Analyse orientée corpus d'universaux de Greenberg sur Universal Dependencies

Hee-Soo Choi, Bruno Guillaume, Karën Fort

Analyse orientée corpus d'universaux de Greenberg sur Universal Dependencies

Hee-Soo Choi[1, 2] Bruno Guillaume[1] Karën Fort[1, 3]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) ATILF, Université de Lorraine & CNRS, F-54000 Nancy, France
(3) Sorbonne Université F-75006 Paris, France
hee-soo.choi@loria.fr, Bruno.Guillaume@loria.fr, karen.fort@loria.fr

MOTS-CLÉS : universaux linguistiques, typologie linguistique, ordre des mots, multilinguisme, Universal Dependencies, corpus arborés, GREW.

KEYWORDS: language universals, linguistic typology, word order, multilinguism, Universal Dependencies, treebanks, GREW.

1 Entre typologie linguistique et TAL

Avec l'expansion des données numériques, de plus en plus de ressources multilingues voient le jour et constituent des traces durables de nos pratiques langagières. Ces dernières sont particulièrement précieuses dans le domaine du Traitement Automatique des Langues (TAL) dont l'un des objectifs est de traiter toutes les langues naturelles. Dans nos travaux, nous nous sommes intéressés à la vérification d'universaux de Greenberg (Greenberg, 1966) de manière automatique et empirique sur les corpus d'Universal Dependencies (UD) (de Marneffe et al., 2021). L'étude des universaux repose fortement sur une tradition empirique et typologique mais également sur des connaissances tirées d'ouvrages de référence. À travers nos expériences, nous fournissons des résultats fondés uniquement sur de grandes quantités de données avec un échantillon de 141 corpus, soit 74 langues d'UD 2.7. Avec l'outil GREW (Guillaume, 2021), nous avons ainsi déterminé trois ordres de mots (l'ordre sujet - verbe - objet, l'ordre adposition - nom et l'ordre adjectif - nom) et vérifié quatre universaux. En faisant le choix de traiter chaque corpus individuellement, nous avons pu évaluer l'homogénéité entre corpus d'une même langue et analyser les raisons des possibles divergences. Enfin, notre étude sur 74 langues permet également de soulever des incohérences interlinguistiques liées au schéma d'annotations.

2 Contraintes méthodologiques

La vérification des universaux linguistiques ont demandé de faire des choix dans notre approche en raison de certaines contraintes. Tout d'abord, un travail de tri sur les universaux a dû être effectué. Les universaux de Greenberg, établis sur la base de 30 langues de différentes familles de langues, sont au nombre de 45 et sont relatifs à l'ordre des mots, la syntaxe et la morphologie. En raison des annotations d'UD qui ne nous permettent pas d'obtenir certaines informations notamment au niveau morphologique, nous avons choisi de traiter les universaux relatifs à l'ordre des mots. Concernant nos données, en mettant à disposition un schéma d'annotations universel, le projet UD s'inscrit dans

l'objectif de traiter un maximum de langues et de favoriser ainsi les recherches multilingues. Les corpus d'UD constituent donc des données à première vue optimales pour notre tâche de vérification d'universaux. Toutefois, cette tâche reste ambitieuse et est mise à mal par les différences notables qui résident entre les langues et qui se traduisent par des incohérences dans les annotations. Le caractère universel des annotations est donc à prendre avec précaution. Enfin, nous avons choisi de traiter 74 langues, qu'évidemment nous ne maîtrisons pas toutes. Cette contrainte nous impose donc de nous intéresser à des caractéristiques basiques et universelles afin d'éviter davantage de biais. Par ailleurs, nous faisons le choix de nous positionner au niveau du corpus et non au niveau de la langue, ce qui revient à traiter, en réalité, non pas 74 langues mais 141 corpus.

3 Des résultats en accord avec Greenberg et WALS

Avant d'entreprendre la vérification des universaux, nous avons déterminé trois ordres de mots pour les 141 corpus : i) l'ordre sujet - verbe - objet, ii) l'ordre adposition - nom et iii) l'ordre adjectif - nom. Nous avons utilisé GREW (Guillaume, 2021), un outil de réécriture de graphes permettant de faire des requêtes sur les corpus et d'extraire les occurrences d'une construction linguistique particulière.

Pour déterminer l'ordre dominant de manière quantitative, nous utilisons le même critère que WALS (Dryer and Haspelmath, 2013) qui considère un ordre comme dominant s'il présente une fréquence d'apparition au moins deux fois plus grande que le deuxième ordre le plus fréquent (Dryer, 2013). Nous avons donc calculé le ratio entre les deux ordres les plus fréquents. Si le ratio est supérieur ou égal à deux, on considère l'ordre le plus fréquent comme l'ordre dominant, sinon on considère que le corpus n'a pas d'ordre dominant (noté NDO pour *No Dominant Order*). Dans le cas où seulement deux ordres sont possibles (par exemple, adjectif - nom / nom - adjectif), si le ratio est supérieur à 2, la fréquence de l'ordre le plus fréquent est supérieure à $\frac{2}{3}$. Utiliser cette mesure nous a permis de comparer les résultats obtenus pour les trois ordres avec les données présentes dans WALS.

Nos résultats sont majoritairement en accord avec ceux de WALS pour les trois ordres, mis à part certaines exceptions pour lesquelles nous fournissons une analyse plus détaillée. Si les ordres adjectif - nom et adposition - nom présentent des résultats relativement tranchés avec une tendance marquée pour un des deux cas, l'ordre sujet - verbe - objet présente plus d'hétérogénéité entre les corpus d'une même langue. Pour cet ordre, nous avons soulevé plusieurs facteurs expliquant l'hétérogénéité : le genre des corpus (corpus oraux, corpus de romans, corpus de journaux, corpus de textes bibliques…), la période des corpus (pour les langues mortes notamment), les erreurs d'annotations et certaines spécificités des langues (topicalisation du sujet par exemple) (Choi et al., 2021).

Suite à la classification selon les trois ordres précédemment décrits, quatre universaux de Greenberg ont été vérifiés :

Universel 1 *Dans les phrases déclaratives avec un sujet nominal et un objet nominal, l'ordre dominant est presque toujours un ordre dans lequel le sujet précède l'objet.*

Dans nos résultats, sur 141 corpus, 91 sont SVO, 24 sont SOV, 4 sont VSO et 22 sont NDO, ce qui confirme l'universel de Greenberg pour 119 corpus et 59 langues. Concernant les corpus sans ordre dominant, si nous considérons uniquement l'ordre le plus fréquent sans calculer le ratio, tous les corpus présentent un des trois ordres où le sujet précède l'objet, à l'exception de deux corpus : l'`Amharic-ATT` et le `Latin-LLCT`.

Universel 3 *Les langues d'ordre dominant VSO sont toujours prépositionnelles.*

Universel 17 *Avec une fréquence largement supérieure à la normale, les langues d'ordre dominant VSO ont l'adjectif après le nom.*

Nous avons traité ces universaux ensemble dans la mesure où ils concernent tous deux les langues d'ordre dominant VSO. Le tableau 1 donne ainsi les corpus d'ordre dominant VSO avec leurs fréquences, ainsi que la proportion de prépositions et de l'ordre Nom - Adjectif. Les universaux sont vérifiés mais sur seulement quatre corpus.

Corpus	VSO	Prep	Nom - Adj
Arabic-NYUAD	54,56 %	99,97 %	99,69 %
Irish-IDT	99,14 %	99,78 %	98,91 %
Scottish_Gaelic-ARCOSG	97,49 %	100 %	84,82 %
Welsh-CCG	78,57 %	100 %	82,54 %

TABLE 1 – Proportions de prépositions et d'ordre Nom - Adjectif dans les corpus VSO.

Universel 4 *Avec une fréquence largement supérieure à la normale, les langues d'ordre normal SOV sont postpositionnelles.*

D'après nos résultats, 24 corpus sont d'ordre dominant SOV, ce qui correspond à 15 langues. Dix langues sont postpositionnelles : le bambara, le basque, le coréen, le hindi, le japonais, le kazakh, l'ouïghour, le ourdou, le telugu et le turc.

Les cinq langues restantes sont l'afrikaans, l'allemand, le latin, le perse et le sanskrit. L'afrikaans et le perse présentent des corpus fortement SOV mais sont prépositionnels. L'exception du perse est également soulignée par Greenberg bien que cette langue ne soit pas dans son échantillon. L'allemand et le latin sont des langues multicorpus mais elles ne comptent qu'un seul corpus d'ordre dominant SOV, leurs autres corpus étant considérés comme sans ordre dominant. Nous supposons que la formulation de Greenberg « ordre normal SOV » permet de ne pas prendre en compte ce type de langues dans son universel. Et enfin, pour le corpus du sanskrit, nous n'avons trouvé aucune occurrences de postposition ou préposition. En effet, le motif GREW utilisé détecte les postpositions et prépositions annotées avec l'étiquette ADP (adposition). Or le corpus du sanskrit présente l'étiquette PART (particule). Nous pouvons noter que nous retrouvons le même phénomène sur un des corpus du coréen, le `Korean-PUD`.

4 Les limites du schéma d'annotations universel

Si UD propose un noyau d'annotations universel, les annotateurs sont libres dans leurs choix d'annotations, ce qui provoque une certaine hétérogénéité entre les langues mais aussi entre corpus d'une même langue. Pour extraire les constructions linguistiques décrites précédemment, nous devions identifier précisément les annotations codant pour ces constructions. Lors de cette étape, nous avons relevé plusieurs incohérences, notamment un conflit entre deux étiquettes de partie du discours : ADP (adposition) et PART (particule). La distinction entre une particule et une adposition est délicate à définir, en particulier dans les langues agglutinantes. Les adpositions, les conjonctions de coordination et de subordination sont considérées comme des particules dans les directives d'UD, mais celles-ci demandent à privilégier l'étiquette la plus précise possible.

Par ailleurs, nous avons remarqué que certains corpus présentent un nombre conséquent de dépendants non-nominaux pour les relations `nsubj` et `obj`, ce qui est contradictoire avec les directives d'UD. Dans le même esprit, la relation `case` implique normalement des gouverneurs nominaux, ce qui n'est pas respecté par certains corpus.

Le schéma UD se base initialement sur un schéma d'annotations créé pour l'anglais, le Stanford Dependencies (de Marneffe et al., 2014), ce qui a orienté certains choix d'annotations. Pour les langues présentant une structure différente de l'anglais, adapter les annotations amène les créateurs des corpus à faire des choix d'annotations qui ne sont pas forcément en accord avec les autres corpus de la langue. Cela impacte la cohérence entre les langues ainsi qu'entre corpus d'une même langue.

5 Conclusion

Nos résultats confirment les observations linguistiques sur de grandes quantités de données. Notre travail peut ainsi compléter les bases de données typologiques comme WALS qui présente des brèches pour sept langues que nous avons traitées : l'afrikaans, le féroïen, le galicien, le kazakh, le maltais, le naija et le slovaque. En outre, nous nous sommes efforcés de fournir des analyses pour expliquer les incohérences détectées dans les annotations d'UD, soit en examinant les documentations relatives aux corpus, soit en faisant appel à des locuteurs natifs autour de nous. L'aspect collaboratif du projet nous a ensuite permis de faire des retours et ainsi contribuer à l'amélioration des corpus concernés.

Références

Choi, H.-S., Guillaume, B., Fort, K., and Perrier, G. (2021). Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In *Recent Advances in Natural Language Processing (RANLP2021)*, en ligne, Bulgarie.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies : A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Islande. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2) :255–308.

Dryer, M. S. (2013). Determining dominant word order. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J. H., editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.

Guillaume, B. (2021). Graph matching and graph rewriting : GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.