Rapport de projet
Lexicalisation de l'analyseur

Tanguy Tiran, Aymene Mohammed Bouayed, Raissa Camelo Salhab et Julien Antigny

Décembre 2021

1 Introduction

Un analyseur de dépendances par transition permet de trouver les opérations nécessaires pour construire la structure de dépendances d'une phrase écrite en langue naturelle.

Un analyseur syntaxique de dépendances en transition utilisant uniquement des étiquettes morpho-syntaxiques en entrée tel que celui initialement fourni pour ce projet peut être vu comme limité. A travers les parties de discours (POS), un modèle n'accède en effet pas à la forme spécifique de chaque mot. De ce fait, enrichir l'entrée de l'analyseur afin qu'il ait accès à des informations lexicales plus détaillées pourrait améliorer les prédictions faites par cet analyseur. Deux formes de lexicalisation parmi plusieurs ont été envisagées dans ce projet :

— L'ajout d'affixes (suffixes seuls, préfixes seuls et suffixes + préfixes).
— L'ajout des plongements de mots (word embeddings).

Explorer ces deux directions nous permettra d'aborder deux points principaux :

— La lexicalisation de l'analyseur permet-elle d'obtenir de meilleurs résultats que l'analyseur initialement fourni ? Il est possible que les informations de forme soient en partie redondantes avec les informations apportées par les étiquettes morpho-syntaxiques : il sera intéressant de tester les affixes et les word embeddings avec et sans les parties de discours.
— Les word embeddings sont des représentations gourmandes. Moins coûteuse en temps de calcul, l'utilisation d'affixes peut-elle s'avérer être aussi efficace que celle des embeddings ?

Tout en tenant de répondre à ces questions, nous nous demanderons si, de par leur caractéristiques distinctes, certaines langues seront plus facilement prédites que d'autres et si les features interagiront différemment avec elles. Nous effectueront nos expérimentations pour 4 langues : l'anglais, le français, le portugais et l'italien.

Les parties 2 et 3 détaillent les intérêts des deux formes de lexicalisation ainsi que la manière avec laquelle nous les avons implémentées. La partie 4 présente les résultats de nos expérimentations. Une comparaison plus approfondie des différentes approches est fournie dans la cinquième partie.

2 Affixes

L'ajout d'affixes est une forme de lexicalisation de l'analyseur. L'information contenue dans ces affixes présente un intérêt syntaxique : prenons le cas du français, des suffixes particuliers sont utilisés par les verbes (le "-er" de l'infinitif, ou le "-ons" de la deuxième personne du pluriel...), le pluriel ("-s"), les adverbes ("-ment"), les adjectifs ("-eux" de courageux, "-if" de craintif...). Dans de nombreuses langues, les cas grammaticaux s'illustrent à travers l'ajout de préfixes ou de suffixes.

Plusieurs méthodes permettent de représenter les affixes afin de les fournir à l'analyseur. Dans un premier temps, nous avons directement encodé les N premiers et les M derniers caractères de chaque mot. Cette approche ne s'est pas révélée concluante (une possibilité est qu'elle nécessiterait un très grand nombre

d'exemples en entrée pour fonctionner). Une autre technique moins gourmande est de ne conserver uniquement que les affixes les plus fréquents dans chaque langue, et de les représenter à l'aide d'un vecteur one-hot où chaque bit équivaut à 1 si l'affixe associé est présent.

Nous avons estimé l'identité des affixes les plus fréquents dans chaque langue à partir des corpus fournis.

Deux paramètres doivent donc être choisis : la taille des affixes et le nombre d'affixes fréquents représentés dans vecteurs one-hot.

2.1 Choix des paramètres

Les affixes de plus grande taille peuvent contenir des informations plus complètes et de meilleur qualité pour chaque mot. Prenons par exemple le cas des suffixes en français : "-ent" peut souvent indiquer la 3ème personne du pluriel ("ils mangent, "ils jouent"), mais également des adverbes ("différemment, joliment...") Choisir des suffixes de 4 lettres plutôt que 3 permettra en partie de distinguer ce type de cas. Moins de mots seront en revanche représentés par des affixes de grande taille.

Augmenter le nombre d'affixes conservés ne devrait pas augmenter le temps de préparation des données et ne devrait jamais diminuer les performances de l'analyseur. En fournissant des vecteurs de plus grande taille à nos modèles, les temps d'apprentissage seront toutefois rallongés (mais resteront raisonnables). Nous pouvons donc nous demander si les affixes moins fréquents (rangs 50 à 100 et 100 à 150) seront suffisamment représentés pour pouvoir apporter un gain de performances qui vaille la peine d'augmenter les temps d'apprentissage.

Afin de choisir les paramètres qui seront conservés, nous avons testé les performances de l'analyseur dans nos quatre langues avec des longueurs de préfixes et de suffixes différentes. Les paramètres optimaux pour les préfixes et suffixes ont ainsi été évalués séparément, en faisant varier la taille (2, 3 et 4 lettres) ainsi que le nombre d'affixes fréquents conservés (50, 100 et 150 affixes). Les features models utilisés sont similaires à celui de la baseline (seules les features ont été manipulées).

Les performances de l'analyseur dans les différentes conditions ont été évaluées en prenant la moyenne des résultats obtenus dans les 4 langues. Ces résultats sont reportés ci-dessous pour les préfixes puis les suffixes.

2.1.1 Préfixes

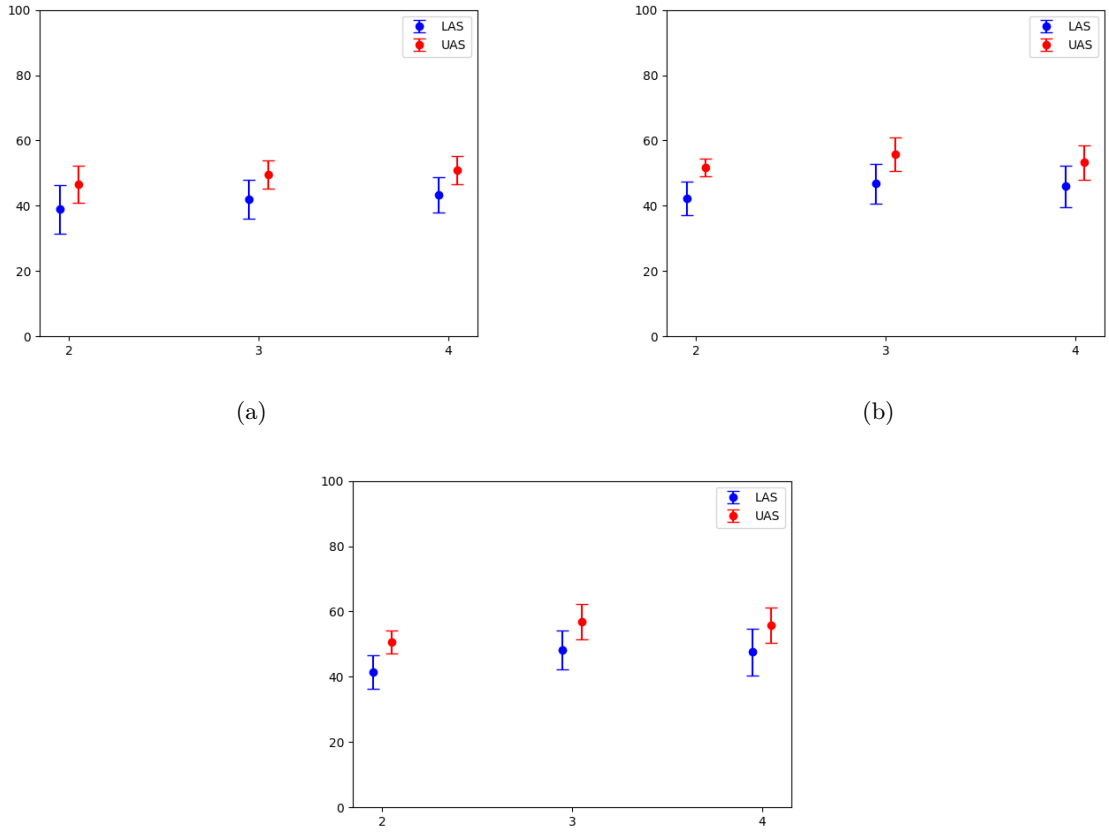
La figure 5 présente, pour les préfixes, les effets des deux paramètres sur les performances de l'analyseur. Ces effets semblent limités, mais augmenter le nombre de préfixes utilisés permet d'obtenir de meilleurs résultats. De plus, la différence entre les préfixes de taille 3 et de taille 4 est minime, mais les préfixes de taille 3 obtiennent tout de même de meilleurs résultats.

Nous avons donc opté pour des préfixes de 3 lettres et des vecteurs de taille 150.

2.1.2 Suffixes

La figure 2 présente, pour les suffixes, les effets des deux paramètres sur les performances de l'analyseur. Tout comme pour les préfixes, ces effets semblent ici être limités.

Nous avons ici opté pour des suffixes de 4 lettres et des vecteurs de taille 150.

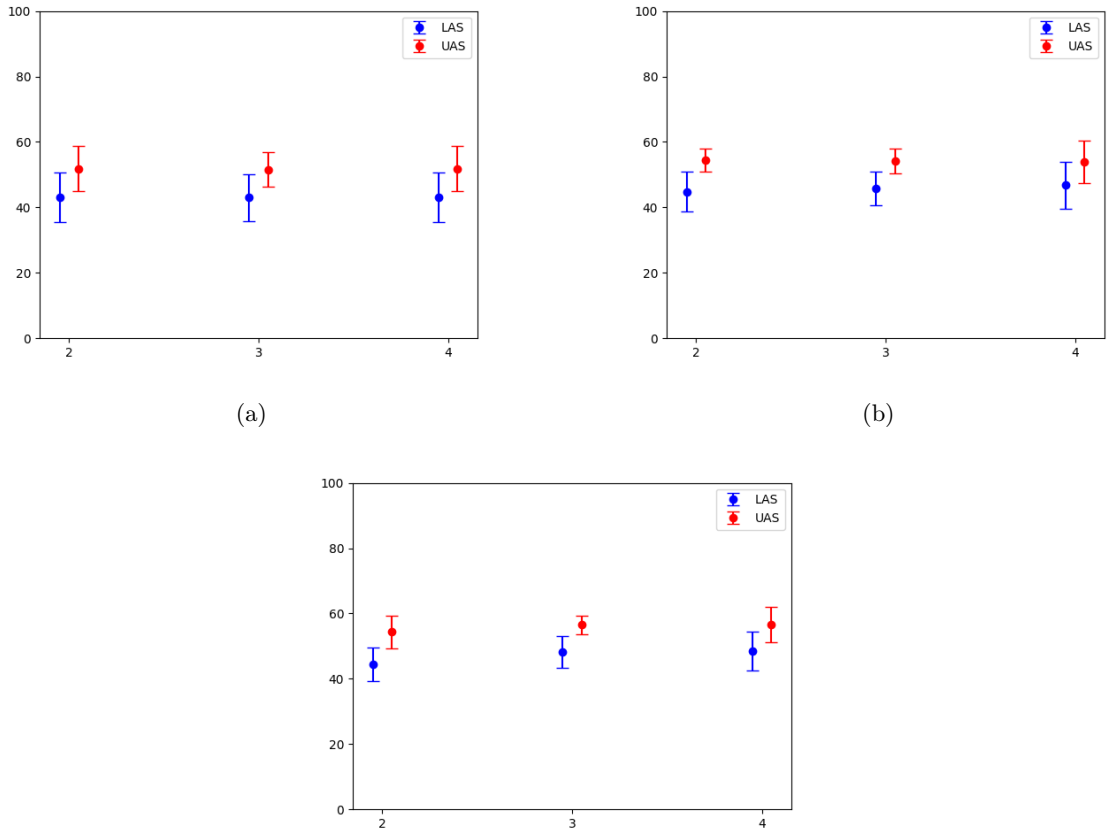


(a)

(b)

(c)

FIGURE 1 – **Effets de la taille des préfixes et du nombres de préfixes fréquents conservés sur les performances de l'analyseur.** Nombre de préfixes utilisés : 50 (a), 100 (b) et 150 (c).



(a)

(b)

(c)

FIGURE 2 – **Effets de la taille des suffixes et du nombres de suffixes fréquents conservés sur les performances de l'analyseur.** Nombre de suffixes utilisés : 50 (a), 100 (b) et 150 (c).

3 Word embeddings

L'approche de word embeddings se base sur l'ajout de vecteurs d'encodage de mots de taille n au vecteur d'entrée du réseau de neurones qui prédit les opérations que l'oracle doit faire. Ceci diffère de l'approche simple qui consiste en la concaténation d'un vecteur one-hot de taille $|v|$ pour chaque mot sélectionné depuis la configuration (où v est le vocabulaire, $|v|$ est le nombre de mots dans le vocabulaire, avec $|v| >> n$) (voir figure 3).

Les word embeddings permettent de représenter de façon compacte la forme des mots tout en incluant un mélange d'informations sémantiques, syntaxiques et lexicales (tel que le genre). En rajoutant aux vecteurs d'entrée du réseau les embeddings, l'analyseur pourra donc avoir une meilleure idée des opérations à effectuer, et pourra prédire correctement les dépendances.

Afin d'implémenter les word embeddings dans l'analyseur initial, nous avons utilisé la bibliothèque `FastText` [1]. Cette dernière fournit des word embeddings de dimension 300 pour différentes langues dont l'anglais, le français, l'italien et le portugais. Par la suite, nous avons réduit la dimension de ces vecteurs à 100 en utilisant la fonction `reduce_model` de `FastText` afin de ne pas avoir une entrée au réseau de neurones trop volumineuse.

1. `https://fasttext.cc/`

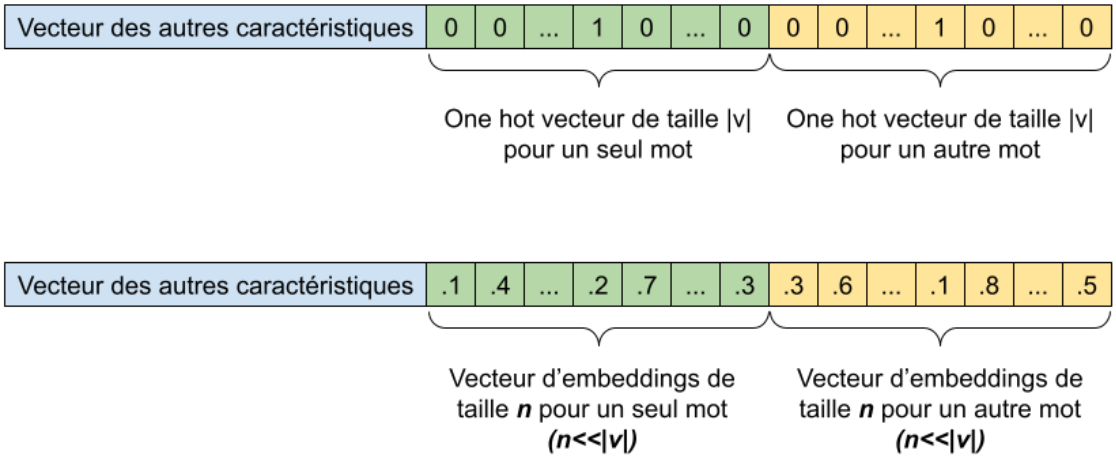


FIGURE 3 – Idée des word embeddings.

4 Résultats

Les performances de l'analyseur rapportées dans cette section ont été évaluées après utilisation d'affixes (préfixes seuls, suffixes seuls, préfixes et suffixes) et de word embeddings. Ces expérimentations ont été réalisées avec et sans étiquettes morpho-syntaxiques (POS). Pour tous les modèles, 100 000 et 700 exemples ont respectivement été utilisés en phase d'apprentissage et de test. L'architecture du réseau utilisé (la même pour toutes les conditions) ainsi que les différents feature models sont disponibles en annexe.

Pour chaque langue et condition, 3 apprentissages et tests différents ont été effectués. Les moyennes et écart-types des scores obtenus ont été calculés pour chaque langue et rapportés dans chacun des tableaux suivants.

	POS	
	LAS	UAS
EN	67.73 (1.25)	73.02 (1.09)
FR	68.40 (1.12)	73.41 (1.14)
PT	69.46 (1.53)	74.28 (1.85)
IT	74.00 (0.17)	79.90 (0.84)

TABLE 1 – Performances de l'analyseur pour notre baseline (POS seulement).

4.1 Affixes

De manière surprenante, peu de différences de scores ont été observées après utilisation des préfixes seuls, suffixes seuls, et des préfixes et suffixes en simultané (voir tableau 7 dans l'annexe C). Il aurait pourtant été raisonnable de supposer :

— Que les suffixes et préfixes auraient obtenus des résultats différents (si, par exemple, les suffixes apportent plus d'informations syntaxiques que les préfixes).

— Que l'utilisation simultanée des suffixes et préfixes aurait permis d'obtenir des performances significativement supérieures à celles des suffixes seuls et préfixes seuls.

Seuls sont donc reportés dans la table 2 les résultats pour les suffixes et préfixes en simultané (très légèrement supérieurs à ceux des suffixes et préfixes isolés), avec et sans parties de discours.

Les résultats obtenus par les affixes seuls sont nettement inférieurs à ceux de la baseline (Tableau 1). Environ 50% des dépendances y restent tout de même correctement prédites, preuve que l'information contenue dans les suffixes et préfixes présente tout de même un intérêt dans le cadre de ce projet.

	S&P		S&P + POS	
	LAS	UAS	LAS	UAS
EN	44.33 (3.92)	53.39 (5.32)	67.39 (1.46)	72.82 (0.95)
FR	47.71 (2.14)	57.54 (3.61)	72.62 (1.71)	76.92 (1.68)
PT	56.81 (1.44)	63.96 (0.90)	71.62 (1.52)	76.16 (1.60)
IT	50.84 (2.70)	61.29 (3.03)	75.78 (0.87)	81.25 (1.52)

TABLE 2 – Performances de l'analyseur à partir d'affixes (S&P), avec et sans parties de discours.

Après utilisation combinée des affixes et des parties de discours, les scores obtenus en français, portugais et italien semblent légèrement supérieurs à ceux de la baseline (environ 2 à 3 % de différence).

De grandes différences existent cependant entre les quatre langues, en particulier lorsque les affixes ne sont pas utilisés en conjonction avec les parties de discours. Plus de 12 points de pourcentage de différence peuvent ainsi être observés entre la LAS en anglais et en portugais dans la condition affixes seuls, ce qui explique peut-être pourquoi l'utilisation combinée d'affixes et de parties de discours ne permet pas d'améliorer les résultats en anglais par rapport à la baseline. En conséquence, nous concluons que les affixes ne sont pas des prédicteurs efficaces en anglais dans le cadre de cette tâche, mais qu'ils peuvent s'avérer informatif pour les autres langues.

4.2 Word embeddings

Les moyennes des résultats obtenus pour les word embeddings (avec et sans partie de discours) ainsi que les écarts-types sont rapportés dans le tableau 3.

	WE		WE + POS	
	LAS	UAS	LAS	UAS
EN	52.81 (0.96)	62.65 (1.19)	68.85 (0.09)	74.95 (0.22)
FR	56.08 (2.47)	65.68 (3.07)	73.36 (0.96)	78.46 (0.93)
PT	64.19 (1.02)	70.15 (1.44)	75.29 (0.65)	79.09 (1.10)
IT	59.43 (1.24)	66.76 (1.13)	76.21 (1.69)	81.44 (1.09)

TABLE 3 – Performances de l'analyseur à partir des word embdedings, avec et sans parties de discours.

On remarque que l'utilisation des WE résulte en des performances qui sont notablement plus basses que la baseline où on utilise les POS seulement (Tableau 1). Cependant en combinant les deux on arrive à augmenter la précision LAS et UAS (cette augmentation est même notable pour certaines langues dont le français et le portugais). Nous pouvons également apprécier le fait que la combinaison WE+POS permet de réduire l'écart type.

De plus, en comparant les f-scores obtenu en utilisant la baseline avec les POS uniquement et en utilisant les WE uniquement (voir figures 5a, 5c en annexe D) on remarque que malgré le fait que l'utilisation des POS seuls soit plus performante que les WE seuls, la majorité des dépendances restent correctement prédites dans les deux cas.

Ainsi, en interprétant les résultats obtenus dans le tableau 3 et les figures de f-score en annexe, nous pouvons conclure qu'il y a des différences entre les POS et les WE en dépit de la redondance d'informations apportées par ces deux features. Ces différences peuvent être observées indirectement lorsqu'on utilise les POS avec les WE, ce qui augmente les scores dans toutes les langues.

Par ailleurs, en analysant les résultats en LAS et UAS, on observe de grands écarts de performance entre les langues lorsque les embeddings sont utilisés sans les parties de discours. Cet écart diminue et passe de 12% à environ 8% lorsque l'on rajoute les POS, restant ainsi tout de même important. Dans les deux cas, certaines langues paraissent donc plus difficiles à analyser que d'autres : cette difficulté pourrait être liée aux corpus utilisés ou à des caractéristiques intrinsèques des différentes langues.

5 Comparaisons

Deux points principaux ont été mentionnés en introduction :
— Lexicaliser l'analyseur permet-il d'obtenir de meilleurs résultats qu'avec les parties de discours ?
— Les word embeddings (plus coûteux en temps de calcul et en ressources) permettent-ils d'obtenir de meilleurs résultats que les affixes ?

Nous avons, afin de répondre à ces questions, comparé les performances des analyseurs utilisant les affixes ou les embeddings, avec ou sans parties de discours.

5.1 Sans parties de discours

Le tableau 4 représente côte à côte les résultats en LAS et UAS obtenus en utilisant seulement les affixes (S&P) ou bien seulement les word embeddings (WE).

	S&P		WE	
	LAS	UAS	LAS	UAS
EN	44.33 (3.92)	53.39 (5.32)	52.81 (0.96)	62.65 (1.19)
FR	47.71 (2.14)	57.54 (3.61)	56.08 (2.47)	65.68 (3.07)
PT	56.81 (1.44)	63.96 (0.90)	64.19 (1.02)	70.15 (1.44)
IT	50.84 (2.70)	61.29 (3.03)	59.43 (1.24)	66.76 (1.13)

TABLE 4 – Performances de l'analyseur sans partie de discours.

Pour toutes les langues on observe une augmentation de ∼9% en LAS et ∼7% en UAS lorsqu'on utilise les embeddings plutôt que les affixes. Les embeddings sont donc, en l'absence des parties de discours, notablement plus utiles et informatifs que les affixes pour prédire les dépendances et permettent d'avoir moins de variance sur certaines langues.
Il est probable que la variance élevée observée pour les affixes soit en grande partie due à la manière avec laquelle nous les avons implémentés : de nombreux mots ne voient pas leurs affixes (car peu fréquents) conservés par le modèle, et leur forme devient donc totalement invisible aux yeux de l'analyseur (ce qui ajoute un degré de variabilité supplémentaire).

5.2 Avec parties de discours

Le tableau 5 représente côte à côte les résultats en LAS et UAS obtenus en utilisant les affixes et les embeddings avec parties discours. La baseline y est également représentée (POS).

	POS		S&P + POS		WE + POS	
	LAS	UAS	LAS	UAS	LAS	UAS
EN	67.73 (1.25)	73.02 (1.09)	67.39 (1.46)	72.82 (0.95)	68.85 (0.09)	74.95 (0.22)
FR	68.40 (1.12)	73.41 (1.14)	72.62 (1.71)	76.92 (1.68)	73.36 (0.96)	78.46 (0.93)
PT	69.46 (1.53)	74.28 (1.85)	71.62 (1.52)	76.16 (1.60)	75.29 (0.65)	79.09 (1.10)
IT	74.00 (0.17)	79.90 (0.84)	75.78 (0.87)	81.25 (1.52)	76.21 (1.69)	81.44 (1.09)

TABLE 5 – Performances de l'analyseur après utilisation des POS (baseline), affixes + POS et embeddings + POS.

Comparée à la baseline, l'utilisation combinée d'affixes et de parties de discours semble améliorer les résultats en français, en portugais et en italien (entre 1.5 et 4 points de pourcentage d'augmentation). Cette amélioration est néanmoins plus marquée pour les embeddings, où les scores sont supérieurs pour toutes les langues par rapport aux 2 autres conditions.

Les différences de scores entre les différentes conditions restent tout de même peu prononcées (maximum + 6% environ). Nous avons vu dans les parties précédentes que les affixes et embeddings étaient, par eux-même, des prédicteurs moins puissants que les parties de discours. Pour que ces deux features améliorent

fortement le modèle en étant utilisées en conjonction avec les parties de discours, il faudrait qu'elles comblent les lacunes de la baseline en prédisant mieux certaines dépendances par rapport aux parties de discours.

Les f-scores représentés sur la figure 4 montrent que quelques soient les dépendances les plus fréquentes, les parties de discours obtiennent de meilleurs résultats que les embeddings, qui obtiennent eux-mêmes de meilleurs résultats que les affixes. Pour l'analyseur, les informations contenues par ces différentes features semblent donc en grande partie redondantes (ce qui explique pourquoi les modèles les combinant n'améliorent les scores que dans des proportions mesurées).

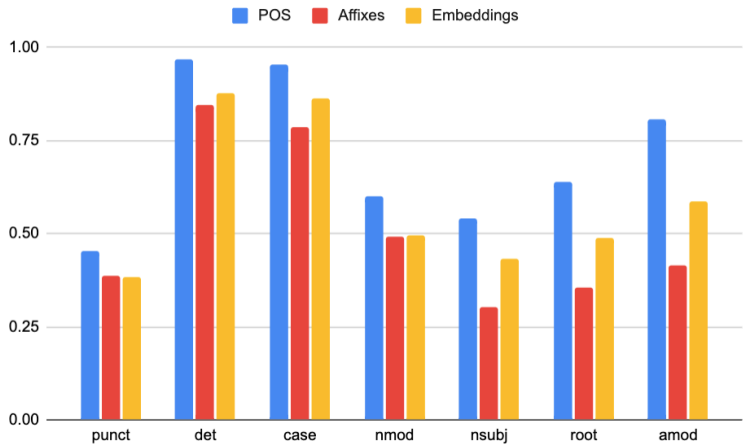


FIGURE 4 – **F-scores des dépendances les plus fréquentes après utilisation des parties de discours, des affixes (sans POS) et des embeddings (sans POS).** Les types de dépendances présents sur ce graphique sont triés dans l'ordre décroissant de fréquence et représentent plus de 2/3 des dépendances utilisées dans les quatre langues.

6 Conclusion

Les expérimentations réalisées au cours de ce projet nous aident à répondre aux questions formulées en introduction. Si les affixes et les embeddings se sont montrés par eux-mêmes moins performants que les parties de discours, il a été possible, en les combinant avec la baseline, d'améliorer les résultats de l'analyseur en transition.
De plus, les embeddings se sont révélés être de meilleurs prédicteurs que les affixes. Notre conclusion est que, malgré leur coût onéreux en ressources, les embeddings doivent être la méthode privilégiée de lexicalisation de l'analyseur. Les embeddings permettent de représenter la forme des mots tout en apportant des informations riches sémantiquement et syntaxiquement : il serait intéressant de les tester avec un très grand nombre de données d'apprentissage afin d'évaluer précisément leur limite.

Le code de nos expérimentations est publiquement disponible sur `https://github.com/aymene98/Transition_based_parser`.

A Fichiers fm

Après plusieurs tests, nous avons conservé des fichiers .fm contenant les caractéristiques suivantes :

Mot	POS	Prefixes	Suffixes	WE
W B -2	✓	✓	✓	-
W B -1	✓	✓	✓	✓
W B 0	✓	✓	✓	✓
W B 1	✓	✓	✓	✓
W B 2	✓	✓	✓	-
W S 0	✓	✓	✓	✓
W S 1	✓	✓	✓	✓

TABLE 6 – Contenu des fichiers fm selon les features utilisées.

B Architecture du réseau de neurones

Après avoir affiné les hyperparamètres du réseau de neurones utilisé dans le Transition based parser, nous avons abouti sur l'achitecture suivante :

```
    model = Sequential()
model.add(Dense(units=500, input_dim=inputSize, kernel_regularizer=L2(0.001)))
model.add(Dropout(0.2))
model.add(BatchNormalization())
model.add(PReLU())

   model.add(Dense(units=500 ))
model.add(Dropout(0.2))
model.add(BatchNormalization())
model.add(Activation('relu'))
   model.add(Dense(units=200))
model.add(Dropout(0.2))
model.add(BatchNormalization())
model.add(Activation('relu'))

   model.add(Dense(units=outputSize, activation='softmax'))

   model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

   model.fit(x_train, y_train, epochs=20, batch_size=32, validation_data=(x_dev,y_dev), verbose=2)
```

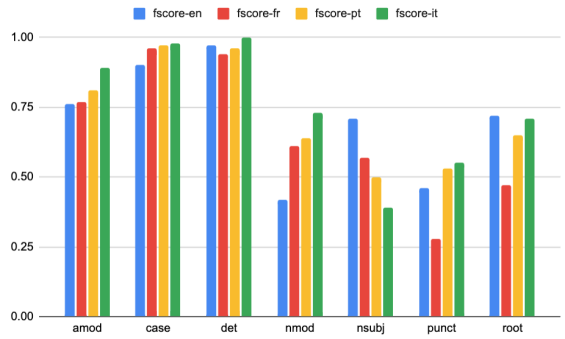
C Tableaux supplémentaires

Depuis le tableau 7 on remarque que la performance obtenues en utilisant les suffixes seuls, les suffixes et les préfixes sont très proches que ce soit avec les POS ou sans les POS. Par contre, pour l'anglais, le français et le portugais on voit bien que les préfixes obtiennent une performance légèrement plus basses que les autres configurations. Ceci suggère que les préfixes ont moins d'influence sur l'inférence des dépendances.

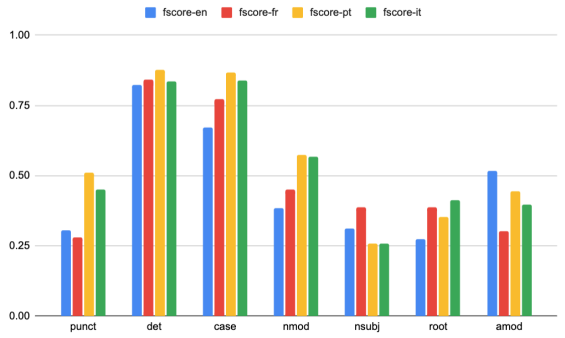
	Suffixes		Préfixes		S&P	
	LAS	UAS	LAS	UAS	LAS	UAS
EN	43.80 (0.99)	52.86 (2.12)	40.84 (2.05)	53.10 (1.40)	44.33 (3.92)	53.39 (5.32)
FR	45.37 (1.83)	54.40 (1.69)	41.85 (1.36)	51.59 (0.94)	47.71 (2.14)	57.54 (3.61)
PT	56.63 (0.65)	63.45 (0.32)	55.39 (0.76)	62.59 (0.28)	56.81 (1.44)	63.96 (0.90)
IT	48.92 (1.25)	58.61 (1.12)	50.36 (2.28)	60.00 (2.64)	50.84 (2.70)	61.29 (3.03)
	Suffixes + POS		Préfixes + POS		S&P + POS	
	LAS	UAS	LAS	UAS	LAS	UAS
EN	67.40 (0.17)	72.34 (0.60)	67.15 (1.62)	72.29 (1.32)	67.39 (1.46)	72.82 (0.95)
FR	72.24 (1.54)	76.97 (1.70)	71.72 (1.63)	76.36 (1.04)	72.62 (1.71)	76.92 (1.68)
PT	72.63 (1.20)	76.84 (0.79)	71.98 (0.98)	76.16 (1.35)	71.62 (1.52)	76.16 (1.60)
IT	76.45 (2.47)	81.48 (2.99)	75.92 (0.88)	81.48 (1.30)	75.78 (0.87)	81.25 (1.52)

TABLE 7 – Performances de l'analyseur à partir de suffixes, de préfixes, de suffixes et de préfixes, avec et sans parties de discours.

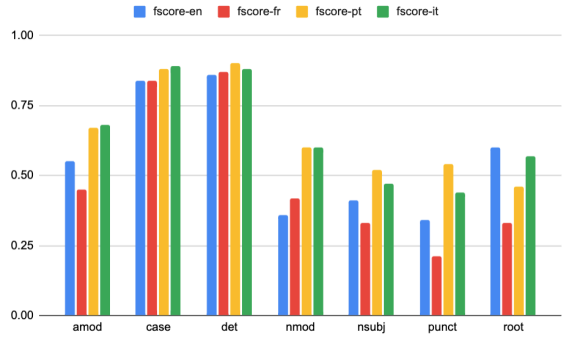
D Figures supplémentaires



(a) F-score des dépendances les plus fréquentes après utilisation des POS.



(b) F-score des dépendances les plus fréquentes après utilisation des affixes.



(c) F-score des dépendances les plus fréquentes après utilisation des embeddings.

FIGURE 5 – F- score obtenu sur les dépendances les plus fréquentes en utilisant les POS seuls, les affixes seuls et les word embeddings seuls.