Vérification d'universaux linguistiques sur des corpus multilingues annotés en syntaxe de dépendances

Mémoire de Master 2 Langue et Informatique

Présenté par :
Hee-Soo CHOI

Sous la direction de :
Karën FORT (Sorbonne Université, LORIA)
Bruno GUILLAUME (LORIA)

Table des matières

Table des figures

Liste des tableaux

Remerciements

En premier lieu, je tiens à remercier mon encadrante et professeure, Karën Fort, de m'avoir ouverte au monde de la recherche et d'avoir cru en moi dès le premier stage. Merci pour nos échanges qui m'ont tant apporté à une période où tout était incertain, tant sur le plan professionnel que personnel.

Je remercie également mon encadrant au Loria, Bruno Guillaume, pour sa confiance et sa bienveillance tout au long des deux stages.

Je voudrais ensuite remercier toute l'équipe Sémagramme à commencer par les membres permanents : Philippe De Groote, chef de l'équipe Sémagramme, de m'avoir accueillie dans l'équipe à deux reprises. Merci à Maxime Amblard sans qui je n'aurais pas pu faire ce stage financé. Merci à Michel Musiol pour son aide si précieuse dans la préparation au concours de contrats doctoraux. Merci à Guy Perrier d'avoir pris le temps de faire des retours à chacune de mes sollicitations.

Je remercie tous les doctorants du bureau 230 de m'avoir si bien intégrée à l'équipe et d'avoir répondu à mes doutes sur le doctorat : Maria Boritchev, Pierre Ludmann, Chuyuan Li, Samuel Buchel, Amandine Lecomte et Siyana Pavlova.

Un grand merci également à Pierre Lefebvre, ingénieur de l'équipe, pour sa patience et sa disponibilité face aux stagiaires en détresse.

Ce mémoire a été l'occasion de me plonger dans l'immensité que représentent les langues. Je remercie tous les locuteurs vers qui je me suis tournée durant mes recherches : Maria et Karën pour le russe, Chuyuan pour le chinois, Siyana pour le bulgare, Guy pour le polonais, Yutao Li pour le cantonais, Eunjo Lee et Lou Guillaume pour le japonais, Alma Guillaume pour le grec ancien et le latin, Igor Sadovnikov pour le biélorusse, Sashi Narayan et Lydie Lemoine pour l'hindi, Vincent Vandeghinste pour le néerlandais et Hilda Mock pour l'arabe. Je remercie particulièrement Kim Gerdes pour ses retours sur l'allemand et d'avoir partagé volontiers ses scripts de typométrie.

Les recherches présentées dans ce mémoire ont été menées durant un stage financé par le projet de recherche OLKi de Lorraine Université d'Excellence. Je remercie pour finir les responsables du projet OLKi, Maxime Amblard et Aurore Coince, de m'avoir accordé ce financement.

Introduction

Le nombre exact de langues parlées dans le monde reste encore aujourd'hui difficile à définir. Il est actuellement estimé à plus de 7 000, ETHNOLOGUE en comptant 7 139[1]. 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 un des objectifs est de couvrir toutes les langues naturelles existantes.

Dans ce mémoire, nous abordons la notion d'universalité linguistique. Plus précisément, nous cherchons à vérifier des universaux linguistiques de manière automatique et empirique sur des corpus multilingues. Notre étude se situe à la frontière entre deux domaines : la typologie linguistique et le TAL. Au-delà de s'ancrer dans ces domaines, elle présente surtout des enjeux pour leur développement respectif.

D'une part, l'étude des universaux est fortement liée à la typologie linguistique, qui permet de classer les langues selon des traits communs. Si la typologie renferme une approche empirique traditionnelle, elle se base également sur des ouvrages de référence. À travers nos expériences, nous fournissons des résultats fondés sur de grandes quantités de données que nous exploitons grâce à des outils automatiques. Ces observations représentent des informations empiriques qui sont susceptibles de confirmer voire de compléter les bases de données typologiques.

D'autre part, il a été démontré qu'inclure des connaissances linguistiques dans les systèmes de TAL permet de les rendre plus performants, vis-à-vis d'une langue mais également pour une approche davantage universelle. En effet, traiter toutes les langues *via* des systèmes universels et ainsi réduire l'hétérogénéité dans les langues représentées en TAL, constitue actuellement un des défis majeurs du domaine. Par ailleurs, ce mémoire est l'occasion d'utiliser un outil de TAL basé sur la réécriture de graphes, GREW dont les fonctionnalités permettent non seulement d'explorer les données mais également de pallier certaines de leurs limites.

La vérification des universaux linguistiques nécessite de la rigueur, notamment en termes de définition des notions linguistiques, mais aussi un certain recul sur la quantité de données à manipuler. Il est donc important de délimiter dès le début de notre projet, les contraintes méthodologiques auxquelles nous devons faire face.

C1 : Des universaux pauvres en définitions Dans nos expériences, nous choisissons de vérifier certains universaux de Joseph H. Greenberg (Greenberg, 1966a). Greenberg est un linguiste considéré comme un des pionniers de la typologie linguistique moderne et son travail d'élaboration de 45 universaux constitue, encore aujourd'hui, une référence. Malgré des analyses empiriques approfondies, nous pouvons reprocher un certain manque de définitions des concepts linguistiques auxquels Greenberg s'intéresse. Nos expériences

1. Voir : https://www.ethnologue.com/, septembre 2021.

sont donc susceptibles de traiter des concepts linguistiques dont les définitions diffèrent de celles de Greenberg.

C2 : Des annotations « universelles » Les corpus sur lesquels nous vérifions les universaux sont issus du projet Universal Dependencies (UD). En mettant à disposition un schéma d'annotations universel, ce projet collaboratif 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.

C3 : Une approche multilingue et multicorpus La troisième contrainte que nous rencontrons est le fait d'explorer 74 langues, qu'évidemment nous ne maîtrisons pas toutes. Cette contrainte nous impose, d'une part, de faire un travail rigoureux dans la définition des concepts linguistiques que nous traitons et d'autre part, 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.

Le premier chapitre de ce mémoire décrit dans un premier temps, les définitions des concepts que nous aborderons et dans un second temps, l'état de l'art dans les travaux relatifs à la typologie linguistique et au TAL.

Nous présentons dans le deuxième chapitre, les données d'Universal Dependencies, l'outil que nous utilisons pour nos expériences, Grew et enfin les mesures que nous avons choisies pour permettre une analyse quantitative de nos résultats.

Nous consacrons ensuite le troisième chapitre à la description de nos expériences, à commencer par le choix des universaux à tester, suivi par les travaux préalables nécessaires à la vérification de ces universaux.

Le quatrième et dernier chapitre détaille les résultats des universaux testés, les raisons de l'hétérogénéité entre des corpus d'une même langue et enfin les différentes observations relatives à UD que nous avons pu relever à travers l'analyse simultanée de 141 corpus.

Toutes les informations relatives aux corpus que nous utilisons dans ce mémoire sont présentées dans les *README* de chaque corpus et sont disponibles sur le *github* du projet Universal Dependencies : https://github.com/universaldependencies/. Nous mettons également à disposition tous nos scripts, motifs Grew, tableaux de résultats et figures obtenues sur https://gitlab.inria.fr/ud-greenberg/experiments.

Chapitre 1

Définitions et État de l'art

Sommaire

1.1 Universaux linguistiques

1.1.1 Définitions

En linguistique, un universel correspond à une propriété, une tendance ou une relation considérée comme commune à toutes les langues ou du moins, à un grand nombre de langues. Malgré la diversité évidente des langues, des similarités entre elles ont incité les linguistes à s'intéresser à une possible universalité.

D'un point de vue terminologique, il est commun de voir apparaître « language universals » mais aussi « linguistic universals », que nous traduirons respectivement par « universaux langagiers » et « universaux linguistiques ». Selon Cooreman and Goyvaerts (1980), l'étude des universaux langagiers concerne principalement les analyses typologiques de données interlinguistiques, tandis que l'étude des universaux linguistiques se concentre davantage sur les théories et les grammaires du langage. Toutefois, il n'est pas rare de voir les deux formulations pour désigner le même objet d'étude, la distinction n'étant pas toujours faite de manière explicite dans la littérature. En français, l'adjectif « linguistique » relatif à l'étude des langues est aussi souvent employé dans le sens « relatif au langage » [1]. Dans ce mémoire, nous utiliserons le terme d'universaux linguistiques pour désigner le concept en général et les « language universals » de Greenberg (Greenberg, 1966a).

Dans les faits, la distinction entre universaux langagiers et universaux linguistiques se calque sur l'opposition entre deux approches dans l'étude des universaux : l'approche empirique, représentée dans les travaux de Joseph H. Greenberg et l'approche rationaliste, portée par les idées de Noam Chomsky. C'est sous l'influence de ces deux linguistes qu'un fort intérêt pour les universaux se développe à partir des années 60.

Comrie (1989) décrit plusieurs paramètres intervenant dans l'opposition entre ces deux approches, notamment les données sur lesquelles les recherches sont menées, le degré d'abstraction de l'analyse et les explications de l'existence des universaux. Croft (2003) fait également la distinction entre les travaux de Greenberg, portés par une méthode typologique et ceux de Chomsky qui sont inhérents à une approche générative des universaux. Il décrit toutefois que ces approches présentent des similitudes : elles sont toutes deux universalistes, s'intéressent à la structure du langage et questionnent le concept de « langue

1. Voir : https://www.cnrtl.fr/definition/langagier, août 2021.

humaine ».

L'approche empirique Dans l'étude des universaux linguistiques, l'approche empirique considère qu'il faut s'appuyer sur des données provenant d'un grand nombre de langues. Les empiristes, dont Greenberg fait partie, considèrent que les langues ne varient pas indéfiniment et que les limites de la variation déterminent des universaux linguistiques (Croft, 2003).

L'approche empirique nécessite donc de se baser sur des données provenant de différentes langues afin d'observer l'étendue de la variation des phénomènes linguistiques. La difficulté majeure de cette approche se situe dans la détermination d'un échantillon de langues. Il est évidemment difficile de couvrir toutes les langues du monde mais nous pouvons questionner le degré de représentativité de l'échantillon par rapport à la réalité. Comrie (1989) considère qu'il serait insensé de considérer une observation attestée sur un grand nombre de langues comme fausse du fait d'une seule exception. L'approche empirique implique alors de définir les universaux en prenant en compte le nombre de langues concernées. C'est pourquoi, les universaux non-absolus sont généralement plus fréquents que les universaux absolus.

L'approche rationaliste L'approche rationaliste est principalement représentée dans les travaux de Chomsky et sa grammaire générative (Chomsky, 1982). Contrairement à l'approche empirique, les rationalistes défendent l'idée qu'il y a un caractère inné dans les universaux linguistiques. En effet, Chomsky considère que les humains naissent avec des règles de grammaire de base qui permettent l'acquisition du langage. Cela expliquerait pourquoi un enfant apprend naturellement à parler.

De plus, Chomsky affirme que toutes les langues partagent une structure grammaticale commune, démontrée par le partage de certains concepts linguistiques comme les noms ou les verbes. De ce fait, il considère qu'il n'est pas nécessaire de dégager des universaux à partir d'un large échantillon de langues mais plutôt à partir d'une analyse profonde d'une langue ou d'un petit nombre de langues.

L'approche rationaliste justifie l'existence des universaux linguistiques par une origine génétique commune. Selon Comrie (1989), le facteur génétique commun est une hypothèse cohérente mais qui a le défaut d'être uniquement spéculative et difficilement vérifiable. Dans ce mémoire, le fait de disposer de corpus multilingues de grande taille nous amène à adopter une approche empirique qui s'inscrit dans la lignée des travaux de Greenberg.

1.1.2 Types d'universaux

Il existe différents types d'universaux, en termes de structures logiques mais aussi de domaines auxquels ils sont rattachés (phonologie, morphologie, ordre des mots...). Dans cette partie, nous détaillons la structure logique des universaux.

La classification la plus classique distingue d'une part les universaux absolus et non-absolus (ou tendances) et d'autre part les universaux implicationnels et non-implicationnels. Comrie (1989) précise qu'il peut exister des universaux absolus implicationnels et non-implicationnels, comme des tendances implicationnelles et non implicationnelles.

Les universaux absolus Les universaux absolus sont des caractéristiques qui sont valables pour toutes les langues. On considère alors qu'il n'y a aucune exception. Ces

universaux sont davantage relatifs aux propriétés générales du langage humain comme par exemple le fait que toutes les langues présentent la double-articulation, le caractère vocal et une syntaxe. Un universel absolu connu du domaine phonologique est que toutes les langues ont des voyelles.

Greenberg et al. (1966) les désignent par une autre appellation, les universaux non restreints. Dans ces universaux, ils incluent également des universaux présentant des limites numériques comme : « Toutes les langues ont un nombre de phonèmes supérieur à 10 ou inférieur à 70. ».

Les universaux non absolus ou les tendances Les universaux non absolus sont des universaux valables sur toutes les langues à quelques exceptions près. En réalité, il est difficile de faire la distinction entre un universel absolu et une forte tendance dans le sens où un universel peut être considéré comme absolu jusqu'à ce qu'une exception apparaisse. Par conséquent, il n'y a aucune certitude qu'un universel absolu le soit réellement.

Comrie (1989) propose une approche plus statistique pour faire une distinction moins radicale entre les universaux absolus et les tendances. Il donne l'exemple de l'universel suivant, qui revient régulièrement comme un universel absolu : « Dans un ordre de mots basique, le sujet précède l'objet ». Des exceptions sont connues comme le Malagasy qui est d'ordre VOS et le Hixkaryana qui est OVS, mais elles représentent un très faible pourcentage sur le nombre total de langues (moins d'un pourcent). Selon Comrie, il convient de prendre en compte la proportion d'exceptions et la proportion de langues qui valident l'universel et non considérer un universel comme non absolu dès qu'une exception existe.

Greenberg et al. (1966) parlent également d'universaux statistiques pour parler des « quasi-universaux ». Un universel statistique peut prendre cette forme : « La probabilité qu'une langue ait au moins une consonne nasale est supérieure (dans ce cas, largement supérieure) à celle qu'elle en soit dépourvue. ». Il existe effectivement très peu de langues qui n'ont aucune consonne nasale : le quileute et quelques langues voisines du salish, qui sont des langues amérindiennes.

Les universaux implicationnels Comme leur nom l'indique, les universaux implicationnels sont des universaux qui mettent en jeu deux caractéristiques p et q, la présence de p impliquant la présence de q. Cependant, la réciproque n'est pas forcément vérifiée. La majorité des universaux de Greenberg sont implicationnels, par exemple :

« Si l'objet pronominal suit le verbe, alors l'objet nominal le suit également. » (Universel 25).

Les implications mutuelles ou équivalences entre deux caractéristiques non universelles et indépendantes sont relativement rares. « Si une langue a un clic latéral, elle a toujours un clic dental. » est une équivalence, mais elle n'est pas universelle puisqu'elle ne s'applique qu'à certaines langues du sud et de l'est de l'Afrique (Greenberg et al., 1966).

Les universaux non-implicationnels Les universaux non-implicationnels sont des caractéristiques attestées sur toutes les langues sans référence à une autre caractéristique. Voici deux exemples :

— « Toutes les langues ont des voyelles » est un universel non-implicationnel et absolu,
— « Presque toutes les langues ont des consonnes nasales » est un universel non-implicationnel et non-absolu.

1.1.3 Les universaux de Greenberg

Joseph H. Greenberg (1915-2001) est un linguiste américain connu principalement pour ses travaux en typologie linguistique et en classification génétique des langues. Il est considéré comme un des pionniers de la typologie linguistique moderne (Bender, 2009).

Dans le cadre de la *Conference on Language Universals* de 1961, il lui a été demandé ainsi qu'à James J. Jenkins, psychologue et Charles E. Osgood, linguiste, d'écrire un *memorandum* sur les universaux linguistiques qui servirait de base à l'étude de ce sujet. À la suite de cela, plusieurs auteurs ont proposé leurs articles, dont Greenberg, qui présenta ses 45 universaux (Greenberg, 1966a). Il édite ensuite *Universals of Language* (Greenberg, 1966b) qui réunit les articles de la conférence traitant des universaux linguistiques.

Les 45 universaux de Greenberg sont des universaux linguistiques, majoritairement implicationnels, relatifs à l'ordre des mots, à la syntaxe et à la morphologie [2]. Pour réaliser ce travail, Greenberg se base sur un échantillon de 30 langues :

- Sept langues d'Europe : le basque, le serbe, le gallois, le norvégien, le grec moderne, l'italien et le finnois.
- Sept langues d'Afrique : le yoruba, le nubien, le swahili, le peul (Fulani), le masaï, le songhai et le berbère.
- Neuf langues d'Asie : le turc, l'hébreu, le bourouchaski, l'hindi, le kannada, le japonais, le thaï, le birman et le malais.
- Deux langues d'Océanie : le maori et le loritja.
- Cinq langues amérindiennes : le maya, le zapotèque, le quechua, le chibcha et le guarani.

Ces langues ont été choisies majoritairement pour des raisons pratiques, soit parce que Greenberg avait déjà une certaine connaissance de celles-ci, soit parce qu'il disposait de grammaires qu'il jugeait « adéquates ». Par ailleurs, il s'est efforcé d'utiliser des langues de différentes familles afin d'avoir « une couverture génétique et géographique la plus large possible », tout en étant conscient de la présence de biais. Il avance toutefois qu'une affirmation valable pour cet échantillon de 30 langues a une forte probabilité d'être valable pour toutes les langues.

Avant de traiter les universaux, Greenberg a établi une typologie d'ordre de base en s'appuyant sur trois critères :

- l'opposition entre prépositions (Pr) et postpositions (Po),
- l'ordre du sujet (S), du verbe (V) et de l'objet (O) dans les phrases déclaratives avec un sujet nominal et un objet nominal,
- l'ordre de l'adjectif qualificatif par rapport au nom (A/N [3]).

Les universaux de Greenberg sont en grande partie implicationnels et font intervenir ces classifications, notamment dans les universaux relatifs à l'ordre des mots et à la syntaxe. [4] Concernant l'ordre sujet, verbe et objet, Greenberg remarque qu'une langue n'a qu'un seul ordre dominant, même si elle peut en présenter plusieurs. Sur les six ordres possibles (SVO, SOV, VSO, VOS, OSV, OVS), les trois ordres où le sujet précède l'objet sont majoritairement dominants : SVO, SOV et VSO, les trois autres n'apparaissant que très rarement. Cette observation permet à Greenberg de réduire sa classification à ces trois ordres et ainsi attribuer chaque langue à l'une des 12 classes du tableau 1.1.

2. La liste des 45 universaux est disponible en annexe 4.3.2
3. A désigne le cas où l'adjectif précède le nom et N désigne le cas où le nom précède l'adjectif.
4. Les universaux morphologiques requièrent d'autres caractéristiques que nous n'aborderons pas dans nos travaux.

	I (VSO)	II (SVO)	III (SOV)
Po-A	0	1	6
Po-N	0	2	5
Pr-A	0	4	0
Pr-N	6	6	0

TABLEAU 1.1 – Tableau croisé des 30 langues classées selon les trois critères de typologie d'ordre de base (Tableau 1 de Greenberg (1966a)).

Greenberg observe des liens de corrélation entre deux types d'extrêmes : VSO/SOV d'une part et Pr-N/Po-A d'autre part. D'après le tableau 1.1, les langues VSO ont toutes des prépositions et le nom avant l'adjectif (Pr-N) et les langues SOV ont toutes des postpositions mais l'adjectif peut être avant ou après le nom. De plus, l'ordre SVO est plus corrélé à Pr-N que Po-A. Greenberg déduit qu'il y a une corrélation forte entre la prépositionnalité/postpositionnalité et les trois types d'ordre sujet - verbe - objet. Quant à l'ordre adjectif - nom, il juge qu'il est moins lié à l'ordre sujet - verbe - objet.

Nous pouvons souligner que l'analyse de Greenberg ne s'arrête pas aux 30 langues sur lesquelles il se base. En effet, le linguiste est conscient que des langues font exception à certains de ses universaux, même si elles ne sont pas incluses dans son échantillon.

1.2 Typologie et TAL

1.2.1 Un aperçu de la typologie linguistique

La typologie est une discipline qui étudie les variations de phénomènes linguistiques dans les langues du monde à travers des analyses comparatives (Comrie, 1989; Croft, 2003). Elle établit des classifications des langues selon certaines caractéristiques communes. La typologie linguistique est fortement liée à l'étude des universaux. Comme le décrit Comrie (1989), si la typologie linguistique s'intéresse aux variations et aux différences entre les langues, l'étude des universaux concerne plutôt les limites de ces variations. La frontière est donc mince entre ces deux thématiques.

Le travail des typologistes revient à répondre à la problématique « what's where why ? » de Nichols (1992) que nous traduirons par « quoi, où, pourquoi ? ». Plus précisément, le « quoi » fait référence aux différents phénomènes linguistiques, le « où » aux langues dans lesquelles apparaissent ces phénomènes, tant sur le plan géographique que génétique (familles de langue) et enfin répondre au « pourquoi » consiste à trouver des explications à l'existence de ces phénomènes et à la distribution observée. Pour y répondre, Bickel (2007) distingue trois sous-branches de la typologie :

— La typologie qualitative : la définition des similarités et des différences de caractéristiques au sein des langues et entre les langues.
— La typologie quantitative : la mesure de la variation des caractéristiques en se basant sur des données empiriques.
— La typologie théorique : l'analyse et l'explication de ces caractéristiques.

La typologie linguistique s'intéresse à différents types de traits linguistiques : phonologiques, sémantiques, lexicales et morphosyntaxiques. Les observations issues des travaux de typologie linguistique sont ensuite représentées dans des bases de données typologiques

Name	Type	Coverage	Notes
World Atlas of Language Structures (WALS)	Phonology Morphosyntax Lexicosemantics	2,676 languages; 192 features; 17% of features have values	Defines language features and provides values for a large set of languages; originally intended for study of areal distribution of features
Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages; 148 features; 45% of features have values	Similar to WALS, but differs in being fully open to public editing (Wikipedia-style), and by the addition of numerous example sentences for each feature
Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology Morphosyntax Lexicosemantics	76 languages; 130 features; 18,526 examples	Designed to allow comparison with WALS
PHOIBLE Online	Phonology	1,672 languages; 2,160 segments	Collates and standardises several phonological segmentation databases, in addition to new data
Lyon-Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages	Documents a broader range of features than PHOIBLE, including syllable structures and tone systems; provides bibliographic information and links to recorded samples
URIEL Typological Compendium	Phonology Morphosyntax Lexicosemantics	8,070 languages and dialects; 284 features; approximately 439,000 feature values	Collates features from WALS, SSWL, PHOIBLE, and 'geodata'(e.g. language names, ISO codes, etc.) from sources such as Glottolog and Ethnologue; includes cross-lingual distance measures based on typological features; provides estimates for empty feature values

FIGURE 1.1 – Vue d'ensemble des bases de données typologiques publiquement accessibles (Tableau 1 de O'Horan et al. (2016)).

dans lesquelles un phénomène linguistique admet plusieurs valeurs. Les langues sont alors classées selon les valeurs qu'elles présentent. En figure 1.1, O'Horan et al. (2016) présentent une vue d'ensemble des bases de données typologiques en précisant le type de caractéristiques traitées, la couverture en termes de langues et de caractéristiques et quelques informations supplémentaires.

Les bases de données sont très hétérogènes dans la couverture des langues et des caractéristiques qu'elles abritent. Nous pouvons remarquer que seulement 17 % des caractéristiques de WALS (Dryer and Haspelmath, 2013) possèdent des valeurs. De plus, la classification tend à masquer la variation intralinguistique dans les valeurs de certains traits. Par exemple, WALS présente le français comme étant une langue ayant un ordre nom - adjectif. Dans les faits, l'ordre adjectif - nom est également possible mais cette information n'est pas explicitée.

Malgré une couverture hétérogène, les bases de données typologiques constituent des recueils d'informations importants. Dans la partie suivante, nous présentons le potentiel que peuvent avoir ces données dans l'amélioration des systèmes de Traitement Automatique des Langues.

1.2.2 Enjeux pour le multilinguisme

Les avancées récentes en Traitement Automatiques des Langues (TAL) ouvrent de plus en plus l'horizon au multilinguisme et à l'ambition de traiter toutes les langues du monde. En réalité, il existe encore aujourd'hui une forte hétérogénéité dans les langues représentées dans le domaine. Joshi et al. (2020) décrivent l'état actuel de la situation en prenant en compte le type de langue, les ressources disponibles et la représentation de la

langue dans les conférences de TAL. Six classes de langues sont établies selon la quantité de ressources annotées et non annotées disponibles, allant de 0 pour désigner la classe des langues les plus pauvres à 5, classe des langues les plus riches en termes de ressources.

La figure 1.2 fait état d'un important écart entre les deux extrêmes : 88,19 % du nombre total de langues n'ont que très peu de ressources disponibles tandis que 0,28 % en disposent d'un grand nombre. Cette disparité dans les ressources se retrouve également en termes de représentation des langues dans les conférences puisque les langues les plus riches en ressources sont les plus représentées.

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

FIGURE 1.2 – Nombre de langues, nombre de locuteurs et pourcentage du total des langues pour chaque classe (Tableau 1 de Joshi et al. (2020)).

Pallier cette hétérogénéité constitue un des défis majeurs du TAL et incite au développement de systèmes indépendants de la langue. Une des possibilités est d'utiliser des systèmes qui ne requièrent aucune spécification linguistique mais seulement de grandes quantités de données sur lesquelles ils seraient entraînés. Bender (2009) insiste toutefois sur la nécessité d'intégrer des connaissances linguistiques afin d'éviter de modéliser des systèmes codant des biais spécifiques à une langue. C'est le cas notamment des modèles n-grammes, à première vue dépourvus de connaissances linguistiques mais qui sont en réalité plus performants sur des langues présentant un ordre de mots relativement fixe et peu de morphologie flexionnelle. Ils ne sont donc pas totalement « language-independent ». Par connaissances linguistiques, Bender fait référence aux données que nous fournit la typologie linguistique et dont l'intégration aux systèmes de TAL est peu coûteuse. Les informations linguistiques obtenues grâce aux classifications des typologistes sont d'ailleurs de plus en plus utilisées pour améliorer les systèmes de TAL. Ponti et al. (2019) parcourent à grande échelle les interactions notables entre ces deux domaines et s'intéressent en particulier aux caractéristiques typologiques utiles pour le TAL et aux types de systèmes qui ont profité de l'intégration de ces caractéristiques.

Quelles caractéristiques typologiques ? Les caractéristiques typologiques relatives à l'ordre des mots sont les plus récurrentes dans les études de TAL utilisant des données typologiques. En effet, ces caractéristiques sont particulièrement utiles dans des tâches morphosyntaxiques comme l'analyse syntaxique en dépendances (Naseem et al., 2012). La figure 1.3 montre les données de WALS relatives à l'ordre des mots qui ont été les plus utilisées dans les travaux sur l'analyse syntaxique en dépendances. Nous pouvons voir que les ordres impliquant le nom et ses dépendants sont les plus exploités, tout comme ceux impliquant le verbe.

En outre, les bases de données typologiques fournissent également des informations en termes de phonologie. Tsvetkov et al. (2016) ont notamment produit un modèle polyglotte entraîné pour prédire des séquences de phones dans différentes langues en intégrant des caractéristiques phonologiques d'URIEL (Littell et al., 2017).

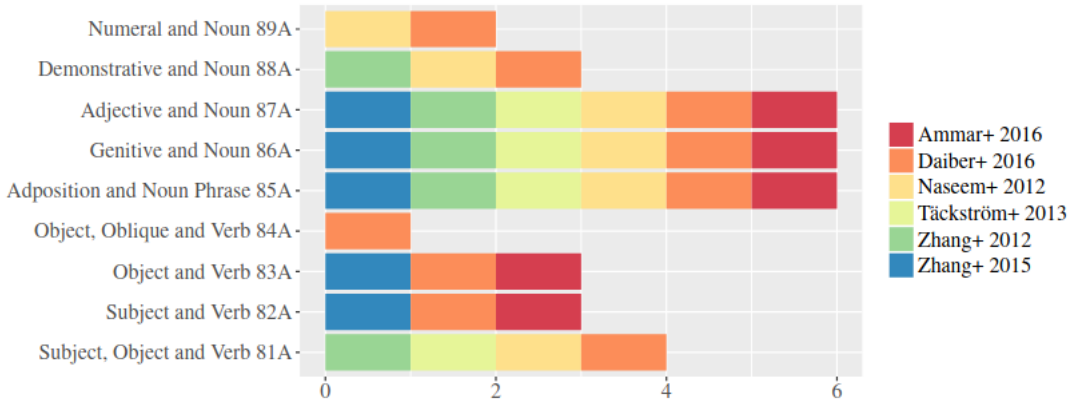

FIGURE 1.3 – Caractéristiques de l'ordre des mots utilisées dans des travaux d'analyse syntaxique en dépendances (Figure 4 de Ponti et al. (2019)).

Dans leurs perspectives, Ponti et al. (2019) voient un fort potentiel des caractéristiques sémantiques fournies dans les bases de données typologiques pour améliorer des tâches telles que la levée d'ambiguïté lexicale et l'analyse des sentiments.

Quels systèmes de TAL ? Pour surmonter les limites du manque de ressources dans certaines langues, la communauté du TAL s'est orientée vers des méthodes d'apprentissage non supervisé. Deux méthodes s'avèrent être prometteuses pour considérer les langues peu dotées : le *language transfer* (transfert linguistique) et le *multilingual joint learning* (apprentissage multilingue conjoint).

Le *language transfer* consiste à transférer des informations linguistiques des langues riches en ressources à celles qui en sont pauvres. De manière générale, les informations provenant des langues sources doivent être adaptées à certaines propriétés des langues cibles. Naseem et al. (2012) ont introduit la notion de *selective sharing* dans leur modèle de *language transfer*. Le modèle a pour but de faire l'analyse syntaxique de phrases d'une langue cible en se basant sur plusieurs langues sources. Si les relations tête-modifieur entre les parties du discours sont universelles, l'ordre des parties du discours est spécifique à une langue. Les relations de dépendances universelles sont donc apprises sur toutes les langues sources tandis que l'ordre des parties du discours est appris à partir des données typologiques intégrées.

Le principe des modèles de *multilingual joint learning* est d'apprendre conjointement sur des données de plusieurs langues et ainsi traiter des tâches multilingues, comme la traduction automatique neuronale. Un autre exemple de tâche utilisant l'apprentissage multilingue conjoint est l'analyse syntaxique sur plusieurs langues avec un seul *parser*. Ammar et al. (2016) développent un *parser* entraîné sur différentes données multilingues : des plongements lexicaux, des annotations de parties de discours à grain fin et des caractéristiques typologiques entre autres. Les résultats montrent que leur système surpasse en moyenne les performances d'un *parser* monolingue, le système étant entraîné sur davantage de données (présentant toutefois plus de bruit) (Ponti et al., 2019). Dans la même lignée, Scholivet et al. (2019) utilisent un *parser* délexicalisé et testent leur système sur 40 langues. Ils démontrent qu'intégrer les descriptions typologiques de WALS permet au *parser* d'apprendre des généralisations interlinguistiques et ainsi améliorer l'analyse syntaxique.

1.2.3 Vers une typologie quantitative

La typologie étant portée par une forte tradition empirique, le développement du TAL multilingue et des ressources numériques a permis des études sur l'ordre des mots dans différentes langues. En termes de ressources numériques, les *treebanks*, ou corpus arborés, constituent des données précieuses et largement exploitées, notamment dans des travaux relatifs à la syntaxe et la sémantique.

Liu (2010) se base sur des *treebanks* de 20 langues différentes pour mesurer les fréquences des dépendances à droite (*head-initial*) et à gauche (*head-final*) mais aussi pour extraire les ordres des mots sujet-verbe, verbe-objet et adjectif-nom. Ses résultats sont en accord avec ceux de WALS, ce qui l'amène à conclure que les ressources de TAL peuvent effectivement être exploitées pour la typologie linguistique.

Cependant, ses travaux datant d'avant la création d'Universal Dependencies, les corpus utilisés sont annotés en syntaxe de dépendances mais dans des schémas différents. Liu utilise les versions converties aux schémas CoNLL-X'06 (Buchholz and Marsi, 2006) et CoNLL-X'07 (Nivre et al., 2007). La plupart des corpus sont maintenant présents dans UD. L'auteur précise lui-même dans ses perspectives qu'il serait intéressant d'utiliser des corpus parallèles annotés selon le même schéma d'annotations.

C'est ce qu'a entrepris Östling (2015) en utilisant des corpus parallèles du Nouveau Testament traduit et aligné sur 986 langues. Les textes sources sont annotés avec les étiquettes des parties du discours *Universal PoS Tags* (Petrov et al., 2012) et la structure de dépendances est au format *Universal Dependency Treebank* (McDonald et al., 2013). Les textes sont ensuite alignés avec un outil d'alignement multilingue (Östling, 2014). Östling s'est intéressé à cinq ordres de mots, présentés en figure 1.4, et a comparé ses résultats avec ceux de WALS. Dans cette étude, Östling cherche plus à évaluer les performances de son système d'alignement qu'à établir un ordre des mots dominant des langues considérées.

Feature	Languages	Types	Tokens	Most common
81A: Subject, Object, Verb (Dryer, 2013e)	342	85.4%	85.7%	SOV: 43.3%
82A: Subject, Verb (Dryer, 2013d)	376	89.4%	90.4%	SV: 79.8%
83A: Object, Verb (Dryer, 2013c)	387	96.4%	96.4%	VO: 54.8%
85A: Adposition, Noun Phrase (Dryer, 2013b)	329	94.8%	95.1%	Prep: 50.4%
87A: Adjective, Noun (Dryer, 2013a)	334	85.9%	88.0%	AdjN: 68.9%

FIGURE 1.4 – Accord entre WALS et les résultats d'Östling (2015). La fréquence de l'ordre le plus commun est donnée en comparaison (Tableau 2 de Östling (2015)).

Dès l'apparition du projet Universal Dependencies, les corpus ont été largement utilisés dans des travaux typologiques multilingues traitant de l'ordre des mots.

Futrell et al. (2015) se sont intéressés à la notion de liberté de l'ordre qu'ils tentent de mesurer quantitativement sur des corpus d'UD en 34 langues. Ils utilisent pour cela une mesure d'entropie qui se révèle être fiable pour estimer la variabilité dans la direction d'une relation syntaxique. Ils démontrent également une corrélation entre la liberté de l'ordre du sujet et de l'objet et la présence du marquage de cas nominatif-accusatif. Berdicevskis and Piperski (2020) examinent ce phénomène en russe et en allemand et confirment que l'absence de marqueur de cas entraîne un ordre du sujet, du verbe et de l'objet plus rigide.

Dans le même esprit que Futrell et al. (2015), Levshina (2019) utilise les corpus d'UD et la mesure d'entropie pour évaluer la variabilité de l'ordre des mots mais adopte une approche plus centrée sur les relations de dépendance syntaxique. Elle propose ainsi une classification des langues selon l'entropie dans l'ordre des mots.

Chen and Gerdes (2017) démontrent que les corpus d'UD permettent d'obtenir une autre classification en se basant uniquement sur des structures délexicalisées des arbres. En effet, leurs expériences ne prennent en compte que les types, les fréquences et les directions des relations de dépendances. Ils proposent alors une classification de 43 langues en utilisant la mesure des *Directional Dependency Distances* (DDD) qui montre notamment à quel point une langue est centripète (tête-finale) ou centrifuge (tête-initiale). La figure 1.5 présente les langues centripètes avec des valeurs négatives et les langues centrifuges avec des valeurs positives. Les auteurs remarquent que cette classification apporte de nouvelles informations sur certaines langues, comme par exemple pour le chinois qui est fortement centripète, caractéristique qui n'est pas explicite dans la classification en ordre SVO, SOV etc.

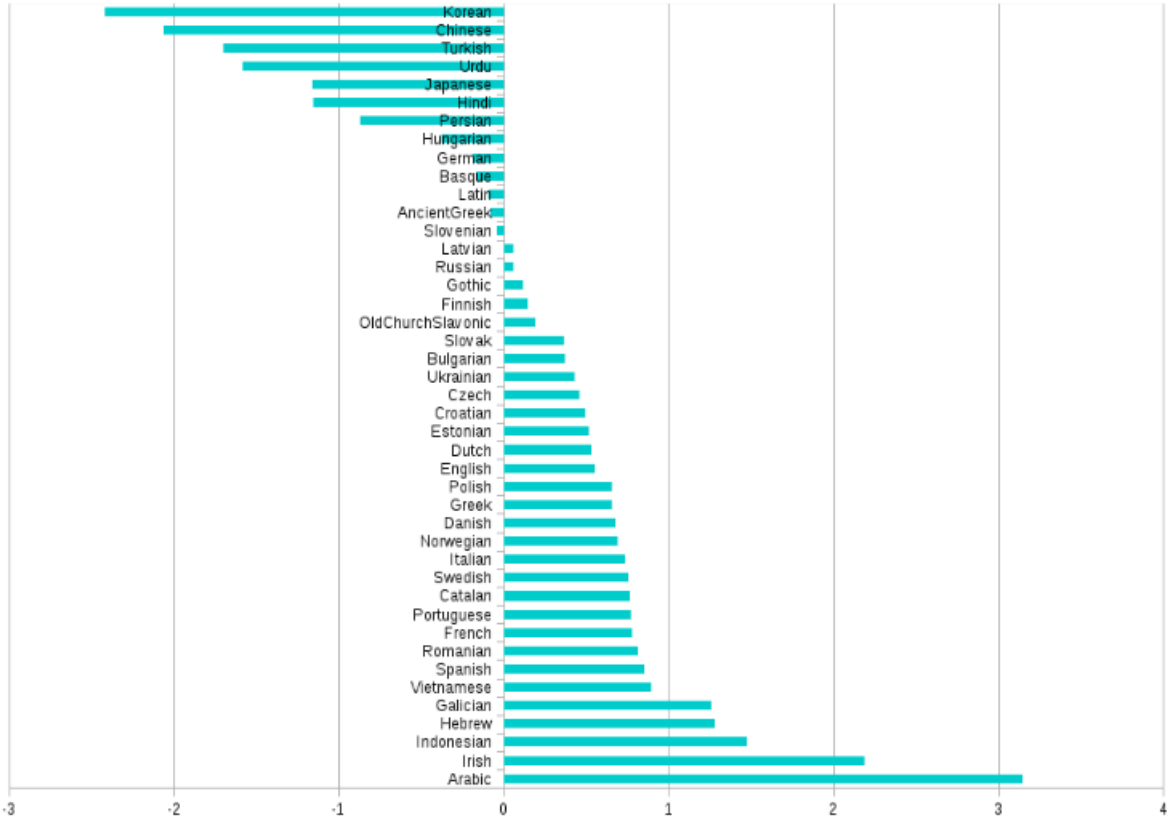

FIGURE 1.5 – Langues classées selon la distance de dépendance (Figure 1 de Chen and Gerdes (2017)).

Par ailleurs, Chen and Gerdes (2017) établissent une autre classification mais cette fois-ci par corpus afin de mesurer la cohérence entre les différents corpus d'une même langue. Bien que les corpus d'une même langue présentent des valeurs de distance relativement proches, les auteurs remarquent certains écarts et les relient à la présence d'incohérences entre les corpus. Leurs méthodes peuvent donc être exploitées dans la détection de similarités ou d'erreurs dans les corpus.

Dans une moindre mesure, Alzetta et al. (2018) utilisent des corpus d'UD de trois

langues, l'anglais, l'espagnol et l'italien, pour identifier des caractéristiques typologiques sur l'ordre sujet - verbe et adjectif - nom. En utilisant un algorithme d'évaluation de plausibilité aux corpus d'UD, ils parviennent à reconstruire les ordres des mots attestés dans la littérature.

Enfin, Gerdes et al. (2021) s'intéressent à l'exploration quantitative d'universaux de Greenberg sur les corpus SUD, *Surface-Syntactic Universal Dependencies*, qui constituent des corpus annotés en syntaxe de dépendances dans un schéma dérivé d'UD (Gerdes et al., 2018, 2019). L'objectif de leur étude est de mettre en lumière le caractère quantitatif de certains universaux à travers des diagrammes de « typométrie » et ainsi déterminer ce qu'ils appellent des « universaux quantitatifs ». Les auteurs proposent deux types de diagramme, un diagramme à une dimension et un à deux dimensions dans lesquels ils représentent les langues selon les fréquences d'apparition des phénomènes considérés.

La figure 1.6 présente le diagramme à deux dimensions du pourcentage de l'ordre verbe - objet pronominal (VOpron) en fonction de l'ordre verbe - objet nominal (VOnom) qui correspond à l'universel 25 de Greenberg :

Universal 25 : If the pronominal object follows the verb, so does the nominal object[5].

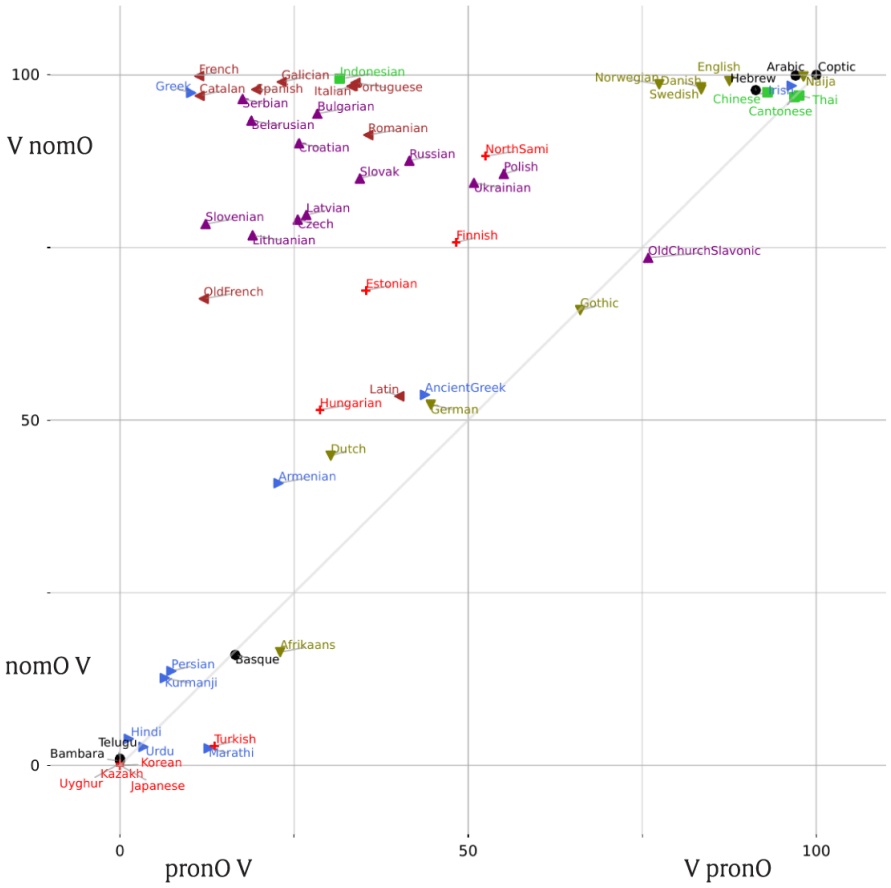

FIGURE 1.6 – Graphique en nuage de points du pourcentage de VOpron en fonction du pourcentage de VOnom (Figure 2 de Gerdes et al. (2021)).

Les auteurs décident de traiter quantitativement l'universel 25, qui est un universel absolu et implicationnel, en ajoutant une notion de seuil :

5. Si l'objet pronominal suit le verbe, alors l'objet nominal le suit également.

Universal 25' : For every language, if the percentage of pronominal objects on the right of the verb is greater than 75%, so is the percentage of nominal objects on the right of the verb[6].

Avec cet universel, ils remarquent que les diagrammes présentent des zones vides qui varient en fonction du seuil, comme le montre la figure 1.7. La négation d'une implication $A \rightarrow B$ étant $A \wedge \neg B$, l'universel 25' signifie donc qu'il n'y a aucune langue avec un ordre VOpron $\geq$ 75 % et un ordre VOnom $<$ 75 %, ce qui est traduit par la partie grise vide sur la figure à gauche en 1.7.

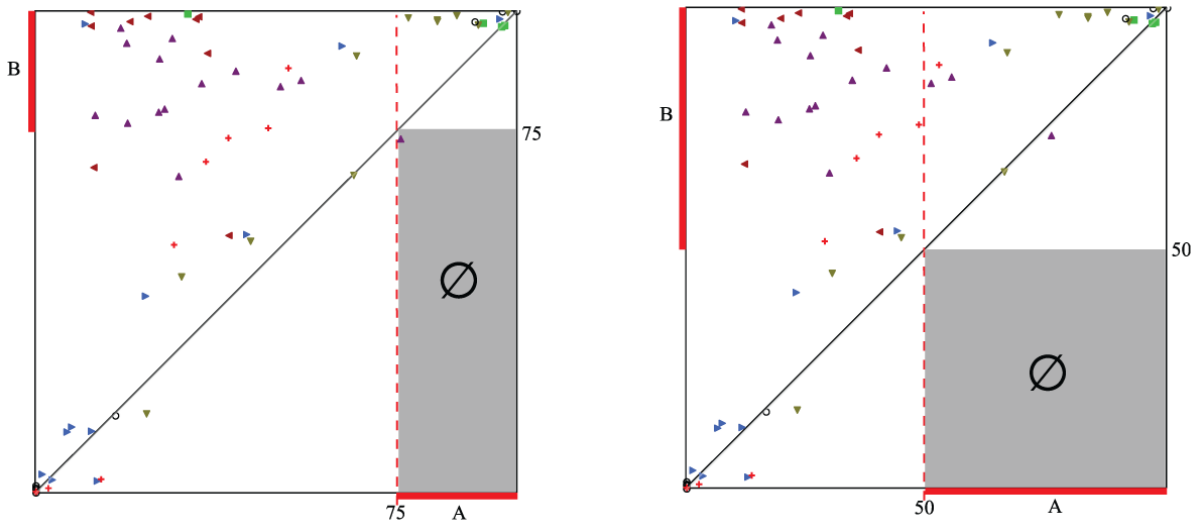

FIGURE 1.7 – Universel 25'
a. VOpron $\geq$ 75 % $\rightarrow$ VOnom $\geq$ 75 %
b. VOpron $\geq$ 50 % $\rightarrow$ VOnom $\geq$ 50 % (Figure 3 de Gerdes et al. (2021)).

En utilisant la représentation en diagramme des pourcentages observés dans les corpus, Gerdes et al. (2021) montrent que les universaux implicationnels absolus présentent des zones vides dans les diagrammes tandis que les universaux implicationnels statistiques (non-absolus) présentent des zones presque vides. Cette approche typométrique a pour perspective d'inférer de nouveaux universaux en s'appuyant sur des données continues et non catégorielles comme le font les universaux traditionnels.

Bien que nos expériences s'inscrivent dans la lignée des travaux de Gerdes et al. (2021), nos expériences se distinguent des leurs dans le choix de traiter les corpus individuellement et non de regrouper les corpus d'une même langue ensemble.

6. Pour toutes les langues, si le pourcentage d'objets pronominaux à la droite du verbe est supérieur à 75 %, alors le pourcentage d'objets nominaux à la droite du verbe l'est aussi.

Chapitre 2

Méthodologie

Sommaire

2.1 Universal Dependencies

2.1.1 Des corpus pour le TAL multilingue

Universal Dependencies (UD) est un projet collaboratif de création de corpus annotés en syntaxe de dépendances suivant un schéma d'annotations universel. Ce schéma d'annotations a été pensé pour être applicable à un grand nombre de langues et cohérent dans l'annotation de constructions similaires, tout en permettant des extensions propres à chaque langue.

Par ce schéma, le projet a pour but de faciliter le développement d'analyseurs syntaxiques (*parser*) multilingues, l'apprentissage interlinguistique et la recherche sur l'analyse syntaxique dans une perspective de typologie linguistique (Nivre et al., 2020). Pour cela, le projet constitue un compromis entre six points :

1. UD doit être satisfaisant en termes d'analyse linguistique pour chaque langue.
2. UD doit permettre la typologie linguistique, fournir une base appropriée pour mettre en évidence le parallélisme interlinguistique entre les langues et les familles de langue.
3. UD doit permettre une annotation rapide et cohérente pour un annotateur humain.
4. UD doit être facilement compréhensible et utilisable par un non-linguiste.
5. UD doit permettre une analyse syntaxique automatique précise.
6. UD doit supporter des tâches de compréhension du langage comme l'extraction de relations, la traduction automatique. . ..

La difficulté d'UD consiste à assurer ces points tout en allant vers une constante amélioration du schéma d'annotations et une expansion du nombre de corpus et de langues. Depuis 2014, les corpus et les langues disponibles dans UD augmentent de manière soutenue, grâce à une mise à jour des versions, deux fois par an en mai et en novembre. La figure 2.1 présente les différentes versions en termes de nombre de corpus et de langues.

L'aspect collaboratif du projet permet une certaine liberté dans le sens où aucune contrainte concernant la langue ou la taille du corpus n'est posée. Par conséquent, nous

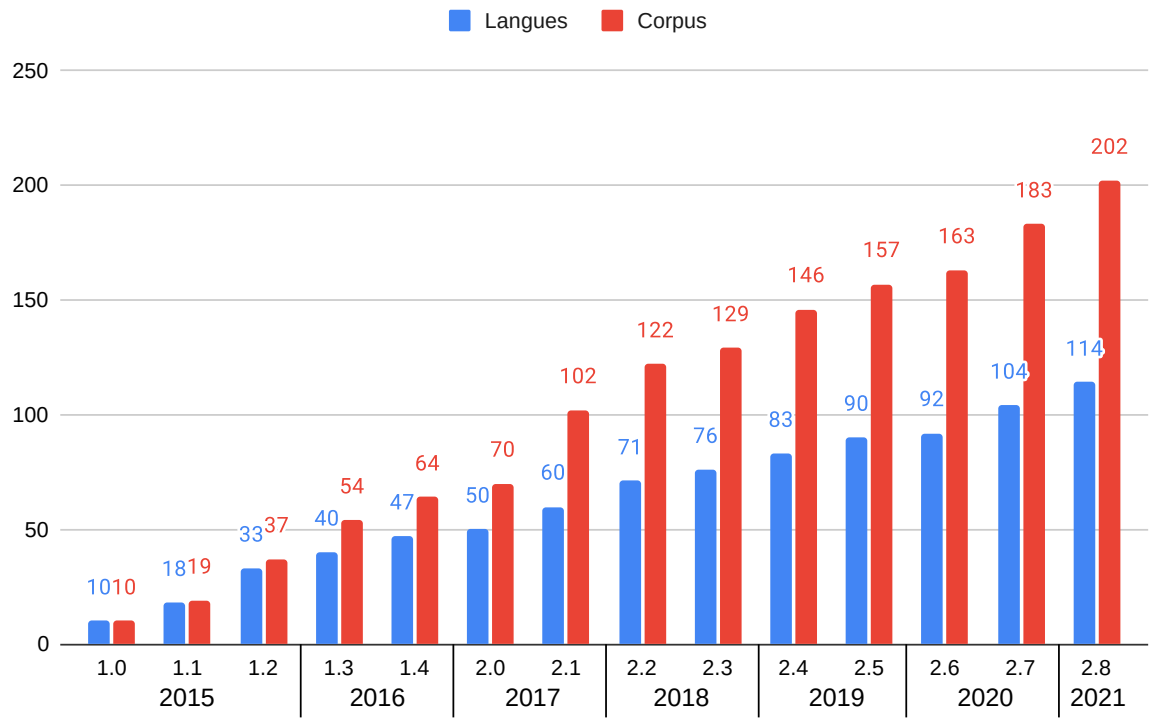

FIGURE 2.1 – Évolution des versions d'UD en nombre de corpus et de langues.

pouvons observer une hétérogénéité entre les corpus à plusieurs niveaux : la taille, le genre de texte, la quantité d'annotations, la manière dont ils ont été annotés (manuellement, conversion automatique depuis un corpus pré-existant...). De plus, les familles de langues sont représentées de manière disparate, la famille indo-européenne étant largement plus présente que les autres familles. En effet, 62 langues sur 114 sont de la famille indo-européenne dans UD 2.8.

La taille UD ne pose aucune limite de taille pour soumettre un corpus. Dans UD 2.8, version la plus récente, le plus petit corpus, `Soi-AHA` [1], compte huit phrases tandis que le plus grand, le `German-HDT` en compte 189 928. Sur les 202 corpus d'UD 2.8, 52 d'entre eux comptent moins de 1 000 phrases, soit 25 % du nombre total de corpus. Les corpus de petites tailles sont généralement des corpus de langues peu dotées en termes de ressources. La constitution du corpus est dans ce cas faite manuellement, donc sur un nombre limité de phrases.

Le genre Les corpus d'UD présentent un éventail varié de genres. Les textes peuvent provenir de journaux, de textes de la bible, d'œuvres de fiction, de textes de loi, de Wikipedia ou d'exemples de grammaire etc. Par ailleurs, certains corpus sont des transcriptions de corpus oraux, ce qui permet de mettre en valeur les spécificités de la langue parlée.

La quantité d'annotations Comme le montrent les figures 2.2 et 2.3, la quantité d'annotations est très variable selon les corpus, UD n'exigeant comme annotations obligatoires que les dépendances syntaxiques et les étiquettes des parties du discours. Des annotations

1. Le soi (ou sohi) est une langue iranienne.

supplémentaires peuvent être ajoutées facultativement telles que la lemmatisation et des traits de morphologie (le genre, le nombre, le temps et le mode des verbes...).

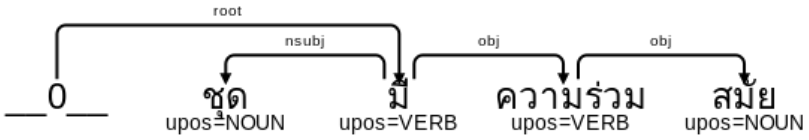

FIGURE 2.2 – Annotations du corpus `Thai-PUD`.

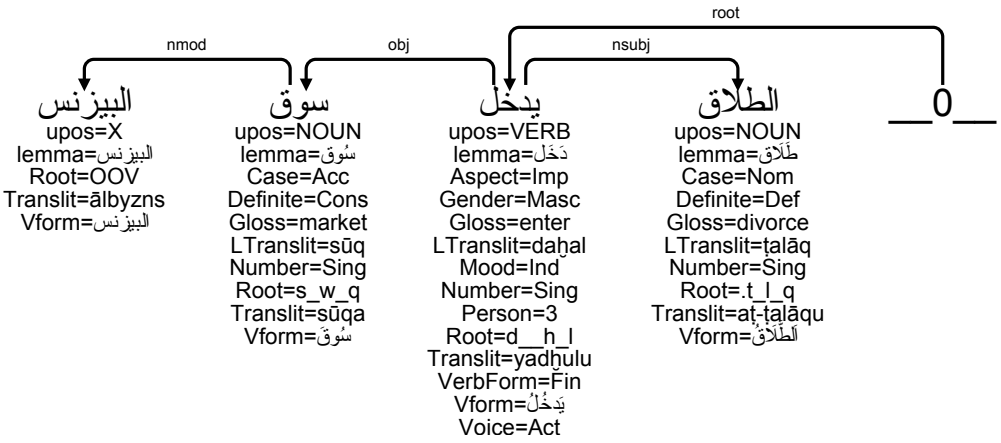

FIGURE 2.3 – Annotations du corpus `Arabic-PADT`.

L'annotation au schéma UD Les corpus sont majoritairement des corpus annotés selon d'autres schémas qui ont été convertis vers le schéma UD. À titre d'exemple, les corpus `PUD` (Parallel Universal Dependencies) ont été annotés avec le schéma d'annotations de McDonald et al. (2013) et convertis ultérieurement. D'autres, relativement récents, ont été annotés selon le schéma UD dès leur constitution. Ce paramètre peut influencer sur la qualité des annotations et peut expliquer l'hétérogénéité entre des corpus d'une même langue qui ne présentent pas exactement les mêmes annotations.

2.1.2 Un schéma d'annotations « universel »

Le schéma d'annotations UD est le résultat de l'évolution de trois projets : Stanford Dependencies (De Marneffe et al., 2006), les étiquettes des parties du discours de Google universal (Petrov et al., 2012) et les étiquettes morphosyntaxiques d'Interset (Zeman, 2008). Ce schéma universel permet d'annoter des corpus de différentes langues en syntaxe de dépendances et ainsi faciliter les analyses interlinguistiques.

La syntaxe de dépendances La syntaxe constitue la branche de la linguistique qui décrit les règles selon lesquelles les mots se combinent pour former des énoncés dans une langue. Il existe différentes façons de modéliser la syntaxe, cependant la syntaxe de

dépendances est largement influencée par la notion de constituants, que nous retrouvons également dans le modèle syntagmatique de la syntaxe (Bonfante et al., 2018).

La syntaxe de dépendances trouve son origine dans les travaux de Tesnière (1959) qui utilise les dépendances pour représenter les fonctions reliant les mots entre eux, telles que la fonction sujet, objet ou déterminant. L'architecture correspond alors à des triplets, un mot-source (le gouverneur) et un mot-cible (le dépendant) reliés par une fonction.

Quant à la syntaxe syntagmatique ou de constituants de Chomsky (1957), elle se base sur des divisions binaires en syntagmes. Un syntagme correspond à un ensemble de mots formant une unité syntaxique et sémantique d'une phrase. Il est composé d'une tête (ou noyau) qui détermine sa fonction dans la phrase. Un syntagme peut donc être de plusieurs types : nominal (SN), verbal, (SV), adjectival (SA), prépositionnel (SP) et adverbial (SADV).

Ces deux syntaxes peuvent être représentées sous forme d'arbres. Les figures 2.4 et 2.5 présentent respectivement l'arbre syntagmatique et l'arbre de dépendances de la phrase suivante :

(1) Les champignons produisent de puissants antibiotiques. [2]

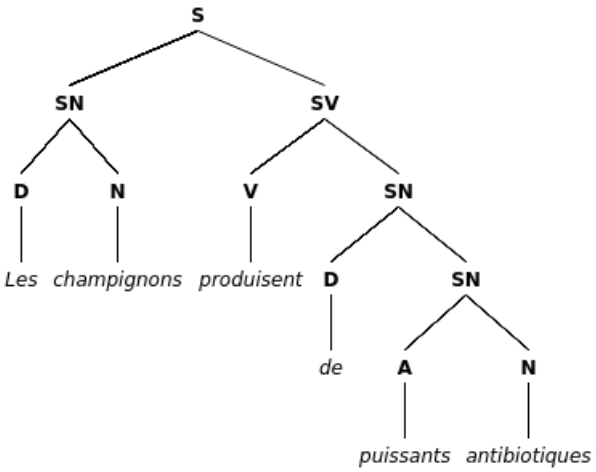

FIGURE 2.4 – Arbre syntagmatique de la phrase (1).

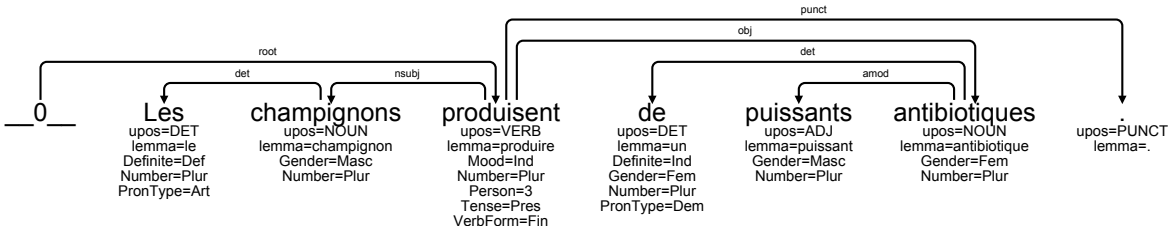

FIGURE 2.5 – Arbre de dépendances de la phrase (1) selon le schéma UD.

Dans une représentation en dépendances, la flèche traduit une relation syntaxique qui part d'un gouverneur vers un dépendant. Des algorithmes reposant sur la notion de tête d'un syntagme, permettent de passer d'un arbre syntagmatique à un arbre de dépendances.

2. [fr_partut-ud-888] extraite du corpus French-ParTUT.

Nous précisons que l'arbre de dépendances en 2.5 est la version corrigée de l'arbre de dépendances présent dans UD, qui se trouve en figure 2.6. En effet, nous nous sommes rendu compte que les annotations ne respectaient pas les règles d'UD avec l'adjectif « puissants » annoté comme un oblique et non un adjectif.

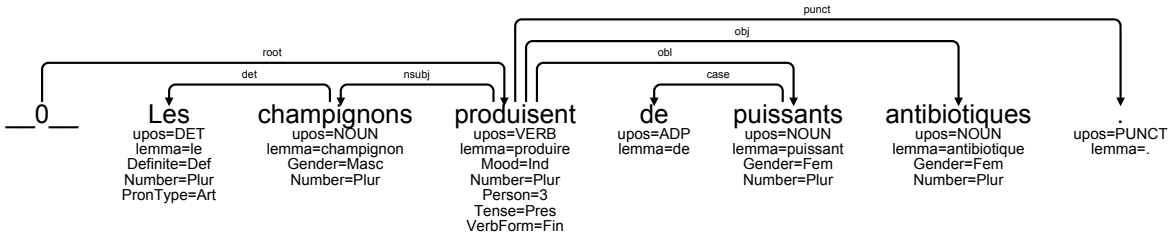

FIGURE 2.6 – Arbre de dépendances de la phrase (1) dans le corpus `French-ParTUT`.

Vers un parallélisme entre les langues L'annotation au schéma UD se fait en plusieurs étapes : la segmentation en *tokens*, l'étiquetage en parties du discours et en traits morphologiques et la liaison entre les *tokens* par des relations syntaxiques.

L'annotation en dépendances implique la segmentation des corpus en *tokens* et plus précisément en « mots syntaxiques ». Le schéma UD demande de segmenter les *tokens* présentant des clitiques comme en espagnol « dámelo = da me lo » et les contractions comme en français « au = à le ». UD précise que les méthodes de segmentation peuvent varier selon les spécificités des langues et que chaque corpus doit être documenté sur la manière de segmenter [3].

Concernant l'étiquetage des *tokens*, UD met à disposition 17 étiquettes de parties du discours [4] et 24 traits morphologiques universels [5] présentés dans le tableau 2.1 et en figure 2.7 [6]. Il existe également d'autres traits morphologiques non définis dans les directives d'UD mais qui ont été ajoutées dans certaines langues. Chaque étiquette présente une définition universelle relativement large mais il est possible d'ajouter une documentation plus précise selon les langues.

Nous précisons que toutes les informations ont été prises telles qu'elles sont présentées dans la documentation d'UD. Les classifications sont évidemment discutables, notamment la division en classe de mots ouverte ou fermée. UD indique également que la division en caractéristiques lexicales et flexionnelles est approximative, tout comme la distinction entre étiquettes nominales et verbales.

La figure 2.8 détaille les 37 relations syntaxiques d'UD. Les relations syntaxiques sont classées selon plusieurs catégories :

— les catégories fonctionnelles en relation avec la tête : arguments centraux des prédicats propositionnels, dépendants non-centraux des prédicats propositionnels et dépendants des nominaux (en lignes),

3. Voir : https://universaldependencies.org/u/overview/tokenization, septembre 2021.
4. Voir : https://universaldependencies.org/u/pos/index, septembre 2021.
5. Voir : https://universaldependencies.org/u/dep/index, septembre 2021.
6. Le détail des traits morphologiques est disponible en annexe 4.3.2.

Classe de mots ouverte	Classe de mots fermée	Autres
ADJ : adjectif	ADP : adposition	PUNCT : ponctuation
ADV : adverbe	AUX : auxiliaire	SYM : symbole
INTJ : interjection	CCONJ : conjonction de coordination	X : autre
NOUN : nom	DET : déterminant	
PROPN : nom propre	NUM : numéral	
VERB : verbe	PART : particule	
	PRON : pronom	
	SCONJ : conjonction de subordination	

TABLEAU 2.1 – Parties du discours du schéma d'annotations UD.

Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

FIGURE 2.7 – Traits morphologiques universels du schéma d'annotations UD (Figure extraite de la documentation d'UD).

— les catégories structurelles du dépendant : nominaux, propositions, mots modificateurs et mots fonctionnels (en colonnes).

Les relations dans la partie inférieure du tableau correspondent à celles qui ne sont pas des relations de dépendances au sens strict.

Certaines relations peuvent présenter des extensions comme `poss` dans `det:poss` qui spécifie un déterminant possessif. Comme les traits morphologiques, des extensions sont décrites universellement mais chaque langue peut ajouter des extensions spécifiques [7].

Pour permettre une meilleure analyse entre les langues, les annotations d'UD donnent davantage d'importance aux mots lexicaux (les noms, les verbes, les adjectifs et certains adverbes) qui sont plus stables que les mots grammaticaux (Guillaume et al., 2019).

Dans la figure 2.9, une même phrase est représentée en français et en russe, les relations en trait continu portant sur les mots lexicaux et les relations en pointillés sur les mots grammaticaux. Nous pouvons observer que les annotations sur les mots lexicaux permettent de conserver les relations syntaxiques en trait continu et donc de faire un parallèle entre ces deux langues.

7. L'annexe 4.3.2 présente le détail des relations syntaxiques et des extensions universelles.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

FIGURE 2.8 – Relations syntaxiques dans le schéma d'annotations UD (Figure extraite de la documentation d'UD).

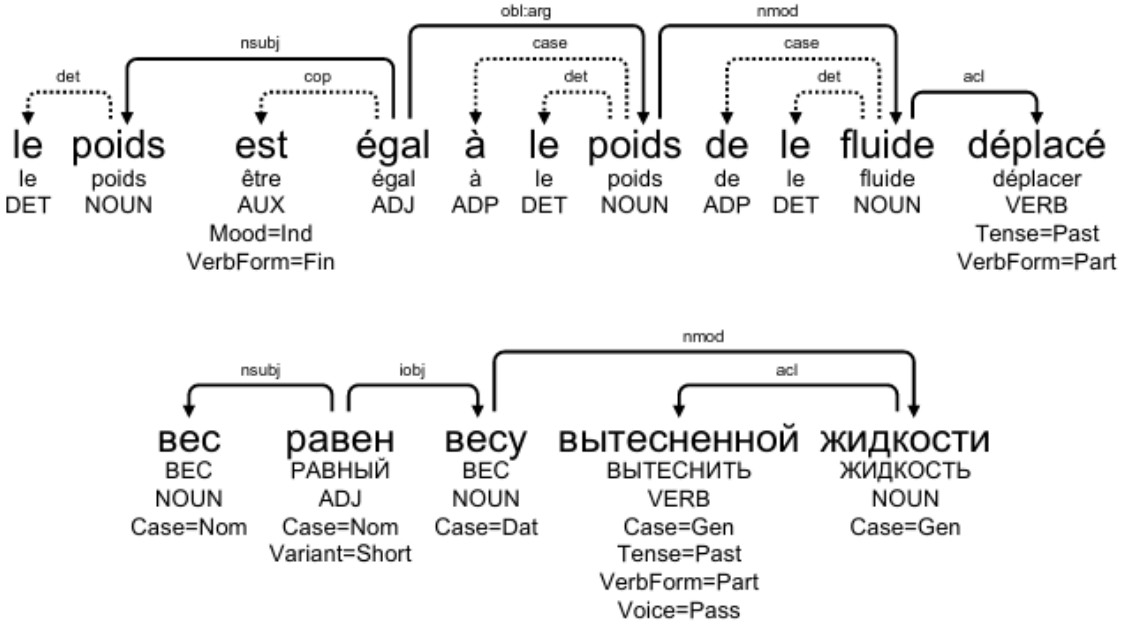

FIGURE 2.9 – Annotations UD pour la même phrase en français et en russe (Figure 1 de Guillaume et al. (2019)).

Toutefois, ce choix n'est pas sans conséquence et certaines constructions linguistiques en pâtissent. C'est notamment le cas des verbes impliquant une prépositions comme « compter sur quelqu'un » ou « dépendre de quelqu'un ». Avec le schéma UD, la préposition se retrouve dépendante de « quelqu'un » alors qu'elle est, en réalité, inhérente au verbe.

2.1.3 De UD 2.7 à UD 2.7_{1K}

Dans notre étude, nous utilisons la version 2.7 d'UD, sortie en novembre 2020. Cette version était la plus récente lors du début de nos expériences, nous avons donc décidé de nous y tenir bien que la version 2.8 soit sortie en mai 2021.

La version 2.7 d'UD contient 104 langues et 183 corpus. Nos expériences impliquant l'extraction d'occurrences et des comparaisons de fréquences, nous choisissons d'éliminer les corpus de moins de 1 000 phrases que nous considérons comme trop petits pour être représentatifs. Une fois ce filtre appliqué, nous obtenons un ensemble de 74 langues et 141 corpus que nous appelons UD 2.7_{1K} [8].

Nous décidons également de ne pas regrouper les corpus d'une même langue en un seul corpus afin d'examiner la cohérence entre les corpus. Sur les 74 langues, 29 langues présentent plus d'un corpus et 45 langues n'ont qu'un seul corpus. La figure 2.10 résume la composition d'UD 2.7 et d'UD 2.7_{1K}.

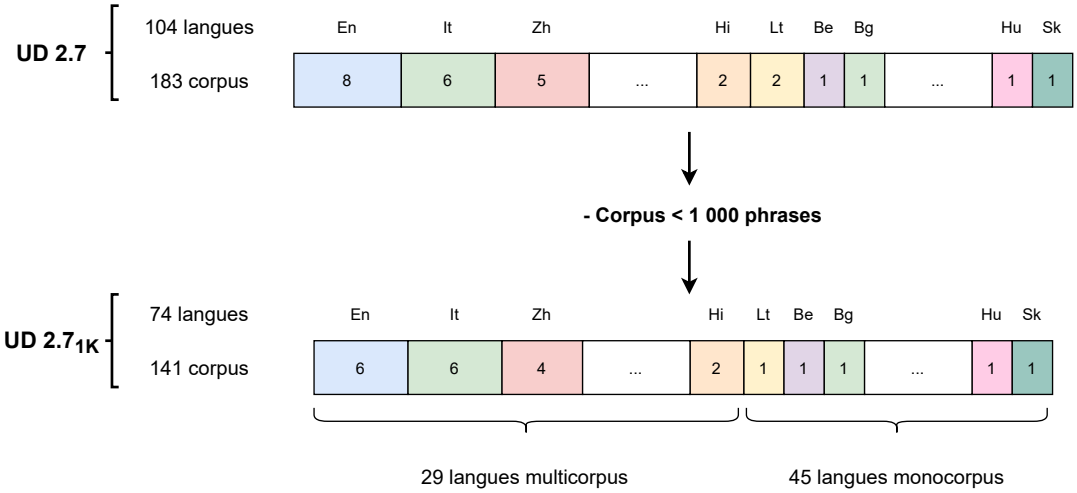

FIGURE 2.10 – Composition en corpus et en langues d'UD 2.7_{1K}.

Sur les 30 langues de l'échantillon de Greenberg, 14 langues sont présentes l'ensemble UD 2.7_{1K} :

— Le basque, le serbe, le gallois, le norvégien, le grec moderne, l'italien et le finnois (Europe).
— Le yoruba (Afrique).
— Le turc, l'hébreu, l'hindi, le japonais et le thaï (Asie).
— Le guarani (Amérique indienne).

Nous remarquons que toutes les langues européennes de l'échantillon de Greenberg sont dans UD 2.7_{1K}, ce qui n'est pas le cas des autres familles de langues. En effet, il réside une forte hétérogénéité dans les langues représentées dans UD en général mais également dans notre échantillon. Le tableau 2.2 montre que 76 % des phrases et 65 % des corpus d'UD 2.7_{1K} appartiennent à la famille indo-européenne.

Nous notons également que notre échantillon contient deux corpus *code-switching*, le `Hindi_English-HIENCS` et le `Turkish_German-SAGT`. Le *code-switching* ou l'alternance codique en français, désigne de manière générale l'utilisation de deux langues dans un

8. La liste complète des langues et des corpus est disponible en annexe 4.3.2.

même discours ou dans une même phrase. Les caractéristiques relatives à chaque langue sont donc susceptibles d'apparaître dans des proportions variables dans ces corpus.

Famille de langue	Phrases	Corpus	Langues
Indo-européenne	1 121 320	92	43
Ouralienne	79 053	9	5
Japonique	66 099	3	1
Sino-tibétaine	61 436	6	3
Afro-asiatique	41 438	8	6
Coréenne	34 702	3	1
Turque	23 810	6	3
Créole	9 242	1	1
Basque	8 993	1	1
Austronésienne	7 623	3	1
Code-switching	3 789	2	2
Austro-asiatique	3 000	1	1
Nigéro-congolaise	2 107	1	1
Dravidienne	1 328	1	1
Tupi	1 046	1	1
Mandées	1 026	1	1
Tchouktches-kamtchadales	1 004	1	1
Taï-Kadaï	1 000	1	1

TABLEAU 2.2 – Taille des familles de langue d'UD 2.7_{1K} en nombre de phrases, de corpus et de langues.

Choisir le « bon » échantillon de langues pour la recherche d'universaux est une tâche délicate. Si Greenberg a choisi des langues de différents continents pour obtenir une couverture génétique plus large, l'échantillon reste toutefois petit avec seulement 30 langues. Dans notre cas, nous disposons d'un plus grand échantillon en termes de taille avec 74 langues, cependant l'hétérogénéité dans les familles est plus marquée. Nous sommes évidemment conscients des biais de cette hétérogénéité mais nous choisissons néanmoins de traiter le maximum de corpus et de langues que nous propose UD.

Le projet UD représente une source de données profitable aux travaux multilingues mais il convient de garder à l'esprit qu'il existe différents biais.

Un premier exemple est le biais lié au genre des corpus. En effet, UD précise le genre des textes dans la documentation de chaque corpus. D'après ces informations, nous remarquons que les corpus peuvent être de genres très différents et parfois même au sein d'un même corpus. Il existe des corpus contenant des extraits de journaux, de Wikipedia, de romans, de poésie, de réseaux sociaux et également de langue orale.

Le second biais se situe dans les différences d'annotations. Bien qu'un schéma universel existe, les annotations peuvent varier selon les langues, du fait de définitions différentes de certains concepts linguistiques par exemple. Cette divergence peut également être présente entre les corpus d'une même langue. Les corpus n'étant pas annotés par les mêmes équipes d'annotateurs, des choix linguistiques différents peuvent entraîner des incohérences.

En raison de ces biais, nous décidons de considérer chaque corpus et non de regrouper des corpus d'une même langue en un seul ensemble. Cette approche nous permet d'une part, de constater une variation de phénomènes linguistiques selon les caractéristiques

propres à un corpus et d'autre part, de mesurer l'homogénéité et détecter des possibles incohérences entre les corpus d'une même langue.

2.2 Grew

Pour interroger les corpus, nous avons utilisé Grew[9], *Graph Rewriting for NLP*, un outil de réécriture de graphes dédié aux applications de TAL telles que l'analyse syntaxique ou le passage d'une annotation syntaxique à une annotation sémantique. Grew présente l'avantage de reconnaître des motifs sur les corpus d'UD sans avoir à manipuler le code de reconnaissance en lui-même. Avec des scripts Python, il aurait fallu changer le code à chaque nouvelle requête, processus chronophage et source d'erreurs.

2.2.1 La réécriture de graphes

La réécriture de graphes permet de reconnaître un motif dans un graphe et de le transformer selon des règles de réécriture.

Dans les expériences présentées dans ce mémoire, nous utilisons uniquement la fonction de reconnaissance de motifs de Grew mais nous avons également tenté de pallier certaines limites de la syntaxe de dépendances avec des règles de réécriture, notamment le cas de la coordination dans l'extraction de l'ordre sujet - verbe - objet (Choi et al., 2021). En effet, la syntaxe de dépendances ne permet pas à un sujet d'être relié à plusieurs verbes. Dans l'exemple en figure 2.11, deux cas d'ordre SVO sont présents : *He obtains these things* et *(He) loses the ability*. Cependant, le deuxième cas n'est pas détecté car *loses* n'est pas relié au sujet. Nous avons alors ajouté une nouvelle relation, appelée `isubj` pour *implicit subject*, avec une règle de réécriture. Le motif de reconnaissance est ensuite adapté pour prendre en compte cette nouvelle relation.

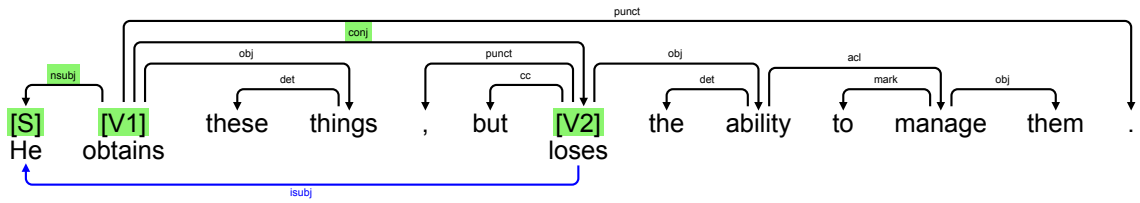

Figure 2.11 – Ajout d'une relation `isubj` avec la réécriture de graphes.

D'autres types d'annotations peuvent être ajoutés grâce à la réécriture de graphes. Guillaume and Perrier (2012) l'ont notamment utilisée pour annoter automatiquement le `French-Treebank` en dépendances sémantiques à partir des annotations syntaxiques de surface.

Outre l'annotation de corpus, une autre utilité de la réécriture de graphes est la conversion entre différents schémas (Guillaume, 2021). Nous pouvons citer l'exemple du corpus `Sequoia` (Candito and Seddah, 2012). Ce corpus, annoté en dépendances selon un schéma différent d'UD, a été converti avec Grew et donc par réécriture de graphes au format UD.

Un autre exemple de conversion est celui entre le schéma UD et le schéma SUD, *Surface-Syntactic Universal Dependencies*. SUD est un autre schéma d'annotations en

9. https://grew.fr

dépendances, qui contrairement à UD, se base sur des critères syntaxiques favorisant les têtes fonctionnelles. De plus, les relations sont définies sur des bases distributionnelles et fonctionnelles (Gerdes et al., 2018, 2019). La réécriture de graphes permet de convertir de manière bidirectionnelle les corpus d'UD en SUD. Ces derniers sont d'ailleurs disponibles sur Grew-match.

2.2.2 Grew, mode d'emploi

Dans cette partie, nous présentons synthétiquement le fonctionnement de Grew en commençant par la syntaxe permettant de faire des requêtes sur les corpus UD. Ensuite, nous expliquerons les différentes fonctionnalités de Grew et Grew-match que nous avons utilisées.

La syntaxe de Grew Les motifs de reconnaissance peuvent s'appliquer à toutes les annotations disponibles, une étiquette de partie de discours, de trait morphologique ou une relation syntaxique.

La figure 2.12 donne un exemple basique d'un motif de reconnaissance et le résultat obtenu dans un corpus annoté au schéma UD. Des nœuds sont fixés avec des noms de notre choix, ici `S` et `V`. Ils sont décrits de la manière suivante :

- `S [upos=PRON, Gender="Fem"]` : le nœud `S` correspond à un pronom féminin,
- `V [upos=VERB, lemma="faire"]` : le nœud `V` correspond à un verbe ayant pour lemme « faire »,
- `V -[nsubj]-> S` : les nœuds `S` et `V` sont reliés par une relation de sujet nominal.

Il est également possible de filtrer sur des différences et non des égalités. Par exemple, la requête `S [upos=PRON, Gender<>"Fem"]` renvoie les pronoms qui ne sont pas féminins.

```
pattern {
  S [upos=PRON, Gender="Fem"];
  V [upos=VERB, lemma="faire"];
  V -[nsubj]-> S;
}
```

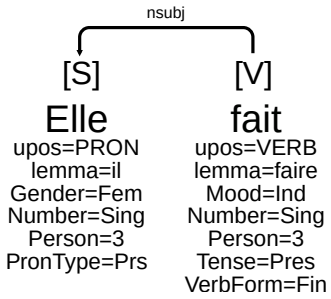

Figure 2.12 – Exemple de motif de reconnaissance Grew.

Grew permet également de faire des requêtes avec des conditions d'exclusion comme le montre la figure 2.13. Cette requête renvoie uniquement les verbes non reliés à une entité `S` par une relation `nsubj`.

```
pattern {
  V [upos=VERB]
}
without {
  V -[nsubj]-> S
}
```

Figure 2.13 – Motif Grew impliquant une condition exclusive.

Dans nos expériences sur l'ordre des mots, GREW est un outil adéquat puisqu'il nous permet de contrôler l'ordre d'apparition des nœuds définis. La requête en figure 2.14 renvoie les cas où un pronom `P` précède un nom `N`. Deux chevrons << signifient que `P` se trouve avant `N`, tandis qu'un chevron simple < signifie que `P` se trouve **juste avant** `N`.

```
pattern {
  N [upos=NOUN];
  P [upos=PRON];
  P << N
}
```

FIGURE 2.14 – Motif GREW impliquant une condition sur l'ordre des mots.

Sa syntaxe étant relativement intuitive, GREW nous a permis de faire un grand nombre de requêtes-tests sur les corpus d'UD lors du choix des motifs de reconnaissance. En effet, le fait d'appliquer un même motif de reconnaissance sur des corpus de différentes langues exige une certaine précision pour éviter d'extraire des données bruitées.

GREW en ligne de commandes GREW est un outil qui s'utilise en ligne de commandes. Plusieurs fonctionnalités sont disponibles, celle de réécriture `transform` et celle de reconnaissance de motifs `count` entre autres.

Dans nos expériences, nous utilisons uniquement la commande `count` qui nous permet de compter les occurrences d'un ou plusieurs motifs sur tous les corpus d'UD 2.7_{1K}. Cette commande nous renvoie alors un tableau CSV avec le nom du corpus, le nombre de phrases de chaque corpus et les occurrences du ou des motif(s). Le tableau 2.3 donne les occurrences des ordres adjectif - nom et nom - adjectif que nous détaillons en section 3.2.3 sur les trois premiers corpus d'UD par ordre alphabétique.

Corpus	# sentences	adj_nom	nom_adj
Afrikaans-AfriBooms	1934	2555	457
Akkadian-RIAO	1799	68	1664
Amharic-ATT	1074	49	4

TABLEAU 2.3 – Tableau des occurrences obtenu avec la commande `count` de GREW.

Une autre fonctionnalité de GREW a été particulièrement utile pour nos expériences, le *clustering* qui permet de trier les résultats selon une certaine condition. Par exemple, si nous prenons le motif `pattern { A -[nsubj]-> B }` et que nous ajoutons le *clustering* `B.upos`, la commande nous renvoie les différentes parties du discours de B vérifiant ce motif ainsi que leurs occurrences. Le tableau 2.4 présente les résultats sur les trois premiers corpus d'UD.

Corpus	NOUN	PRON	PROPN	...	SYM	VERB	X
Afrikaans-AfriBooms	1068	1847	24	...	3	0	33
Akkadian-RIAO	175	171	140	...	0	2	0
Amharic-ATT	307	958	162	...	0	3	0

TABLEAU 2.4 – Résultat d'une requête GREW avec *clustering*.

Cette commande a été utilisée à plusieurs reprises pour comprendre certaines incohérences entre les annotations et les instructions d'UD mais aussi entre les corpus d'une

même langue.

Dans nos expériences, nous faisons le choix d'appliquer le même motif à tous les corpus afin d'exploiter les annotations universelles. Bien que certains corpus présentent des informations plus précises que d'autres, adapter un motif aux spécificités d'un corpus constituerait un biais supplémentaire dans notre analyse.

GREW-MATCH La fonctionnalité de reconnaissance de graphes est proposée *via* le service web GREW-MATCH [10]. L'interface graphique est présentée dans la figure 2.15.

Par rapport à GREW en ligne de commandes, GREW-MATCH a l'avantage de permettre à l'utilisateur de modifier le motif directement et d'obtenir des exemples illustrés par les arbres de dépendances. Cependant, la requête ne peut être lancée que sur un seul corpus et une limite de 1 000 résultats est posée, ce qui ne nous permet d'obtenir le total des occurrences.

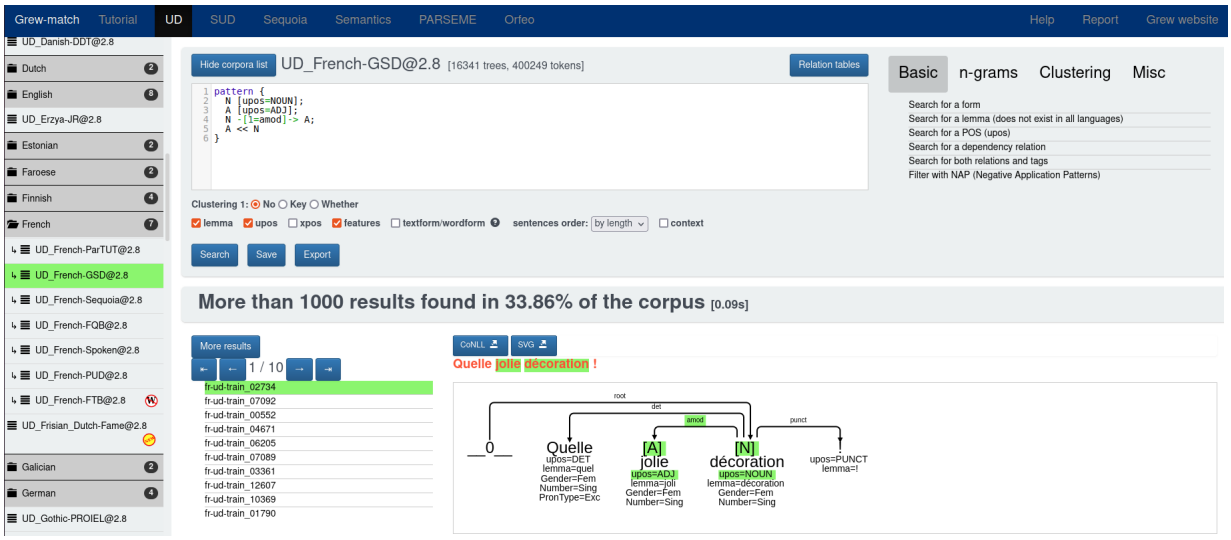

FIGURE 2.15 – Interface graphique de GREW-MATCH.

GREW-MATCH représente un outil intéressant pour explorer des corpus et récupérer certaines données grâce aux fonctionnalités suivantes :

— l'export des arbres de dépendances au format SVG,
— l'export des CoNLL-U, format tabulaire des annotations,
— l'export des résultats avec les contextes droit et gauche,
— la sauvegarde des requêtes.

Dans nos expériences, nous avons utilisé conjointement GREW et GREW-MATCH. D'une part, GREW nous a permis d'extraire les occurrences des motifs sur les 141 corpus et d'autre part, nous avons recouru à GREW-MATCH pour approfondir nos analyses sur certaines langues et visualiser les arbres de dépendances.

2.3 Quantifier des concepts qualitatifs

Notre étude se concentre sur l'analyse de phénomènes linguistiques et plus précisément sur la fréquence d'apparition de ces phénomènes dans des corpus de différentes langues.

10. http://match.grew.fr

Nous avons donc réfléchi à un moyen de mesurer ces fréquences selon des critères quantitatifs afin de permettre une comparaison des résultats. Dans cette partie, nous définissons, dans un premier temps, la notion d'ordre dominant et dans un second temps, nous détaillons les différentes mesures auxquelles nous avons pensé pour évaluer l'homogénéité entre les corpus d'une même langue.

2.3.1 Mesurer l'ordre dominant

La notion d'ordre dominant est omniprésente dans les travaux de Greenberg, en particulier dans l'ordre sujet - verbe - objet. Dans ses travaux, Greenberg ne considère pas la notion de dominance comme basée sur la fréquence d'apparition mais en lien avec des relations harmoniques et dysharmoniques. Il illustre son propos en prenant l'exemple de l'universel 25 :

« Si l'objet pronominal suit le verbe, alors l'objet nominal le suit également ».

Cet universel est une implication que nous écrivons comme suit :

$$\text{VOpron} \rightarrow \text{VOnom}$$

Les conséquences de l'implication sont :

1. L'objet nominal suit le verbe, que l'objet pronominal suive ou précède le verbe.
2. L'objet nominal précède le verbe **si et seulement si** l'objet pronominal précède le verbe.

Greenberg dit alors que « l'ordre VO est dominant sur l'ordre OV puisque l'ordre OV ne se produit que dans des conditions spécifiques, à savoir lorsque l'objet pronominal précède également, alors que l'ordre VO n'est pas soumise à de telles limitations. » (p.97). Il ajoute également que « l'ordre OnomV est harmonique avec OpronV mais dysharmonique avec VOpron puisqu'il ne se produit pas avec lui. De même, l'ordre VOnom est harmonique avec le VOpron et dysharmonique avec le OpronV ». Il reformule ces observations ainsi :

« Un ordre dominant peut toujours se produire, mais son opposé, le récessif, ne se produit que lorsqu'une construction harmonique est également présente. »

Jakobson (1966) confirme que l'idée de dominance n'est pas basée sur la fréquence d'apparition et qu'elle constitue en fait un critère stylistique. Il donne l'exemple des six possibles ordres sujet - verbe - objet qui apparaissent en russe. Jakobson décrit l'ordre SVO comme le seul « stylistiquement neutre ». Par opposition, les autres ordres, les récessifs, « sont vécus par les locuteurs et les auditeurs natifs comme des changements emphatiques divers »(p.269).

Notre approche étant basée sur de grandes quantités de données, en termes de langues considérées et de taille des corpus, nous ne sommes pas en mesure de fournir une analyse stylistique de l'ordre dominant. En revanche, nous pouvons apporter à la notion d'ordre dominant un aspect plus quantitatif en se basant sur les fréquences d'apparition.

Selon WALS (Dryer, 2013a), la notion d'ordre dominant d'une langue peut avoir deux sens :

— soit l'ordre est le seul possible pour la langue,
— soit la langue présente plusieurs ordres différents et l'un d'eux est plus fréquemment utilisé.

Pour déterminer l'ordre dominant de manière quantitative, nous utilisons le même critère que WALS 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, 2013a). Nous calculons donc le ratio entre les deux premiers ordres les plus fréquents :

— Si le ratio est supérieur ou égal à 2, l'ordre le plus fréquent est l'ordre dominant.
— Si le ratio est strictement inférieur à 2, nous considérons qu'il n'y a pas d'ordre dominant, noté NDO (*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 supérieur à $\frac{2}{3}$.

Utiliser le même critère que WALS nous permet également de comparer nos résultats avec les leurs. Ces derniers étant basés sur des ressources descriptives comme des grammaires de référence, il est intéressant d'observer si nos résultats empiriques confirment leurs analyses.

2.3.2 Mesurer l'homogénéité entre corpus

Dans nos expériences, nous avons fait le choix de travailler au niveau du corpus afin de comparer les corpus d'une même langue et d'évaluer leur cohérence. L'homogénéité a été mesurée principalement dans les résultats relatifs à l'ordre sujet - verbe - objet, les distributions des six ordres étant plus difficiles à comparer que les caractéristiques à deux éléments (adjectif - nom et adposition - nom). Nous avons donc exploré différentes mesures pour trouver celle qui est la plus adaptée à nos données.

Entropie de Shannon L'entropie de Shannon mesure la quantité d'information contenue ou délivrée par une source d'information (Shannon, 1948). En TAL, elle est régulièrement utilisée pour calculer la variabilité dans l'ordre des mots (Levshina, 2019; Futrell et al., 2015). Autrement dit, elle permet de calculer le degré de flexibilité d'une langue ou d'un corpus par rapport à une propriété. Dans notre cas, la propriété est l'ordre sujet - verbe - objet. La formule d'entropie de Shannon est décrite comme suit :

$$H = -K \sum_{i=1}^{n} \mathrm{p_i} \log \mathrm{p_i}$$

où K est une constante positive.

Shannon décrit la constante K comme équivalente au choix d'une unité de mesure et simplifie la formule par :

$$H = -\sum_{i=1}^{n} \mathrm{p_i} \log \mathrm{p_i}$$

Nous avons fait une première expérience dans laquelle nous calculons l'entropie des corpus du roumain. L'ordre sujet - verbe - objet permettant six ordres, la variable aléatoire X peut prendre six valeurs différentes. Nous avons utilisé une entropie normalisée par le

	SVO	SOV	VSO	VOS	OSV	OVS	H
Romanian-NonStandard	**38,07 %**	**31,87 %**	9,66 %	3,97 %	1,71 %	14,72 %	0,80
Romanian-RRT	**85,32 %**	7,76 %	1,12 %	0,70 %	1,18 %	3,91 %	0,33
Romanian-SiMoNERo	**97,61 %**	0,97 %	0,09 %	0,09 %	0,13 %	1,10 %	0,08

TABLEAU 2.5 – Distribution des ordres sujet - verbe - objet et valeurs d'entropie (H) des corpus du roumain.

maximum $\frac{\log(6)}{\log(2)} \approx 2,58$, la probabilité de chaque ordre étant de $\frac{1}{6}$. Le tableau 2.5 présente les corpus du roumain avec la distribution des ordres sujet - verbe - objet et leurs entropies.

Plus l'entropie est proche du maximum, plus le corpus présente une distribution des ordres homogène et donc n'a pas d'ordre dominant. On parle également de notion de « désordre ». Plus l'entropie est proche de 0, plus le corpus présente un seul ordre avec une fréquence d'apparition élevée.

En roumain, nous remarquons que les corpus `Romanian-RRT` et `Romanian-SiMoNERo`, très fortement SVO, ont effectivement une entropie faible, respectivement à 0,33 et 0,08. Par opposition, le `Romanian-NonStandard` a une entropie proche de 1 à 0,80, qui s'explique par le fait que le corpus présente deux ordres aux proportions proches, le SVO et SOV.

Une faible entropie indique qu'une valeur de X est dominante mais ne nous dit pas laquelle est dominante. Par conséquent, si un corpus est fortement SVO et un corpus fortement SOV, leurs entropies seront proches. Cette mesure ne nous permet pas de comparer les distributions en elles-mêmes mais plutôt de calculer l'homogénéité d'une distribution. Puisque cette information nous est déjà donnée avec le ratio, nous décidons de ne pas utiliser l'entropie dans nos prochaines expériences.

Divergence de Kullback-Leibler La divergence de Kullback-Leibler (ou divergence K-L) est une mesure liée à la théorie de l'information et calcule la dissimilarité entre deux distributions de probabilités (Kullback and Leibler, 1951). Elle correspond à une entropie relative.

Si nous considérons les fréquences de six ordres sujet - verbe - objet, nous utilisons des probabilités discrètes. Considérons P et Q deux distributions de probabilités discrètes, P représente une distribution de données observées tandis que Q représente une distribution théorique.

$$D_{\mathrm{KL}}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Bien qu'elle permette de comparer deux distributions, la divergence K-L implique d'avoir une distribution théorique de référence. Dans nos expériences, nous n'avons que des distributions observées. Nous avons alors pensé à, soit choisir arbitrairement un corpus de référence (le plus gros par exemple), soit calculer toutes les valeurs de divergence K-L en mettant chaque corpus en distribution de référence.

Par ailleurs, la divergence K-L n'est pas une mesure symétrique, ce qui ne facilite pas l'interprétation des résultats. Du fait de ces inconvénients, nous avons conclu que cette mesure n'était pas adaptée à nos expériences.

Similarité cosinus La similarité cosinus consiste à comparer deux vecteurs de dimensions n et de calculer leur similarité selon le cosinus de leur angle. Cette mesure est fréquemment exploitée pour mesurer la similarité de documents textuels.

$$\text{Soit deux vecteurs A et B, } cos\theta = \frac{A \cdot B}{\parallel A \parallel \parallel B \parallel}$$

Cette mesure implique de représenter les corpus sous forme de vecteurs. Dans nos expériences, nous utilisons les fréquences des six ordres sujet - verbe - objet de chaque corpus pour obtenir un vecteur de dimensions six.

La similarité cosinus est la mesure que nous avons choisi de retenir puisqu'elle nous permet de comparer les distributions de deux corpus de manière symétrique, ce qui n'est pas le cas de la divergence K-L. Les calculs de similarité cosinus sont détaillés dans la partie sur les résultats en section 4.2.

Chapitre 3

Expériences

Sommaire

3.1 Universaux sélectionnés

Avant d'entreprendre nos expériences, il a fallu faire un travail de tri sur les universaux de Greenberg pour dégager ceux vérifiables avec les données et les outils à notre disposition. Dans un premier temps, nous avons éliminé les universaux morphologiques, les corpus ne présentant pas tous des annotations morphologiques complètes. Les universaux les plus abordables pour notre étude sont ceux relatifs à l'ordre des mots.

Ordre dominant sujet - verbe - objet Dans un premier temps, nous nous intéressons à deux universaux traitant l'ordre sujet - verbe - objet :

Universal 1 : In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object. [1]

Universal 6 : All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order. [2]

Les universaux suivants nécessitent d'établir une première classification selon l'ordre sujet - verbe - objet. Par la suite, d'autres classifications sont déterminées selon l'ordre adjectif - nom et la présence de prépositions ou de postpositions. Les universaux implicationnels font la corrélation entre ces différentes caractéristiques.

Corrélation ordre sujet - verbe - objet et prépositions/postpositions Greenberg remarque une forte corrélation entre l'ordre sujet - verbe - objet et la présence de prépositions ou de postpositions, avec l'opposition entre les deux extrêmes : SOV-Postpositions et VSO-Prépositions.

Universal 3 : Languages with dominant VSO order are always prepositional. [3]

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.
2. Toutes les langues d'ordre dominant VSO ont l'ordre SVO comme alternative ou comme seul ordre alternatif.
3. Les langues d'ordre dominant VSO sont toujours prépositionnelles.

Universal 4 : With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional. [4]

Les formulations de Greenberg sont plus ou moins catégoriques en raison des exceptions qui existent. En effet, Greenberg est conscient de l'existence de langues d'ordre SOV mais à prépositions (le iraqw et le perse notamment).

Corrélation prépositions/postpositions et ordre génitif - nom Sur les 30 langues de l'échantillon de Greenberg, 14 langues ont des prépositions, 14 ont des postpositions et deux langues ne présentent pas cette information dans la classification de Greenberg (le turc et le songhai). Toutes les langues à postpositions ont le génitif qui précède le nom qui le gouverne et sur les 14 langues à prépositions, 13 d'entre elles ont le génitif qui suit le nom, la seule exception étant le norvégien. Cela donne lieu à l'universel 2 :

Universal 2 : In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes. [5]

Autres corrélations Greenberg remarque que l'ordre de l'adjectif par rapport au nom suit les mêmes tendances que l'ordre génitif - nom.

Universal 5 : If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun. [6]

Par extension des corrélations précédentes, il détermine également des liens entre l'ordre VSO et l'ordre de l'adjectif et du nom avec son 17$^{\text{ème}}$ universel.

Universal 17 : With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun. [7]

3.2 Travaux préalables

De la même manière que Greenberg, nous avons préalablement classé les corpus selon les trois critères de base [8] :

1. l'ordre sujet - verbe - objet,
2. la présence de prépositions/postpositions,
3. l'ordre adjectif - nom.

Nous avons également comparé nos résultats à la classification de WALS (Dryer and Haspelmath, 2013), référence en typologie des langues.

4. Avec une fréquence largement supérieure à la normale, les langues d'ordre normal SOV sont postpositionnelles.
5. Dans les langues à prépositions, le génitif suit presque toujours le nom qui le gouverne, tandis que dans les langues à postpositions, il le précède presque toujours.
6. Si une langue a un ordre dominant SOV et le génitif qui suit le nom qui le gouverne, alors l'adjectif suit également le nom.
7. Avec une fréquence largement supérieure à la normale, les langues d'ordre dominant VSO ont l'adjectif après le nom.
8. Les résultats détaillés pour chaque corpus sont présentés en annexe 4.3.2.

3.2.1 Ordre Sujet (S) - Verbe (V) - Objet (O)

Le premier universel de Greenberg traite de l'ordre sujet - objet - verbe dans les phrases déclaratives avec un sujet nominal et un objet nominal. Pour établir cette première classification, nous avons utilisé six *patterns* correspondant aux six ordres possibles : SVO, SOV, VSO, VOS, OSV, OVS. La figure 3.1 représente le motif pour l'ordre SVO.

D'après les instructions d'UD, la relation `nsubj` (*nominal subject*) encode la relation sujet entre un verbe et un sujet nominal. De même, la relation `obj` (*object*) relie le verbe à un objet nominal. Les relations `nsubj` et `obj` impliquent donc par essence des sujets et des objets nominaux, c'est-à-dire des noms (`NOUN`), des pronoms (`PRON`) ou des noms propres (`PROPN`). Malgré une certaine redondance, nous faisons le choix de fixer les étiquettes morphosyntaxiques du sujet et de l'objet après avoir remarqué des incohérences dans les annotations de certains corpus que nous détaillons dans la section 4.3.

Les annotations dans les corpus ne nous permettent pas de filtrer de manière précise les phrases déclaratives. Nous nous sommes alors basés sur la ponctuation en éliminant toutes les phrases dont le verbe est relié à un point d'interrogation ou un point d'exclamation (qui correspond à la partie *without* du motif). Cette méthode reste toutefois fragile étant donné qu'il existe des corpus sans ponctuation et que nous n'avons aucune certitude que le système de ponctuation est le même dans toutes les langues traitées. En outre, nous n'avons pas pu filtrer les phrases impératives car certains corpus n'indiquent pas le mode des verbes.

```
pattern {
  V [upos=VERB];
  V -[1=nsubj]-> S; S[upos=PROPN|PRON|NOUN];
  V -[1=obj]-> O; O[upos=PROPN|PRON|NOUN];
  S << V; V << O;
}
without {
  P [lemma="?"|"!"];
  V -[punct]-> P;
}
```

FIGURE 3.1 – Motif de l'ordre SVO.

Après classé les ordres selon leurs fréquences d'apparition, le ratio entre les fréquences des deux premiers ordres est calculé. Les corpus NDO correspondent aux corpus pour lesquels le ratio est inférieur à 2. Sur 141 corpus :

— 91 sont SVO,
— 24 sont SOV,
— 4 sont VSO,
— 22 sont NDO.

Sur les 29 langues multi-corpus, tous les corpus d'une même langue présentent le même ordre dominant, à l'exception de six langues : l'allemand, l'arabe, le grec ancien, le latin, le néerlandais et le roumain. Nous expliquons en détails ces exceptions dans la section 4.2.

3.2.2 Prépositions/Postpositions

Greenberg emploie les termes « prepositionnal language » ou « language with prepositions » mais n'explique pas la manière dont il détermine qu'une langue présente des prépositions ou des postpositions.

En français, nous ne parlons que de prépositions, mais nous désignerons ici les prépositions et les postpositions comme des adpositions au terme large. Pour pouvoir comparer nos résultats avec ceux de WALS [9], nous décidons de prendre en compte les adpositions reliées à un groupe nominal. En UD, cela revient à dire que le mot gouverneur est un nom commun, un pronom ou un nom propre. La figure 3.2 présente les deux motifs correspondants. Nous fixons une entité A étiquetée comme étant une adposition (ADP), reliée au groupe nominal GN par une relation `case` de marqueur de cas ou de complément par une adposition [10]. De plus, nous décidons d'exclure les cas où l'adposition fait partie d'une expression polylexicale, marquée avec les relations `fixed` et `flat` dans UD.

```
pattern {
  A [upos=ADP];
  GN [upos=NOUN|PRON|PROPN];
  GN -[1=case]-> A;
  A << GN;
}
without {
  A -[1=fixed|flat]-> X
}
```

```
pattern {
  A [upos=ADP];
  GN [upos=NOUN|PRON|PROPN];
  GN -[1=case]-> A;
  GN << A;
}
without {
  A-[1=fixed|flat]-> X
}
```

FIGURE 3.2 – Motifs Pr - GN à gauche, GN - Po à droite

Les figures 3.3 et 3.4 présentent les deux ordres possibles : préposition (Pr) - groupe nominal (GN) en français et groupe nominal - postposition (Po) en coréen. Le phrase en coréen se traduit littéralement par : *L'essence du film < Le Parrain > est la rédemption et le salut.*

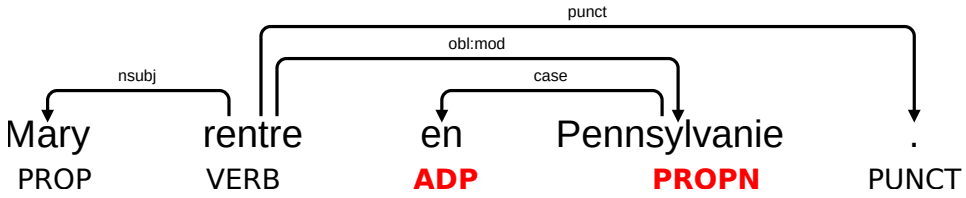

FIGURE 3.3 – Phrase Pr - GN du corpus `French-GSD`.

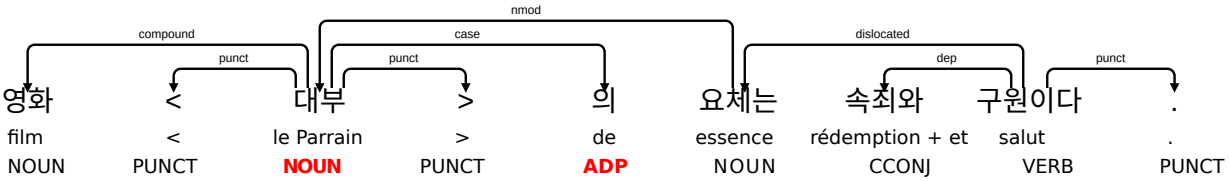

FIGURE 3.4 – Phrase GN - Po du corpus `Korean-KAIST`.

9. Voir : https://wals.info/feature/85A, août 2021.
10. Voir : https://universaldependencies.org/u/dep/case, août 2021.

Les résultats sont relativement tranchés avec une tendance marquée pour un des deux types d'adposition.

Ainsi, sur les 141 corpus :

— 108 ont un ordre dominant Pr - nom,
— 30 ont un ordre dominant nom - Po,
— Un corpus ne présente pas d'ordre dominant : `Chinese-PUD`,
— Deux corpus n'ont montré aucune occurrence des deux cas : `Korean-PUD` et le `Sanskrit-Vedic`.

Tous les corpus d'une même langue présentent le même type d'adposition sauf ceux du chinois et du coréen du fait des deux corpus atypiques.

Le chinois Pour comprendre le cas du `Chinese-PUD`, nous avons comparé les résultats obtenus dans les corpus du chinois, résultats présentés dans le tableau 3.1. Nous observons que les trois autres corpus du chinois ont très largement des prépositions [11]. Nous notons également que WALS considère le chinois comme ne présentant pas d'ordre dominant dans l'ordre adposition - groupe nominal.

Corpus	Pr	Post
Chinese-GSD	99,92 %	0,08 %
Chinese-GSDSimp	99,92 %	0,08 %
Chinese-HK	87,40 %	12,60 %
Chinese-PUD	**64,57 %**	**35,43 %**

TABLEAU 3.1 – Proportions de prépositions et de postpositions dans les corpus du chinois.

Corpus	acl	appos	case	case:loc	conj	mark
Chinese-GSD	**98,41 %**	0,10 %	0,20 %	0,00 %	0,20 %	1,09 %
Chinese-GSDSimp	**98,42 %**	0,10 %	0,20 %	0,00 %	0,20 %	1,09 %
Chinese-HK	0,00 %	0,00 %	9,68 %	**90,32 %**	0,00 %	0,00 %
Chinese-PUD	0,00 %	0,00 %	0,00 %	**99,39 %**	0,00 %	0,61 %

TABLEAU 3.2 – Distribution des types de relation entre le groupe nominal et la postposition dans les corpus du chinois.

D'après le tableau 3.2, nous remarquons que les postpositions ne sont pas annotés avec une relation `case` dans les corpus de `GSD` mais avec une relation `acl`. Par conséquent, les postpositions ne sont pas couvertes par le motif que nous avons utilisé.

Par opposition, le `Chinese-PUD` a 99 % des postpositions reliées par une relation `case:loc` qui est un sous-type de la relation `case` et donc couverte par le motif. De même, le `Chinese-HK` présente la relation `case:loc` à 90,32 %. Cependant, le `Chinese-HK` a une proportion de postpositions relativement faible à 12,60 %, ce qui s'explique simplement par le fait que le corpus présente peu de phrases de cette construction.

Nous pourrions penser que les relations `case` et `acl` impliquent des postpositions de nature différente. Cependant, tous les corpus du chinois utilisent des postpositions qui correspondent généralement à des postpositions indiquant la localisation comme :

11. Nous précisons que le `Chinese-GSDSimp` constitue un corpus du chinois simplifié du GSD mais ces deux corpus présentent toutefois des chiffres légèrement différents. Nous les considérons donc comme deux corpus à part entière.

— 上 (shàng = au-dessus/sur),
— 中 (zhōng = milieu),
— 下 (xià = en-dessous/sous).

Il n'y a donc aucune raison que la relation soit annotée différemment. Ici, nous avons simplement affaire à un choix d'annotation différent entre les équipes d'annotateurs.

Si nous prenons en compte tous les cas de postposition reliée par la relation `acl` dans les corpus GSD, nous obtenons une distribution de 72,87 % de prépositions et 27,13 % de postpositions, ce qui concorde avec les distributions du `Chinese-PUD` et du `Chinese-HK`.

Le guide d'annotation d'UD décrit la relation `acl` comme une relation de complément phrastique d'un nom. En chinois, les compléments peuvent précéder le nom et être formés ou non avec la particule 的. Ils peuvent également suivre le nom, auquel cas ils seront juxtaposés au nom [12]. De manière universelle, la relation `acl` doit lier un nom et la tête du complément phrastique. Il est assez rare de trouver une adposition en dépendant d'une relation `acl`, cela étant contraire aux indications d'UD. Nous supposons qu'il y a une incohérence dans le choix d'annotations dans ces corpus, incohérence que nous avons signalée en ouvrant une discussion sur le *Git* du corpus [13].

Le coréen Le coréen est une langue agglutinante utilisant des postpositions à la fin des mots pour désigner leurs fonctions. Ces postpositions sont parfois désignées comme étant des particules. Dans le schéma d'annotations d'UD, il est indiqué que l'étiquette PART désigne un mot fonctionnel et ne doit être utilisée que lorsque le mot ne correspond pas aux définitions des autres mots fonctionnels comme les adpositions et les conjonctions [14].

Cependant, dans le corpus du `Korean-PUD`, l'étiquette ADP n'est pas utilisée et toutes les adpositions sont annotées avec l'étiquette PART. De plus, dans la version 2.7 du corpus, les adpositions ne sont pas reliées avec une relation `case` mais avec une relation `dep:prt` [15].

La distinction entre une particule et une adposition est délicate à définir. 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.

Une autre difficulté se pose dans le cas des langues agglutinantes, au niveau de la segmentation en mots (*tokenization*). Les corpus de coréen étant majoritairement segmentés selon les espaces, une postposition qui est collée au mot qui la précède forme un seul *token*, ce qui n'est pas détectable par la requête. Les postpositions détectées par notre motif le sont pour une des raisons suivantes :

— la postposition est une postposition qui permet un espace avec le groupe nominal,
— la postposition et le groupe nominal sont séparés par une ponctuation (un chevron, un guillemet, une parenthèse…), comme dans l'exemple 3.4.

Le nombre de postpositions détectées est donc largement plus faible que le nombre de postpositions présentes en réalité, que nous ne pouvons pas couvrir dans le cadre de cette étude.

Le sanskrit Comme le `Korean-PUD`, le `Sanskrit-VEDIC` ne présente aucune occurrence des *patterns* utilisés car l'adposition est annotée en tant que particule. Avec cette an-

12. Voir : https://universaldependencies.org/zh/dep/acl, août 2021.
13. Voir : https://github.com/UniversalDependencies/UD_Chinese-GSD/issues/4.
14. Voir : https://universaldependencies.org/u/pos/PART, août 2021.
15. Dans la version 2.8, la relation `dep:prt` a été remplacée par la relation `case`.

notation, nous obtenons 79,76 % de postpositions et donc 20,24 % de prépositions. Le sanskrit étant une langue morte, elle n'est pas répertoriée dans WALS. Sans spécialiste du sanskrit, nous n'avons pas pu pousser l'analyse plus loin.

Comparaison avec WALS WALS présente l'ordre adposition - nom sous la *feature* 85A (Dryer, 2013c). En comparant avec leurs résultats, sur 74 langues :

— 50 langues présentent le même ordre que WALS,
— 21 langues ne sont pas dans WALS (essentiellement les langues mortes) ou ne présentent pas la *feature* 85A,
— 3 langues ne présentent pas le même ordre : le chinois, le cantonais et l'amharique qui sont considérés comme NDO par WALS.

Le `Cantonese-HK` présente une proportion de prépositions à 87,84 %. Le cantonais et le chinois ayant une syntaxe relativement proche, nous pouvons supposer que le cantonais peut effectivement avoir des prépositions et des postpositions mais que ce corpus a soit peu de constructions avec des postpositions, soit les annotations ne nous permettent pas de couvrir tous les cas avec notre motif.

Quant au corpus de l'amharique, l'`Amharic-ATT` a une proportion de prépositions à 83,81 %. Greenberg considère également que l'amharique est une langue prépositionnelle.

3.2.3 Ordre Adjectif - Nom

La détermination de l'ordre adjectif - nom se trouve être plus simple, dans la mesure où les concepts sont présents dans toutes les langues et sont précisément annotés. Nous utilisons le motif présenté en figure 3.5, où nous fixons un nom N et un adjectif A reliés entre eux par la relation `amod` (*adjectival modifier*).

```
pattern {
  N [upos=NOUN];
  A [upos=ADJ];
  N -[1=amod]-> A;
  A << N
}
```

FIGURE 3.5 – Motif de l'ordre adjectif - nom.

Sur les 141 corpus :

— 84 présentent l'ordre adjectif - nom,
— 43 présentent l'ordre nom - adjectif,
— 14 n'ont pas d'ordre dominant : un corpus du français, trois corpus de l'italien, deux corpus du polonais, les deux corpus du grec ancien, les quatre corpus du latin, le corpus de vieux russe et le corpus du gotique.

Tous les corpus d'une même langue présentent le même ordre, à l'exception du français, de l'italien et du polonais.

Cas du français, de l'italien et du polonais Dans ces trois langues, il est possible de trouver l'adjectif avant le nom mais aussi après le nom. Généralement, la place de l'adjectif ne change pas le sens de la phrase dans ces langues, sauf cas particuliers [16]. Il

16. Par exemple, « une ancienne maison » a un sens différent de « une maison ancienne ».

existe toutefois des règles telles que placer les adjectifs courts avant le nom ou placer les adjectifs de couleur après le nom en français et en italien. Il est également possible de changer la place de l'adjectif pour créer une emphase et insister sur la qualité portée par l'adjectif.

Le français Le `French-Spoken` est le seul corpus du français n'ayant pas d'ordre dominant. Ce corpus est le seul corpus de français oral parmi les sept corpus du français. Comme le montre le tableau 3.3, les corpus écrits du français présentent des proportions très similaires avec environ 70 % de nom - adjectif, tandis que le `French-Spoken` présente les deux ordres de manière relativement homogène.

Corpus	adjectif - nom	nom - adjectif
French-FQB	29,31 %	70,69 %
French-FTB	30,86 %	69,14 %
French-GSD	29,95 %	70,05 %
French-ParTUT	27,10 %	72,90 %
French-PUD	31,13 %	68,87 %
French-Sequoia	24,24 %	75,76 %
French-Spoken	**46,71 %**	**53,29 %**

TABLEAU 3.3 – Proportions d'ordres adjectif - nom et nom - adjectif dans les corpus du français.

Cette différence peut être expliquée par le fait que les corpus écrits proviennent majoritairement d'articles de journaux et de Wikipedia. Les phrases sont composées d'adjectifs de domaines plus scientifiques qui ont tendance à être placés après le nom. Par ailleurs, le corpus du français oral montre une sur-représentation d'adjectifs courants tels que « petit » et « grand » qui se placent avant le nom.

L'italien En italien, trois corpus ont un ordre dominant nom - adjectif et trois n'ont pas d'ordre dominant. Sur les trois corpus sans ordre dominant, deux d'entre eux sont des corpus de *tweets*, l'`Italian-PoSTWITA` et l'`Italian-TWITTIRO`. Le troisième corpus qui n'a pas d'ordre dominant est l'`Italian-ParTUT` mais cela est dû à un effet de seuil, la distribution étant de 33,99 % et 66,01 %.

Comme en français, les trois corpus d'ordre dominant nom - adjectif ainsi que le corpus d'`Italian-ParTUT` ont des textes extraits d'articles de journaux et de Wikipédia. La distribution des ordres est similaire dans ces corpus avec une proportion autour des 70 % pour l'ordre nom - adjectif.

Les résultats montrent que le genre des corpus a une influence sur le type d'adjectifs employés et donc sur la place de l'adjectif par rapport au nom.

Le polonais En polonais, les adjectifs peuvent être placés avant ou après le nom, sauf pour les adjectifs courts qui sont toujours placés avant. Placer un adjectif après le nom permet parfois de mettre l'emphase sur celui-ci.

Contrairement au français et à l'italien, nous ne pouvons pas expliquer les différences de distribution par le genre des textes. Les corpus sont tous composés de genres variés : des articles de journaux, de fiction, de Wikipédia.

Corpus	adjectif - nom	nom - adjectif
Italian-ISDT	29,89 %	70,11 %
Italian-ParTUT	**33,99 %**	**66,01 %**
Italian-PoSTWITA	**41,31 %**	**58,69 %**
Italian-PUD	31,04 %	68,96 %
Italian-TWITTIRO	**51,69 %**	**48,31 %**
Italian-VIT	31,76 %	68,24 %

TABLEAU 3.4 – Proportions d'ordres adjectif - nom et de nom - adjectif dans les corpus de l'italien.

Corpus	adjectif - nom	nom - adjectif
Polish-LFG	71,30 %	28,69 %
Polish-PDB	**64,48 %**	**35,52 %**
Polish-PUD	**59,73 %**	**40,27 %**

TABLEAU 3.5 – Proportions d'ordres adjectif - nom et de nom - adjectif dans les corpus du polonais.

Les langues mortes Les corpus du latin, de grec ancien, du vieux russe et du gotique n'ont pas d'ordre dominant dans l'ordre de l'adjectif et du nom. Les proportions entre les deux ordres sont homogènes, avec des ratios très proches de 1 pour ces langues. Nous pouvons supposer que ces langues permettent les deux ordres, dans la mesure où elles présentent une grande liberté dans l'ordre des mots (Levshina, 2019).

Comparaison avec WALS L'ordre adjectif - nom est présent dans WALS sous la *feature* 87A (Dryer, 2013b).

Sur les 74 langues :

— 54 langues présentent le même ordre que ceux obtenus dans les expériences décrites plus haut,

— 17 langues ne sont pas dans WALS ou n'ont pas la *feature* 87A,

— 3 langues sont définies comme sans ordre dominant dans nos résultats, l'italien, le français et le polonais. Cependant, les corpus où l'ordre est défini pour ces langues sont tous cohérents avec les résultats de WALS.

Si nous ne comptons pas les langues mortes et les deux langues *code-switching* qui ne sont pas répertoriées dans WALS, six langues ne présentent pas la *feature* 87A : l'afrikaans, le féroïen, le galicien, le kazakh, le maltais et le slovaque.

3.2.4 Génitif

La notion de génitif a été particulièrement délicate à traiter dans nos expériences en raison des différentes formes qu'il peut prendre selon les langues. De ce fait, nous n'avons pas réussir à poser une définition du génitif de manière rigoureuse.

Dans son étude, Greenberg ne donne aucune définition de ce qu'il considère comme étant un génitif. Il ne parle que de l'ordre entre le génitif et le nom auquel il est rattaché. D'autre part, il a fallu déterminer si UD permettait d'annoter le génitif. Dans son schéma d'annotations, UD propose la relation `nmod` (*nominal modifier*) pour annoter un attribut

ou un complément génitif entre deux noms ou groupes nominaux [17]. Par ailleurs, UD met à disposition une étiquette morphologique `Case` qui permet d'indiquer le cas de certains mots, notamment pour les langues à cas. Cependant, tous les corpus des langues à cas ne fournissent pas cette information.

Si le génitif est connu pour exprimer la possession, il est aussi utilisé pour exprimer d'autres relations sémantiques comme dans l'exemple : « la construction de la maison » qui s'apparente plutôt à une relation entre un objet, « la maison » et un verbe « construire » représenté par le substantif « construction » (Dryer, 2013d). Il est très souvent associé à la notion de complément du nom.

Dans un premier temps, nous avons utilisé la relation `nmod` entre deux noms, et dans les langues à cas, nous avons observé si le dépendant de la relation `nmod` était annoté en tant que génitif avec l'étiquette `Case`. La fonction *clustering* de Grew nous permet d'obtenir, sur 1 000 phrases, la distribution des cas du dépendant dans le corpus du russe `Russian-GSD` (Figure 3.6).

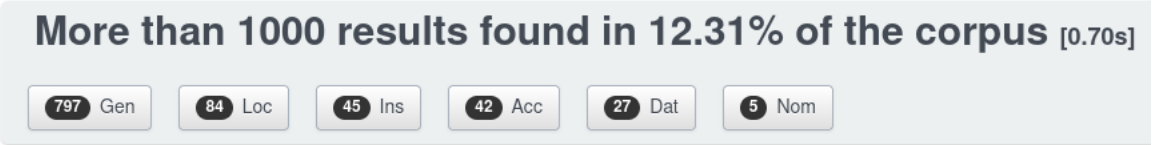

Figure 3.6 – Distribution des cas du dépendant de la relation `nmod` sur 1 000 phrases du `Russian-GSD`.

Nous remarquons dans la figure 3.6 qu'environ 20 % des noms dépendants de la relation `nmod` sont annotés avec un cas différent du génitif. Cela nous amène à penser que la relation `nmod` couvre des constructions qui vont au-delà d'une simple construction de génitif.

En outre, l'annotation `Case` seule ne peut être exploitée pour extraire le génitif, dans la mesure où certaines règles grammaticales imposent de décliner les noms au génitif sans pour autant coder une relation sémantique particulière. C'est le cas en russe où le génitif est obligatoire après certaines prépositions et certains nombres cardinaux.

Toutes ces contraintes ne nous permettent donc pas d'extraire le génitif de manière rigoureuse, non seulement sur une langue, mais d'autant plus dans notre démarche qui vise à traiter 74 langues automatiquement. Par conséquent, nous avons décidé de n'établir aucune classification impliquant le génitif et de ne pas traiter les deux universaux associés : l'universel 2 et l'universel 5.

17. Voir : https://universaldependencies.org/u/dep/nmod, août 2021.

Chapitre 4

Résultats

Sommaire

4.1 Universaux testés

4.1.1 Universel 1

Universal 1 : In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

Les trois ordres où le sujet précède l'objet sont : SVO, SOV et VSO. Les résultats obtenus montrent que sur 141 corpus :

— 91 sont SVO,
— 24 sont SOV,
— 4 sont VSO,
— 22 sont NDO.

Nos résultats confirment donc 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`. Le tableau 4.1 présente les fréquences des six ordres pour ces corpus.

Corpus	SVO	SOV	VSO	VOS	OSV	OVS	Ratio
Amharic-ATT	4,70 %	**28,86 %**	8,95 %	0,45 %	12,97 %	**44,07 %**	1,53
Latin-LLCT	**30,18 %**	29,00 %	4,31 %	0,93 %	**32,65 %**	2,91 %	1,08

TABLEAU 4.1 – Distribution des ordres sujet - verbe - objet et le ratio dans les corpus `Amharic-ATT` et `Latin-LLCT`.

L'amharique Pour l'amharique, l'ordre le plus fréquent est l'ordre OVS à 44,07 %. Il est suivi de l'ordre SOV à 28,86 %, ces deux ordres ayant un ratio de 1,53. L'ordre OVS reste un cas assez rare d'ordre dominant. Pour avoir un ordre de grandeur, sur les 1 376 langues présentant la caractéristique pour l'ordre dominant sujet - verbe - objet dans

WALS, seules 11 ont l'ordre OVS comme ordre dominant [1] (Dryer, 2013a).

D'après WALS et Greenberg, l'amharique est une langue SOV, ce qui correspond au 2ème ordre le plus fréquent dans nos résultats. Sans locuteur de l'amharique, nous pouvons seulement supposer qu'il y a soit :

— des erreurs d'annotations dans le corpus,
— la possibilité d'employer l'ordre OVS dans certaines phrases,
— une influence du genre des textes du corpus, qui est très hétérogène puisque les phrases peuvent être des exemples de grammaires (`Turkish-GB`), extraites de textes de fiction (`Czech-FicTree`), de la bible (`Gothic-PROIEL`) et de journaux (`Hungarian-Szeged`) entre autres.

Le corpus `Latin-LLCT` Ce corpus du latin présente une distribution relativement homogène entre trois ordres SVO, SOV et OSV qui sont fréquents à 30 % environ. Le latin étant une langue libre dans l'ordre des mots, cela peut expliquer l'absence d'ordre dominant. De plus, les textes proviennent de différents siècles, ce qui a une certaine influence sur la manière de construire les phrases.

4.1.2 Universaux 3 et 17

Les universaux 3 et 7 concernent les langues d'ordre dominant VSO, nous les traitons donc ensemble dans cette partie.

Universal 3 : Languages with dominant VSO order are always prepositional.

Universal 17 : With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.

Nous avons affaire ici à deux types d'universaux différents. L'universel 3 est un universel absolu avec la formulation « always prepositional ». À l'inverse, l'universel 17 est un universel statistique avec l'expression « With overwhemingly more than chance frequency », qui accepte donc quelques exceptions. Le tableau 4.2 donne les corpus d'ordre dominant VSO avec leurs fréquences, ainsi que la proportions de prépositions et de l'ordre Nom - Adjectif.

Corpus	VSO	Pr	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 %

Tableau 4.2 – Proportions de prépositions et d'ordre Nom - Adjectif dans les corpus VSO.

Nos résultats confirment les universaux de Greenberg, toutefois ils ne sont vérifiés que sur quatre corpus. Concernant l'arabe, deux autres corpus sont disponibles mais ne présentaient pas d'ordre dominant, avec un conflit entre les ordres VSO et SVO. Nous

1. Les 11 langues OVS sont principalement des langues d'Amérique du Sud et d'Océanie. Une langue vient d'Afrique, le Päri, parlée au Soudan du Sud.

observons tout de même leurs chiffres sur les proportions de prépositions et de l'ordre Nom - Adjectif dans le tableau 4.3.

Corpus	VSO	Pr	Nom - Adj
Arabic-PADT	31,20 %	99,93 %	99,78 %
Arabic-PUD	49,45 %	99,46 %	99,93 %

TABLEAU 4.3 – Proportions de prépositions et d'ordre Nom - Adjectif dans les corpus VSO.

Les deux autres corpus de l'arabe sont largement prépositionnels et présentent l'ordre Nom - Adjectif à presque 100 %. Il est intéressant de remarquer que le corpus `Arabic-PADT` présente ces caractéristiques alors qu'il a une fréquence de SVO plus élevée que celle de VSO (48,15 %). Ce résultat est cohérent avec l'observation de Greenberg, selon laquelle les langues SVO qui sont davantage corrélées à la prépositionnalité et l'ordre Nom - Adjectif que la postpositionnalité et l'ordre Adjectif - Nom.

4.1.3 Universel 4

De même que l'universel 17, l'universel 4 est également un universel statistique. Greenberg considère qu'une grande majorité des langues d'ordre SOV ont des postpositions.

Universal 4 : With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.

D'après nos résultats, nous avons 24 corpus d'ordre dominant SOV, ce qui correspond à 15 langues. Pour visualiser cet universel, nous utilisons les graphiques de Gerdes et al. (2021) en traitant chaque corpus et non chaque langue à la différence de ceux-ci.

La figure 4.1 présente les corpus d'ordre SOV avec le pourcentage en fonction de leurs proportions de postpositions. Les couleurs correspondent à différentes classes de langues :

- les langues indo-européennes représentées par un triangle :
  - les langues romanes en brun,
  - les langues germaniques en olive,
  - les autres langues indo-européennes en bleu,
- les langues agglutinantes en rouge,
- les autres langues en noir.

Nous pouvons observer un groupement de corpus en haut à droite de la figure correspondant aux corpus très fortement SOV et postpositionnels. Les langues représentées sont : le bambara, le coréen, le hindi, le japonais, le kazakh, le ourdou, le telugu et le turc.

Il y a ensuite un groupe de corpus fortement SOV mais ayant très peu de postpositions : l'`Afrikaans-AfriBooms`, le `Persian-Seraji`, le `Persian-PADT` et le `Korean-PUD`. L'afrikaans et le perse sont deux langues d'ordre dominant SOV mais elles sont effectivement prépositionnelles. L'exception du perse est soulignée par Greenberg avec l'amharique qui n'est pas SOV dans nos résultats. L'ordre dominant de l'afrikaans peut cependant être remis en doute sachant que c'est une langue issue du néerlandais, qui lui n'a pas d'ordre dominant selon nos résultats.

Le corpus `Korean-PUD` ne présente pas de postpositions mais cela est dû aux problèmes d'annotations et de couverture de motifs détaillés en section 3.2.2. Nous pouvons d'ailleurs noter que les deux autres corpus du coréen ont très largement des postpositions.

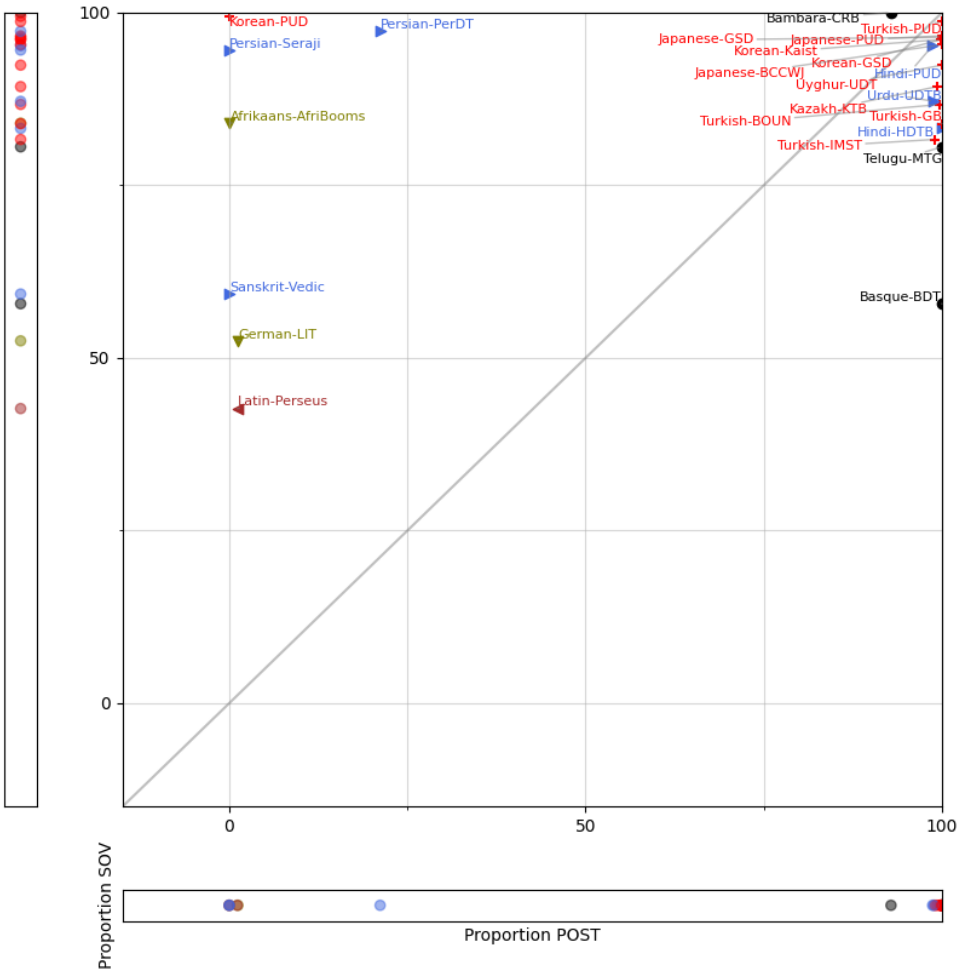

FIGURE 4.1 – Corpus d'ordre SOV en fonction de leurs proportions de postpositions.

Par ailleurs, trois corpus se détachent avec un pourcentage de SOV autour des 50 % : le `German-LIT`, le `Latin-Perseus` et le `Sanskrit-Vedic`. Le corpus de l'allemand et du latin ont un ordre dominant SOV mais comme nous le détaillons en section 4.2, ces deux langues sont considérées comme n'ayant pas d'ordre dominant. Nous supposons que la formulation de Greenberg « normal SOV order » permet de ne pas prendre en compte ce type de langues dans son universel. Concernant le sanskrit, nous sommes dans le même cas de figure que le `Korean-PUD`. Notre motif n'a pas permis de trouver d'occurrences de postpositions ni de prépositions.

Enfin, le corpus du basque se retrouve isolé à cause d'un pourcentage de SOV relativement faible à 57 %, mais suffisant pour être considéré comme l'ordre dominant, le second ordre étant l'ordre SVO à 19 %. WALS considère également le basque comme une langue SOV et postpositionnelle.

Nos résultats sont majoritairement en accord avec l'universel 4 de Greenberg. La présence d'exceptions est également confirmée avec les corpus du perse.

4.2 Hétérogénéité entre corpus d'une même langue

Dans nos expériences, pour les langues présentant plusieurs corpus, nous décidons de traiter les corpus séparément et non de considérer tous les corpus comme une seule ressource pour la langue. Cela nous permet d'observer s'il y a une hétérogénéité entre les corpus d'une même langue, les possibles variations et en comprendre les raisons.

L'homogénéité entre les corpus d'une même langue est variable selon les différentes classifications observées. Elle est particulièrement visible dans l'ordre sujet - verbe - objet où six ordres différents sont possibles, contrairement aux ordres composés de deux éléments, adjectif - nom et adposition - nom. Le but ici est de trouver une mesure pour déterminer quantitativement l'homogénéité dans les distributions des ordres sujet - verbe - objet dans les corpus d'une même langue. Pour ce faire, nous avons choisi d'utiliser une similarité cosinus que nous appliquons à chaque paire de corpus qui sont représentés sous la forme de vecteurs de dimensions six, les six composantes correspondant aux six ordres.

Le tableau 4.4 montre un exemple des distributions des ordres sujet - verbe - objet pour les corpus du roumain. En roumain, deux corpus sont considérés comme SVO, le `Romanian-SiMoNERo` et le `Romanian-RRT`, tandis que le `Romanian-NonStandard` n'a pas d'ordre dominant. À première vue, les deux corpus SVO présentent des distributions proches, mais diffèrent fortement avec le `Romanian-NonStandard`.

Corpus	SVO	SOV	VSO	VOS	OSV	OVS
Romanian-NonStandard	**38,07 %**	**31,87 %**	9,66 %	3,97 %	1,71 %	14,72 %
Romanian-RRT	**85,32 %**	7,76 %	1,12 %	0,70 %	1,18 %	3,91 %
Romanian-SiMoNERo	**97,61 %**	0,97 %	0,09 %	0,09 %	0,13 %	1,10 %

TABLEAU 4.4 – Distribution des ordres sujet - verbe - objet des corpus du roumain.

Nous calculons donc les similarités cosinus entre toutes les paires de corpus, représentées dans une *heatmap* en figure 4.2. Plus la valeur de cosinus est proche de 1, plus les distributions des corpus sont similaires et donc homogènes. La figure 4.2 montre clairement que le corpus du `Romanian-NonStandard` se distingue des deux autres avec des valeurs de cosinus en dessous de 0,80 tandis que les deux corpus SVO ont un cosinus très proche de 1, à 0,9965.

Nous effectuons le même calcul sur les langues multicorpus et nous isolons la valeur de cosinus la plus faible entre toutes les paires de corpus d'une même langue. Pour le roumain, le cosinus minimum est donc de 0,7458. Ensuite, nous faisons un classement de ces langues selon leur valeur de cosinus minimum. Le classement est présenté en figure 4.3.

Sur les 29 langues multicorpus, 23 langues ont une forte homogénéité avec un cosinus minimum au dessus de 0,91. À l'exception du tchèque, ces corpus ont même un cosinus minimum supérieur à 0,95. Près de 80 % des langues multicorpus ont une forte homogénéité dans leurs distributions de l'ordre sujet - verbe - objet.

Nous remarquons ensuite que les six langues présentant les plus faibles cosinus minimum, en vert sur la figure 4.3 correspondent aux six langues où au moins un corpus ne présente pas le même ordre dominant que les autres. Nous avons poussé l'analyse plus loin pour ces six langues afin de tenter de comprendre les raisons de leur hétérogénéité.

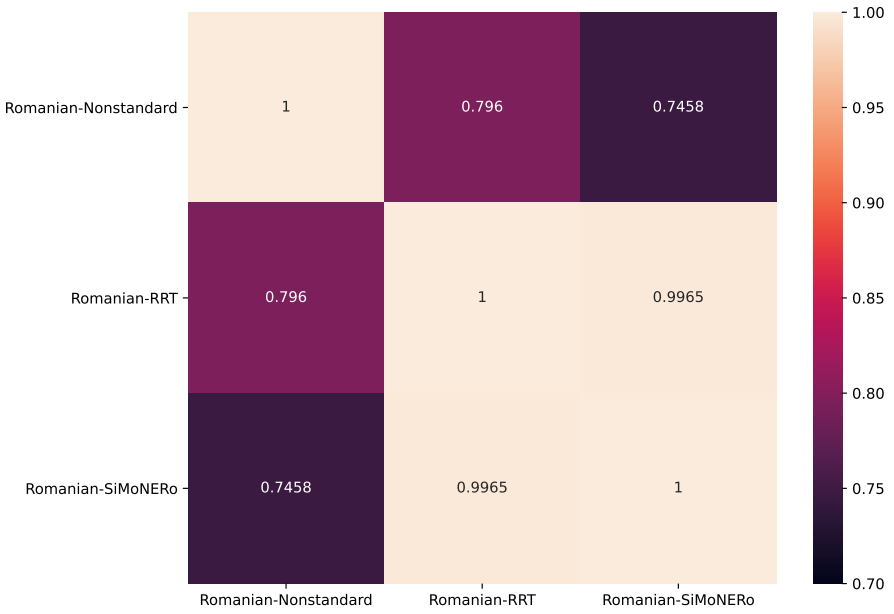

FIGURE 4.2 – Valeurs de cosinus pour les trois corpus du roumain.

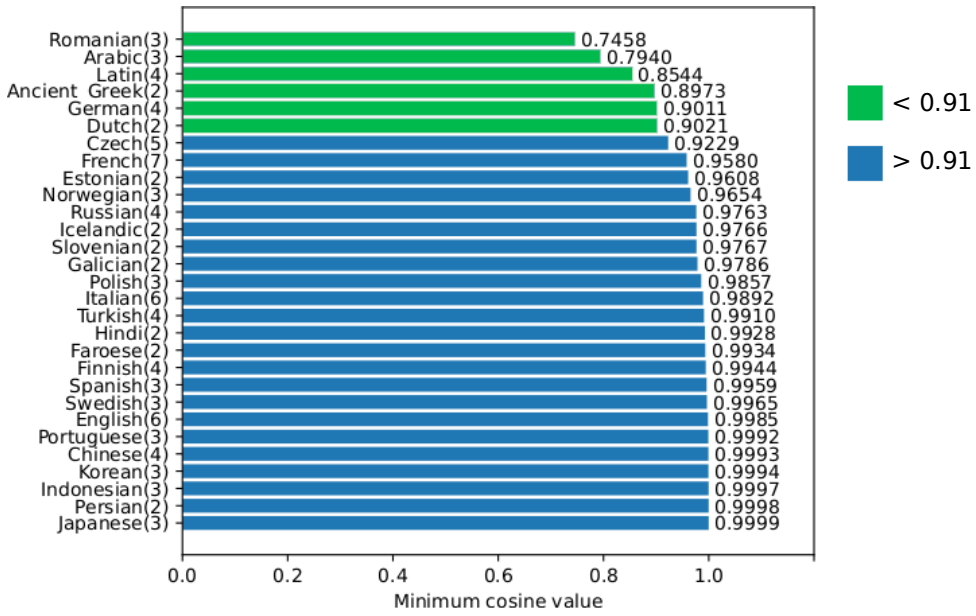

FIGURE 4.3 – Langues multicorpus (nombre de corpus pour la langue entre parenthèses) classées selon la valeur de cosinus minimum.

4.2.1 L'influence du genre

Nous avons supposé que le genre des textes des corpus pouvait avoir une influence dans la distribution de l'ordre sujet - verbe - objet. En effet, d'un corpus à un autre, les textes peuvent provenir de différents genres. Il est également possible d'avoir des textes de différents genres dans un même corpus.

Le roumain Le genre a une forte influence dans le cas du roumain. WALS considère le roumain comme une langue SVO, ce qui est cohérent avec les résultats trouvés sur les corpus `Romanian-RRT` et `Romanian-SiMoNERo`. Quant au `Romanian-NonStandard`, c'est un

corpus atypique qui contient des textes très variés : des extraits du Nouveau Testament, des folklores en vers, de ballades roumaines et moldaves, de chroniques et de poésie. Par opposition, les deux autres corpus du roumain ont des textes extraits d'articles de Wikipedia, de journaux, du domaine médical et académique et de fiction. Nous pouvons supposer que ces corpus présentent des formulations plus standards et conventionnelles, tandis que les genres du `Romanian-NonStandard` donnent plus de liberté à la manière de tourner les phrases.

L'allemand Certaines langues admettent plusieurs ordres selon les constructions. C'est le cas de l'allemand qui présente l'ordre SVO dans les propositions principales et l'ordre SOV dans les propositions subordonnées. De ce fait, WALS considère l'allemand comme n'ayant pas d'ordre dominant.

Le tableau 4.5 présente les distributions des corpus de l'allemand. Nous notons que trois corpus n'ont pas d'ordre dominant avec l'ordre SOV en plus fréquent et SVO en deuxième plus fréquent. Seul le `German-LIT` a un ordre dominant SOV. Pour le `LIT`, nous supposons qu'il y une influence du genre, ce corpus étant le seul à présenter des textes du premier romantisme (fin du 18$^{\text{ème}}$ siècle) qui sont courts et sous forme d'aphorisme. Ces textes traitent des questions philosophiques concernant notamment l'art et la beauté. Les trois autres corpus de l'allemand ont des textes provenant majoritairement de journaux et de Wikipedia.

Corpus	SVO	SOV	VSO	VOS	OSV	OVS	Ratio
German-GSD	**32,82 %**	**33,64 %**	20,73 %	3,58 %	6,29 %	2,93 %	1,02
German-HDT	**24,68 %**	**47,22 %**	14,42 %	2,02 %	7,30 %	4,37 %	1,91
German-LIT	22,58 %	**52,47 %**	8,28 %	2,37 %	12,80 %	1,50 %	2,32
German-PUD	**32,23 %**	**45,35 %**	13,79 %	2,66 %	5,32 %	0,66 %	1,40

TABLEAU 4.5 – Distribution des ordres sujet - verbe - objet et le ratio dans les corpus de l'allemand.

Si nous observons les cosinus des corpus de l'allemand en figure 4.4, les valeurs de cosinus sont relativement élevées, supérieures à 0,95 sauf entre le corpus `German-LIT` et `German-GSD` qui est à 0,9011. Le `German-GSD` a une distribution plus homogène entre l'ordre SVO et SOV, comme l'indique le ratio à 1,02. En revanche, le `German-LIT` a un ratio plus élevé, à 2,32 pour ces deux ordres. Par ailleurs, nous pouvons noter que le troisième ordre le plus fréquent est généralement l'ordre VSO, sauf pour le `German-LIT` qui a l'ordre OSV à 12,80 %.

Si les corpus `German-GSD`, `German-HDT` et `German-PUD` sont homogènes dans leurs distributions, nous supposons que le corpus `German-LIT` est un cas atypique en raison du genre de ses textes.

En observant les valeurs de cosinus, nous pouvons noter que le `German-GSD` est le moins homogène avec les autres corpus. Si nous ne prenons pas en compte le cosinus minimum entre le `German-LIT` et le `German-GSD`, le cosinum minimum est celui entre le `German-GSD` et le `German-HDT` à 0,9517. Cette observation nous amène à penser que les ratios et les valeurs de cosinus permettent d'obtenir des informations différentes sur les corpus.

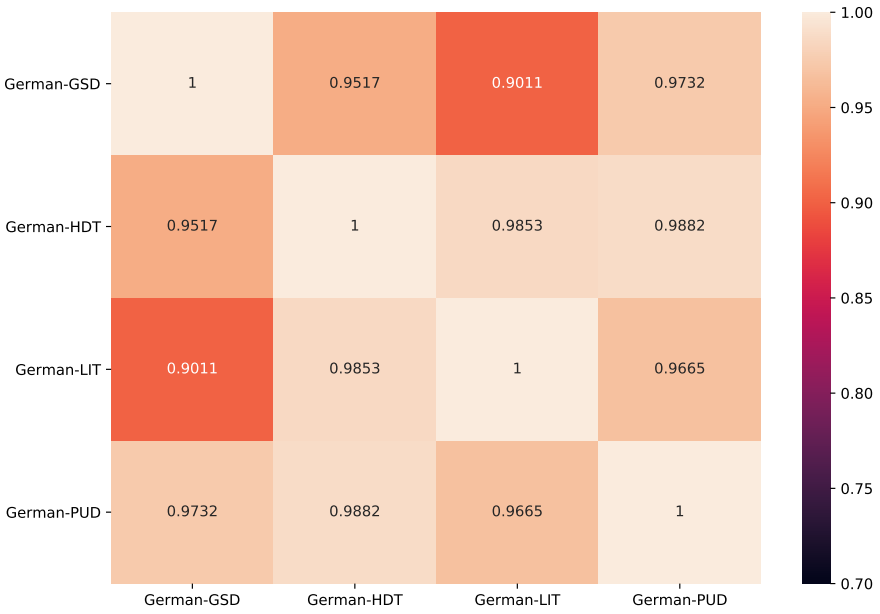

FIGURE 4.4 – Valeurs de cosinus pour les quatre corpus de l'allemand.

4.2.2 Le facteur temporel

Pour les langues mortes, le grec ancien et le latin, les textes des corpus s'étendent sur différents siècles.

Le latin Dans le cas du latin, le `Latin-Perseus` présente des textes datant du 1er siècle avant J.-C. tandis que les textes du `Latin-ITTB` sont du latin médiéval du 13ème siècle. La figure 4.5 montre effectivement que le cosinus minimum est obtenu entre ces deux corpus.

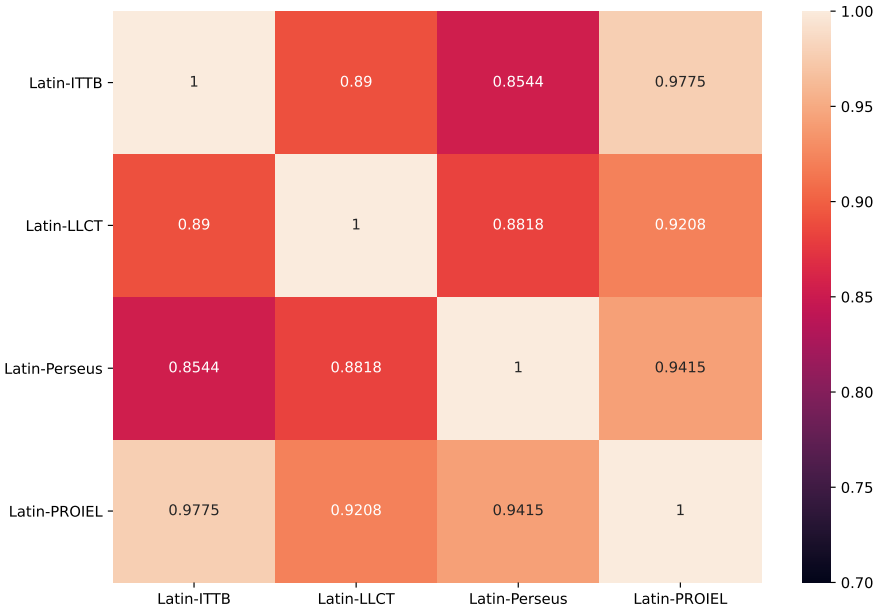

FIGURE 4.5 – Valeurs de cosinus pour les quatre corpus du latin.

Le grec ancien Le grec ancien présente deux corpus, chacun présentant un mélange de textes de différentes périodes. Le `Ancient_Greek-PROIEL` contient des textes d'Hérodote (5ème siècle) et de la bible et le `Ancient_Greek-Perseus` présente des textes allant d'Homère (8ème siècle avant J.-C) à Athénée de Naucratis (fin du 2ème siècle). D'après le tableau 4.6, nous pouvons voir que tous les ordres sont présents avec des proportions plus importantes pour les ordres SVO et SOV.

Corpus	SVO	SOV	VSO	VOS	OSV	OVS	Ratio
AncientGreek-Perseus	**24,96 %**	**30,89 %**	3,82 %	5,42 %	16,02 %	18,88 %	1.24
AncientGreek-PROIEL	**41,50 %**	**21,24 %**	8,25 %	7,82 %	11,58 %	9,61 %	1.95

TABLEAU 4.6 – Distribution des ordres sujet - verbe - objet et le ratio dans corpus du grec ancien.

L'hétérogénéité pour les langues mortes est en réalité d'autant plus significative qu'elles sont très libres dans l'ordre des mots qui peut varier selon les époques mais aussi selon les auteurs.

4.2.3 Les spécificités des langues

Les deux dernières langues présentant différents ordres dominants sont le néerlandais et l'arabe. Leurs hétérogénéités peuvent être expliquées par certaines particularités de ces langues.

Le néerlandais Comme l'allemand, le néerlandais est une langue admettant l'ordre SVO et l'ordre SOV selon le type de propositions. Deux corpus sont disponibles pour le néerlandais et leurs distributions sont présentées dans le tableau 4.7.

Corpus	SVO	SOV	VSO	VOS	OSV	OVS	Ratio
Dutch-Alpino	**29,00 %**	**51,28 %**	10,02 %	0,06 %	7,38 %	2,26 %	1,77
Dutch-LassySmall	**41,08 %**	**32,43 %**	22,72 %	0,00 %	3,03 %	0,74 %	1,27

TABLEAU 4.7 – Distribution des ordres sujet - verbe - objet et le ratio dans les corpus du néerlandais.

Les deux corpus n'ont pas d'ordre dominant mais l'`Dutch-Alpino` a l'ordre SOV en ordre le plus fréquent, tandis que le `Dutch-Lassy-Small` a l'ordre SVO. Tout comme pour l'allemand, WALS considère que le néerlandais n'a pas d'ordre dominant. Ici, nous concluons simplement que les fréquences dépendent des proportions de constructions présentant l'ordre SVO et de celles présentant l'ordre SOV.

L'arabe En arabe, nous avons trois corpus dans notre échantillon. WALS considère l'arabe moderne comme étant VSO mais nous obtenons des résultats hétérogènes entre les corpus. D'après le tableau 4.8, un corpus est d'ordre VSO et deux corpus n'ont pas d'ordre dominant mais avec un premier ordre différent : l'ordre SVO pour le `Arabic-PADT` et l'ordre VSO pour le `Arabic-PUD`.

Bien que l'ordre VSO soit l'ordre le plus courant, l'arabe permet de déplacer le sujet avant le verbe pour mettre l'accent sur le sujet. C'est le phénomène de topicalisation. Le corpus `Arabic-PADT` qui présente environ 48 % de SVO contient énormément de phrases

Corpus	SVO	SOV	VSO	VOS	OSV	OVS	Ratio
Arabic-NYUAD	16,96 %	0,32 %	**54,56 %**	22,63 %	0,28 %	5,24 %	2,41
Arabic-PADT	**48,15 %**	0,00 %	**31,20 %**	20,55 %	0,02 %	0,08 %	1,54
Arabic-PUD	**35,71 %**	0,00 %	**49,45 %**	9,34 %	0,00 %	5,49 %	1,38

TABLEAU 4.8 – Distribution des ordres sujet - verbe - objet et le ratio dans les corpus de l'arabe.

qui s'avèrent être des titres de journaux et dans lesquels la topicalisation est très fréquente. C'est pourquoi ce corpus a un fort taux de SVO. La figure 4.6 présente un exemple avec la phrase : « Les compagnies aériennes luttent contre la propagation du SRAS ».

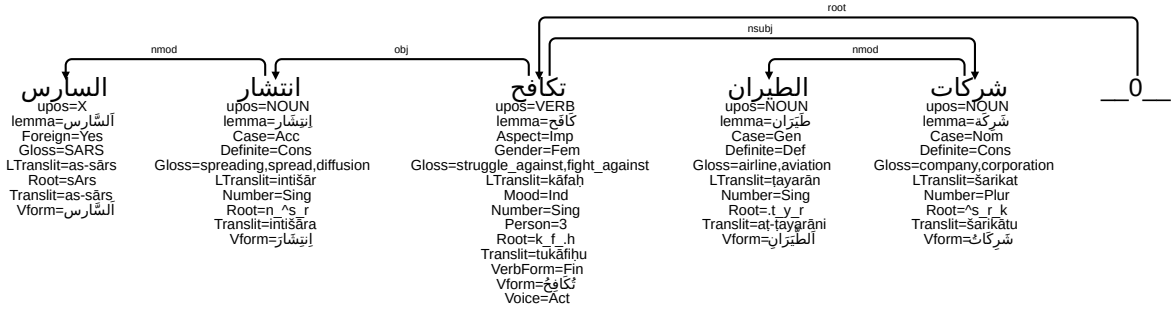

FIGURE 4.6 – Phrase du corpus `Arabic-PADT` présentant une topicalisation du sujet.

Les distributions des trois corpus sont très hétérogènes, avec les trois ordres les plus fréquents, VSO, SVO et VOS mais qui n'apparaissent pas dans les mêmes proportions selon les corpus. Cette observation est confirmée avec les cosinus dans la figure 4.7. Nous remarquons que le `Arabic-PUD` a un cosinus autour de 0,92 avec les deux autres corpus, ce qui est largement au dessus du cosinus minimum mais qui n'est pas si élevé comparé aux cosinus minimum des autres langues. Par ailleurs, la figure nous montre que le `Arabic-PADT` et le `Arabic-NYUAD` ont un cosinus faible à 0,794, ces deux corpus présentant les distributions les plus opposées.

4.3 Observations et retours sur UD

Dans cette partie, nous présentons différents points marquants sur les annotations UD que nous avons pu remarquer à travers l'analyse de 141 corpus simultanément. Nous traitons les relations de sujet nominal `nsubj`, d'objet nominal `obj` et complément par une adposition `case`. Ces relations mettent en jeu des groupes nominaux selon les instructions d'UD.

4.3.1 Les relations à dépendants nominaux

Dans nos expériences pour déterminer un ordre sujet - verbe - objet, nous devions considérer uniquement les sujets et objets nominaux. Ces fonctions sont annotées respectivement avec les relations `nsubj` et `obj`.

Nous avons lancé une première expérience sans fixer les parties du discours du sujet et de l'objet en tant que nom, pronom ou nom propre, pour éviter une redondance

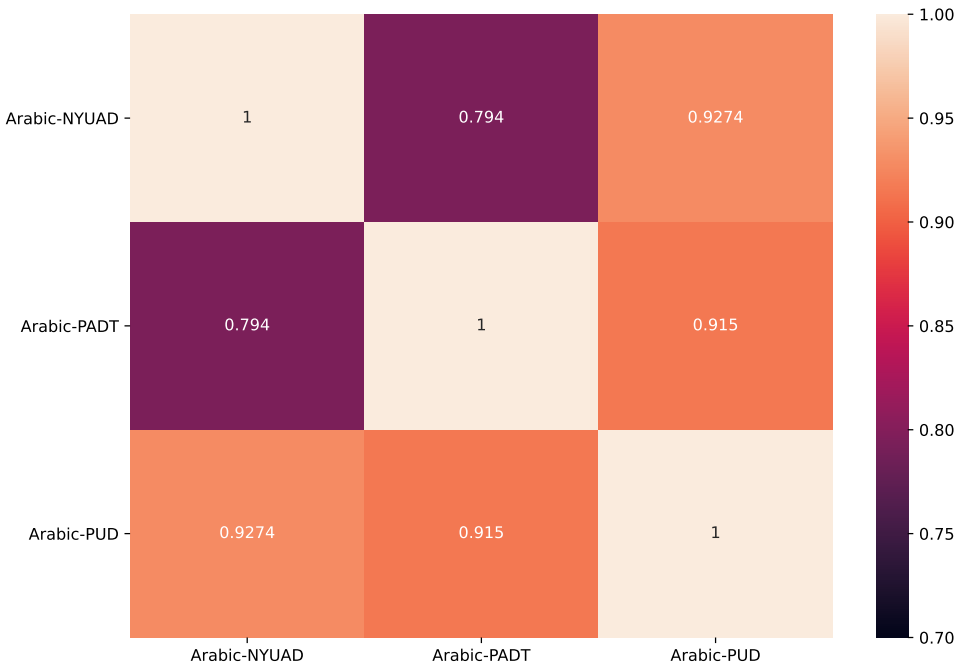

FIGURE 4.7 – Valeurs de cosinus pour les trois corpus de l'arabe.

dans les motifs. Nous avons obtenu des résultats hétérogènes notamment entre les corpus d'une même langue. En observant de plus près certains corpus, nous retrouvions à plusieurs reprises et sur des corpus de langues différentes que le sujet et l'objet n'étaient pas nominaux, ce qui est incohérent avec les définitions des relations `nsubj` et `obj`.

Avec GREW et les motifs en figure 4.8, nous avons calculé la proportion de sujets et d'objets non nominaux lorsqu'ils sont reliés à un verbe.

```
pattern {
  V [upos=VERB];
  S [upos<>PROPN|PRON|NOUN];
  V -[1=nsubj]-> S;
}
```

```
pattern {
  V [upos=VERB];
  O [upos<>PROPN|PRON|NOUN];
  V -[1=obj]-> O;
}
```

FIGURE 4.8 – Motifs sujet non nominal à gauche, objet non nominal à droite.

D'après la figure 4.9, nous pouvons observer que sur 141 corpus, 118 corpus présentent moins de 10 % de sujets non nominaux dans la relation `nsubj` avec un verbe. 17 corpus en présentent entre 10 % et 20 % et six en présentent à plus de 20 %. Nous retrouvons quasiment les mêmes proportions pour les objets non nominaux. Toutefois, ce ne sont pas obligatoirement les mêmes corpus qui présentent le plus de sujets et d'objets non nominaux.

Plus de 80 % des corpus sont cohérents avec les instructions d'UD sur la relation `nsubj` et `obj`. Cependant, six corpus présentent des sujets non nominaux à plus de 20 % et quatre corpus ont plus de 20 % d'objets non nominaux. Les tableaux 4.9 et 4.10 détaillent les proportions relatives à chaque corpus ainsi que les parties du discours les plus représentées.

Nous remarquons que les sujets non nominaux des trois corpus `PROIEL` sont majoritairement des adjectifs entre 72 % et 78 %. Le `Thai-PUD` et le `Slovenian-SST` ont une forte proportion de déterminants en sujet et sur les 25 % de sujets non nominaux de

FIGURE 4.9 – Distribution des corpus selon leurs proportions de sujets non nominaux à gauche et d'objets non nominaux à droite.

l'`Arabic-PADT`, 50 % sont annotés X, étiquette utilisée lorsque les autres étiquettes de parties du discours ne sont pas adaptées.

Quant aux quatre corpus présentant des objets non nominaux à plus de 20 %, pour trois corpus, les objets non nominaux sont des verbes : le `Turkish-IMST`, le `Hindi-HDTB` et le `Urdu-UDTB`. Le dernier corpus est l'`Ancient_Greek-PROIEL` qui présente des adjectifs en objets.

Corpus	Sujets non nominaux	VERB	ADJ	DET	X
Old_Church_Slavonic-PROIEL	27,06 %	20, 94 %	**73,89 %**	0,00 %	0,00 %
Thai-PUD	25,86 %	14,40 %	0,28 %	**83,10 %**	0,00 %
Arabic-PADT	25,15 %	0,34 %	10,72 %	31,88 %	**50,34 %**
Ancient_Greek-PROIEL	23,67 %	17,55 %	**78,55 %**	0,00 %	0,00 %
Gothic-PROIEL	21,82 %	22,42 %	**72,17 %**	0,00 %	0,00 %
Slovenian-SST	20,24 %	0,00 %	9,46 %	**79,28 %**	1,35 %

TABLEAU 4.9 – Proportions et parties du discours des sujets non nominaux pour les six corpus ayant le plus de sujets non nominaux.

Corpus	Objets non nominaux	VERB	ADJ
Turkish-IMST	32,80 %	**65,72 %**	26,03 %
Hindi-HDTB	26,43 %	**89,00 %**	9,55 %
Urdu-UDTB	24,87 %	**88,83 %**	8,14 %
Ancient_Greek-PROIEL	20,96 %	19,12 %	**76.95 %**

TABLEAU 4.10 – Proportions et parties du discours des objets non nominaux pour les quatre corpus ayant le plus d'objets non nominaux.

Une forte proportion de sujets ou d'objet non nominaux peut avoir une influence sur la détermination de l'ordre sujet - verbe - objet. C'est le cas notamment dans deux corpus : le `Slovenian-SST` et le `Hindi-HDTB`.

Le `Slovenian-SST` présente 20,24 % de sujets non nominaux et 19,01 % d'objets non nominaux. Lorsque nous ne fixons pas les parties du discours pour le sujet et pour l'objet,

nous n'obtenons aucun ordre dominant à cause d'un ratio à 1,71 entre les ordres SVO et SOV. En ajoutant les parties du discours dans nos *patterns* d'extraction, le corpus devient SVO avec un ratio de 2,27. Nous gagnons également en homogénéité avec le deuxième corpus du slovène qui est SVO, le `Slovenian-SSJ`, avec un cosinus passant de 0,94 à 0,97.

Nous remarquons le même phénomène pour l'`Hindi-HDTB` qui a 26,43 % d'objets non nominaux. Le corpus est considéré comme n'ayant pas d'ordre dominant (SOV/SVO) avec un ratio de 1,18 sans les parties de discours fixées. Ce résultat se trouve être incohérent avec le deuxième corpus d'hindi, le `Hindi-PUD` qui est SOV avec un ratio à 23,04. En demandant l'aide d'une locutrice de l'hindi, nous nous rendons compte que certains objets sont annotés en tant que verbes dans le corpus. Lorsque nous fixons les parties du discours en tant qu'entités nominales, le `Hindi-HDTB` devient SOV avec un ratio de 8,39[2]. Les résultats gagnent fortement en cohérence avec un cosinus entre les deux corpus qui passe de 0,7676 à 0,9928.

4.3.2 La relation `case`

La même réflexion s'est posée lors de la détermination des prépositions et des postpositions avec la relation `case`. Selon UD, la relation `case` permet d'annoter un élément marqueur de cas ou une adposition. Ces éléments sont des dépendants des noms auxquels ils sont rattachés[3].

La relation `case` implique donc des gouverneurs nominaux. De la même manière que pour les relations `nsubj` et `obj`, nous avons calculé la proportion de relations `case` à gouverneurs non nominaux dans les 141 corpus avec le motif 4.10.

```
pattern {
        A [upos=ADP];
        GOV [upos<>NOUN|PRON|PROPN];
        GOV -[case]-> A;
}
```

FIGURE 4.10 – Motif de la relation `case` impliquant un gouverneur non nominal.

La figure 4.11 montre une répartition similaire aux relations `nsubj` et `obj` avec et 110 corpus à moins de 10 % de gouverneurs non nominaux, 24 corpus entre 10 % et 20 % et sept corpus à plus de 20 %. Ces derniers sont présentés dans le tableau 4.11.

Dans la suite de cette partie, nous explorons les possibles raisons de ces observations en turc, en chinois et en coréen dans la mesure où nous pouvons comparer les résultats entre les différents corpus. Par conséquent, nous ne traiterons pas les deux langues monocorpus, le basque et l'amharique dans notre analyse.

Le turc Nous pouvons remarquer que les pourcentages de gouverneurs non nominaux sont élevés pour deux corpus du turc avec 45,95 % et 37,49 %. En observant les parties du discours des gouverneurs non nominaux, nous pouvons voir que plus de la moitié sont des verbes. En comparaison, les deux autres corpus du turc présentent des proportions faibles de gouverneurs non nominaux : le `Turkish-GB` est à 6,23 % et le `Turkish-PUD` est à 4,97 %.

2. Le `Hindi-PUD` reste SOV avec un ratio de 22,37.
3. Voir : https://universaldependencies.org/u/dep/case, septembre 2021.

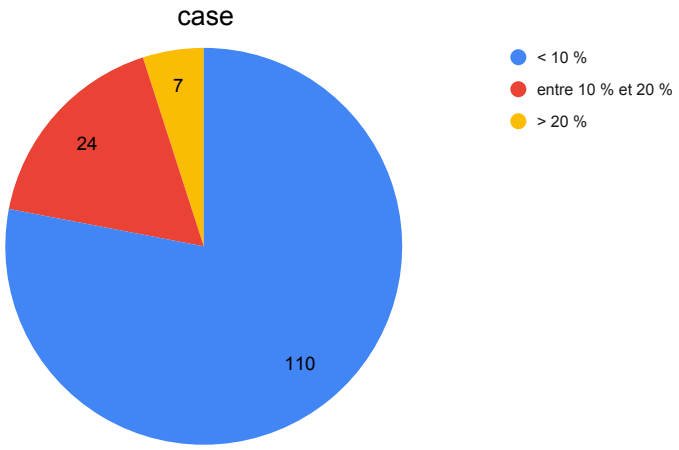

FIGURE 4.11 – Distribution des corpus selon leurs proportions de compléments non nominaux introduits par une adposition.

Corpus	Gouverneurs non nominaux	VERB	NUM	ADJ	DET	PART
Turkish-BOUN	45,95 %	**54,73 %**	0,96 %	17,46 %	1,33 %	0,00 %
Turkish-IMST	37,49 %	**55,40 %**	6,89 %	29,46 %	3,51 %	0,00 %
Chinese-GSD	28,77 %	**60,76 %**	1,02 %	6,31 %	0,09 %	27,27 %
Chinese-GSDSimp	28,55 %	**61,21 %**	1,03 %	6,35 %	0,09 %	27,29 %
Basque-BDT	28,27 %	0,00 %	19,85 %	24,95 %	**38,94 %**	0,00 %
Amharic-ATT	22,47 %	**75,41 %**	1,64 %	7,38 %	5,74 %	0,00 %
Korean-Kaist	20,59 %	19,51 %	**56,50 %**	5,69 %	0,00 %	0,00 %

TABLEAU 4.11 – Proportions et parties du discours des gouverneurs non nominaux reliés à une adposition.

La documentation nous précise que les relations de dépendances du `Turkish-BOUN` ont été annotées manuellement dans le schéma UD, tandis que le `Turkish-IMST` est le résultat d'une conversion semi-automatique du IMST Treebank (Sulubacak et al., 2016). En l'absence d'un locuteur du turc, nous ne pouvons formuler que des hypothèses. Le corpus `Turkish-BOUN` étant annoté manuellement par des natifs, nous pouvons supposer qu'ils ont fait un choix d'annotations différent des autres corpus mais aussi contradictoire avec les instructions d'UD. Pour le `Turkish-IMST`, cela peut être dû au processus de conversion semi-automatique.

Dans la version 2.8 d'UD, quatre nouveaux corpus ont été ajoutés et il est indiqué que des mises à jour ont été effectuées afin de gagner en cohérence dans tous les corpus du turc [4]. Cependant, nos résultats demeurent identiques concernant le `Turkish-BOUN` et le `Turkish-IMST` de la version 2.8.

Le chinois Les deux corpus `GSD` du chinois présentent des pourcentages autour de 28 % de gouverneurs non nominaux dans la relation `case`. En réalité, le corpus `Chinese-GSDSimp` est une version du `Chinese-GSD` en chinois simplifié obtenue par conversion automatique puis correction manuelle. Il n'est donc pas surprenant que nous obtenions des résultats similaires dans ces deux corpus.

4. Voir : https://github.com/UniversalDependencies/UD_Turkish-BOUN, septembre 2021.

Les gouverneurs non nominaux de ces corpus sont majoritairement des verbes à environ 60 %. Du fait de la structure de la langue chinois, la définition de la relation `case` est légèrement différente de la définition universelle. En effet, la relation `case` est utilisée sur des particules marquant des relations comme le génitif, des prépositions y compris les « co-verbes », les marques de valence et les comparatifs [5].

Les « co-verbes » correspondent à des particules ou adpositions liées aux verbes comme :
— 在+ 落在 : « tomber sur »,
— 向+ 奔向 : « courir vers ».

Dans les corpus `GSD`, ils sont annotés en tant qu'adposition et sont donc reliés aux verbes avec la relation `case`. Dans les deux autres corpus du chinois, le `Chinese-PUD` et le `Chinese-HK`, les pourcentages de gouverneurs non nominaux sont respectivement de 2,29 % et 5,22 %. En explorant ces corpus avec GREW-MATCH, nous observons que les liens entre une adposition et un verbe sont annotés avec la relation `mark`. En chinois, elle est utilisée sur un mot fonctionnel marquant une proposition subordonnée [6]. Dans la définition universelle, cette relation met en jeu une conjonction de subordination annotée `SCONJ` plutôt qu'une adposition mais le chinois emploie les deux étiquettes.

Pour le chinois, nous sommes face à plusieurs difficultés :
— l'utilisation des relations `case` et `mark` pour annoter les co-verbes,
— le conflit entre les étiquettes `ADP`, `PART` et `SCONJ`

Comme nous l'avons précédemment observé en partie 3.2.2, les corpus `Chinese-GSD` et `Chinese-GSDSimp` se distinguent des deux autres corpus à cause de certaines annotations incohérentes avec le schéma UD.

Le coréen Le corpus `Korean-Kaist` présente un pourcentage de 20,59 % de gouverneurs non nominaux dans la relation `case` et 56,50 % de ces gouverneurs sont des nombres. La figure 4.12 montre un exemple d'une phrase avec une construction `case` et un gouverneur numéral.

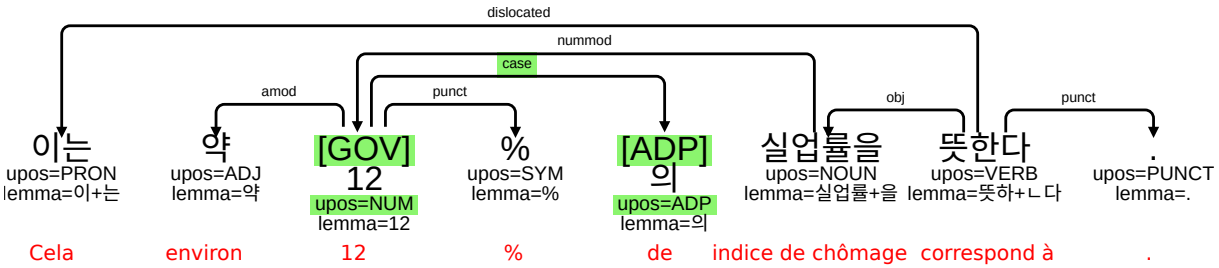

FIGURE 4.12 – Phrase du corpus `Korean-Kaist` impliquant une relation `case` et un gouverneur numéral.

La phrase se traduit par : « Cela correspond à environ 12 % d'indice de chômage. ». Le `Korean-GSD` présente également cette annotation mais il contient peu de phrases impliquant un numéral suivie d'un signe pour cent [7]. En effet, le `Korean-Kaist` en présente

5. Voir : https://universaldependencies.org/zh/dep/case, septembre 2021.
6. Voir : https://universaldependencies.org/zh/dep/mark, septembre 2021.
7. Comme nous le détaillons dans la partie 3.2.2, le `Korean-PUD` ne présente pas l'étiquette `ADP`.Nous ne faisons donc pas de comparaison ici.

beaucoup plus car une grande partie des phrases provient de corpus scientifiques.

Pour annoter les nombres suivis de symbole, il n'y a pas de réel consensus entre les langues sur l'entité qui doit porter la relation syntaxique. En français, comme le montre la figure 4.13, le numéral et le symbole sont reliés par une relation `nummod` et le symbole est relié au nom auquel il se rattache par une relation `nmod`. De plus, nous remarquons que le signe % est annoté en tant que nom et non un symbole.

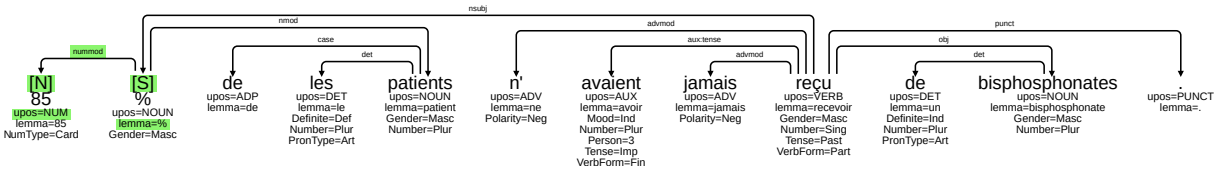

FIGURE 4.13 – Phrase du corpus `French-Sequoia` impliquant un numéral et un symbole.

Bien qu'UD décide que la relation `case` implique des gouverneurs nominaux, d'autres langues présentent des relations `case` entre une adposition et un numéral. En français par exemple, le corpus `French-Sequoia` présente cette construction dans les dates de la forme « en 1980 ».

Conclusion

Dans ce mémoire, nous avons profité des avantages apportés par le TAL, en termes de ressources et d'outils, pour vérifier empiriquement des universaux linguistiques sur plus de 70 langues. Notre étude s'inscrit dans le même esprit que le travail de Greenberg, à la différence que nous l'avons entreprise sur de grandes quantités de données.

Les résultats obtenus s'avèrent être majoritairement cohérents, d'une part avec les observations de Greenberg, d'autre part avec les informations contenues dans la base de données typologiques WALS. En outre, nous nous sommes efforcés de fournir des analyses pour expliquer les incohérences détectées, soit en examinant les documentations relatives aux corpus, soit en faisant appel à des locuteurs natifs autour de nous. Nos résultats constituent des informations typologiques nouvelles qui peuvent compléter des brèches dans les bases de données. Nous pouvons notamment citer six langues pour lesquelles les ordres dominants sujet - verbe - objet et adjectif - nom ne sont pas fournis dans WALS : l'afrikaans, le féroïen, le galicien, le kazakh, le maltais et le slovaque.

De plus, une partie importante de notre étude a été consacrée à l'évaluation de la cohérence entre les corpus d'une même langue. Après avoir parcouru les corpus, nous nous sommes rendu compte qu'il existait une grande variation dans la représentation des réalisations langagières : la langue orale, la langue écrite dans les journaux, dans les réseaux sociaux (corpus de *tweets*), la poésie, les romans, les grammaires, etc. Considérer les corpus individuellement a permis de préserver leurs spécificités et d'évaluer leurs impacts sur nos résultats.

Au-delà de confirmer les universaux linguistiques testés, notre étude soulève plusieurs problématiques liées à Universal Dependencies, et plus particulièrement au caractère universel de son schéma d'annotations. Si les finalités qui sous-tendent la création du schéma UD sont louables, son application sur toutes les langues reste une tâche délicate. Il faut noter que 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 qu'ils doivent justifier et qui ne sont pas forcément en accord avec les autres corpus de la langue. Nous nous retrouvons donc avec une cohérence moindre entre les langues mais également entre corpus d'une même langue.

Ce phénomène est en réalité une contrepartie de l'aspect collaboratif d'UD. L'évolution du projet est possible grâce à la participation bénévole de la communauté internationale du TAL. Bien qu'UD pose certaines exigences avant d'accepter un corpus, aucune autorité n'existe pour vérifier de manière approfondie le respect du guide d'annotation. Toutefois, l'aspect collaboratif nous a permis de contribuer à l'amélioration des corpus en partageant nos observations *via* les dépôts *Git* du projet. Par ailleurs, une partie de nos expériences a fait l'objet d'un article appelé *Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting* (Choi et al., 2021), publié à RANLP 2021.

Annexes

Universaux de Greenberg

1. In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.
2. In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.
3. Languages with dominant VSO order are always prepositional.
4. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.
5. If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun.
6. All languages with dominant VSO order have SVO as an alternative or as the only alternative basic order.
7. If in a language with dominant SOV order there is no alternative basic order, or only OSV as the alternative, then all adverbial modifiers of the verb likewise precede the verb. (This is the 'rigid' subtype of III.)
8. When a yes-no question is differentiated from the corresponding assertion by an intonational pattern, the distinctive intonational features of each of these patterns are reckoned from the end of the sentence rather than from the beginning.
9. With well more than chance frequency, when question particles or affixes are specified in position by reference to the sentence as a whole, if initial, such elements are found in prepositional languages, and, if final, in postpositional.
10. Question particles or affixes, when specified in position by reference to a particular word in the sentence, almost always follow that word. Such particles do not occur in languages with dominant order VSO.
11. Inversion of statement order so that verb precedes subject occurs only in languages where the question word or phrase is normally initial. This same inversion occurs in yes-no questions only if it also occurs in interrogative word questions.
12. If a language has dominant order VSO in declarative sentences, it always puts interrogative words or phrases first in interrogative word questions ; if it has dominant order SOV in declarative sentences, there is never such an invariant rule.
13. If the nominal object always precedes the verb, then verb forms subordinate to the main verb also precede it.
14. In conditional statements, the conditional clause precedes the conclusion as the normal order in all languages.
15. In expressions of volition and purpose, a subordinate verbal form always follows the main verb as the normal order except in those languages in which the nominal object always precedes the verb.

16. In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb.
17. With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.
18. When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise.
19. When the general rule is that the descriptive adjective follows, there may be a minority of adjectives which usually precede, but when the general rule is that descriptive adjectives precede, there are no exceptions.
20. When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite.
21. If some or all adverbs follow the adjective they modify, then the language is one in which the qualifying adjective follows the noun and the verb precedes its nominal object as the dominant order.
22. If in comparisons of superiority the only order, or one of the alternative orders, is standard-marker-adjective, then the language is postpositional. With overwhelmingly more than chance frequency if the only order is adjective-marker-standard, the language is prepositional.
23. If in apposition the proper noun usually precedes the common noun, then the language is one in which the governing noun precedes its dependent genitive. With much better than chance frequency, if the common noun usually precedes the proper noun, the dependent genitive precedes its governing noun.
24. If the relative expression precedes the noun either as the only construction or as an alternate construction, either the language is postpositional, or the adjective precedes the noun or both.
25. If the pronominal object follows the verb, so does the nominal object.
26. If a language has discontinuous affixes, it always has either prefixing or suffixing or both.
27. If a language is exclusively suffixing, it is postpositional ; if it is exclusively prefixing, it is prepositional.
28. If both the derivation and inflection follow the root, or they both precede the root, the derivation is always between the root and the inflection.
29. If a language has inflection, it always has derivation.
30. If the verb has categories of person-number or if it has categories of gender, it always has tense-mode categories.
31. If either the subject or object noun agrees with the verb in gender, then the adjective always agrees with the noun in gender.
32. Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number.
33. When number agreement between the noun and verb is suspended and the rule is based on order, the case is always one in which the verb precedes and the verb is in the singular.

34. No language has a trial number unless it has a dual. No language has a dual unless it has a plural.
35. There is no language in which the plural does not have some nonzero allomorphs, whereas there are languages in which the singular is expressed only by zero. The dual and the trial are almost never expressed only by zero.
36. If a language has the category of gender, it always has the category of number.
37. A language never has more gender categories in nonsingular numbers than in the singular.
38. Where there is a case system, the only case which ever has only zero allomorphs is the one which includes among its meanings that of the subject of the intransitive verb.
39. Where morphemes of both number and case are present and both follow or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case.
40. When the adjective follows the noun, the adjective expresses all the inflectional categories of the noun. In such cases the noun may lack overt expression of one or all of these categories.
41. If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system.
42. All languages have pronominal categories involving at least three persons and two numbers.
43. If a language has gender categories in the noun, it has gender categories in the pronoun.
44. If a language has gender distinctions in the first person, it always has gender distinctions in the second or third person, or in both.
45. If there are any gender distinctions in the plural of the pronoun, there are some gender distinctions in the singular also.

Relations syntaxiques dans UD

— `acl` : clausal modifier of noun (adnominal clause)
— `acl:relcl` : relative clause modifier
— `advcl` : adverbial clause modifier
— `advmod` : adverbial modifier
— `advmod:emph` : emphasizing word, intensifier
— `advmod:lmod` : locative adverbial modifier
— `amod` : adjectival modifier
— `appos` : appositional modifier
— `aux` : auxiliary
— `aux:pass` : passive auxiliary
— `case` : case marking
— `cc` : coordinating conjunction
— `cc:preconj` : preconjunct
— `ccomp` : clausal complement
— `clf` : classifier
— `compound` : compound
— `compound:lvc` : light verb construction
— `compound:prt` : phrasal verb particle
— `compound:redup` : reduplicated compounds
— `compound:svc` : serial verb compounds
— `conj` : conjunct
— `cop` : copula
— `csubj` : clausal subject
— `csubj:pass` : clausal passive subject
— `dep` : unspecified dependency
— `det` : determiner
— `det:numgov` : pronominal quantifier governing the case of the noun
— `det:nummod` : pronominal quantifier agreeing in case with the noun
— `det:poss` : possessive determiner
— `discourse` : discourse element
— `dislocated` : dislocated elements
— `expl` : expletive
— `expl:impers` : impersonal expletive
— `expl:pass` : reflexive pronoun used in reflexive passive
— `expl:pv` : reflexive clitic with an inherently reflexive verb
— `fixed` : fixed multiword expression
— `flat` : flat multiword expression
— `flat:foreign` : foreign words
— `flat:name` : names
— `goeswith` : goes with
— `iobj` : indirect object
— `list` : list
— `mark` : marker
— `nmod` : nominal modifier
— `nmod:poss` : possessive nominal modifier
— `nmod:tmod` : temporal modifier

— `nsubj` : nominal subject
— `nsubj:pass` : passive nominal subject
— `nummod` : numeric modifier
— `nummod:gov` : numeric modifier governing the case of the noun
— `obj` : object
— `obl` : oblique nominal
— `obl:agent` : agent modifier
— `obl:arg` : oblique argument
— `obl:lmod` : locative modifier
— `obl:tmod` : temporal modifier
— `orphan` : orphan
— `parataxis` : parataxis
— `punct` : punctuation
— `reparandum` : overridden disfluency
— `root` : root
— `vocative` : vocative
— `xcomp` : open clausal complement

Traits morphologiques universels dans UD

— `Abbr` : abbreviation
— `AdpType` : adposition type
— `AdvType` : adverb type
— `Animacy` : animacy
— `Aspect` : aspect
— `Case` : case
— `Clusivity` : clusivity
— `Clusivity[obj]` : clusivity agreement with object
— `Clusivity[psor]` : possessor's clusivity
— `Clusivity[subj]` : clusivity agreement with subject
— `ConjType` : conjunction type
— `Definite` : definiteness or state
— `Degree` : degree of comparison
— `Deixis` : relative location encoded in demonstratives
— `DeixisRef` : person to which deixis is relative
— `Echo` : is this an echo word or a reduplicative ?
— `Evident` : evidentiality
— `Foreign` : is this a foreign word ?
— `Gender` : gender
— `Gender[dat]` : gender agreement with the dative argument
— `Gender[erg]` : gender agreement with the ergative argument
— `Gender[obj]` : gender agreement with object
— `Gender[psor]` : possessor's gender
— `Gender[subj]` : gender agreement with subject
— `Hyph` : hyphenated compound or part of it
— `Mood` : mood
— `NameType` : type of named entity
— `NounClass` : noun class
— `NounType` : noun type
— `NumForm` : numeral form
— `NumType` : numeral type
— `NumValue` : numeric value
— `Number` : number
— `Number[abs]` : number agreement with absolutive argument
— `Number[dat]` : number agreement with dative argument
— `Number[erg]` : number agreement with ergative argument
— `Number[obj]` : number agreement with object
— `Number[psed]` : possessed object's number
— `Number[psor]` : possessor's number
— `Number[subj]` : number agreement with subject
— `PartType` : particle type
— `Person` : person
— `Person[abs]` : person agreement with the absolutive argument
— `Person[dat]` : person agreement with the dative argument
— `Person[erg]` : person agreement with the ergative argument
— `Person[obj]` : person agreement with object

— `Person[psor]` : possessor's person
— `Person[subj]` : person agreement with subject
— `Polarity` : polarity
— `Polite` : politeness
— `Polite[abs]` : politeness agreement with absolutive argument
— `Polite[dat]` : politeness agreement with dative argument
— `Polite[erg]` : politeness agreement with ergative argument
— `Poss` : possessive
— `PrepCase` : case form sensitive to prepositions
— `PronType` : pronominal type
— `PunctSide` : which side of paired punctuation is this ?
— `PunctType` : punctuation type
— `Reflex` : reflexive
— `Style` : style or sublanguage to which this word form belongs
— `Subcat` : subcategorization
— `Tense` : tense
— `Typo` : is this a misspelled word ?
— `VerbForm` : form of verb or deverbative
— `VerbType` : verb type
— `Voice` : voice

Classification des corpus selon les caractéristiques testées

Les tableaux suivants présentent les valeurs de corpus pour trois caractéristiques :
— l'ordre sujet (S) - verbe (V) - objet (O),
— la présence de prépositions (Pr) ou postpositions (Po),
— l'ordre adjectif - nom (A) ou nom - adjectif (N).

Corpus	S – V – O	Pr/Po	A/N
Afrikaans-AfriBooms	SOV	Pr	A
Akkadian-RIAO	NDO (SOV/OSV)	Pr	N
Amharic-ATT	NDO (OVS/SOV)	Pr	A
Ancient_Greek-Perseus	NDO (SOV/SVO)	Pr	NDO (A/N)
Ancient_Greek-PROIEL	NDO (SVO/SOV)	Pr	NDO (A/N)
Arabic-NYUAD	VSO	Pr	N
Arabic-PADT	NDO (SVO/VSO)	Pr	N
Arabic-PUD	NDO (VSO/SVO)	Pr	N
Armenian-ArmTDP	NDO (SVO/SOV)	Po	A
Bambara-CRB	SOV	Po	N
Basque-BDT	SOV	Po	N
Belarusian-HSE	SVO	Pr	A
Bulgarian-BTB	SVO	Pr	A
Cantonese-HK	SVO	Pr	A
Catalan-AnCora	SVO	Pr	N
Chinese-GSD	SVO	Pr	A
Chinese-GSDSimp	SVO	Pr	A
Chinese-HK	SVO	Pr	A
Chinese-PUD	SVO	NDO (Pr/Po)	A
Chukchi-HSE	NDO (SVO/SOV)	Po	A
Classical_Chinese-Kyoto	SVO	Pr	A
Coptic-Scriptorium	SVO	Pr	N
Croatian-SET	SVO	Pr	A
Czech-CAC	SVO	Pr	A
Czech-CLTT	SVO	Pr	A
Czech-FicTree	SVO	Pr	A
Czech-PDT	SVO	Pr	A
Czech-PUD	SVO	Pr	A
Danish-DDT	SVO	Pr	A
Dutch-Alpino	NDO (SOV/SVO)	Pr	A
Dutch-LassySmall	NDO (SVO/SOV)	Pr	A
English-ESL	SVO	Pr	A
English-EWT	SVO	Pr	A
English-GUM	SVO	Pr	A
English-LinES	SVO	Pr	A
English-ParTUT	SVO	Pr	A
English-PUD	SVO	Pr	A
Erzya-JR	SVO	Po	A
Estonian-EDT	SVO	Po	A
Estonian-EWT	SVO	Po	A

Corpus	S – V – O	Pr/Po	A/N
Faroese-FarPaHC	SVO	Pr	A
Faroese-OFT	SVO	Pr	A
Finnish-FTB	SVO	Po	A
Finnish-OOD	SVO	Po	A
Finnish-PUD	SVO	Po	A
Finnish-TDT	SVO	Po	A
French-FQB	SVO	Pr	N
French-FTB	SVO	Pr	N
French-GSD	SVO	Pr	N
French-ParTUT	SVO	Pr	N
French-PUD	SVO	Pr	N
French-Sequoia	SVO	Pr	N
French-Spoken	SVO	Pr	NDO (A/N)
Galician-CTG	SVO	Pr	N
Galician-TreeGal	SVO	Pr	N
German-GSD	NDO (SOV/SVO)	Pr	A
German-HDT	NDO (SOV/SVO)	Pr	A
German-LIT	SOV	Pr	A
German-PUD	NDO (SOV/SVO)	Pr	A
Gothic-PROIEL	SVO	Pr	NDO (A/N)
Greek-GDT	SVO	Pr	A
Hebrew-HTB	SVO	Pr	N
Hindi_English-HIENCS	NDO (SVO/SOV)	Po	A
Hindi-HDTB	SOV	Po	A
Hindi-PUD	SOV	Po	A
Hungarian-Szeged	NDO (SVO/SOV)	Po	A
Icelandic-IcePaHC	SVO	Pr	A
Icelandic-PUD	SVO	Pr	A
Indonesian-CSUI	SVO	Pr	N
Indonesian-GSD	SVO	Pr	N
Indonesian-PUD	SVO	Pr	N
Irish-IDT	VSO	Pr	N
Italian-ISDT	SVO	Pr	N
Italian-ParTUT	SVO	Pr	NDO (A/N)
Italian-PoSTWITA	SVO	Pr	NDO (A/N)
Italian-PUD	SVO	Pr	N
Italian-TWITTIRO	SVO	Pr	NDO (N/A)
Italian-VIT	SVO	Pr	N
Japanese-BCCWJ	SOV	Po	A
Japanese-GSD	SOV	Po	A
Japanese-PUD	SOV	Po	A
Kazakh-KTB	SOV	Po	A
Korean-GSD	SOV	Po	A
Korean-Kaist	SOV	Po	A
Korean-PUD	SOV	NDO (0)	A
Latin-ITTB	NDO (SVO/SOV)	Pr	NDO (A/N)
Latin-LLCT	NDO (OSV/SVO)	Pr	NDO (N/A)
Latin-Perseus	SOV	Pr	NDO (N/A)
Latin-PROIEL	NDO (SOV/SVO)	Pr	NDO (N/A)
Latvian-LVTB	SVO	Pr	A
Lithuanian-ALKSNIS	SVO	Pr	A

Corpus	S – V – O	Pr/Po	A/N
Maltese-MUDT	SVO	Pr	N
Mbya_Guarani-Dooley	NDO (SVO/SOV)	Po	N
Naija-NSC	SVO	Pr	A
North_Sami-Giella	SVO	Po	A
Norwegian-Bokmaal	SVO	Pr	A
Norwegian-Nynorsk	SVO	Pr	A
Norwegian-NynorskLIA	SVO	Pr	A
Old_Church_Slavonic-PROIEL	SVO	Pr	N
Old_French-SRCMF	NDO (SOV/SVO)	Pr	A
Old_Russian-TOROT	SVO	Pr	NDO (N/A)
Persian-PerDT	SOV	Pr	N
Persian-Seraji	SOV	Pr	N
Polish-LFG	SVO	Pr	A
Polish-PDB	SVO	Pr	NDO (N/A)
Polish-PUD	SVO	Pr	NDO (N/A)
Portuguese-Bosque	SVO	Pr	N
Portuguese-GSD	SVO	Pr	N
Portuguese-PUD	SVO	Pr	N
Romanian-Nonstandard	NDO (SVO/SOV)	Pr	N
Romanian-RRT	SVO	Pr	N
Romanian-SiMoNERo	SVO	Pr	N
Russian-GSD	SVO	Pr	A
Russian-PUD	SVO	Pr	A
Russian-SynTagRus	SVO	Pr	A
Russian-Taiga	SVO	Pr	A
Sanskrit-Vedic	SOV	NDO (0)	A
Scottish_Gaelic-ARCOSG	VSO	Pr	N
Serbian-SET	SVO	Pr	A
Slovak-SNK	SVO	Pr	A
Slovenian-SSJ	SVO	Pr	A
Slovenian-SST	SVO	Pr	A
Spanish-AnCora	SVO	Pr	N
Spanish-GSD	SVO	Pr	N
Spanish-PUD	SVO	Pr	N
Swedish-LinES	SVO	Pr	A
Swedish-PUD	SVO	Pr	A
Swedish-Talbanken	SVO	Pr	A
Telugu-MTG	SOV	Po	A
Thai-PUD	SVO	Pr	N
Turkish_German-SAGT	NDO (SOV/SVO)	Pr	A
Turkish-BOUN	SOV	Po	A
Turkish-GB	SOV	Po	A
Turkish-IMST	SOV	Po	A
Turkish-PUD	SOV	Po	A
Ukrainian-IU	SVO	Pr	A
Urdu-UDTB	SOV	Po	A
Uyghur-UDT	SOV	Po	A
Vietnamese-VTB	SVO	Pr	N
Welsh-CCG	VSO	Pr	N
Wolof-WTB	SVO	Pr	N

UD 2.7_{1K}

Corpus	langue	famille de langue	# sentences	genre
Afrikaans-AfriBooms	Afrikaans	IE,Germanic	1934	legal nonfiction
Akkadian-RIAO	Akkadian	Afro-Asiatic,Semitic	1799	nonfiction government
Amharic-ATT	Amharic	Afro-Asiatic,Semitic	1074	grammar-examples fiction nonfiction bible news
Ancient_Greek-Perseus	Ancient Greek	IE,Greek	13919	fiction
Ancient_Greek-PROIEL	Ancient Greek	IE,Greek	17080	bible nonfiction
Arabic-NYUAD	Arabic	Afro-Asiatic,Semitic	19738	news
Arabic-PADT	Arabic	Afro-Asiatic,Semitic	7664	news
Arabic-PUD	Arabic	Afro-Asiatic,Semitic	1000	news wiki
Armenian-ArmTDP	Armenian	IE,Armenian	2502	blog fiction grammar-examples legal news nonfiction
Bambara-CRB	Bambara	Mande	1026	nonfiction news
Basque-BDT	Basque	Basque	8993	news
Belarusian-HSE	Belarusian	IE,Slavic	23534	fiction legal news nonfiction
Bulgarian-BTB	Bulgarian	IE,Slavic	11138	news legal fiction
Cantonese-HK	Cantonese	Sino-Tibetan	1004	spoken
Catalan-AnCora	Catalan	IE,Romance	16678	news
Chinese-GSD	Chinese	Sino-Tibetan	4997	wiki
Chinese-GSDSimp	Chinese	Sino-Tibetan	4997	wiki
Chinese-HK	Chinese	Sino-Tibetan	1004	spoken
Chinese-PUD	Chinese	Sino-Tibetan	1000	news wiki
Chukchi-HSE	Chukchi	Chukotko-Kamchatkan	1004	spoken
Classical_Chinese-Kyoto	Classical Chinese	Sino-Tibetan	48434	nonfiction
Coptic-Scriptorium	Coptic	Afro-Asiatic,Egyptian	1873	bible fiction nonfiction
Croatian-SET	Croatian	IE,Slavic	9010	news web wiki
Czech-CAC	Czech	IE,Slavic	24709	news nonfiction legal reviews medical
Czech-CLTT	Czech	IE,Slavic	1125	legal
Czech-FicTree	Czech	IE,Slavic	12760	fiction
Czech-PDT	Czech	IE,Slavic	87913	news reviews nonfiction
Czech-PUD	Czech	IE,Slavic	1000	news wiki
Danish-DDT	Danish	IE,Germanic	5512	news fiction spoken nonfiction
Dutch-Alpino	Dutch	IE,Germanic	13578	news
Dutch-LassySmall	Dutch	IE,Germanic	7338	wiki
English-ESL	English	IE,Germanic	5124	Learners-essays
English-EWT	English	IE,Germanic	16622	blog social reviews email
English-GUM	English	IE,Germanic	5961	academic fiction nonfiction news spoken web wiki
English-LinES	English	IE,Germanic	5243	fiction nonfiction spoken
English-ParTUT	English	IE,Germanic	2090	Legal news wiki
English-PUD	English	IE,Germanic	1000	news wiki
Erzya-JR	Erzya	Uralic,Mordvin	1690	fiction
Estonian-EDT	Estonian	Uralic,Finnic	30972	fiction news nonfiction academic
Estonian-EWT	Estonian	Uralic,Finnic	4493	blog web social
Faroese-FarPaHC	Faroese	IE,Germanic	1621	fiction bible nonfiction
Faroese-OFT	Faroese	IE,Germanic	1208	wiki
Finnish-FTB	Finnish	Uralic,Finnic	18723	Grammar-examples
Finnish-OOD	Finnish	Uralic,Finnic	2117	medical web social poetry
Finnish-PUD	Finnish	Uralic,Finnic	1000	news wiki
Finnish-TDT	Finnish	Uralic,Finnic	15136	news wiki blog legal fiction grammar-examples
French-FQB	French	IE,Romance	2289	nonfiction news
French-FTB	French	IE,Romance	18535	news
French-GSD	French	IE,Romance	16341	blog news reviews wiki
French-ParTUT	French	IE,Romance	1020	legal news wiki
French-PUD	French	IE,Romance	1000	news wiki
French-Sequoia	French	IE,Romance	3099	wiki medical news nonfiction
French-Spoken	French	IE,Romance	2806	spoken
Galician-CTG	Galician	IE,Romance	3993	medical legal nonfiction news
Galician-TreeGal	Galician	IE,Romance	1000	news
German-GSD	German	IE,Germanic	15590	news reviews wiki
German-HDT	German	IE,Germanic	189928	news nonfiction web
German-LIT	German	IE,Germanic	1922	nonfiction
German-PUD	German	IE,Germanic	1000	news wiki
Gothic-PROIEL	Gothic	IE,Germanic	5401	bible
Greek-GDT	Greek	IE,Greek	2521	news wiki spoken
Hebrew-HTB	Hebrew	Afro-Asiatic,Semitic	6216	news
Hindi_English-HIENCS	Hindi English	Code-switching	1898	social
Hindi-HDTB	Hindi	IE,Indic	16647	news
Hindi-PUD	Hindi	IE,Indic	1000	news wiki
Hungarian-Szeged	Hungarian	Uralic,Ugric	1800	news
Icelandic-IcePaHC	Icelandic	IE,Germanic	44029	fiction bible nonfiction legal
Icelandic-PUD	Icelandic	IE,Germanic	1000	news wiki
Indonesian-CSUI	Indonesian	Austronesian,Malayo-Sumbawan	1030	nonfiction news
Indonesian-GSD	Indonesian	Austronesian,Malayo-Sumbawan	5593	news blog
Indonesian-PUD	Indonesian	Austronesian,Malayo-Sumbawan	1000	news wiki
Irish-IDT	Irish	IE,Celtic	4910	news fiction web legal government
Italian-ISDT	Italian	IE,Romance	14167	legal news wiki
Italian-ParTUT	Italian	IE,Romance	2090	legal news wiki
Italian-PoSTWITA	Italian	IE,Romance	6713	social
Italian-PUD	Italian	IE,Romance	1000	news wiki
Italian-TWITTIRO	Italian	IE,Romance	1424	soical
Italian-VIT	Italian	IE,Romance	10087	nonfiction news
Japanese-BCCWJ	Japanese	Japanese	57028	news nonfiction fiction blog web
Japanese-GSD	Japanese	Japanese	8071	news blog
Japanese-PUD	Japanese	Japanese	1000	news wiki
Kazakh-KTB	Kazakh	Turkic,Northwestern	1078	wiki fiction news
Korean-GSD	Korean	Korean	6339	news blog
Korean-Kaist	Korean	Korean	27363	news fiction academic
Korean-PUD	Korean	Korean	1000	news wiki
Latin-ITTB	Latin	IE,Latin	26977	nonfiction
Latin-LLCT	Latin	IE,Latin	9023	nonfiction legal
Latin-Perseus	Latin	IE,Latin	2273	fiction nonfiction bible
Latin-PROIEL	Latin	IE,Latin	18411	bible nonfiction
Latvian-LVTB	Latvian	IE,Baltic	13643	news fiction legal spoken academic
Lithuanian-ALKSNIS	Lithuanian	IE,Baltic	3642	news nonfiction legal fiction
Maltese-MUDT	Maltese	Afro-Asiatic,Semitic	2074	news legal nonfiction fiction wiki
Mbya_Guarani-Dooley	Mbya Guarani	Tupian,Tupi-Guarani	1046	fiction
Naija-NSC	Naija	Creole	9242	nonfiction

Corpus	langue	famille de langue	# sentences	genre
North_Sami-Giella	North Sami	Uralic,Sami	3122	nonfiction news
Norwegian-Bokmaal	Norwegian	IE,Germanic	20044	news blog nonfiction
Norwegian-Nynorsk	Norwegian	IE,Germanic	17575	news blog nonfiction
Norwegian-NynorskLIA	Norwegian	IE,Germanic	5250	spoken
Old_Church_Slavonic-PROIEL	Old Church Slavonic	IE,Slavic	6338	bible
Old_French-SRCMF	Old French	IE,Romance	17678	nonfiction legal poetry
Old_Russian-TOROT	Old Russian	IE,Slavic	16944	nonfiction legal
Persian-PerDT	Persian	IE,Iranian	29107	news fiction nonfiction academic web blog
Persian-Seraji	Persian	IE,Iranian	5997	news fiction medical legal social spoken nonfiction
Polish-LFG	Polish	IE,Slavic	17246	fiction nonfiction news spoken social
Polish-PDB	Polish	IE,Slavic	22152	fiction nonfiction news
Polish-PUD	Polish	IE,Slavic	1000	news wiki
Portuguese-Bosque	Portuguese	IE,Romance	9364	news blog
Portuguese-GSD	Portuguese	IE,Romance	12078	blog news
Portuguese-PUD	Portuguese	IE,Romance	1000	news wiki
Romanian-Nonstandard	Romanian	IE,Romance	26224	bible poetry
Romanian-RRT	Romanian	IE,Romance	9524	wiki legal news fiction medical nonfiction academic
Romanian-SiMoNERo	Romanian	IE,Romance	4681	medical
Russian-GSD	Russian	IE,Slavic	5030	wiki
Russian-PUD	Russian	IE,Slavic	1000	news wiki
Russian-SynTagRus	Russian	IE,Slavic	61889	news nonfiction fiction
Russian-Taiga	Russian	IE,Slavic	4964	blog fiction news poetry social wiki
Sanskrit-Vedic	Sanskrit	IE,Indic	3997	nonfiction
Scottish_Gaelic-ARCOSG	Scottish Gaelic	IE,Celtic	3173	nonfiction fiction news spoken
Serbian-SET	Serbian	IE,Slavic	4384	news
Slovak-SNK	Slovak	IE,Slavic	10604	fiction nonfiction news
Slovenian-SSJ	Slovenian	IE,Slavic	8000	news nonfiction fiction
Slovenian-SST	Slovenian	IE,Slavic	3188	spoken
Spanish-AnCora	Spanish	IE,Romance	17680	news
Spanish-GSD	Spanish	IE,Romance	16013	blog news reviews wiki
Spanish-PUD	Spanish	IE,Romance	1000	news wiki
Swedish-LinES	Swedish	IE,Germanic	5243	fiction nonfiction spoken
Swedish-PUD	Swedish	IE,Germanic	1000	news wiki
Swedish-Talbanken	Swedish	IE,Germanic	6026	news nonfiction
Telugu-MTG	Telugu	Dravidian,South-Central	1328	Grammar-examples
Thai-PUD	Thai	Tai-Kadai	1000	news wiki
Turkish_German-SAGT	Turkish German	Code-switching	1891	spoken
Turkish-BOUN	Turkish	Turkic,Southwestern	9761	nonfiction news
Turkish-GB	Turkish	Turkic,Southwestern	2880	Grammar-examples
Turkish-IMST	Turkish	Turkic,Southwestern	5635	nonfiction news
Turkish-PUD	Turkish	Turkic,Southwestern	1000	news wiki
Ukrainian-IU	Ukrainian	IE,Slavic	7060	blog email fiction grammar-examples legal news reviews social web wiki
Urdu-UDTB	Urdu	IE,Indic	5130	news
Uyghur-UDT	Uyghur	Turkic,Southeastern	3456	fiction
Vietnamese-VTB	Vietnamese	Austro-Asiatic,Viet-Muong	3000	news
Welsh-CCG	Welsh	IE,Celtic	1657	Grammar-examples wiki nonfiction fiction news
Wolof-WTB	Wolof	Niger-Congo,Northern-Atlantic	2107	bible wiki

Bibliographie

Alzetta, C., Dell'Orletta, F., Montemagni, S., and Venturi, G. (2018). Universal Dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon. European Language Resources Association (ELRA).

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4 :431–444.

Bender, E. M. (2009). Linguistically naïve != language independent : Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, pages 26–32, Athène, Grèce. Association for Computational Linguistics.

Berdicevskis, A. and Piperski, A. (2020). Corpus evidence for word order freezing in russian and german. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 26–33, Barcelone, Espagne (en ligne).

Bickel, B. (2007). Typology in the 21st century : Major current developments. *Linguistic Typology*, 11 :239–251.

Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*, volume 1 of *Logic, Linguistics and Computer Science Set.* ISTE Wiley.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York, USA. Association for Computational Linguistics.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

Chen, X. and Gerdes, K. (2017). Classifying languages by dependency structure. typologies of delexicalized Universal Dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 54–63, Pise, Italie. Linköping University Electronic Press.

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.

Chomsky, N. (1957). *Éléments de syntaxe structurale.* Mouton, The Hague.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding / Noam Chomsky*. MIT Press Cambridge, Mass.

Comrie, B. (1989). *Language universals and Typology : Syntax and Morphology*. University of Chicago Press.

Cooreman, A. and Goyvaerts, D. (1980). Universals in human language. a historical perspective. *Revue Belge de Philologie Et D'Histoire*, 58(3) :615–638.

Croft, W. (2003). *Typology and Universals*. Cambridge University Press, New York.

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., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Gène, Italie. European Language Resources Association (ELRA).

Dryer, M. S. (2013a). 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. (2013b). Order of adjective and noun. 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. (2013c). Order of adposition and noun phrase. 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. (2013d). Order of genitive and noun. 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.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Suède. Uppsala University, Uppsala, Suède.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*, Bruxelles, Belgique.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2019). Improving Surface-syntactic Universal Dependencies (SUD) : surface-syntactic relations and deep syntactic features. In *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, Paris, France.

Gerdes, K., Kahane, S., and Chen, X. (2021). Typometrics from implicational to quantitative universals in word order typology. *Glossa : a journal of general linguistics*, 6.

Greenberg, J. H. (1966a). 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.

Greenberg, J. H., editor (1966b). *Universals of Human Language*. MIT Press, Cambridge, Mass.

Greenberg, J. H., Osgood, C. E., and Jenkins, J. J. (1966). Memorandum concerning language universals. In Greenberg, J. H., editor, *Universals of Human Language*, pages xv–xxviii. MIT Press, Cambridge, Mass.

Guillaume, B. (2021). Graph Matching and Graph Rewriting : GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*, Kiev/En ligne, Ukraine.

Guillaume, B., de Marneffe, M.-C., and Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL*, 60(2) :71–95.

Guillaume, B. and Perrier, G. (2012). Annotation sémantique du French treebank à l'aide de la réécriture modulaire de graphes (semantic annotation of the French treebank using modular graph rewriting) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 293–306, Grenoble, France. ATALA/AFCP.

Jakobson, R. (1966). Implications of language universals for linguistics. In Greenberg, J. H., editor, *Universals of Human Language*, pages 263–278. MIT Press, Cambridge, Mass.

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, En ligne. Association for Computational Linguistics.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1) :79–86.

Levshina, N. (2019). Token-based typology and word order entropy : A study based on universal dependencies. *Linguistic Typology*, 23(3) :533–572.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). Uriel and lang2vec : Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Liu, H. (2010). Dependency direction as a means of word-order typology : A method based on dependency treebanks. *Lingua*, 120(6) :1567–1578. Contrast as an information-structural notion in grammar.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 92–97, Sofia, Bulgarie. Association for Computational Linguistics.

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*, pages 629–637, Jeju, Corée. Association for Computational Linguistics.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago, 1 edition.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2 : An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (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)*, pages 915–932, Prague, République Tchèque. Association for Computational Linguistics.

O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., and Korhonen, A. (2016). Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 1297–1308, Osaka, Japon.

Östling, R. (2014). Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2 : Short Papers*, pages 123–127, Göteborg, Suède. Association for Computational Linguistics.

Östling, R. (2015). Word order typology through multilingual word alignment. 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)*, pages 205–211, Pékin, Chine. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turquie. European Language Resources Association (ELRA).

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., and Korhonen, A. (2019). Modeling Language Variation and Universals : A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3) :559–601.

Scholivet, M., Dary, F., Nasr, A., Favre, B., and Ramisch, C. (2019). Typological features for multilingual delexicalised dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3919–3930, Minneapolis, Minnesota. Association for Computational Linguistics.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3) :379–423.

Sulubacak, U., Gokirmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., and Eryiğit, G. (2016). Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 3444–3454, Osaka, Japon. The COLING 2016 Organizing Committee.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Librairie C. Klincksieck, Paris.

Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., Black, A. W., Levin, L., and Dyer, C. (2016). Polyglot neural language models : A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1357–1366, San Diego, USA. Association for Computational Linguistics.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Maroc.