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

Manon Scholivet

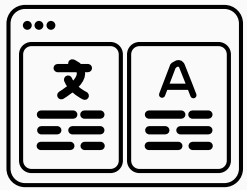
15 octobre 2021

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

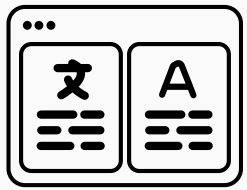
Le Traitement Automatique des Langues (TAL)

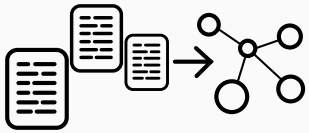
Données dans une langue naturelle



Le Traitement Automatique des Langues (TAL)

Données dans une langue naturelle

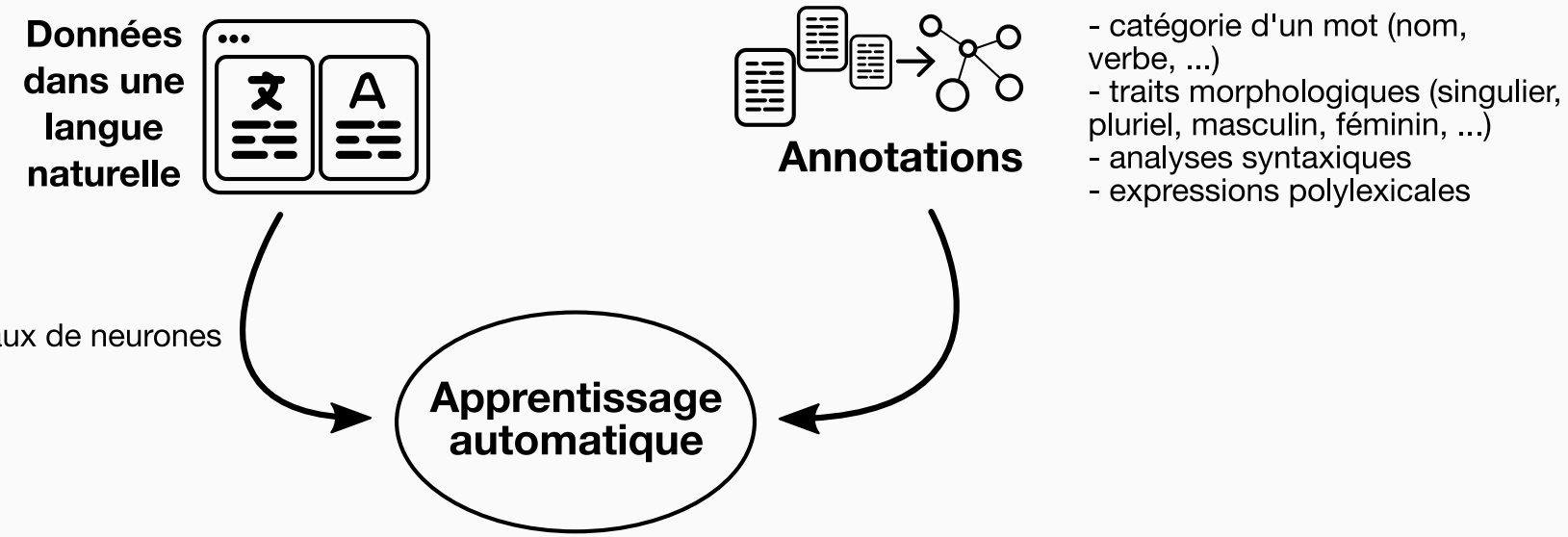




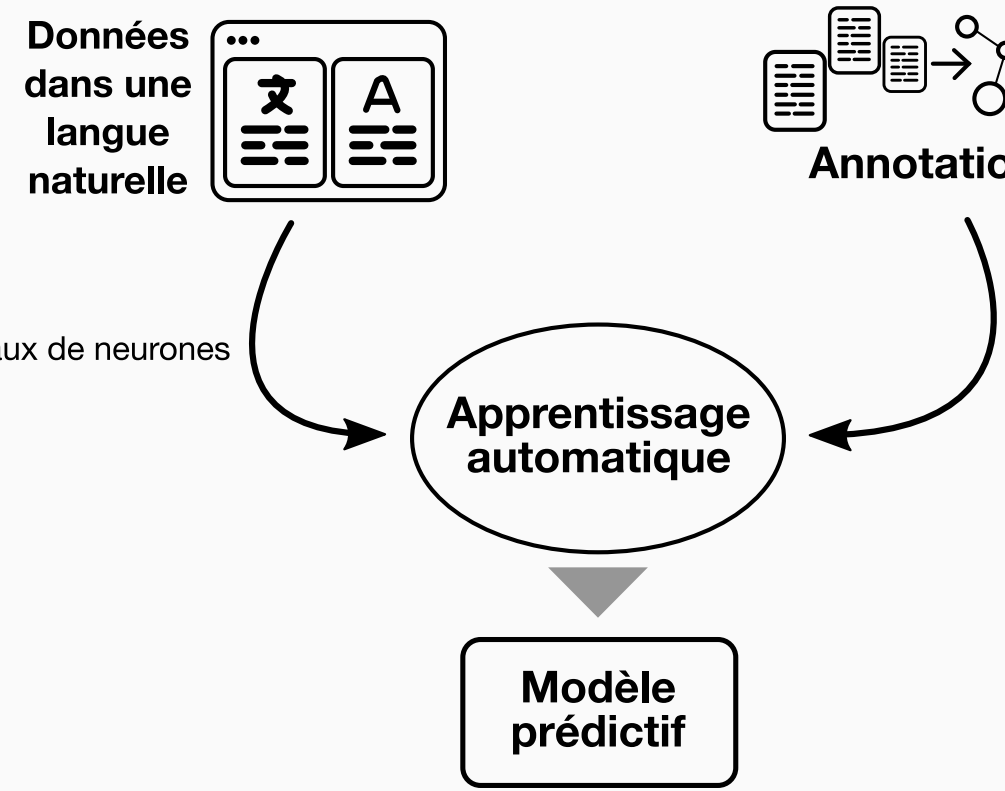
Annotations

- catégorie d'un mot (nom, verbe, ...)
- traits morphologiques (singulier, pluriel, masculin, féminin, ...)
- analyses syntaxiques
- expressions polylexicales

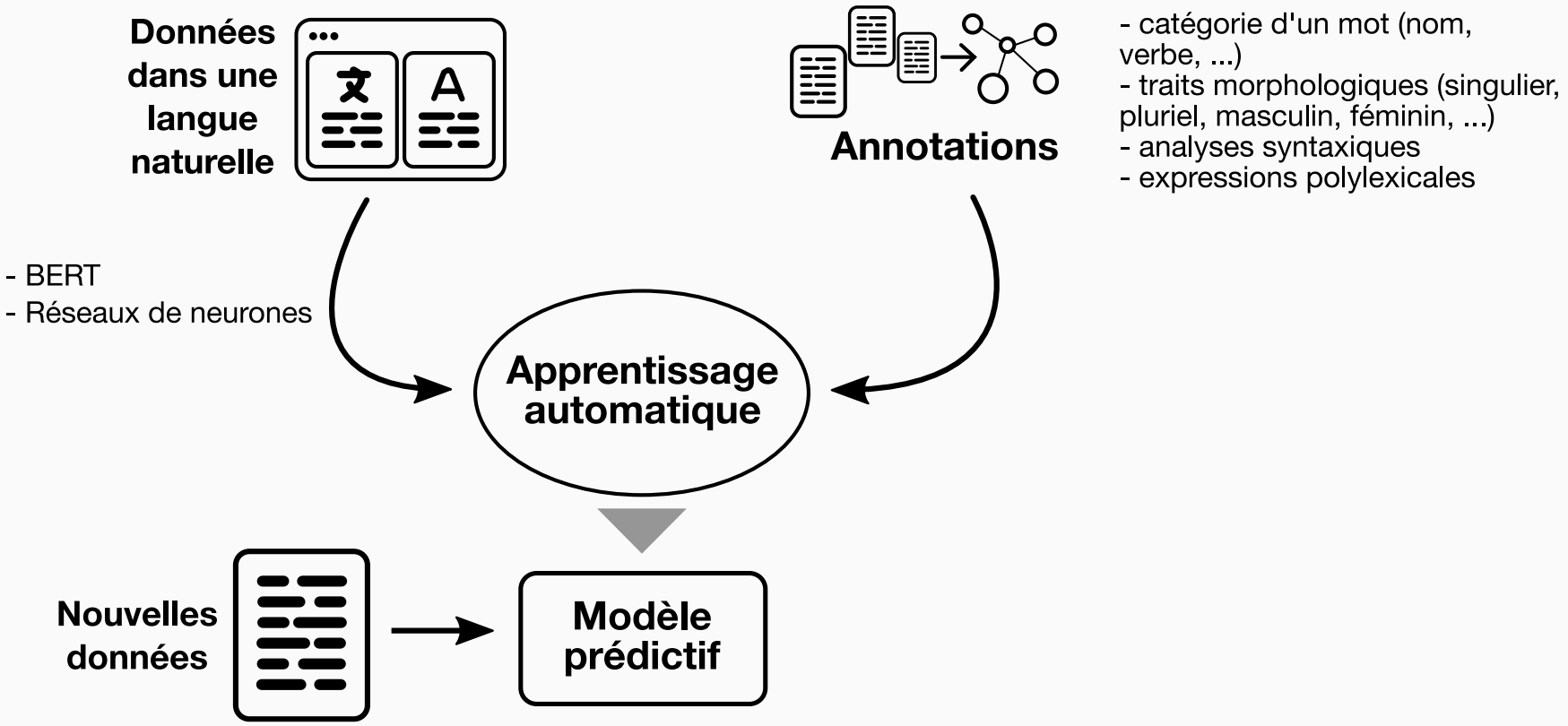
Le Traitement Automatique des Langues (TAL)



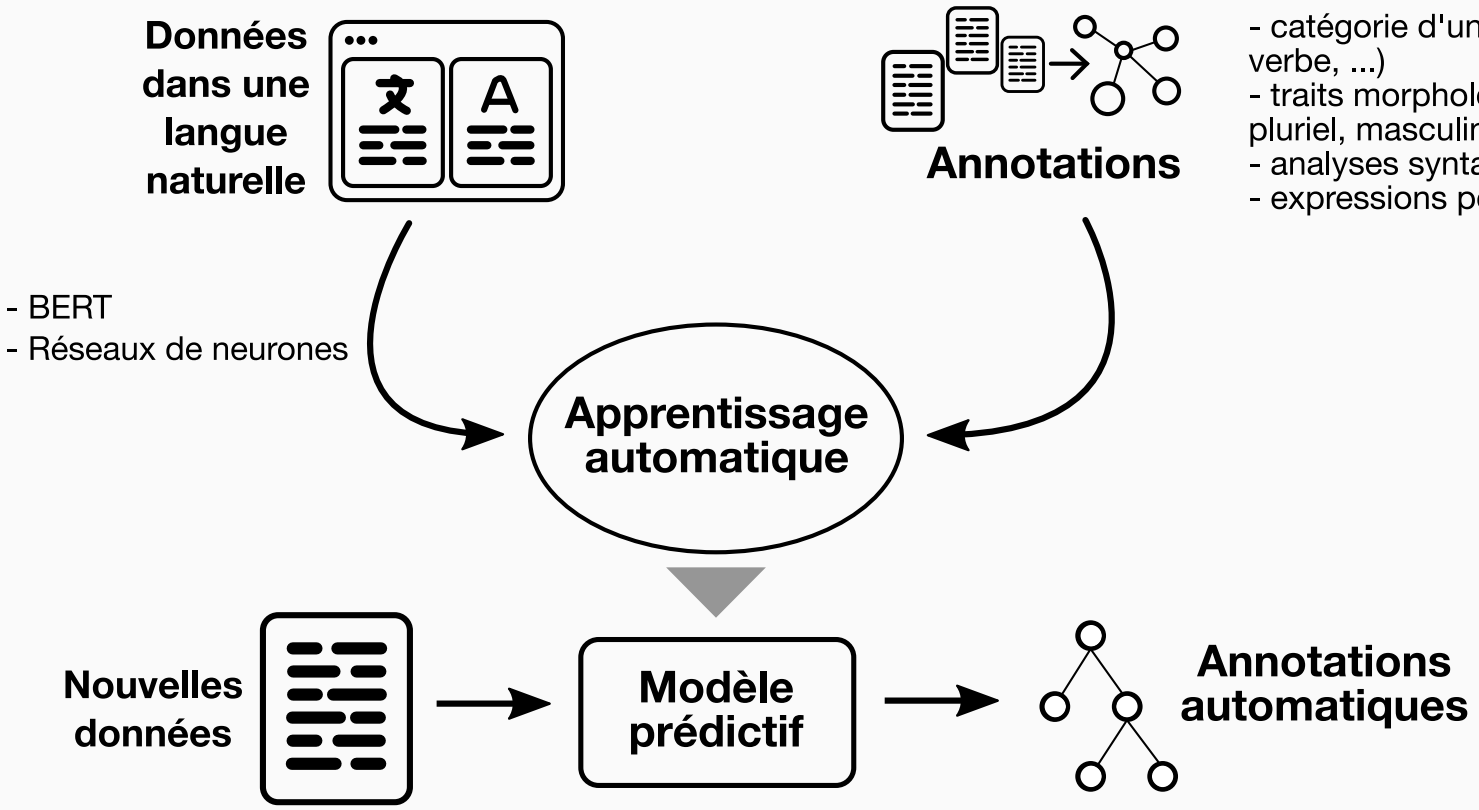
Le Traitement Automatique des Langues (TAL)



Le Traitement Automatique des Langues (TAL)



Le Traitement Automatique des Langues (TAL)

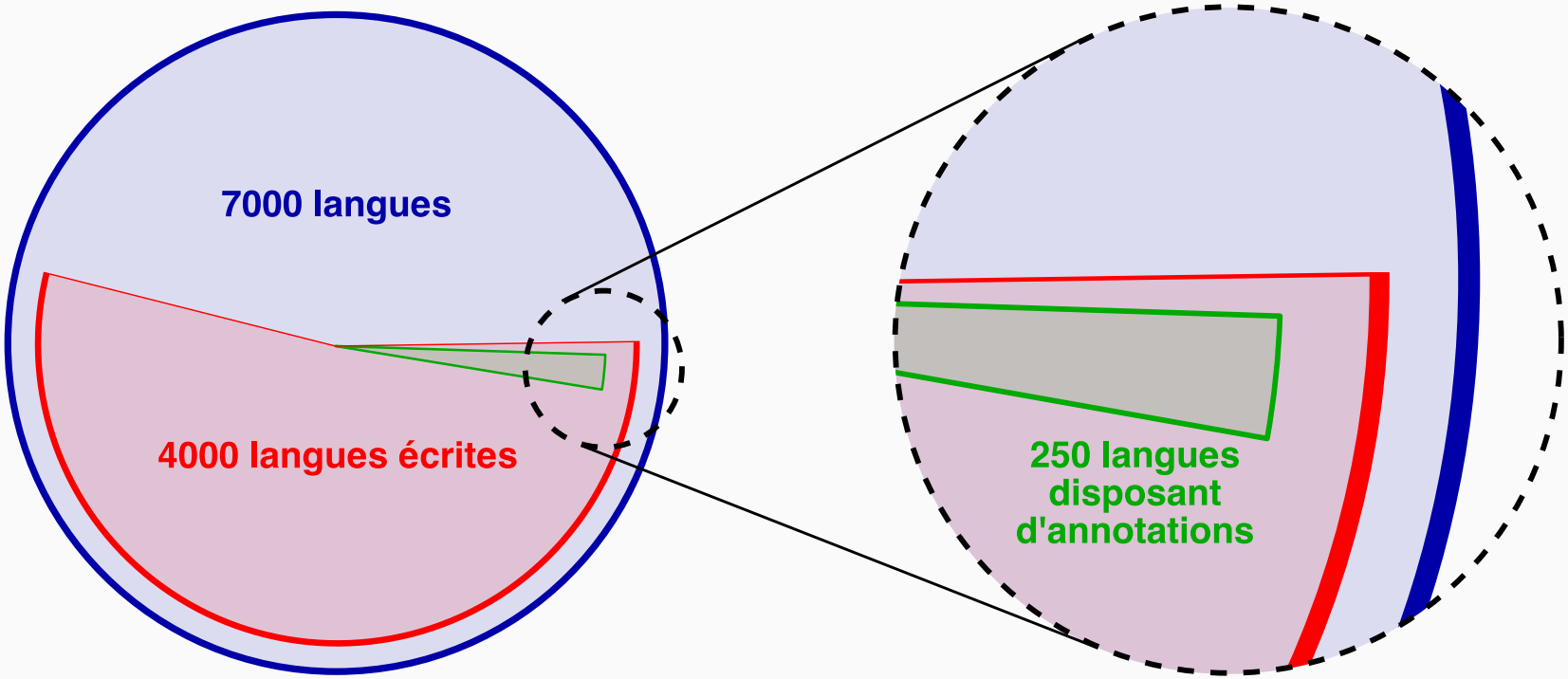


Problématique

- Seules quelques langues disposent d'annotations (Joshi et al. (2020))

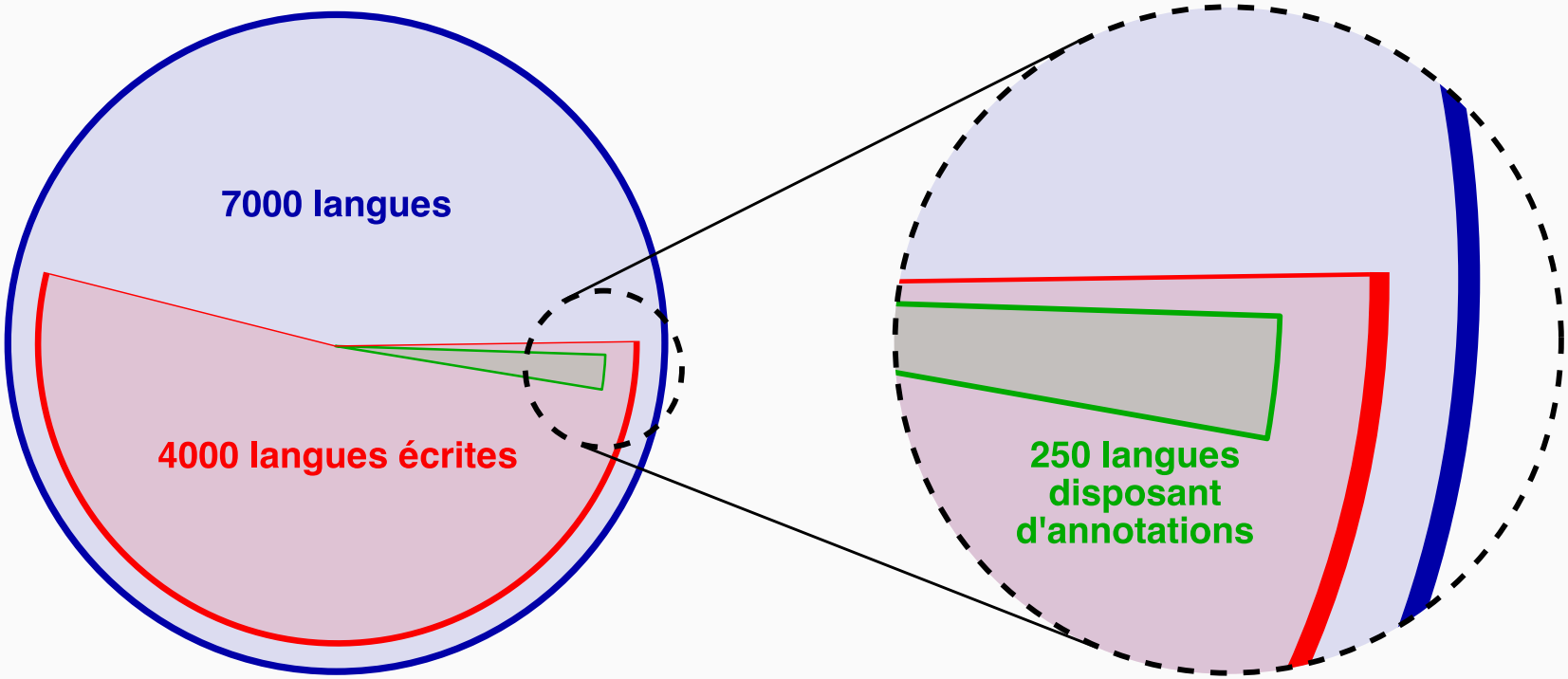
Problématique

- Seules quelques langues disposent d'annotations (Joshi et al. (2020))



Problématique

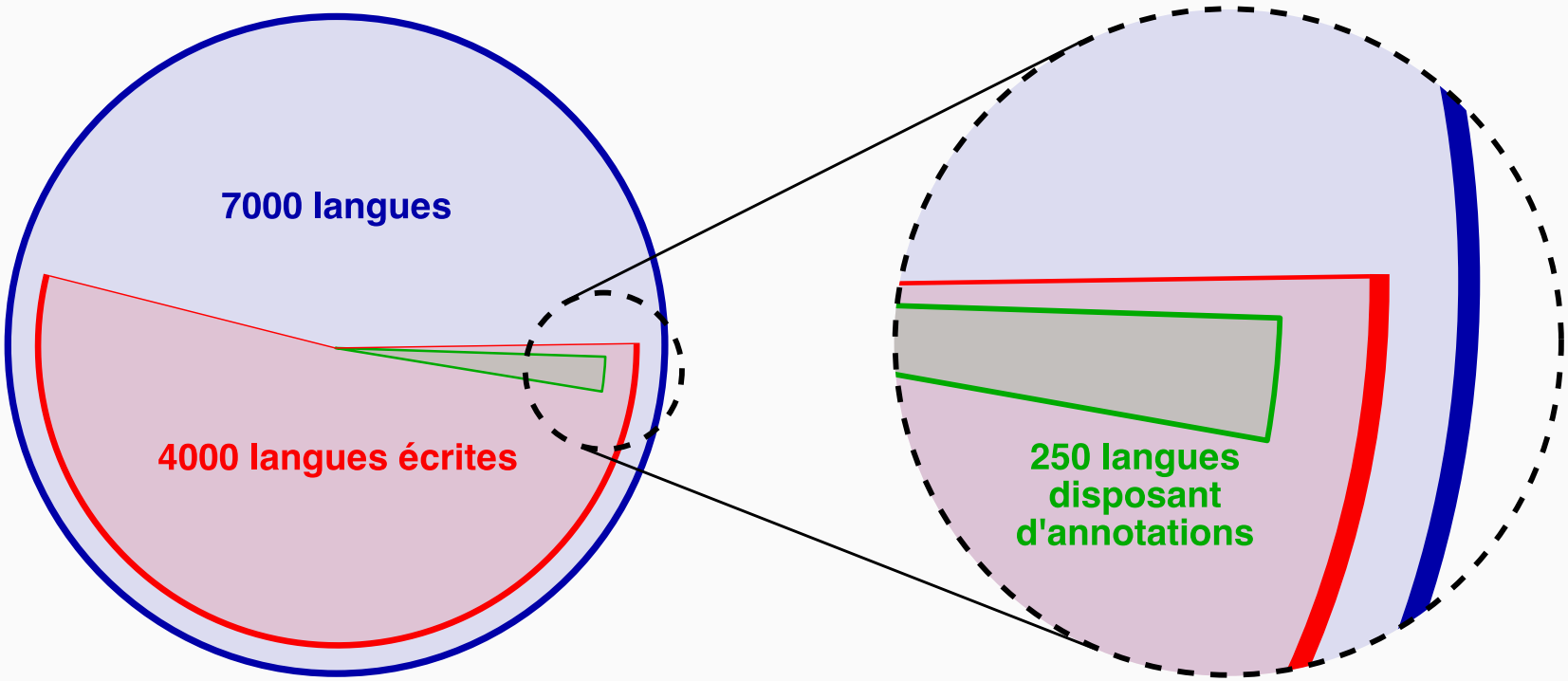
- Seules quelques langues disposent d'annotations (Joshi et al. (2020))



⇒ Est-il possible d'utiliser des ressources d'un ensemble de langues M pour prédire des annotations pour une langue L ?

Problématique

- Seules quelques langues disposent d'annotations (Joshi et al. (2020))

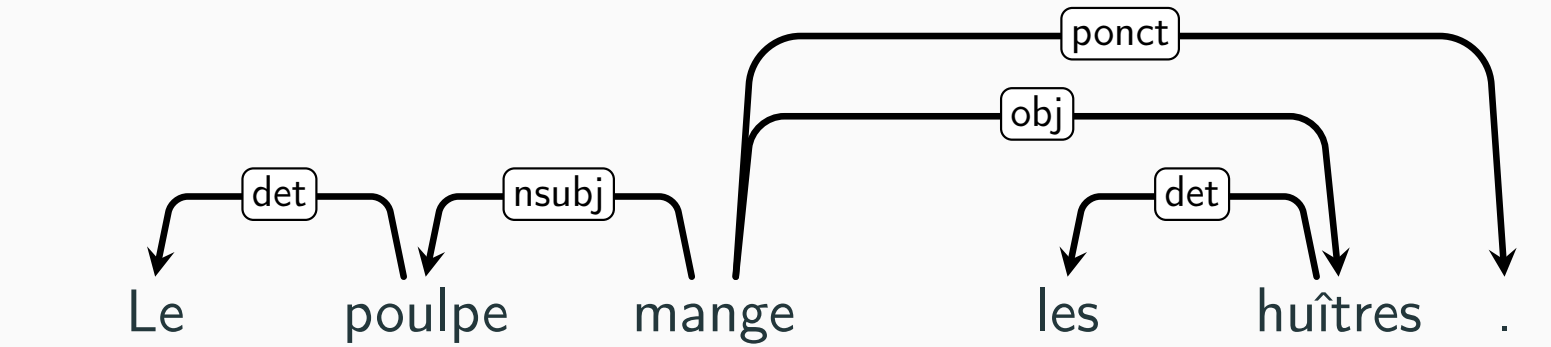


⇒ Est-il possible d'utiliser des ressources d'un ensemble de langues M pour prédire des annotations pour une langue L ?

⇒ Existe-t-il des méthodes facilitant le transfert de connaissances entre langues ?

Une annotation, c'est quoi ?

Arbre syntaxique



POS	DET	NOM	VERB	DET	NOM	PONCT
Morpho	Def|Masc|Sing|Art	Masc|Sing	Ind|Sing|3|Pres|Fin	Def|Masc|Plur|Art	Fem|Plur	_

Quelles solutions possibles ?

- Les Universal Dependencies (UD)
- Représentations multilingues du lexique
- Ressources typologiques (WALS)
- Proximité entre langues

Quelles solutions possibles ?

- Les Universal Dependencies (UD)
- Représentations multilingues du lexique
- Ressources typologiques (WALS)
- Proximité entre langues

Le problème des annotations

- Français : NOM, VRB, DET, ...
- Anglais : NOUN, VERB, DET, ...
- Espagnol : SUST, VERBO, DET, ...

Universal Dependencies (UD) (Nivre et al. (2016))

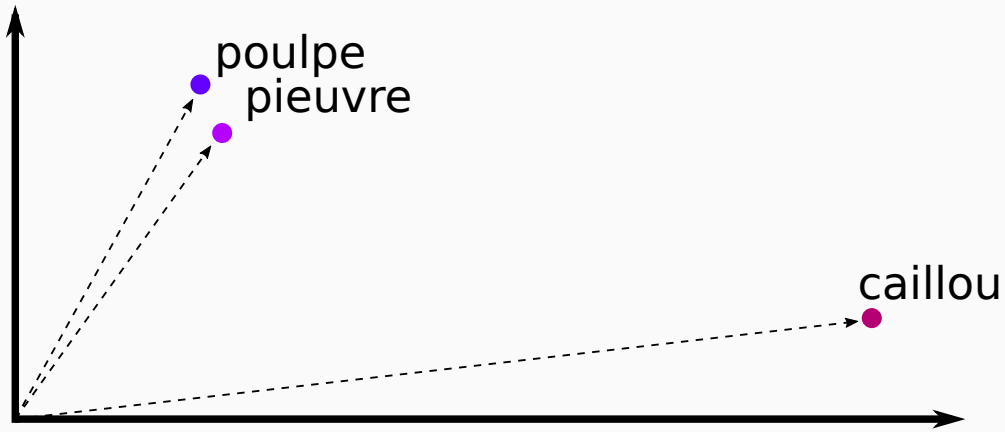
- Représentation qui se veut universelle des relations de dépendances, des parties du discours et des traits morphologiques
- Corpus disponibles pour de nombreuses langues, basés sur un guide d'annotation commun
- Ne résout pas le problème du lexique
- Utilisation de la Version 2.0

Quelles solutions possibles ?

- Les Universal Dependencies (UD)
- Représentations multilingues du lexique
- Ressources typologiques (WALS)
- Proximité entre langues

Représentation des mots

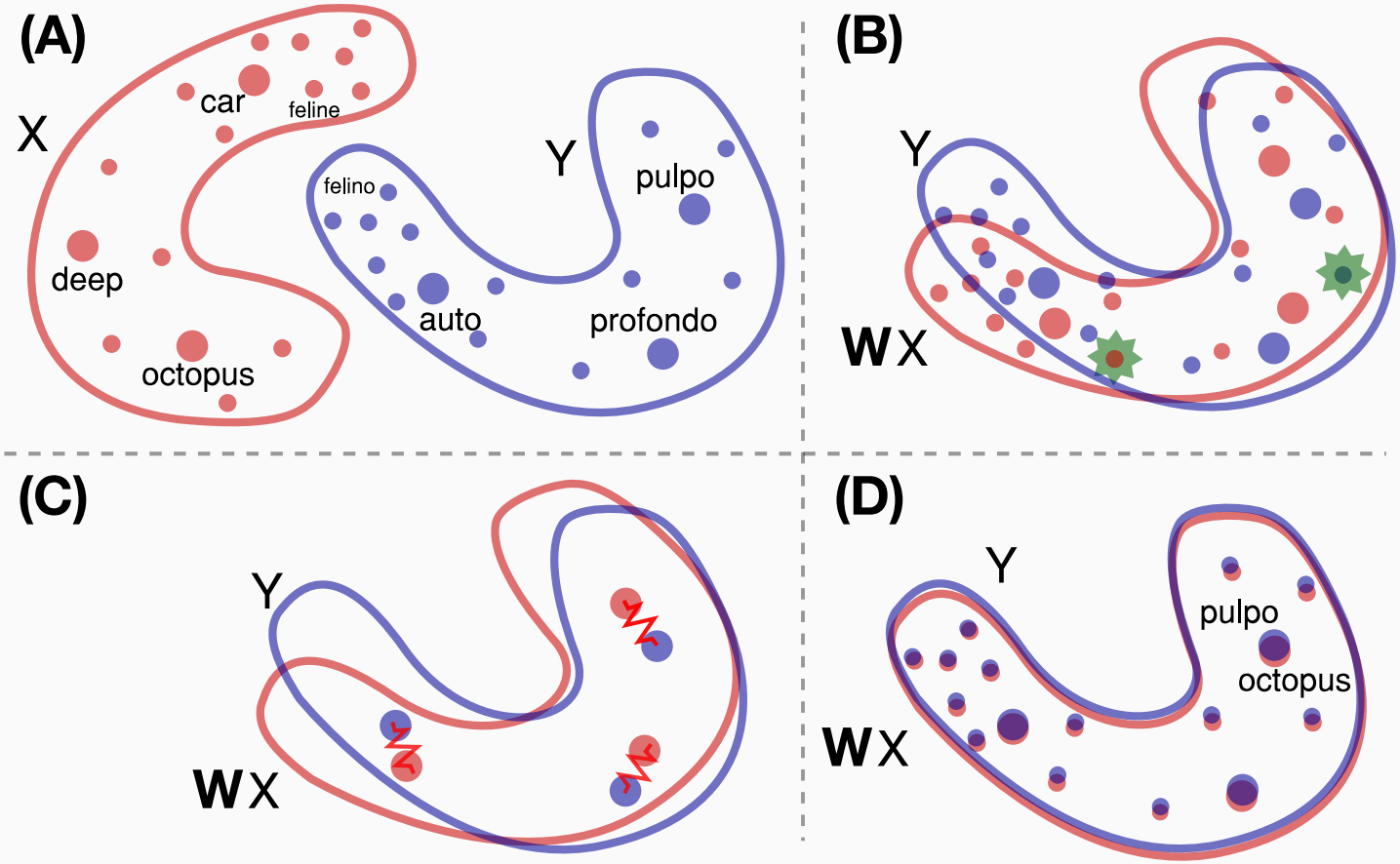
Utilisation de plongements de mots



$\Rightarrow$ Des mots ayant des vecteurs proches ont un sens proche

Représentation des mots multilingues

Il est possible d'aligner les plongements de mots de différentes langues dans un espace commun Mikolov et al. (2013); Conneau et al. (2017)



Représentation des caractères

Utilisation de ce qu'on appelle un modèle de caractères

- Représentation vectorielle à l'aide de plongements
- Construits à partir des préfixes et des suffixes
- Appris lors de l'entraînement (Bi-LSTM)
- Partage possible entre langues :
  - Chat $\Rightarrow$ Chat**s** (français)
  - Cat $\Rightarrow$ Cat**s** (anglais)
  - Gat $\Rightarrow$ Gat**s** (catalan)

Quelles solutions possibles ?

- Les Universal Dependencies (UD)
- Représentations multilingues du lexique
- Ressources typologiques (WALS)
- Proximité entre langues

World Atlas of Language Structures (WALS) Dryer and Haspelmath (2013)

- Typologie des langues
  - $\Rightarrow$ Ordre langue : Sujet-Verbe-Objet (SVO) ?
    Sujet-Objet-Verbe (SOV) ?
- Informations Phonologiques, Grammaticales et Lexicales
- Décrit 2 676 langues grâce à 192 traits
- Un vecteur de 22 traits est extrait pour décrire chaque langue nommé W_{22} (à la manière de Naseem et al. (2012))

World Atlas of Language Structures (WALS) Dryer and Haspelmath (2013)

- Typologie des langues
  - $\Rightarrow$ Ordre langue : Sujet-Verbe-Objet (SVO) ?
    Sujet-Objet-Verbe (SOV) ?
- Informations Phonologiques, Grammaticales et Lexicales
- Décrit 2 676 langues grâce à 192 traits
- Un vecteur de 22 traits est extrait pour décrire chaque langue nommé W_{22} (à la manière de Naseem et al. (2012))

SVO	SV	VO	NOM-ADJ	...

$\Rightarrow$ Le vecteur du français

Quelles solutions possibles ?

- Les Universal Dependencies (UD)
- Représentations multilingues du lexique
- Ressources typologiques (WALS)
- Proximité entre langues

Les expériences

- Analyse syntaxique de la phrase (*Parsing*)
- Étiquetage des POS (*Tagging*)
- Prédiction des POS, de la morphologie ET de l'arbre syntaxique (*TagParsing*)

Analyseur (*Parser*)

Analyseur syntaxique par transitions (Dary and Nasr (2021)) :

- Une pile
- Un buffer
- Un classifieur (réseau de neurones simple)
- Une liste de transitions précédemment réalisées par l'analyseur

Analyseur (*Parser*)

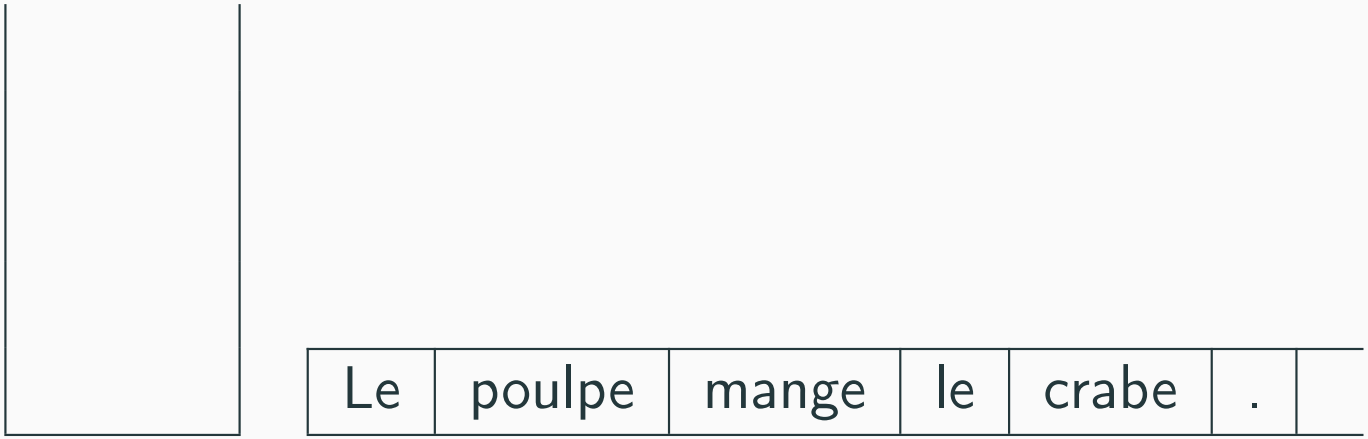
Analyseur syntaxique par transitions (Dary and Nasr (2021)) :

- Une pile
- Un buffer
- Un classifieur (réseau de neurones simple)
- Une liste de transitions précédemment réalisées par l'analyseur

Les transitions possibles :

- `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 empile le mot courant

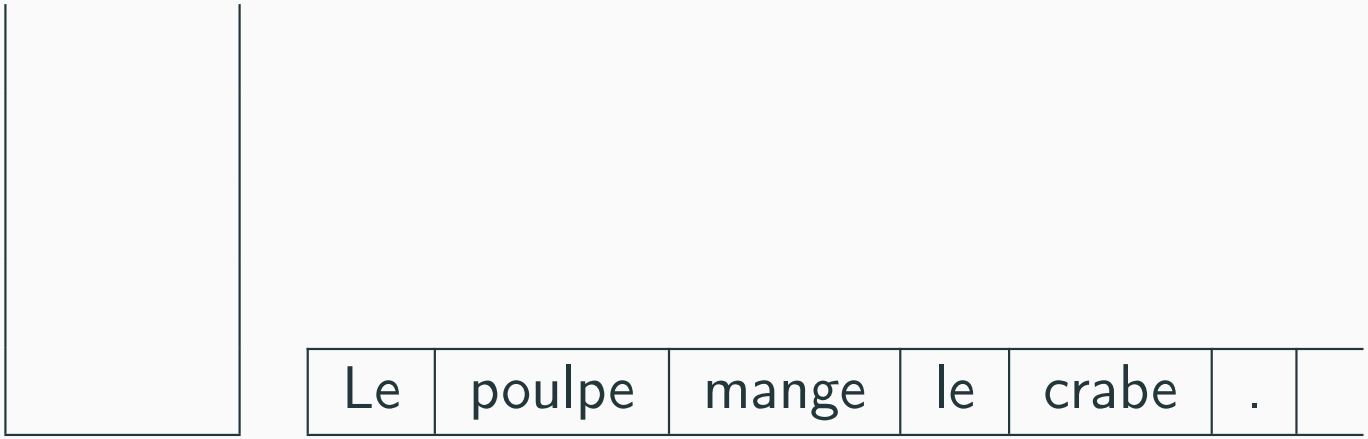
Analyseur (*Parser*)



Le poulpe mange le crabe .

Analyseur (*Parser*)

Classifieur : *SHIFT*



Le poulpe mange le crabe .

Analyseur (*Parser*)

Le poulpe mange le crabe .

Analyseur (*Parser*)

Classifieur : $LEFT_{det}$

Le poulpe mange le crabe .

Analyseur (*Parser*)

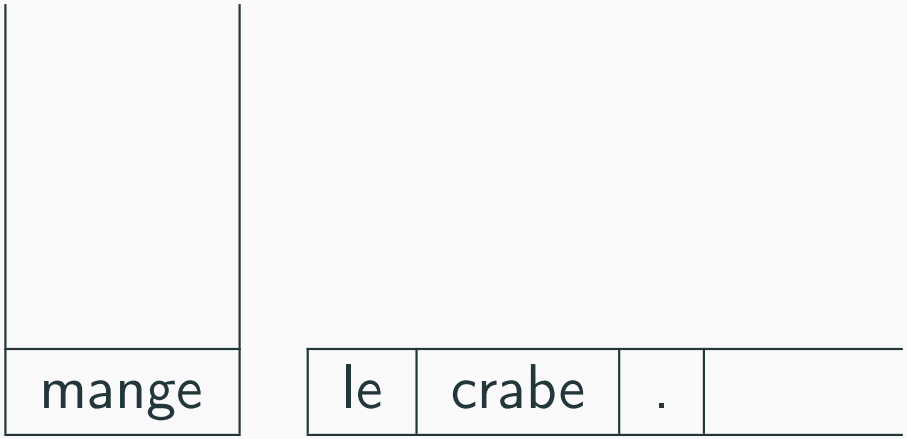
Classifieur : *SHIFT*

Analyseur (*Parser*)

Classifieur : $LEFT_{nsubj}$

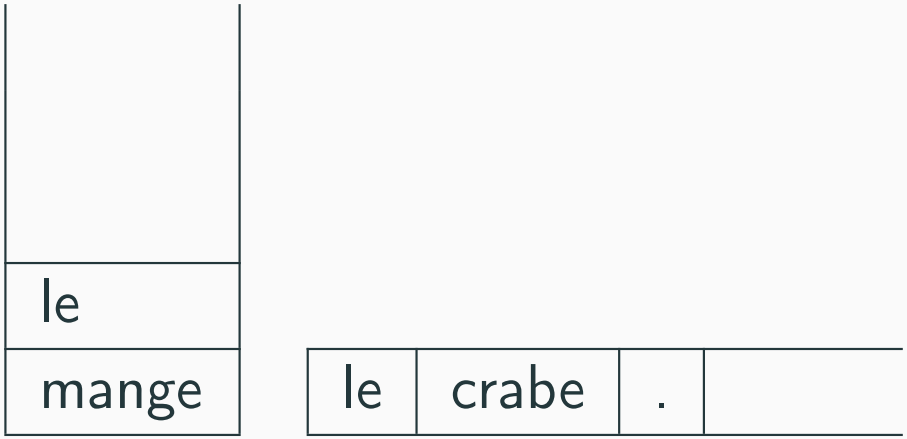
Analyseur (*Parser*)

Classifieur : *SHIFT*



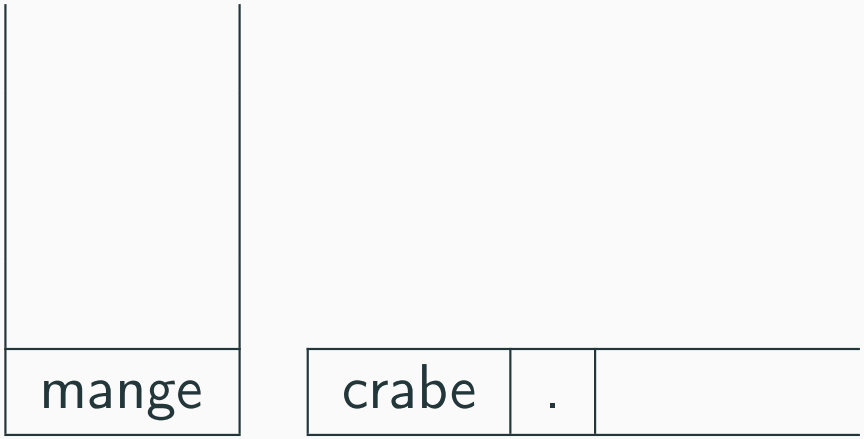
Analyseur (*Parser*)

Classifieur : *SHIFT*

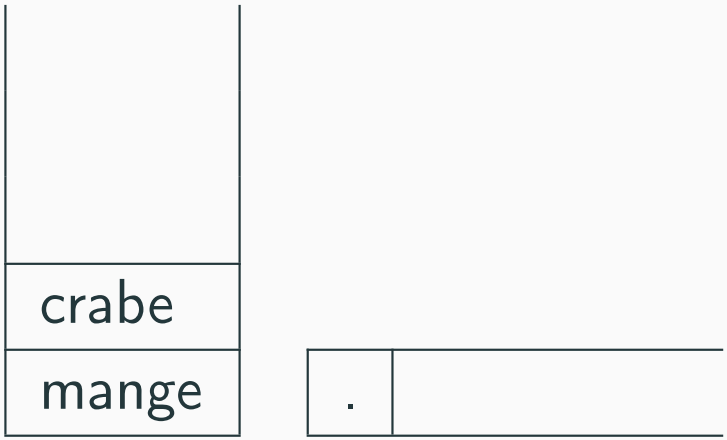


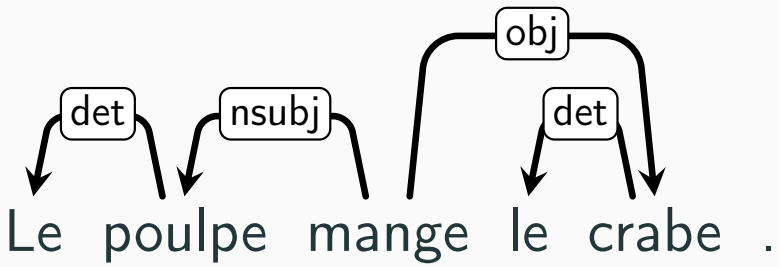
Analyseur (*Parser*)

Classifieur : $LEFT_{det}$

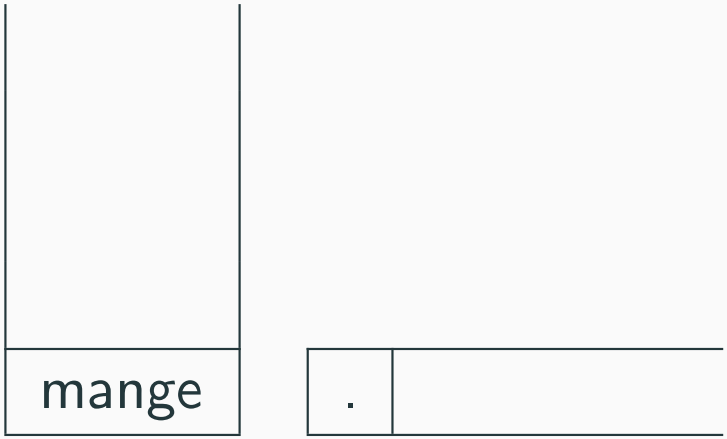


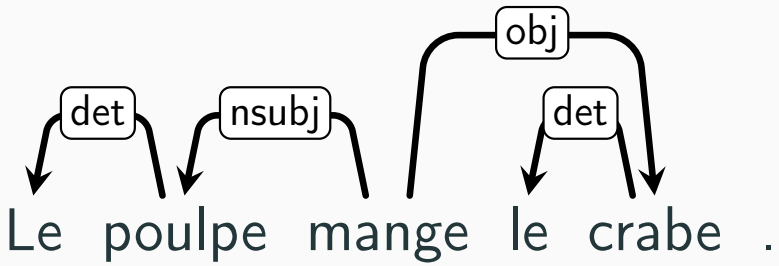
Classifieur : $RIGHT_{obj}$





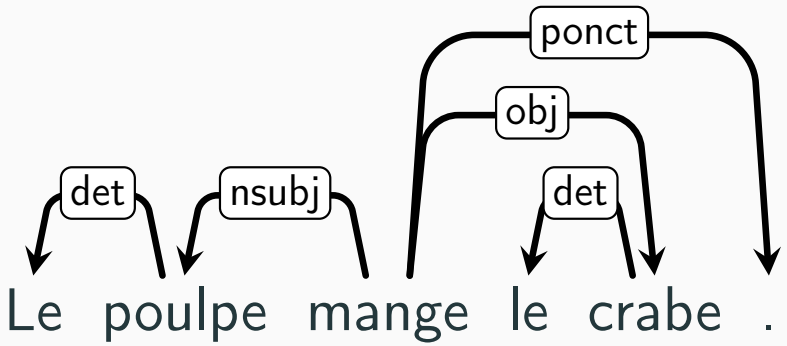
Classifieur : *REDUCE*





Analyseur (*Parser*)

Classifieur : $RIGHT_{ponct}$



Étiqueteur (*Tagger*)

- Idem au parsing, mais sans utiliser de pile :

$\Rightarrow$ Une étiquette est attribuée au mot courant du buffer, puis on passe au suivant

Évaluation Tagging : Exactitude

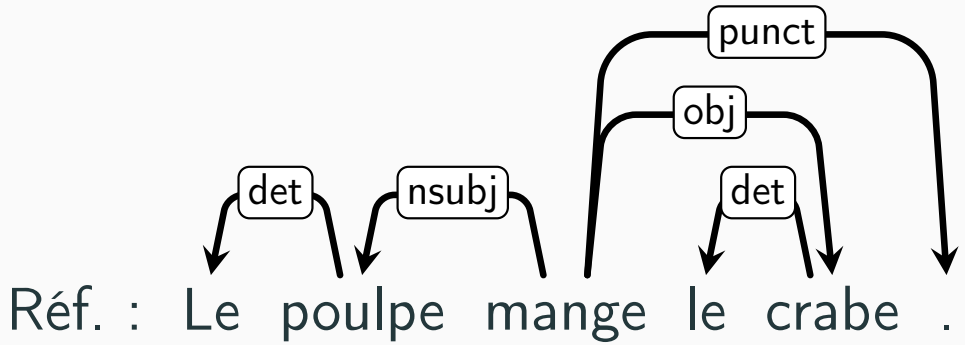
Exactitude : pourcentage de prédictions bien réalisées

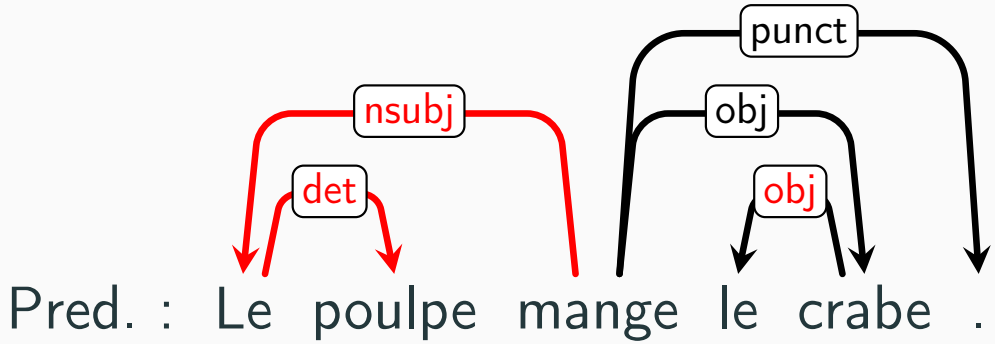
	Le	poulpe	mange	le	crabe	.
Réf.	DET	NOM	VERB	DET	NOM	PONCT
Pred.	VERB	NOM	VERB	NOM	NOM	PONCT

$$\frac{nb\_étiquettes\_bien\_prédites * 100}{nb\_étiquettes\_totales} = \frac{4 * 100}{6} = 66.67\%$$

Évaluation Parsing : LAS

Score d'attachement étiqueté - ou *Labeled Attachment Score* - (LAS) : pourcentage de dépendances bien prédites





$$LAS = \frac{2}{5} = 40\%$$

Évaluation : MACRO Moyenne

MACRO moyenne : moyenne des scores sur toutes les langues

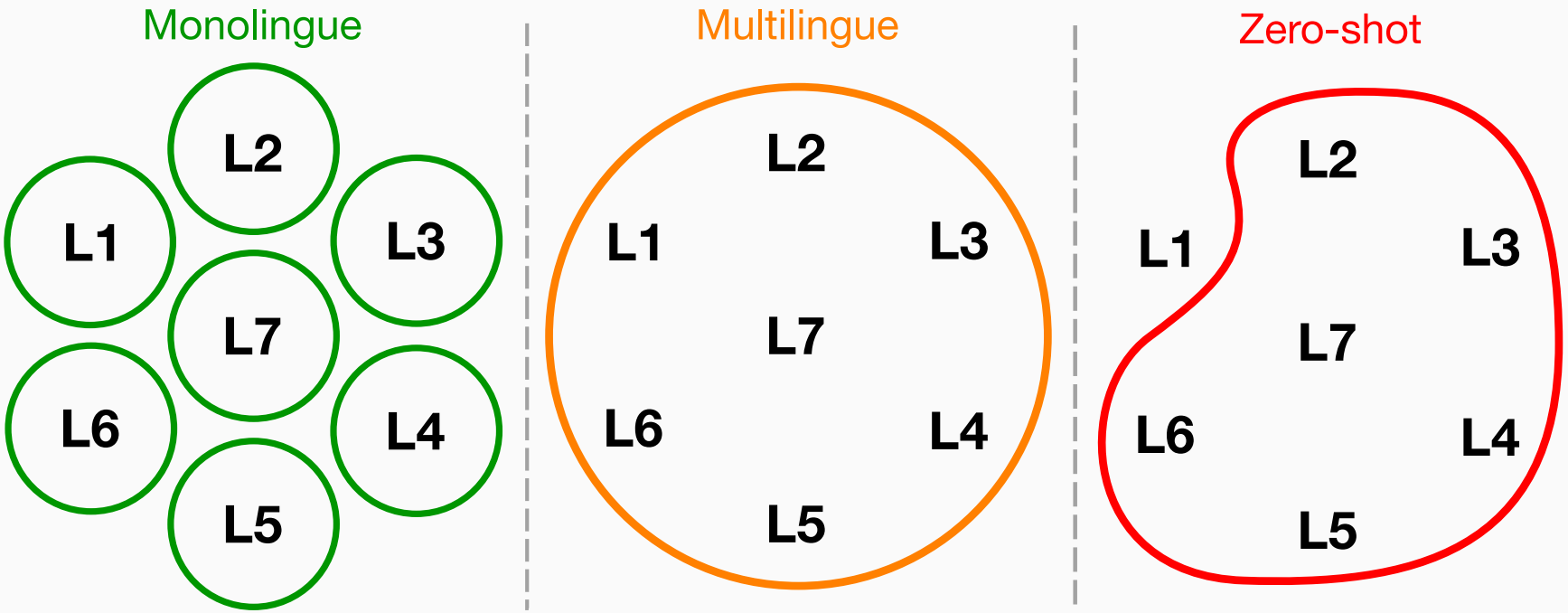
- MACRO LAS
- MACRO Exactitude

Lang.	LAS	Exactitude
Espagnol (es)	90	60
Français (fr)	50	65
Italien (it)	10	50
MACRO Moyenne	50	58.33

Les corpus

- Quelles données d'entraînement ? $\Rightarrow$ Corpus de UD
- 38 langues
- Équilibrage du nombre de tokens : seulement 20 000 par langue

Configurations d'entraînement



- *Mono* : Expériences monolingues
- *Multi* : Expériences multilingue
- ***ZS*** : Expériences *Zero-Shot*

$\Rightarrow$ WALS, plongements de mots, modèle de caractères, ...

Hypothèses

- L'utilisation de traits issus du WALS permettent de partager des connaissances entre langues

Hypothèses

- L'utilisation de traits issus du WALS permettent de partager des connaissances entre langues
- L'utilisation d'un modèle de caractères pourrait permettre d'identifier des marqueurs morphologiques communs entre langues et permettre d'aider au partage de connaissances
  - Chat $\Rightarrow$ Chat**s** (français)
  - Cat $\Rightarrow$ Cat**s** (anglais)
  - Gato $\Rightarrow$ Gato**s** (espagnol)

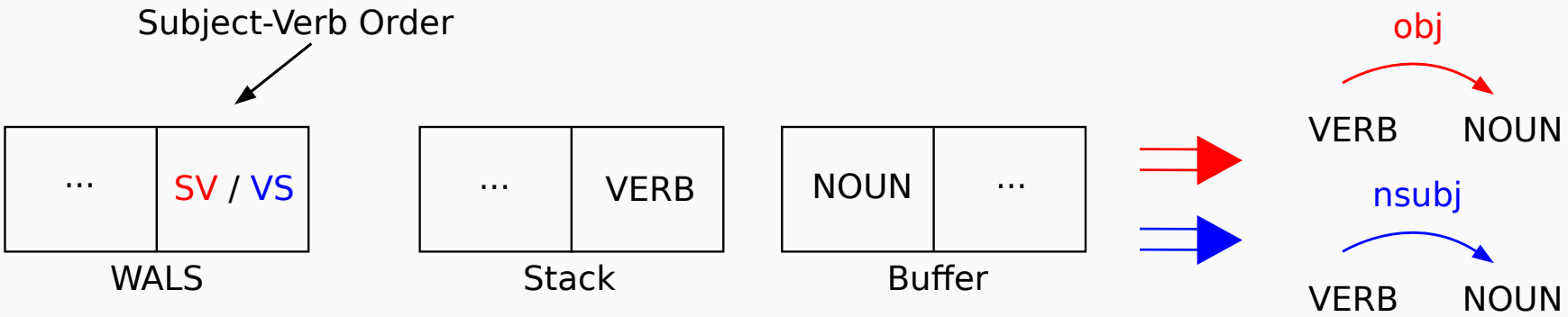
Hypothèses

- L'utilisation de traits issus du WALS permettent de partager des connaissances entre langues
- L'utilisation d'un modèle de caractères pourrait permettre d'identifier des marqueurs morphologiques communs entre langues et permettre d'aider au partage de connaissances
  - Chat $\Rightarrow$ Chat**s** (français)
  - Cat $\Rightarrow$ Cat**s** (anglais)
  - Gato $\Rightarrow$ Gato**s** (espagnol)
- La présence d'une langue proche à l'entraînement est importante dans une situation de zero-shot (***ZS***)

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

La WALS pour aider au partage de connaissances

Le WALS permet-il un meilleur partage de connaissance entre langues ?



Conditions expérimentales

- Analyse syntaxique (Parsing)
  $\Rightarrow$ Scores : LAS
- Délexicalisées (forme du mot inaccessible)
- *Multi* et ***ZS***
- Intégration de W_{22}

Apport du WALS : analyse syntaxique délexicalisée

	Sans W_{22}	Avec W_{22}	Avec ID	Δ	σ
Multi	63.64	67.64	68.25	4.00	2.58
ZS	51.86	53.80	-	1.95	4.60

Apport du WALS : analyse syntaxique délexicalisée

	Sans W_{22}	Avec W_{22}	Avec ID	Δ	σ
Multi	63.64	67.64	68.25	4.00	2.58
ZS	51.86	53.80	-	1.95	4.60

- *Multi* :
  - W_{22} permet d'augmenter la MACRO LAS de 4 points
  - Toutes les langues bénéficient de l'utilisation de W_{22}
  - Moins bon qu'un simple identifiant de la langue

Apport du WALS : analyse syntaxique délexicalisée

	Sans W_{22}	Avec W_{22}	Avec ID	Δ	σ
Multi	63.64	67.64	68.25	4.00	2.58
ZS	51.86	53.80	-	1.95	4.60

- *Multi* :
  - W_{22} permet d'augmenter la MACRO LAS de 4 points
  - Toutes les langues bénéficient de l'utilisation de W_{22}
  - Moins bon qu'un simple identifiant de la langue
- ***ZS*** :
  - W_{22} permet d'augmenter la MACRO LAS de 1.95 points
  - Mais 11 langues ne bénéficient pas de l'utilisation de W_{22}

Apport du WALS : analyse syntaxique délexicalisée

Le WALS est toujours utile en multilingue, mais son utilité en zero-shot est plus ambiguë.

Le WALS est-il un simple identifiant ?

Apport du WALS : analyse syntaxique délexicalisée

Le WALS est toujours utile en multilingue, mais son utilité en zero-shot est plus ambiguë.

Le WALS est-il un simple identifiant ?

Le vecteur du WALS permet-il au classifieur d'avoir accès à l'information de l'identité de la langue en cours de traitement ?

Apport du WALS : analyse syntaxique délexicalisée

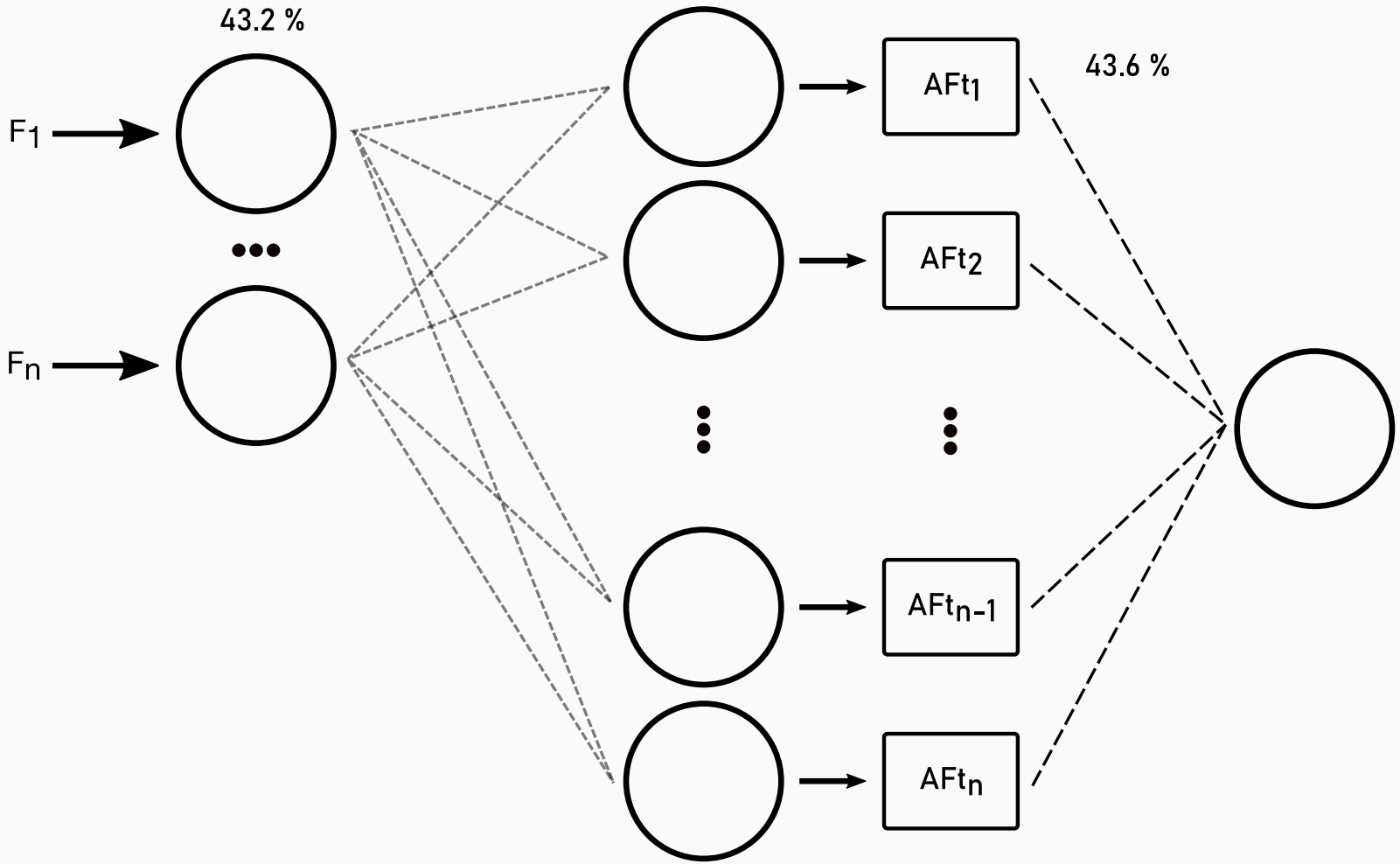
Le WALS est toujours utile en multilingue, mais son utilité en zero-shot est plus ambiguë.

Le WALS est-il un simple identifiant ?

Le vecteur du WALS permet-il au classifieur d'avoir accès à l'information de l'identité de la langue en cours de traitement ?

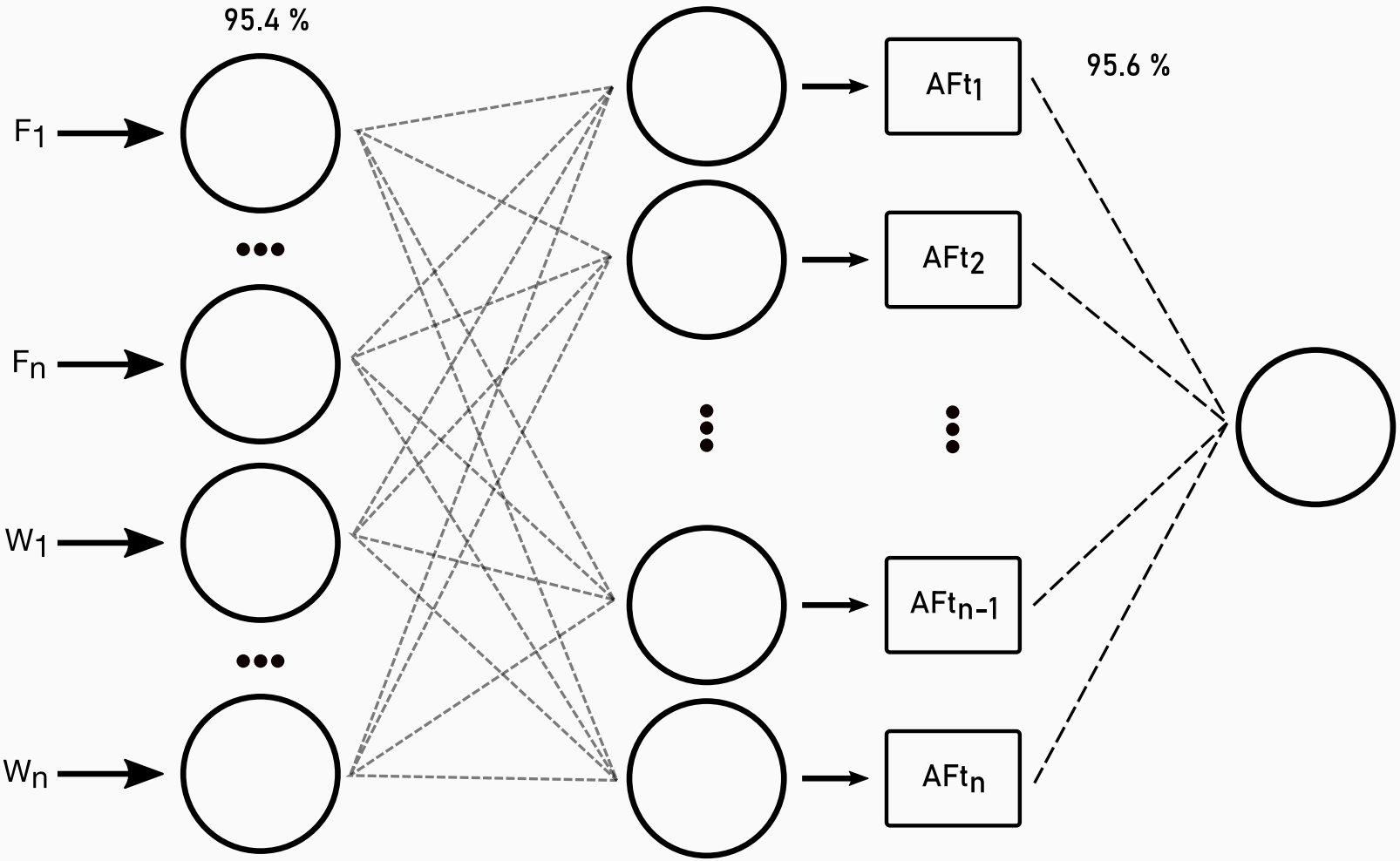
$\Rightarrow$ Probing : Entraînement d'un classifieur pour prédire l'ID de langue

Le WALS : un simple identifiant de la langue ?



⇒ Si l'information de la langue traverse la couche cachée, c'est que cette information est utile

Le WALS : un simple identifiant de la langue ?



⇒ Si l'information de la langue traverse la couche cachée, c'est que cette information est utile

Le WALS : un simple identifiant de la langue ?

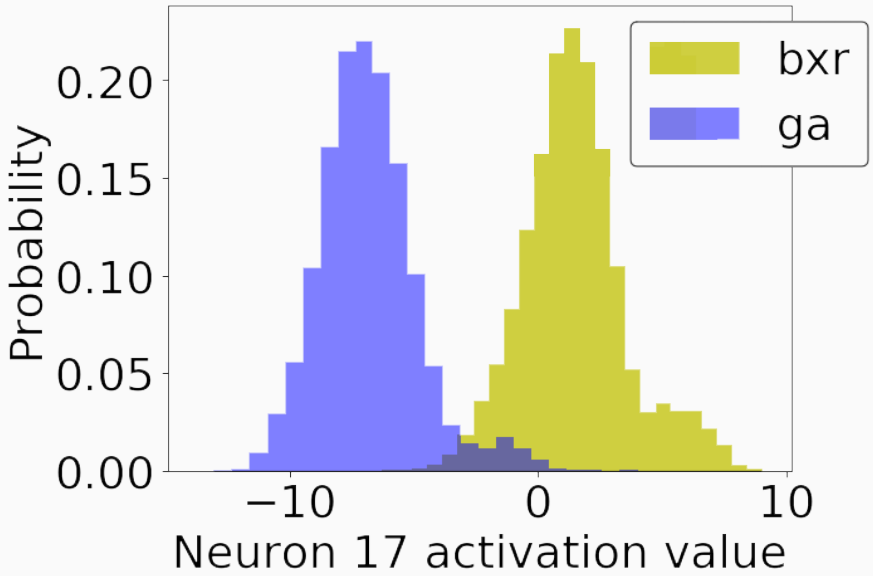
Configuration	Entraîné sur	Précision
Multi	Entrées	43.2
Multi W_{22}	Entrées	95.4
Multi	Activations couche cachée	43.6
Multi W_{22}	Activations couche cachée	95.6

La langue est identifiable au moment de la couche de décision avec W_{22}

$\Rightarrow$ Mais W_{22} n'est-il qu'un ID ?

Le WALS : un simple identifiant de la langue ?

- Mesure des activations après la couche cachée pour chacun des 1 000 neurones, pour chaque langue



- Calcul de la Jensen-Shannon Divergence (JSD) au niveau des neurones entre des paires de langues

$\Rightarrow$ Si la JSD est basse, alors les langues sont proches

Le WALS : un simple identifiant de la langue ?

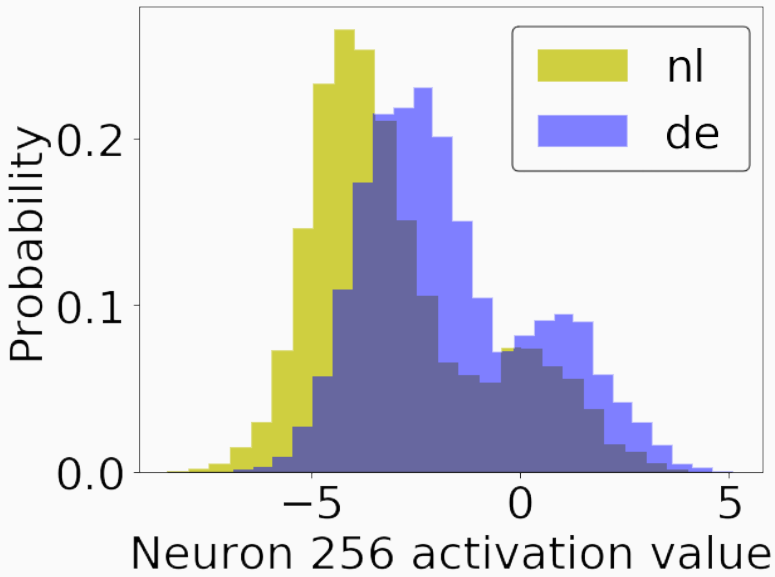
- Néerlandais-Allemand (nl-de) : vecteurs WALS identiques
- Bouriate-Irlandais (bxr-ga) : vecteurs WALS très différents

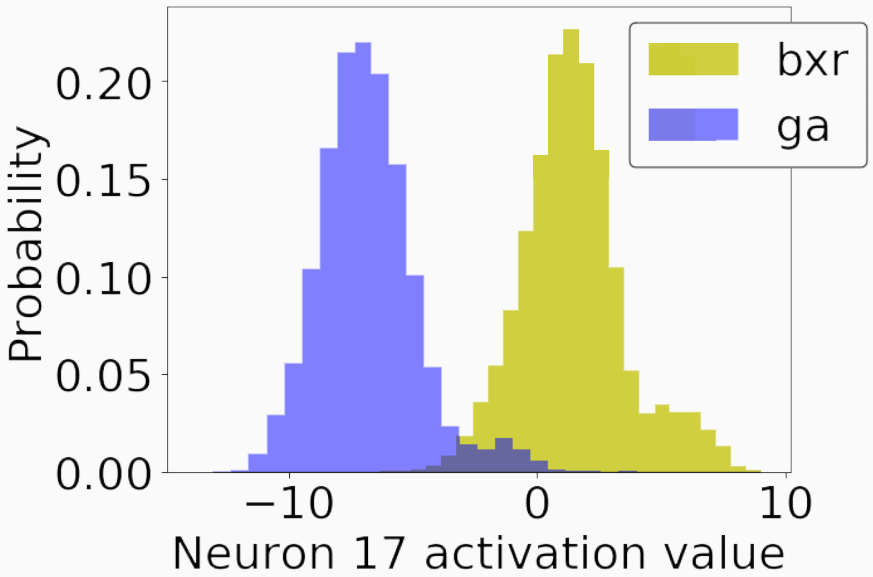
		Moyenne			Max		
L1	L2	*Multi*	*Multi* W_{22}		*Multi*	*Multi* W_{22}	
nl	de	0.860	0.854	↘	1.027	0.940	↘
bxr	ga	0.890	1.160	↗↗	1.600	4.888	↗↗

$\Rightarrow$ Vecteur WALS proches $\rightarrow$ Comportements similaires de l'analyseur

Le WALS : un simple identifiant de la langue ?

- Distributions des activations le neurone ayant la plus grande JSD pour les paires (nl, de) et (bxr, ga)





⇒ Le WALS permet d'amplifier la similarité (ou la différence) du comportement du parser pour des langues proches (ou éloignées)

WALS : conclusions

- Toujours utile en *Multi*
- Utilité moins évidente en ***ZS***
- Sert d'identifiant de la langue : peut être préjudiciable en ***ZS***
- Permet de rapprocher et d'éloigner certaines langues, et permet donc dans une certaine mesure d'aider aux partages de connaissances

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

Modèle de caractères

Le modèle de caractères (MDC) devrait permettre un meilleur partage de connaissances entre langues, en particulier en aidant à identifier des marqueurs morphologiques communs.

- Chat $\Rightarrow$ Chat**s** (français)
- Cat $\Rightarrow$ Cat**s** (anglais)
- Gato $\Rightarrow$ Gato**s** (espagnol)

Conditions expérimentales

- Étiquetage des POS (Tagging)

  $\Rightarrow$ Scores : Exactitude

- Utilisation systématique des plongements de mots
- *Mono*, *Multi* et ***ZS***

Modèle de caractères pour l'étiquetage des POS

	Sans MDC	Avec MDC	Δ	σ
Mono	86.71	92.02	5.31	2.18
Multi	85.36	90.81	5.44	2.01
ZS	65.15	64.16	-0.99	6.63

- *Mono* & *Multi* :
  - Bénéfique pour toutes les langues
  - Les langues morphologiquement riches bénéficient plus du MDC
    - $\Rightarrow$ +6.72 pour le finnois (fi), +7.80 pour le hongrois (hu), +8.34 pour l'estonien (et), +8.90 pour le turc (tr) en *Multi*
    - $\Rightarrow$ Exception : arabe, "seulement" +2.94

Modèle de caractères pour l'étiquetage des POS

	Sans MDC	Avec MDC	Δ	σ
Mono	86.71	92.02	5.31	2.18
Multi	85.36	90.81	5.44	2.01
ZS	65.15	64.16	-0.99	6.63

- ***ZS*** :
  - Bénéfique pour 26/38 langues
  - Mais diminution des résultats en moyenne

Modèle de caractères pour l'étiquetage des POS

Lang.	*ZS* $-c$	*ZS* $+c$	Δ
Hindi (hi)	60.55	35.67	-24.88
Japonais (ja)	45.72	33.28	-12.44
Ourdou (ur)	56.66	38.97	-17.69
Perse (fa)	68.72	56.24	-12.48
Arabe (ar)	66.88	60.23	-6.65
Espagnol (es)	85.81	88.69	2.88
Catalan (ca)	77.39	83.48	6.09
MACRO moy.	65.15	64.16	-0.99

- Les langues souffrant le plus de l'utilisation du MDC sont :
  - Les langues isolées (alphabet unique)
  - Les binômes/trinômes problématiques (ar-ur-fa)
- Les langues qui tirent profit du MDC sont des langues bien représentées (espagnol-catalan)

Modèle de caractères : conclusions

- Toujours bénéfique en *Mono* et *Multi*
- Préjudiciable en ***ZS***
- Très préjudiciable pour les langues isolées (à alphabet unique)
- Bénéfique pour les langues ayant déjà des scores assez haut, et étant bien représentées dans l'ensemble d'entraînement

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

Variabilité des résultats en *ZS*

Grande variabilité des résultats des expériences ***ZS*** selon les langues. Cela pourrait-il venir de l'isolement de la langue testée par rapport à l'ensemble d'entraînement ?

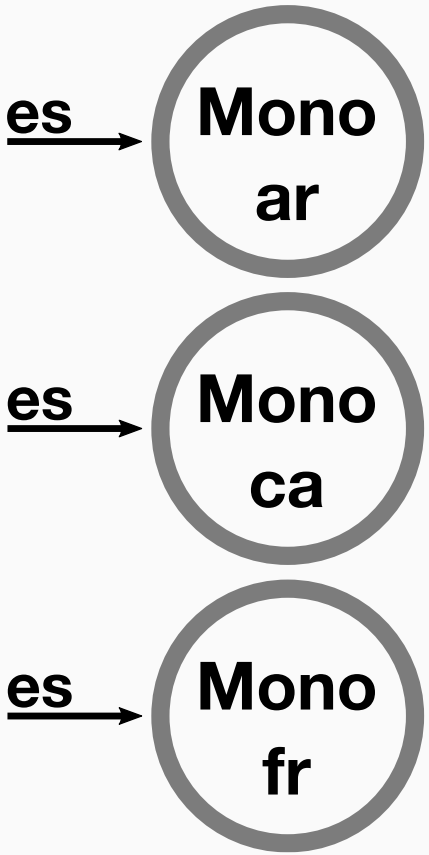
$\Rightarrow$ Comment mesurer l'isolement d'une langue dans un ensemble de langue ?

- Mesure empirique : la Langue la Plus Proche (LPP)
- Mesure théorique : issue du WALS, l'Indice de Connectivité (CI)

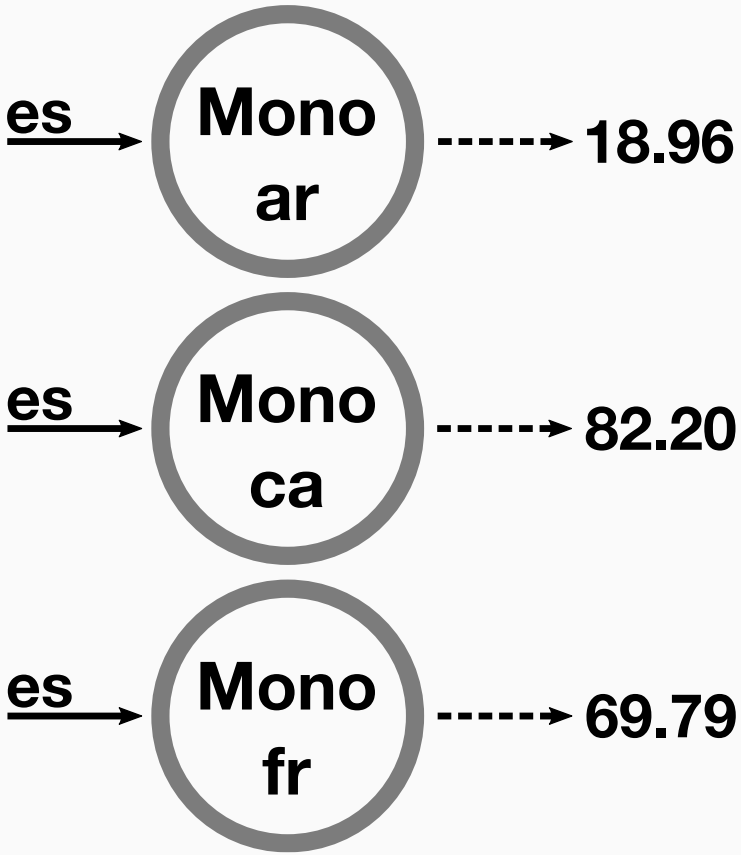
- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
  - Mesure empirique
  - Mesure théorique issue du WALS
- Conclusions et Perspectives

La Langue la Plus Proche (LPP)

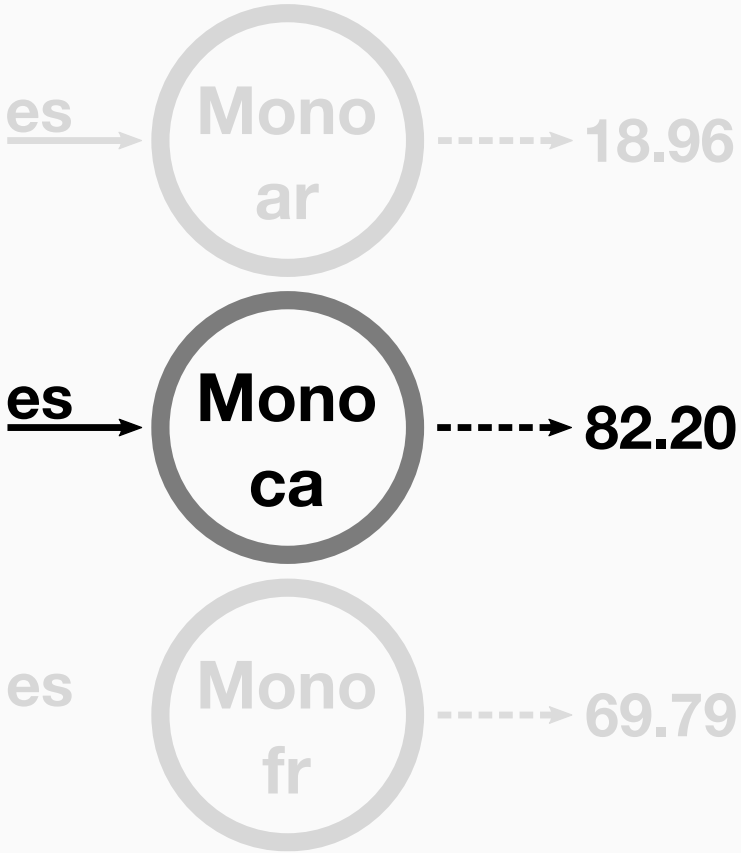
La Langue la Plus Proche (LPP)



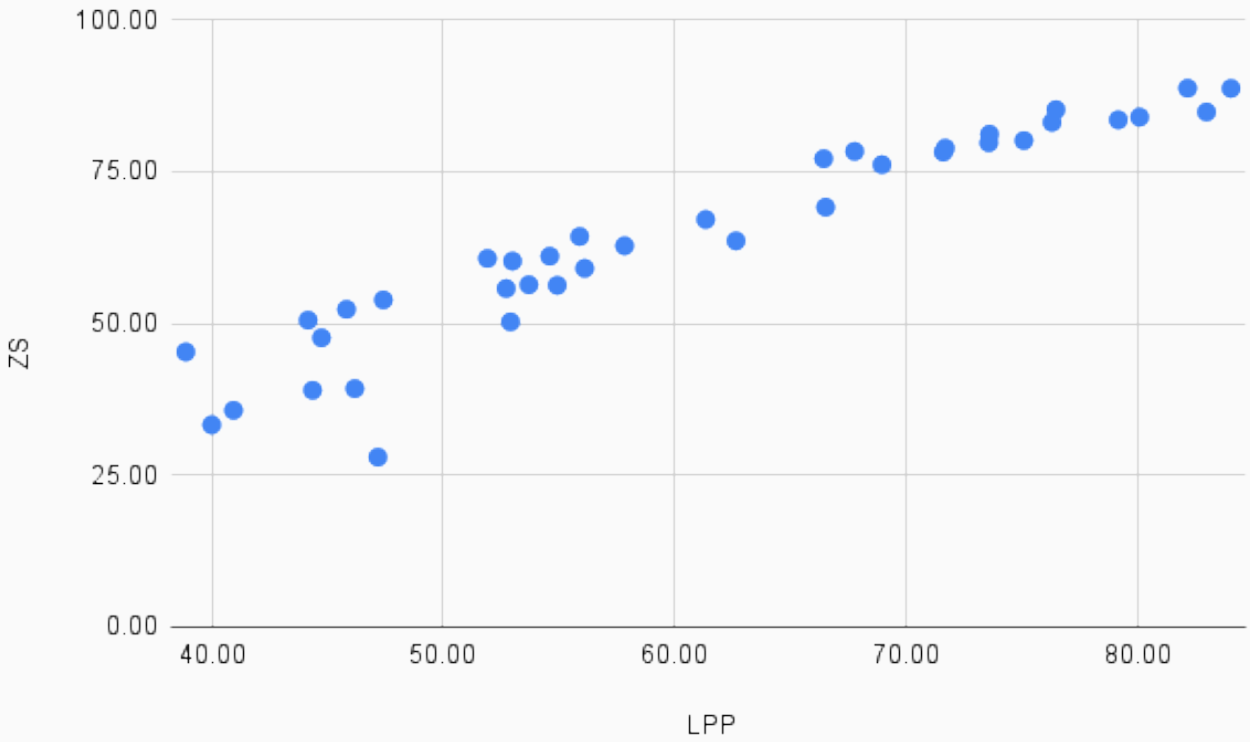
La Langue la Plus Proche (LPP)



La Langue la Plus Proche (LPP)



La Langue la Plus Proche (LPP)



$\Rightarrow$ 0.95 de corrélation entre la LPP et les scores obtenus en ***ZS*** !

- Permet de donner à l'avance une estimation de l'exactitude des prédictions

Problèmes de la LPP

- Nécessite des données annotées pour la calculer
- Coûte cher en termes de temps et de calcul

$\Rightarrow$ On aimerait une mesure indépendante d'annotations, rapide à calculer, et permettant d'estimer la qualité des prédictions dans un cadre de zero-shot

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
  - Mesure empirique
  - Mesure théorique issue du WALS
- Conclusions et Perspectives

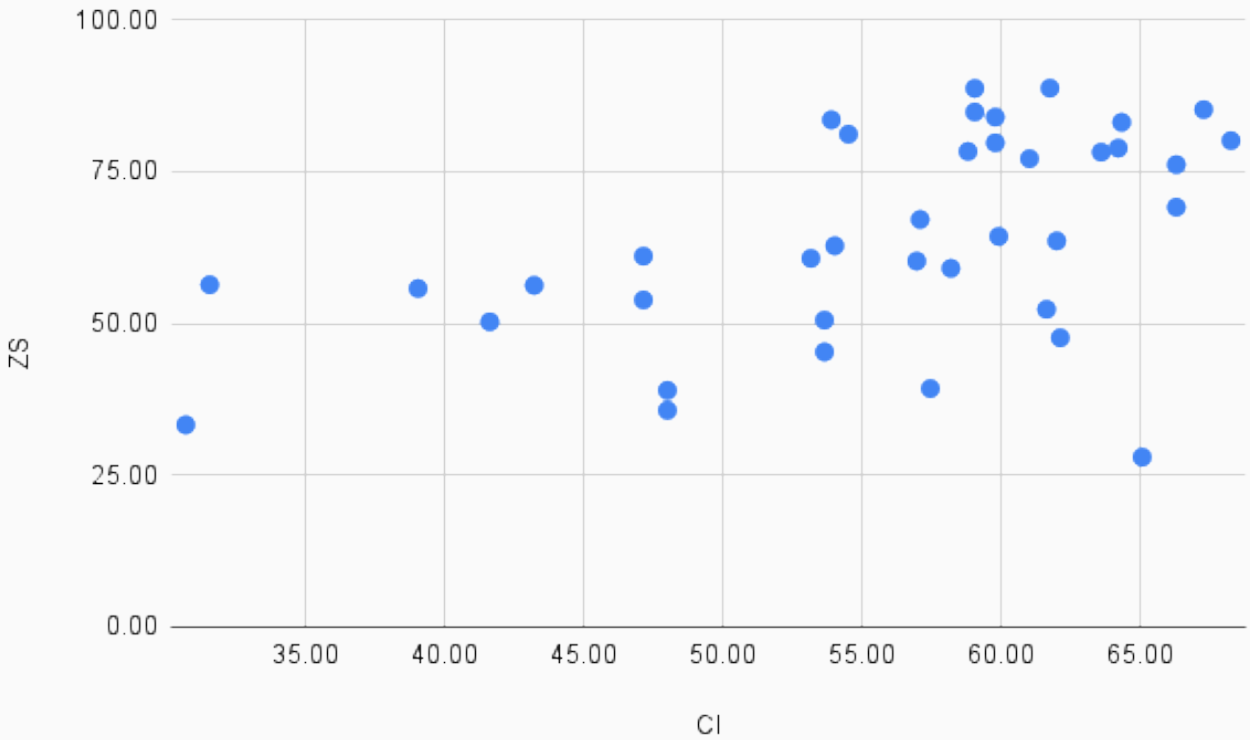
L'indice de connectivité (CI)

- Mesure d'isolement définie à partir du WALS

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

- L'indice de connectivité (CI) d'une langue L est le nombre moyen de valeurs de traits qu'elle partage avec les autres langues
- Plus $CI(L)$ est bas, plus la langue est isolée

L'indice de connectivité (CI)



$\Rightarrow$ 0.50 de corrélation entre le CI et les scores obtenus en ***ZS***

Proximité entre langues : conclusions

- La présence d'une langue "proche" à l'entraînement est LE critère permettant l'obtention de bons résultats en zero-shot
- Difficile d'approximer la mesure empirique à l'aide d'une mesure issue du WALS

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

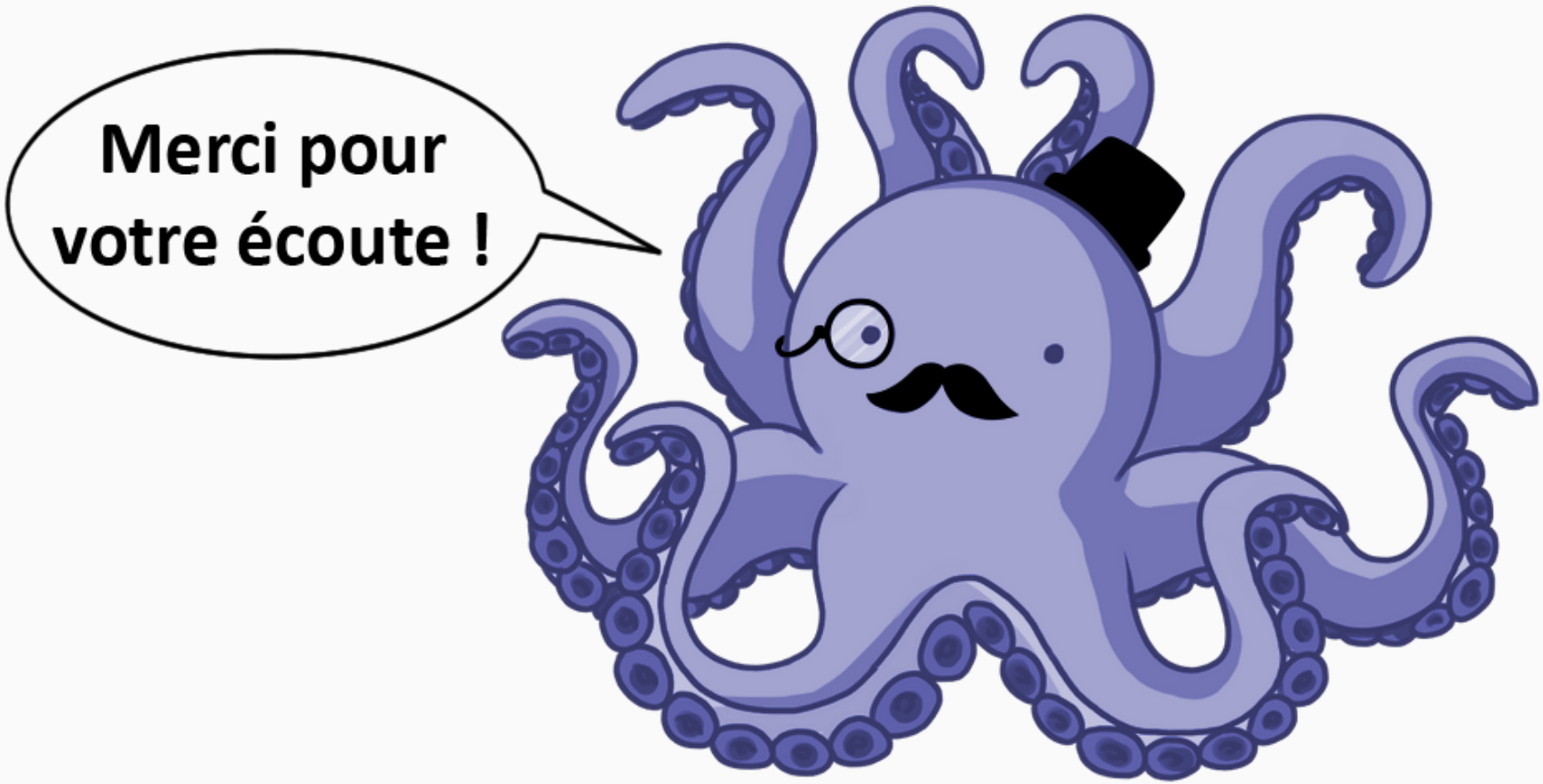
Conclusions

Est-il possible d'aider le partage de connaissances entre langues dans des systèmes d'apprentissages multilingues et zero-shot ?

- ⇒ Le WALS peut permettre un meilleur traitement des langues isolées.
- ⇒ L'utilisation des caractères des mots peut porter préjudice à certaines langues
- ⇒ Importance de la présence d'une langue proche dans l'ensemble d'entraînement
- ⇒ L'hypothèse de Chomsky (Chomsky and Lasnik (2008)) n'a pas pu être démontrée

Perspectives

- Utilisation de M-BERT (Devlin et al. (2019)) pour les représentations multilingues des mots
- Différentes méthodes d'obtentions de vecteurs du WALS
- Utilisation de Morfessor/Wordpiece pour l'obtention de morphèmes universels


Merci pour
votre écoute !

Familles de langues

- Importance de la présence d'une langue "proche"
- Est-il possible de réduire le nombre de langues d'apprentissage ?

$\Rightarrow$ Constructions de familles de langues naturelles et artificielles

Création de familles

- Création de la distance interne moyenne (MID)
- Conservation des ensembles ayant le plus petit MID, le plus grand, et l'ensemble le plus proche de l'ensemble moyen
  - **Romance** : fr, ca, es, it, pt, ro **(MID = 4.13)**
  - **Slavic** : bg, cs, hr, pl, ru, uk, sl **(MID = 4.19)**
  - **German** : de, da, en, nl, no, sv **(MID = 4.47)**
  - **Close** (artificielle) : bg, en, hr, ru, sl, uk **(MID = 2.93)**
  - **Mean** (artificielle) : ar, bg, ca, cs, el, hi **(MID = 9.73)**
  - **Distant** (artificielle) : ca, cs, de, eu, ga, ja **(MID = 14.13)**

Apprentissage par famille de langue (*Multi*)

Famille	*Multi* (fam)	*Multi*	Δ
German	89.41	89.76	-0.35
Slavic	94.01	93.01	1.00
Romance	94.13	93.56	0.57
Mean	94.71	94.07	0.64
Close	92.90	92.03	0.87
Distant	91.88	90.86	1.02

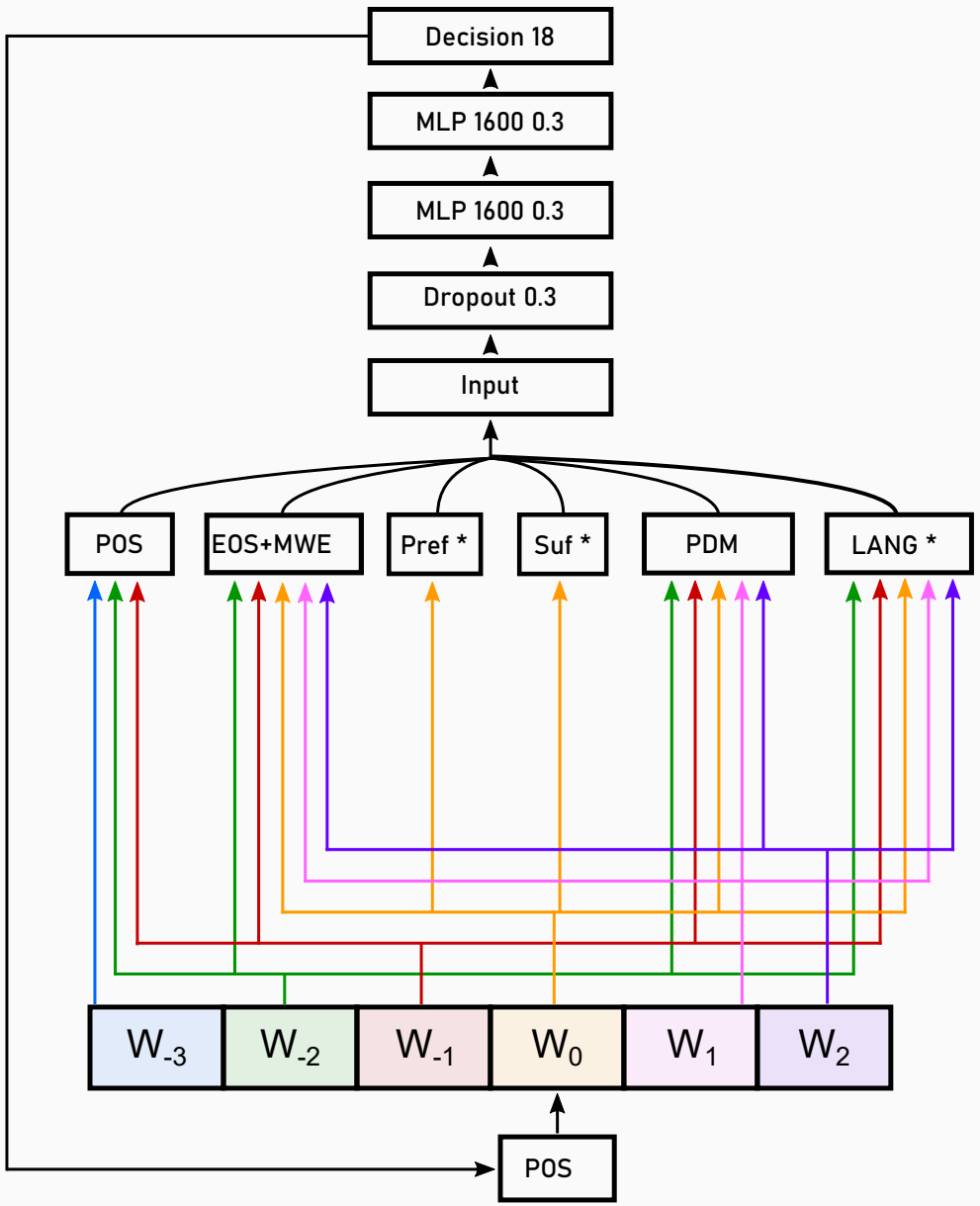
- Augmentation quasi systématique des résultats par rapport au *Multi* classique
- L'ensemble **Distant** plus grosse augmentation
- La réduction du nombre de langues dans les données d'entraînement améliore les résultats ?

Apprentissage par famille de langue (*ZS*)

Famille	*ZS* (fam)	*ZS*	Δ
German	68.39	68.56	-0.17
Slavic	79.31	79.36	-0.05
Romance	76.65	77.99	-1.34
Mean	47.30	64.38	-17.08
Close	92.90	92.03	0.87
Distant	42.16	58.79	-16.63

- Diminution quasi systématique des résultats par rapport au *ZS* classique
- Faible diminution : il est possible d'obtenir presque d'aussi bons résultats avec seulement $\frac{7}{37}^{ème}$ des données
- Importance d'une langue proche à l'entraînement constatée

L'étiqueteur (*tagger*)



Distribution de probabilité des 18 POS

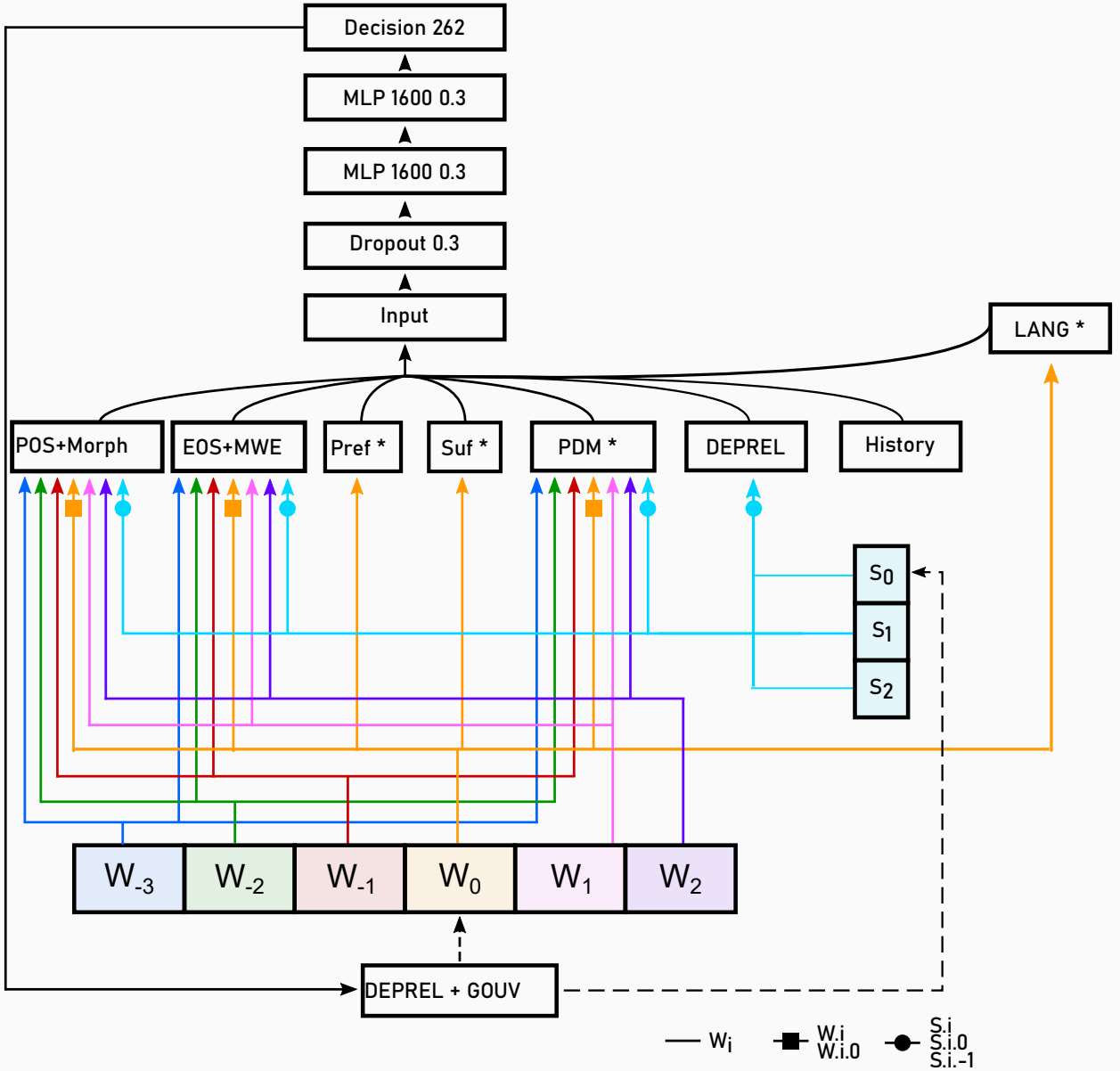
Couches cachées

Concaténation des Bi-LSTMs

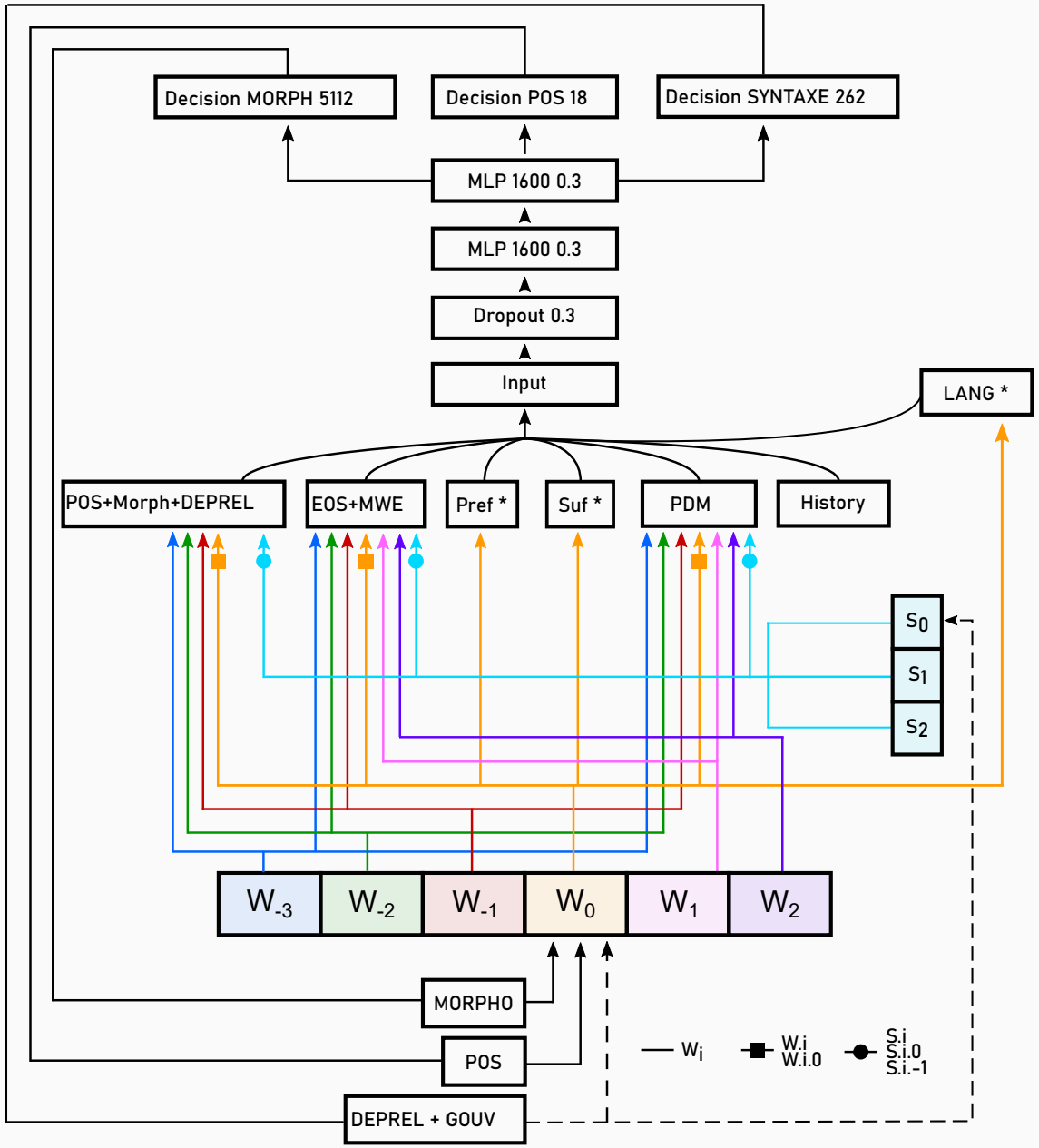
Bi-LSTMs (* optionnels)

Texte

L'analyseur (*parser*)



Le *tagparser*



Apport du WALS : TagParsing

		Sans W_{22}	Avec W_{22}	Avec ID	Δ	σ	min	max
POS	*Multi*	86.94	86.18	87.40	-0.76	7.12	-17.56 hu	16.37 tr
	ZS	61.60	62.82	-	1.23	5.58	-8.58 nno	14.05 ur
Morpho	*Multi*	78.37	75.18	77.68	-0.19	10.97	-39.12 nno	15.24 tr
	ZS	42.24	43.67	-	1.44	9.74	-26.21 nno	32.84 vi
LAS	*Multi*	59.36	58.42	59.76	-0.94	10.28	-23.39 hu	19.06 tr
	ZS	29.54	31.27	-	1.72	6.49	-9.50 nno	16.13 it

- *Multi* :
  - $\approx$ la moitié des langues seulement bénéficient de W_{22}
- *ZS* :
  - De très grosses améliorations des résultats pour certaines langues sont à mettre en parallèle avec de grosses dégradations des résultats

$\Rightarrow$ W_{22} ne semble plus être utile en *Multi*, contrairement aux expériences *ZS*

Apport du WALS : TagParsing

Corrélations de l'apport du WALS entre les différentes prédictions

	POS-Morpho	POS-LAS	Morpho-LAS
Multi	0.72*	0.72*	0.96*
ZS	0.39*	0.40*	0.93*

$\Rightarrow$ Le WALS est utilisé de la même manière pour les prédictions des traits morphologiques et l'analyse syntaxique

$\Rightarrow$ Mais il est utilisé différemment pour la prédiction des POS.

Apport du WALS : TagParsing (conclusions)

- Le WALS ne semble plus être utile en multilingue
- Il est cependant utile en moyenne en zero-shot
- Le WALS semble être utilisé de façon très similaire pour la prédiction des traits morphologiques et pour l'analyse syntaxique.

$\Rightarrow$ Plus d'analyses sont nécessaires pour comprendre ce phénomène

Apport du WALS : étiquetage des POS

	Sans W_{22}	Avec W_{22}	Avec ID	Δ	σ	min	max
Multi	88.16	83.98	91.25	0.42	1.74	-5.99 nno	4.16 lv
ZS	50.51	49.66	-	-0.22	3.64	-13.85 ur	5.21 de

- *Multi* :
  - L'ajout du vecteur W_{22} fait gagner, en moyenne, 0.42 point
  - Augmentation des résultats pour 32/38 langues
- ***ZS*** :
  - L'ajout du vecteur W_{22} fait perdre, en moyenne, 0.22 point
  - Augmentation des résultats pour 13/38 langues

Apport du WALS : étiquetage des POS (*Multi*)

Lang.	*Multi*	*Multi* W_{22}	Δ
nno	90.00	83.62	-6.38
nob	91.41	86.49	-4.92
ru	94.23	92.32	-1.91
es	94.00	92.13	-1.87
pt	93.00	92.39	-0.61
nl	86.64	86.08	-0.56
hi	92.75	92.92	0.17
de	88.94	89.13	0.19
ur	89.53	90.51	0.98
lv	85.84	88.69	2.85
Macro moyenne	90.81	91.23	0.42

- W_{22} semble bénéficier plus aux langues qui avaient le plus souffert du passage en multilingue (0.70 de corrélation entre "l'apport du WALS" et "la chute liée au multilinguisme")
- Langues avec des vecteurs W_{22} identiques : (nno, nob), (nl, de), (hi, ur) $\Rightarrow$ langues indifférenciables ?

Apport du WALS : étiquetage des POS (*ZS*)

Lang.	*ZS*	*ZS* W_{22}	Δ
de	53.84	48.63	-5.21
nl	61.04	56.91	-4.13
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

- Triplet (ar, fa, ur) même alphabet, mais ourdou (ur) très différent de ar et fa, mais proche de hi.

Apport du WALS : étiquetage des POS (En conclusion)

- W_{22} utile en *Multi*
- Utilité ambiguë en ***ZS***
- W_{22} semble jouer un rôle pour rapprocher/éloigner certaines langues

Traits	Description	Valeurs possibles
81A	Ordre du Sujet, de l'Objet et du Verbe	SOV SVO VSO VOS OVS OSV Pas d'ordre dominant
82A	Ordre du Sujet et du Verbe	SV VS Pas d'ordre dominant
83A	Ordre de l'Objet et du Verbe	OV VO Pas d'ordre dominant
85A	Ordre de l'Adposition et de la Phrase nominale	Postpositions Prépositions Inpositions Pas d'ordre dominant Pas d'adpositions
86A	Ordre du génitif et du Nom	Génitif-Nom Nom-Génitif Pas d'ordre dominant
87A	Ordre de l'Adjectif et du Nom	Adjectif-Nom Nom-Adjectif Pas d'ordre dominant Uniquement les clauses relatives à tête interne

Traits	Description	Valeurs possibles
88A	Ordre du démonstratif et du nom	Démonstratif-Nom Nom-Démonstratif Préfixe Démonstratif Suffixe Démonstratif Démonstratifs avant et après le Nom Mixés
89A	Ordre du Nombre et du Nom	Nombre-Nom Nom-Nombre Pas d'ordre dominant Les nombres modifient uniquement les verbes
90A	Ordre de la Clause Relative et du Nom	Nom-Clause Relative Clause Relative-Nom Internally headed Corrélatif Adjoint Doubly headed Mixé
92A	Position des particules de question polaire	Initiale Finale Seconde position Autre position Dans l'une des deux position Aucune particule de question

Traits	Description	Valeurs possibles
94A	Ordre de la subordonné adverbial et de la clause	Mot du subordonné initial Mot du subordonné final Mot subordonné interne Suffixe subordonné Mixé
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
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
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
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

Traits	Description	Valeurs possibles
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
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

Traits	Description	Valeurs possibles
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
143E	Morphèmes négatifs préverbaux	NegV [Neg-V] NegV&[Neg-V] Aucun
143F	Morphèmes négatifs post-verbaux	VNeg [V-Neg] VNeg&[V-Neg] Aucun

Traits	Description	Valeurs possibles
143G	Moyens morpholo- giques mineurs pour signaliser la négation	NegTone NegInfix NegStemChange Aucun
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

- Introduction
- Ressources typologiques (WALS)
- Modèle de caractères
- Proximité entre langues
- Conclusions et Perspectives

Chomsky, N. and Lasnik, H. (2008). The theory of principles and parameters. In *Syntax*, pages 506–569. De Gruyter Mouton.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word Translation Without Parallel Data. *arXiv :1710.04087 [cs]*. arXiv : 1710.04087.

Dary, F. and Nasr, A. (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*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv :1810.04805 [cs]*. arXiv : 1810.04805.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (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*, pages 6282–6293, Online. Association for Computational Linguistics.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv :1309.4168*.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., and Silveira, N. (2016). Universal Dependencies v1 : A Multilingual

Treebank Collection. In *LREC*.