 langues pour lesquelles nos jeux de test sont comparables avec les résultats de UDPipe à campagne d'évaluation CoNLL 2017, 5 de nos modèles obtiennent de meilleurs résultats que UDPipe sur une ou plusieurs tâches : l'hébreu (he), le hongrois (hu), le coréen (ko), l'ukrainien (uk), le vietnamien (vi) et le chinois (zh). L'hébreu (he) et le vietnamien (vi) obtiennent systématiquement de meilleurs résultats avec notre système. En dehors du hongrois (hu) et du vietnamien (vi) [1], ces langues utilisent des alphabets autres que l'alphabet latin. Il est possible que UDPipe ne gère pas de manière optimale les langues utilisant d'autres caractères que ceux issus du latin. Ce système faisant des prédictions de manière séquentielle, si les prédictions sont incorrectes lors de la segmentation en mots, cela impactera toute la chaîne de prédiction qui suit, expliquant potentiellement une partie de ces résultats. C'est d'ailleurs le cas du chinois (zh) et du vietnamien (vi), pour lesquels UDPipe n'obtient qu'un score de 89.55 pour le chinois (zh) et de 84.26 pour le vietnamien (vi) pour la segmentation en tokens, alors qu'il obtient des scores supérieurs à 99% pour la plupart des autres langues. Pour l'hébreu (he), cette langue est reconnue comme étant particulièrement difficile à traiter par les systèmes séquentiels (TSARFATY et al. 2019). UDPipe s'en sort par conséquent moins bien que notre système qui prédit l'analyse syntaxique en même temps que les POS.

Il est évident que notre système *Mono +c* obtient de bien moins bons résultats que l'état de l'art, mais les conditions extrêmes dans lesquelles on place notre système expliquent en partie ces résultats. Ces derniers ne sont cependant pas aberrants, et restent dans l'ordre du raisonnable.

1. Le vietnamien utilise cependant beaucoup de diacritiques, le différenciant fortement des autres langues utilisant l'alphabet latin.

Lang	*Mono +c*		*Mono −c*		*Multi +c*	*Multi −c*	*Multi +c* ID	*Multi −c* ID	*Multi +c* W_{22}	*Multi −c* W_{22}	*ZS +c*	*ZS −c*	*ZS +c* W_{22}	*ZS −c* W_{22}
	acc	std	acc	std	acc	acc	acc	acc	acc	acc	acc	acc	acc	acc
ar	83.25	0.16	78.36	0.66	70.42	74.22	75.20	65.36	82.54	69.07	21.70	28.27	27.53	32.76
bg	88.53	0.10	73.45	0.72	87.57	72.25	80.25	75.20	89.05	68.16	32.77	32.70	29.95	31.49
ca	91.72	0.44	82.77	0.27	84.41	84.15	83.84	85.06	72.73	75.61	62.71	70.22	64.60	70.87
cs	61.30	0.44	48.99	1.35	60.48	45.45	61.71	51.22	56.99	50.48	26.02	23.27	29.23	27.30
da	87.66	0.30	76.72	0.52	84.98	75.53	88.50	79.51	83.90	78.39	47.84	43.03	47.83	41.02
de	74.73	0.31	67.53	0.08	73.89	69.27	67.26	62.26	73.70	69.79	39.35	38.25	29.63	25.65
el	83.87	0.13	75.94	0.24	76.91	72.90	84.81	78.22	84.53	65.87	27.67	26.61	27.82	27.91
en	82.64	0.40	75.84	0.78	83.62	76.19	85.05	69.46	81.12	74.74	33.92	36.00	34.47	36.16
es	89.16	0.12	80.50	1.01	89.74	84.25	85.37	82.54	88.70	81.62	77.71	70.63	79.96	71.73
et	79.16	0.50	63.73	0.12	76.86	58.93	80.08	65.29	78.71	64.23	41.90	43.95	41.96	43.21
eu	75.72	0.05	65.66	0.70	71.85	59.05	74.13	62.21	73.75	58.18	27.09	29.87	36.30	37.63
fa	91.84	0.28	87.32	0.24	91.59	85.56	92.02	87.37	91.33	86.58	21.49	22.69	26.92	29.52
fi	77.48	0.05	65.55	0.14	73.97	56.58	70.31	58.96	74.36	57.37	38.13	41.04	36.24	42.16
fr	79.85	0.34	74.06	0.23	79.86	74.67	79.64	68.77	66.69	74.40	48.42	53.33	48.64	52.80
ga	74.17	0.44	61.39	1.07	74.37	52.19	75.67	57.47	73.81	63.92	28.57	29.50	28.28	29.56
he	86.98	0.46	80.30	0.39	86.39	74.55	85.74	74.07	85.93	79.14	34.18	34.83	35.18	34.09
hi	79.66	0.66	75.93	1.33	67.32	56.68	78.33	65.14	78.13	70.46	34.74	38.46	30.34	42.22
hr	72.92	0.48	60.90	1.36	68.94	50.57	75.22	63.00	73.64	62.33	36.64	31.12	42.25	38.28
hu	84.99	0.59	69.60	0.28	84.06	61.37	85.92	71.22	64.98	66.27	21.77	24.12	27.78	33.43
id	99.33	0.01	99.04	0.02	97.80	98.51	99.25	98.29	99.14	98.90	59.88	60.59	71.98	90.37
it	91.74	0.27	84.95	0.44	91.20	68.30	91.99	83.49	90.77	71.61	57.19	52.99	61.19	45.70
ja	99.90	0.00	99.72	0.02	99.75	97.49	99.90	99.76	99.94	99.74	99.70	97.36	99.24	98.81
ko	99.53	0.01	99.18	0.06	99.08	98.91	99.37	99.35	99.53	99.41	62.26	61.60	45.05	64.41
lv	70.66	0.34	57.05	0.12	68.01	52.39	67.69	54.02	66.29	57.43	35.77	34.38	35.48	34.15
nl	71.86	0.95	64.78	0.21	68.74	62.99	64.70	48.96	64.14	58.75	31.07	32.62	33.29	32.22
nno	79.94	0.04	64.78	0.55	79.85	66.62	58.10	44.92	40.73	39.84	65.96	56.17	39.75	39.86
nob	79.36	0.60	67.31	0.15	76.05	65.02	56.94	46.22	40.70	40.03	58.11	46.06	40.24	40.15
pl	59.22	0.25	49.31	0.11	57.95	48.04	54.46	49.35	53.63	49.18	29.32	31.38	31.98	36.32
pt	80.53	3.37	78.19	0.06	74.79	73.17	55.25	48.06	55.77	53.65	53.83	53.45	54.97	55.16
ro	88.27	0.41	74.56	0.23	80.57	73.66	89.70	50.57	88.37	75.83	18.07	16.71	18.56	18.07
ru	56.77	0.20	48.98	0.03	54.32	45.40	44.35	33.87	41.57	36.68	33.12	31.29	31.32	29.57
sl	74.35	0.77	59.36	0.18	70.30	55.89	72.65	57.86	59.45	58.35	36.85	32.33	37.14	34.72
sv	80.29	0.34	61.30	2.00	81.89	67.52	86.95	75.37	76.13	58.98	35.54	32.86	42.75	39.95
tr	77.92	0.49	63.91	0.72	61.63	60.34	77.59	62.01	76.87	60.97	28.49	28.96	34.11	35.72
uk	54.55	0.08	43.56	0.65	55.07	48.58	52.52	49.81	55.90	41.63	27.63	26.23	31.04	30.40
ur	76.17	0.02	72.75	0.13	76.66	72.69	76.54	70.23	75.98	67.49	16.20	20.73	40.06	37.49
vi	99.39	0.01	99.15	0.02	99.21	97.29	98.49	98.96	99.28	98.10	60.38	89.28	93.22	93.68
zh	98.03	0.09	96.25	0.04	98.07	96.00	96.25	94.17	98.08	96.17	93.02	92.55	93.32	90.99
moy.	81.14	0.38	72.33	0.45	78.37	69.29	77.68	68.09	75.18	67.88	42.24	42.51	43.67	44.62

TABLEAU .18. – Exactitude de la prédiction des traits morphologiques avec les tagparsers pour toutes les langues, pour toutes les expériences.

Lang	*Mono +c*		*Mono −c*		*Multi +c*	*Multi −c*	*Multi +c* ID	*Multi −c* ID	*Multi +c* W_{22}	*Multi −c* W_{22}	*ZS +c*	*ZS −c*	*ZS +c* W_{22}	*ZS −c* W_{22}
	acc	std	acc	std	acc	acc	acc	acc	acc	acc	acc	acc	acc	acc
ar	90.88	0.05	87.38	0.19	78.21	85.55	82.22	76.32	90.36	80.58	55.61	67.03	55.32	64.99
bg	94.12	0.09	88.59	0.22	94.03	88.45	86.47	89.40	94.80	83.66	74.21	78.57	72.41	78.68
ca	93.83	0.40	86.91	0.00	86.63	87.31	85.97	88.24	73.32	78.45	67.88	76.75	70.84	76.31
cs	93.19	0.24	86.62	0.29	92.96	81.02	93.30	86.90	88.85	86.84	70.73	66.22	82.65	78.09
da	89.77	0.26	84.03	0.88	89.64	84.83	89.59	85.22	84.68	84.77	83.03	78.21	83.52	69.68
de	87.39	0.16	78.56	0.09	86.62	79.36	77.25	70.01	86.52	79.72	52.64	55.49	47.49	51.16
el	92.94	0.09	87.20	0.09	86.32	84.41	93.63	88.33	93.29	77.19	46.02	58.80	51.98	61.23
en	87.67	0.05	82.91	0.15	86.96	80.25	87.50	71.91	83.75	78.16	35.46	30.36	29.39	26.27
es	90.59	0.11	85.96	0.54	92.35	89.15	88.50	87.78	91.74	87.25	89.06	84.31	89.13	83.35
et	85.66	0.20	75.46	0.05	83.88	69.98	86.18	76.58	85.53	75.83	65.36	60.95	63.38	62.82
eu	87.68	0.01	81.74	0.27	85.53	76.35	87.55	79.76	87.01	75.15	52.41	52.12	53.88	48.14
fa	92.41	0.29	89.22	0.02	92.19	86.91	92.76	88.29	92.19	88.03	56.00	55.45	53.58	58.38
fi	82.19	0.09	74.25	0.18	79.93	68.90	76.99	70.82	80.21	69.58	68.66	64.50	60.65	60.26
fr	90.75	0.20	85.78	0.21	91.69	86.52	91.88	80.60	76.92	86.98	68.09	76.15	64.28	74.50
ga	84.75	0.34	75.60	1.24	84.81	62.98	85.89	70.25	84.26	76.15	39.43	42.70	47.52	42.84
he	92.22	0.27	88.92	0.16	92.53	85.43	91.38	84.01	92.48	89.34	51.32	54.33	54.29	50.99
hi	91.03	0.28	89.19	0.48	80.95	72.58	91.29	80.68	91.12	86.82	45.13	57.92	45.30	63.31
hr	92.28	0.30	84.91	0.51	88.71	76.24	92.98	85.97	92.34	85.80	71.94	66.37	82.34	77.43
hu	89.56	0.64	80.81	0.24	89.29	73.40	89.84	81.79	71.73	77.67	59.93	58.51	51.36	60.16
id	88.78	0.16	82.70	0.09	79.31	79.58	87.64	73.06	87.52	80.59	46.80	48.58	53.12	55.04
it	91.86	0.54	88.17	0.57	92.27	70.62	92.45	87.17	91.60	74.48	68.77	66.79	78.76	73.19
ja	90.35	0.35	84.64	0.24	89.14	58.48	90.04	83.36	90.34	83.24	32.65	41.74	37.19	41.05
ko	91.77	0.29	81.33	0.40	84.29	77.13	88.19	78.28	90.71	78.87	48.33	54.76	46.26	52.86
lv	86.31	0.02	78.25	0.00	83.83	70.93	82.80	73.92	80.87	77.87	62.63	55.45	61.65	61.49
nl	85.55	0.13	78.62	0.06	84.42	77.65	84.82	64.70	84.24	76.69	59.68	61.73	64.61	61.75
nno	87.10	0.13	79.22	0.23	88.31	79.87	84.50	73.06	78.63	75.93	83.56	77.50	74.98	73.43
nob	88.78	0.59	81.05	0.37	90.72	82.56	86.69	80.70	84.49	79.48	87.68	67.21	84.75	77.87
pl	92.53	0.48	85.00	0.14	91.76	84.29	86.33	85.40	85.44	85.50	77.80	76.59	73.50	76.18
pt	91.56	0.09	86.34	0.39	91.41	86.63	90.85	68.25	91.01	83.52	70.04	79.02	72.65	79.02
ro	91.39	0.28	84.87	0.18	84.52	83.77	92.32	61.52	90.77	85.11	64.30	52.47	68.45	60.69
ru	92.86	0.27	85.61	0.09	88.82	78.35	90.42	67.42	90.38	84.16	76.39	64.83	85.12	72.66
sl	88.98	0.24	79.62	0.01	87.18	77.05	87.45	76.97	74.39	76.96	65.81	63.71	67.06	65.05
sv	89.47	0.15	82.14	1.56	89.26	82.87	90.78	85.05	81.53	67.90	79.34	75.18	80.60	76.42
tr	86.54	0.52	76.81	0.29	68.42	74.31	85.76	75.98	84.79	75.19	55.80	51.63	57.79	57.17
uk	88.72	0.22	76.25	0.70	90.89	82.57	85.40	82.72	91.05	68.87	79.55	75.85	76.71	75.21
ur	87.82	0.13	85.88	0.11	88.34	85.53	88.30	84.45	88.35	79.96	40.29	52.34	54.34	54.88
vi	82.54	0.28	74.78	0.73	81.31	61.63	76.28	71.02	81.34	66.75	39.50	36.60	38.89	28.17
zh	85.40	0.42	77.48	0.08	86.31	76.67	79.16	68.86	86.44	76.18	48.83	48.82	51.52	45.51
moy.	89.40	0.25	82.71	0.32	86.94	78.69	87.40	78.55	86.18	79.61	61.60	61.46	62.82	62.53

TABLEAU .19. – Exactitude de la prédiction des POS avec les tagparsers pour toutes les langues, pour toutes les expériences.

Lang	*Mono +c*		*Mono −c*		*Multi +c*	*Multi −c*	*Multi +c* ID	*Multi −c* ID	*Multi +c* W_{22}	*Multi −c* W_{22}	*ZS +c*	*ZS −c*	*ZS +c* W_{22}	*ZS −c* W
	acc	std	acc	std	acc	acc	acc	acc	acc	acc	acc	acc	acc	acc
ar	75.39	0.64	71.49	0.34	59.14	68.60	65.71	58.51	73.75	62.66	36.05	45.93	38.46	48.98
bg	79.59	0.44	75.12	0.13	80.50	74.00	69.92	75.11	82.45	66.31	55.79	62.13	58.08	66.38
ca	75.45	0.53	69.71	0.05	66.37	72.32	65.39	72.52	54.61	60.52	52.08	63.82	55.78	65.61
cs	71.10	0.68	65.13	0.22	73.31	58.71	73.92	66.48	66.31	67.47	51.01	46.98	62.71	58.52
da	66.25	0.21	62.77	0.23	68.73	65.88	67.97	66.97	62.10	65.59	62.25	58.94	63.33	48.30
de	66.41	0.89	58.92	0.18	66.72	60.49	54.49	49.05	66.38	61.52	44.87	46.09	38.69	44.15
el	76.53	0.28	73.91	0.07	68.78	70.60	79.76	76.81	79.60	59.12	37.42	53.20	50.53	55.74
en	69.43	0.05	64.08	0.32	70.24	62.17	71.09	53.24	64.85	58.75	31.28	29.69	24.85	19.46
es	71.10	0.22	67.51	0.28	75.51	74.03	69.15	72.86	75.79	71.41	72.92	69.57	73.19	69.20
et	62.77	0.46	56.66	1.59	63.15	50.32	66.95	58.96	65.44	57.68	47.85	46.05	45.75	47.26
eu	64.59	0.35	59.49	0.16	60.81	49.52	64.49	54.53	63.71	48.64	30.15	28.90	27.94	24.57
fa	72.27	0.20	71.05	0.01	70.91	64.24	72.05	68.25	70.53	68.38	28.99	24.61	27.27	30.12
fi	61.05	0.35	56.24	0.07	60.58	50.56	54.74	51.76	59.73	50.88	49.26	49.82	42.25	43.60
fr	71.88	0.39	67.47	0.20	74.11	69.35	75.44	61.63	57.70	71.15	53.18	62.12	49.29	58.38
ga	71.23	0.07	64.22	0.94	70.63	49.85	71.35	55.74	69.46	63.77	31.35	33.84	44.90	42.31
he	74.50	0.55	71.94	0.24	74.77	66.98	72.44	64.13	75.56	72.50	46.61	50.79	47.85	47.47
hi	81.16	0.60	80.15	0.88	65.57	55.35	81.38	65.45	81.41	76.59	34.95	38.83	36.56	46.50
hr	71.15	0.32	64.88	0.29	68.97	53.90	75.85	67.08	73.62	67.79	49.48	47.24	63.18	58.97
hu	67.80	1.06	62.17	0.21	68.86	51.20	69.12	63.96	49.66	58.32	46.73	48.53	41.19	48.83
id	73.31	0.48	69.30	0.41	60.10	64.10	72.11	57.11	73.31	66.48	23.56	23.63	35.54	44.04
it	75.42	1.27	71.41	0.39	76.87	51.54	78.38	71.50	77.03	55.94	52.62	50.40	66.35	57.23
ja	83.04	0.34	75.70	0.25	79.68	43.68	82.11	73.26	81.95	73.47	13.02	26.57	11.88	21.00
ko	69.99	0.14	62.39	0.57	56.21	56.67	61.78	57.94	66.91	60.09	36.13	37.14	33.96	36.29
lv	61.34	0.33	56.42	0.70	62.31	51.08	58.06	52.15	56.36	59.76	47.35	38.76	44.46	48.20
nl	62.77	0.04	55.45	0.38	62.31	56.10	63.49	43.41	62.54	55.31	43.17	45.26	49.11	45.74
nno	63.80	0.09	58.45	0.08	68.59	59.96	65.12	54.24	60.27	59.96	63.61	60.24	57.03	59.21
nob	66.22	0.11	61.87	0.41	72.84	64.52	68.22	64.16	65.75	63.29	69.64	47.69	64.88	61.44
pl	79.40	0.38	72.22	0.25	79.00	72.57	70.86	72.59	68.82	73.70	66.73	63.65	60.15	65.42
pt	73.91	0.57	69.99	0.91	75.13	71.66	75.01	50.45	76.39	67.88	56.44	67.36	61.83	67.93
ro	73.80	1.00	68.82	0.47	64.27	67.11	76.03	42.84	73.87	68.79	53.26	38.30	57.60	49.22
ru	76.16	0.25	69.92	0.22	70.19	60.15	63.36	43.36	62.62	58.77	48.06	40.86	56.23	45.70
sl	66.75	0.02	60.34	0.53	67.15	61.34	67.21	60.74	54.20	63.01	54.26	55.05	54.85	57.71
sv	68.30	0.41	62.52	0.70	69.41	63.06	71.68	66.49	59.09	47.21	61.77	58.09	62.82	58.35
tr	59.49	1.44	54.89	0.25	43.65	52.87	60.00	54.01	59.88	53.84	36.59	33.56	41.11	39.53
uk	69.04	0.30	60.17	0.57	74.72	67.65	64.78	67.94	74.79	51.39	61.84	59.32	61.20	62.27
ur	77.58	0.15	76.48	0.02	77.48	74.75	76.82	70.96	77.36	65.35	18.36	24.61	32.66	37.03
vi	54.73	0.02	47.05	0.54	53.84	35.43	49.03	42.11	54.50	39.13	28.43	23.36	29.98	24.09
zh	57.89	0.62	51.37	1.08	60.21	48.23	49.85	40.43	60.72	49.32	23.51	23.63	26.93	25.32
moy.	70.07	0.43	64.94	0.40	67.94	60.28	68.29	60.23	67.34	61.62	45.28	45.38	47.38	48.16

TABLEAU .20. – UAS des tagparsers pour toutes les langues, pour toutes les expériences.

Lang	*Mono +c*		*Mono −c*		*Multi +c*	*Multi −c*	*Multi +c* ID	*Multi −c* ID	*Multi +c* W_{22}	*Multi −c* W_{22}	*ZS +c*	*ZS −c*	*ZS +c* W_{22}	*ZS −c* W
	acc	std	acc	std	acc	acc	acc	acc	acc	acc	acc	acc	acc	acc
ar	66.78	0.57	62.29	0.46	48.77	58.70	55.59	47.27	64.87	52.24	22.48	30.34	19.57	26.97
bg	72.98	0.35	66.92	0.24	73.39	65.40	60.82	67.22	76.27	56.26	38.46	45.62	41.60	50.78
ca	69.87	0.56	62.00	0.09	57.38	63.51	57.03	64.48	43.80	50.54	36.88	49.66	40.69	49.75
cs	64.11	0.87	56.09	0.27	66.03	48.44	67.13	57.81	57.85	58.61	37.75	32.80	50.42	45.42
da	60.23	0.36	55.00	0.04	62.10	58.05	61.43	59.24	54.00	57.88	53.22	48.40	54.27	37.23
de	59.47	0.97	50.48	0.34	59.78	51.52	45.48	38.76	59.50	52.82	28.79	28.16	22.07	23.95
el	71.84	0.48	66.57	0.11	60.40	61.58	74.56	69.54	74.55	48.11	15.40	24.40	24.21	28.72
en	62.27	0.16	55.25	0.53	63.17	52.44	64.32	41.76	56.21	48.47	15.51	17.72	12.66	10.95
es	63.55	0.19	58.69	0.41	68.16	65.65	59.69	63.26	67.63	62.02	63.80	58.81	63.94	58.18
et	52.73	0.25	42.66	1.52	51.83	35.00	56.65	45.41	54.74	44.00	29.66	26.70	27.77	28.78
eu	56.83	0.44	48.96	0.48	51.48	37.92	56.30	43.32	55.06	36.92	13.28	11.70	11.28	7.84
fa	66.33	0.20	63.97	0.13	64.50	56.22	65.72	60.60	64.18	60.86	15.94	13.90	11.79	17.85
fi	51.06	0.36	43.80	0.06	48.52	36.16	42.09	37.56	48.13	36.38	31.79	31.93	24.65	26.50
fr	65.97	0.49	59.61	0.32	68.01	61.49	69.42	52.19	47.98	63.30	38.68	48.53	34.43	44.95
ga	59.58	0.15	50.08	1.77	58.32	34.38	59.61	40.23	56.91	49.71	10.05	10.88	17.65	14.81
he	67.39	0.34	63.78	0.23	67.36	57.92	64.75	55.00	68.05	64.42	19.53	22.60	20.57	20.63
hi	73.72	0.49	71.62	1.05	56.52	44.59	73.52	55.70	73.67	67.01	17.89	25.41	21.56	32.87
hr	62.92	0.25	54.37	0.23	59.65	41.99	67.96	57.71	65.72	58.26	36.65	33.27	51.67	46.54
hu	60.41	1.30	51.31	0.31	60.64	38.11	61.74	52.79	37.25	45.78	22.17	23.52	17.21	24.67
id	65.34	0.30	59.08	0.67	50.25	54.19	63.97	45.51	65.14	56.38	8.96	10.19	17.84	26.23
it	70.35	1.23	64.72	0.62	71.56	41.86	73.24	64.76	71.26	46.30	38.98	37.22	55.11	44.34
ja	76.87	0.32	68.72	0.07	71.60	31.37	74.10	64.16	75.20	64.57	1.95	6.48	3.26	5.77
ko	62.96	0.41	52.14	0.29	46.51	44.06	53.29	45.96	59.02	48.11	17.83	17.54	14.65	18.40
lv	52.95	0.36	45.09	0.38	52.12	37.21	47.73	39.57	45.75	47.65	29.42	21.52	27.36	29.81
nl	53.62	0.02	45.17	0.29	52.66	45.65	53.84	30.61	52.62	44.68	25.80	28.32	34.19	30.79
nno	56.33	0.08	48.85	0.09	61.59	50.79	56.92	43.64	49.68	49.48	55.42	50.18	45.92	47.81
nob	59.12	0.11	52.73	0.13	66.36	55.72	59.96	54.61	56.46	52.83	61.93	36.30	55.12	50.29
pl	71.44	0.61	62.02	0.18	70.61	62.19	61.30	62.59	58.32	63.23	51.06	48.03	43.92	50.06
pt	67.82	0.82	62.51	0.85	68.48	63.91	66.70	39.22	67.71	58.20	40.10	53.37	44.46	53.35
ro	65.76	1.03	58.53	0.47	54.23	55.89	68.19	29.91	65.38	58.49	32.03	20.19	37.69	29.47
ru	69.78	0.26	61.37	0.43	62.53	50.59	53.81	31.15	53.32	48.94	36.09	28.14	46.12	33.63
sl	59.45	0.18	50.83	0.45	58.38	50.59	59.03	49.65	42.73	52.16	38.50	39.21	39.13	41.60
sv	60.93	0.45	52.59	0.93	61.57	53.74	64.04	57.61	48.85	35.19	49.20	45.81	50.09	46.10
tr	49.76	1.12	42.50	0.28	30.67	39.76	49.80	41.19	49.73	41.05	18.36	16.59	21.72	20.71
uk	60.59	0.23	47.89	0.81	67.17	57.70	55.57	57.98	67.58	38.60	47.12	44.63	46.07	48.22
ur	68.69	0.25	67.05	0.13	68.15	64.55	67.09	61.00	68.30	55.12	7.19	13.16	19.17	24.60
vi	44.79	0.15	35.25	0.44	43.31	21.65	37.71	30.08	43.95	26.45	11.21	8.13	12.46	9.32
zh	49.42	0.75	40.48	1.09	51.86	38.26	40.66	30.66	52.56	39.26	3.56	3.31	5.83	4.90
moy.	62.47	0.46	55.29	0.45	59.36	49.70	59.76	49.73	58.42	51.06	29.54	29.28	31.27	31.92

TABLEAU .21. – LAS des tagparsers pour toutes les langues, pour toutes les expériences.

Lang.	% noMorph	Lang.	% noMorph
ar	32	id	91
bg	33	it	41
ca	19	ja	94
cs	22	ko	91
da	29	lv	37
de	45	nl	24
el	28	nno	37
en	32	nob	37
es	21	pl	24
et	31	pt	75
eu	34	ro	15
fa	32	ru	35
fi	26	sl	24
fr	38	sv	32
ga	27	tr	34
he	49	uk	28
hi	14	ur	14
hr	19	vi	89
hu	25	zh	85
moy.	38.5		

TABLEAU .22. – % de mots n'ayant pas de traits morphologiques dans le test.