Traitement du Langage Naturel et Linguistique

LE BELLEGO Victor, MARTZLOFF Alice, MOUSSOX Vincent

M2 IAAA 2021/2022

1 Certaines langues sont elles plus difficiles à analyser que d'autres ?

L'objectif de ce projet est d'analyser et de comprendre les raisons pour lesquelles des analyseurs syntaxiques partageant la même architecture et entraînés sur la même quantité de données obtiennent des performances très différentes sur différentes langues. On peut observer ce phénomène dans la Table 2 qui présente les performances calculées à l'aide des mesures LAS (Labeled Accuracy Score) et UAS (Unlabeled Accuracy Score) obtenues par un analyseur sur 36 langues différentes.

lang	las	uas	lang	las	uas	lang	las	uas
da	66.18	73.06	zh	46.76	55.60	pl	68.76	80.69
hr	58.88	69.67	lv	51.46	59.32	sv	64.52	71.11
id	69.99	75.63	he	64.82	70.06	cs	69.77	77.40
ar	60.88	69.78	ko	46.06	55.00	nl	54.92	63.69
eu	47.94	58.17	ja	71.54	82.71	hu	57.21	69.24
it	74.82	81.01	ca	64.79	71.13	bg	68.08	78.14
fa	47.70	56.17	en	66.57	71.22	vi	48.40	49.62
es	66.09	71.01	pt	70.84	74.55	sl	49.46	61.51
ro	57.23	68.61	et	59.04	73.09	nno	73.58	78.30
de	58.12	64.19	fr	68.40	73.46	nob	66.22	73.93
hi	64.03	72.54	el	69.96	76.58			

Table 1 – LAS/UAS calculées pour les différentes langues (sans les *configurational features*)

1.1 Variables explicatives

1.1.1 Observations effectuées sur le corpus d'apprentissage

Taille du vocabulaire La qualité de l'analyseur pour une langue donnée peut dépendre du genre textuel de ladite langue. Il se traduit entre autre par la longueur des phrases dans cette langue.

Longueur moyenne d'une phrase Le genre textuel d'une langue aura également une influence sur la longueur des phrases dans cette langue. C'est une autre variable qui peut expliquer les capacités d'un analyseur entraîné sur une langue.

Longueur moyenne d'un mot

Taux de projectivité On peut aussi expliquer une telle différence dans les capacités d'analyse de langue par les spécificités de chacune, par exemple, en fonction de son taux de projectivité.

Longueur moyenne des dépendances En ce qui concerne des variables explicatives liées aux dépendances, comme la longueur moyenne des dépendances ou encore le nombre de dépendance, la longueur moyenne de la plus grande dépendance (ci-dessous), ce sont les variables les plus sujettes à relever de la cohérence de l'annotation pour chaque langue.

Nombre moyen de dépendance

Longueur moyenne de la plus grande dépendance Nous évoquerons dans une autre section l'influence de la cohérence des annotations.

1.1.2 Complexité selon M. Parkvall

Dans son papier *The simplicity of creoles in a cross-linguistic perspective* [1] sorti en 2008, Mikael Parkvall s'intéresse à quantifier la complexité des langues. Il part du postulat qu'une expression est d'autant plus complexe qu'elle implique de règles, c'est-à-dire qu'elle requiert une longue description. Ainsi, l'hypothèse de base de l'auteur est la suivante : une langue complexe est une langue avec des constructions plus complexes. Il explore un aspect de complexité structurelle.

Prenons par exemple la voix passive. Lorsqu'elle existe dans une langue, il faut pouvoir définir comment passer de la voie active à la voix passive, ce qui exige une explication de règle supplémentaire. Une langue qui possède une voix passive est donc, en ce qui concerne cette construction spécifique, plus complexe qu'une autre n'en possédant pas. Si on énumère donc un grand nombre de « constructions complexes », la langue la plus complexe sera celle qui en compte le plus grand nombre.

Pour extraire les « constructions complexes » qu'on peut trouver dans une langue, Parkvall utilise le set de données World Atlas of Linguistic Structures (WALS) publié en 2005 par Haspelmath et al.. Il choisit 155 langues parmi plus de 2 500, et 47 caractéristiques parmi plus de 140.

Choix des caractéristiques Parkvall exclut des caractéristiques selon un raisonnement défendu dans son papier et qui s'efforce de mettre la majorité des linguistes d'accord sur le fait qu'une caractéristique apporte ou non de la complexité. Il retient les caractéristiques suivantes :

Caractéristiques du WALS		
Size of consonant inventories	Distance contrast in demonstratives	Morphological imperative
Size of vowel quality inventories	Gender in pronouns	Morphological optative
Phonemic vowel nasalization	Politeness in pronouns	Grammaticalized evidentiality distinctions
Complexity of syllable structure	Person marking on adpositions	Both indirect and direct evidentials
Tone	Comitative ≠ instrumental	Non-neutral marking of full NPs
Overt marking of direct object	Ordinals exist as a separate class beyond 'first'	Non-neutral marking of pronouns
Double marking of direct object	Suppletive ordinals beyond 'first'	Subject marking as both free word and agreement
Possession by double marking	Obligatory numeral classifiers	Passive
Overt possession marking	Possessive classification	Antipassive
Reduplication	Conjunction 'and' ≠ adposition 'with'	Applicative
Gender	Difference between nominal and verbal conjunction	Obligatorily double negation
Number of genders	Grammaticalized perfective/imperfective	Asymetric negation
Non-semantic gender assignment	Grammaticalized past/non-past	Equative copula ≠ Locative copula
Grammaticalized nominal plural	Remoteness distinctions of past	Obligatorily overt equative copula
Definite articles Indefinite articles	Morphological future	
Inclusivity (in either pronouns or verb morphology)	Grammaticalized perfect	

Table 2 – Liste des caractéristiques extraite directement du WALS

Il ajoute à ces caractéristiques, d'autres données « résiduelles » d'auteurs contributeurs au WALS. Ce sont les suivantes :

Enfin, il s'intéresse aussi à une donnée de Harley and Ritter (2002) à laquelle il a eu accès :

Caractéristiques d'auteurs du WALS		
Demonstratives marked for number	Demonstratives marked for gender	Demonstratives marked for case
Total amount of verbal suppletion	Alienability distinctions	

Table 3 – Liste de caractéristiques proposées par les auteurs du WALS

Caractéristique de Harley et Ritter
Number of pronominal numbers

Table 4 – Une caractéristique accessible, inspirée de Harley et al. (2002)

Les valeurs de ces caractéristiques ont toutes été traduite par l'auteur comme des valeurs comprises entre 0 et 1 :

« Oui » ou « non » deviennent 0 ou 1 avec parfois l'introduction de valeurs intermédiaire 0,5. Des valeurs d'intensité comprises entre 1 et 4 sont compréssées en : 0 - 0,25 - 0,5 - 0,75 et 1. Des valeurs catégoriques comme la classification en « simple », « modérément complexe » et « complexe » sont traduites en 0 - 0,5 et 1.

Choix des langues L'auteur s'attèle à choisir des langues dont les annotations sont le moins lacunaires possible pour les caractéristiques décrites ci-dessus. Pour une langue i donnée :

$$\text{Score}_{\,i} = \frac{\sum_{k=1}^{L} \text{contribution}_{\,k}}{L} \tag{1}$$

où k représente une caractéristique. Chaque langue comptant un nombre différent L de caractéristiques (parmi celles choisies par l'auteur) effectivement annontées pour cette langue.

N.B. : en effet, de même que pour les caractéristiques que nous extrayons directement de nos données, Parkvall note que le set de données WALS n'est pas identiquement distribué : les langues ne sont pas identiquement annotées pour les caractéristiques proposées...

1.2 Cohérence des annotations

Dans leur article *Divergences entre annotations dans le projet Universal Dependencies [4] et leur impact sur l'évaluation de l'étiquetage morpho-syntaxique* [2], Guillaume Wisniewski et François Yvon montrent que la dégradation des performances observée lors de l'application d'un analyseur morpho-syntaxique à des données hors domaine comme ici, d'une langue à l'autre, résulte d'incohérences entre les annotations des ensembles de test et d'apprentissage. Ils montrent qu'appliquer le principe de variation des annotations de Dickinson & Meurers [3] permet d'identifier les erreurs d'annotation et donc les incohérences et évaluer leur impact. Nous souhaitions nous inspirer de ces méthodes mais n'avons pu en raison notamment du temps qui nous manque proposer une telle amélioration...

1.3 Conclusion

1.3.1 Est-il possible de connaître a priori les performances de l'analyseur en fonction de certaines caractéristiques de la langue ?

1.3.2 Est-il possible d'utiliser les conclusions de l'étude statistique pour améliorer les performances de l'analyseur ?