Université Paris III - Sorbonne Nouvelle
Master de Sciences du Langages, spécialité Ingénierie Linguistique

Mesures de distances syntaxiques entre langues à partir de treebanks

Marine Courtin

Mémoire dirigé par Sylvain Kahane et Kim Gerdes

Année universitaire 2017-2018

Attestation de non-plagiat

Déclaration sur l’honneur

Je, soussigné (e), déclare avoir rédigé ce travail sans aides extérieures ni sources autres que celles qui sont citées. Toutes les utilisations de textes préexistants, publiés ou non, y compris en version électronique, sont signalées comme telles. Ce travail n’a été soumis à aucun autre jury d’examen sous une forme identique ou similaire, que ce soit en France ou à l’étranger, à l’université ou dans une autre institution, par moi-même ou par autrui.

Date

Signature manuscrite de l’étudiant·e

Abstract

Ce mémoire décrit un travail de recherche qui vise à proposer des mesures de distances syntaxiques entre langues. Nous utilisons des corpus annotés en syntaxe, aussi appelés treebanks, constitués dans le cadre du projet Universal Dependencies. À partir de ces corpus annotés en syntaxe de dépendance, nous extrayons des informations sur la distribution des annotations (parties du discours, dépendances syntaxiques, motifs syntaxiques...). Puis, nous proposons différentes méthodes permettant d'exploiter ces mesures afin de modéliser les relations de distances et de similarités syntaxiques entre les langues de ces corpus. Ainsi, les questions qui motivent ce travail sont les suivantes: les treebanks sont-ils des modèles de langues intéressants pour dégager des connaissances typologiques? Comment mesurer des distances entre langues à partir d'indices syntaxiques? À quels types de connaissances les distances ainsi obtenues nous permettent-elles d'accéder ? Comment le Traitement Automatique des Langues peut-il mettre à profit des mesures qui rendent compte de proximités syntaxiques pour développer de nouveaux systèmes ?

Table des Matières

Liste des Tables

Liste des Illustrations

Chapitre 1

Introduction

1.1 Présentation

Ce mémoire décrit un travail de recherche qui vise à proposer des mesures de distances syntaxiques entre langues. Pour ce faire, nous utilisons des corpus annotés en syntaxe aussi appelés *treebanks* à partir desquels nous extrayons des informations jugées pertinentes pour la description de la syntaxe d'une langue. Nous proposons ensuite différentes méthodes permettant d'exploiter ces mesures afin de modéliser les relations de distance et de similarité entre les langues de notre corpus.

Ainsi, les questions qui motivent ce travail sont les suivantes : les treebanks sont-ils des modèles de langues intéressants pour dégager des connaissances typologiques ? Comment mesurer des distances entre langues à partir d'indices syntaxiques ? À quels types de connaissances les distances ainsi obtenues nous permettent-elles d'accéder ?

1.2 Motivations

1.2.1 Description typologique

La typologie est une branche de la linguistique qui s'intéresse à décrire l'amplitude des variations structurelles observables entre les langues naturelles, ainsi que les limites posées sur ces variations [Comrie, 1981]. Une fois cet espace de variation délimité, il devient alors possible d'envisager une

classification des langues en différents *types* qui possèdent des caractéristiques communes. Dans ce mémoire, nous nous intéressons uniquement aux variations de nature syntaxique, mais il serait également possible de considérer la phonétique, la phonologie, la morphologie, la sémantique et la pragmatique ainsi que les interactions entre ces différents niveaux d'analyse. Étudier les comportement des langues selon certains paramètres syntaxiques nous permettra de rendre compte de similarités et de différences entre langues, que nous proposons de formaliser par des mesures de distances.

Le travail présenté dans ce mémoire fait interagir trois domaines : la typologie linguistique, la linguistique de corpus et le TAL (Traitement Automatique des Langues). Historiquement, la typologie s'est appuyée sur des exemples proposés par des linguistes spécialistes d'une langue ou d'un groupe de langues apparentées. Ainsi, les descriptions typologiques n'étaient pas basées sur des données langagières authentiques mais sur l'intuition des linguistes, et les descriptions proposées sur des variables discrètes, c'est-à-dire avec un nombre de catégories fini et connu d'avance, ne rendant pas toujours compte des nuances internes à chaque langue . Nous pensons que les techniques de TAL ont beaucoup à apporter à la typologie. En effet, des corpus de plus en plus larges, annotés sur différents niveaux d'analyses, et disponibles pour des langues plus variées sont désormais disponibles. Afin de vérifier empiriquement la validité de certaines hypothèses typologiques et d'en proposer de nouvelles, il pourrait être bénéfique de réutiliser les ressources et outils développés pour le TAL, assurant ainsi une meilleure répartition des coûts et davantage d'interactions entre les deux disciplines.

1.2.2 Applications pour le TAL

Du point de vue du TAL, la typologie est un domaine encore assez peu étudié. Quelques travaux se sont intéressés à normaliser et enrichir automatiquement des bases de connaissances typologiques [Littell et al., 2017], participant ainsi à la description typologique de langues pour lesquelles les informations n'étaient pas toujours disponibles.

Plutôt que les recherches sur la description typologique des langues, la majorité des travaux de TAL intégrant des connaissances typologiques semble concerner deux domaines en particulier : le parsing et la traduction automatique . En effet, un intérêt croissant pour le développement d'applications multilingues, et la recherche d'une plus grande indépendance des modèles vis-à-vis

de la langue des textes ont eu une influence majeure sur ces deux domaines [O'Horan et al., 2016].

Dans le domaine du parsing, des initiatives comme UD (Universal Dependencies) [Nivre et al., 2017] ont permis de rendre disponible des treebanks partageant un même schéma d'annotation, ce qui a considérablement facilité l'apprentissage de parsers multilingues. Des défis ou *shared-tasks* comme les CoNLL [Buchholz and Marsi, 2006a] [Nivre et al., 2007] sont régulièrement proposées afin de tester la robustesse de nouveaux outils, et les corpus d'apprentissage et de test incluent désormais un plus large panel de langues, présentant des caractéristiques typologiques variées (morphologie plus ou moins riche, variation de l'ordre des mots...) [Seddah et al., 2014]. Néanmoins, il existe un frein important au développement de ce type de parsers, puisque l'apprentissage de ceux-ci repose majoritairement sur la disponibilité de corpus annotés, corpus qui existent rarement pour les langues moins dotées. L'intégration de connaissances typologiques dans les architectures pourrait pallier le manque de ressources et faciliter l'adaptation des modèles appris sur des langues bien couvertes à de nouvelles langues [Bender, 2009].

Développer des modèles multilingues capables de davantage de généralisation constitue donc désormais un enjeu important pour le TAL. Dans ce contexte, il nous semble que le TAL pourrait s'appuyer sur la typologie afin d'exploiter au mieux des données cross-lingues dans le but de développer des applications multilingues plus robustes face aux variations structurelles présentes dans les langues naturelles.

1.3 Itinéraire

Le chapitre 2 introduira les notions clés du sujet et opérera un balayage sur des travaux antérieurs qui cherchent à positionner les langues selon leurs propriétés syntaxiques. Dans le chapitre 3, nous présenterons le corpus utilisé, ses propriétés, et le schéma d'annotation syntaxique d'Universal Dependencies. Nous tenterons également de dresser une typologie des objets et paramètres syntaxiques qu'il est possible d'extraire depuis des corpus arborés. Le chapitre 4 décrira la sélection des paramètres syntaxiques sur lesquels portera notre étude ainsi que les fonctions de distances sélectionnées. Enfin, dans le chapitre 5, nous nous arrêterons sur les résultats obtenus et discuterons des limites de notre travail.

Chapitre 2

Contextualisation du sujet

2.1 Notions-clés

Structure syntaxique La structure syntaxique est un objet qui explicite la façon dont des unités syntaxiques se combinent pour former un énoncé. Dans ce mémoire, nous étudions des structures d'une nature particulière : des arbres de dépendance syntaxique.

Dépendance syntaxique Une structure de dépendance syntaxique décrit des connexions qui se font directement entre les unités syntaxiques de l'énoncé. Elle ne nécessite donc pas d'introduire des nœuds additionnels dans la structure (contrairement à des structures en constituants par exemple (cf figure 2.3). Ces connexions sont **asymétriques**, c'est à dire que les deux unités en relation (**gouverneur** et **dépendant**) ne jouent pas le même rôle. Pour une introduction aux structures de dépendance voir [Mel'čuk, 1988] [Kahane, 2001]. Afin que le graphe de dépendance introduise une hiérarchie, il doit respecter un critère d'acyclicité orienté [Kahane and Gerdes, 2018], c'est-à-dire qu'il n'y a pas de cycles de dépendances (cf la figure 2.1). Pour le sous-type de graphes de dépendances que nous considérons plus particulièrement, à savoir des arbres de dépendances (cf figure 2.2), cette contrainte se renforce puisqu'aucun type de cycle (orienté ou non) n'est admis. Un arbre de dépendance n'est par défaut pas linéarisé, c'est-à-dire que l'ordre linéaire dans lequel les unités se combinent à l'intérieur de l'énoncé est considéré comme une structure en soi. Le rôle syntaxique joué par chaque unité vis-à-vis de son gouverneur peut être indiqué par typage des relations de dépendances comme sur la figure 2.4.

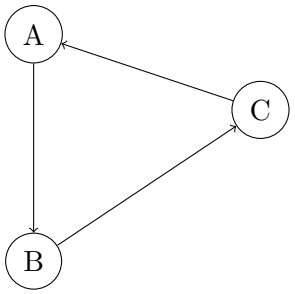


Figure 2.1: Illustration d'un cycle de dépendances

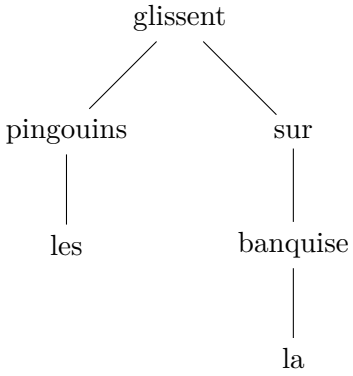


Figure 2.2: Arbre de dépendance syntaxique non-linéarisé et non-typé pour l'énoncé *les pingouins glissent sur la banquise*

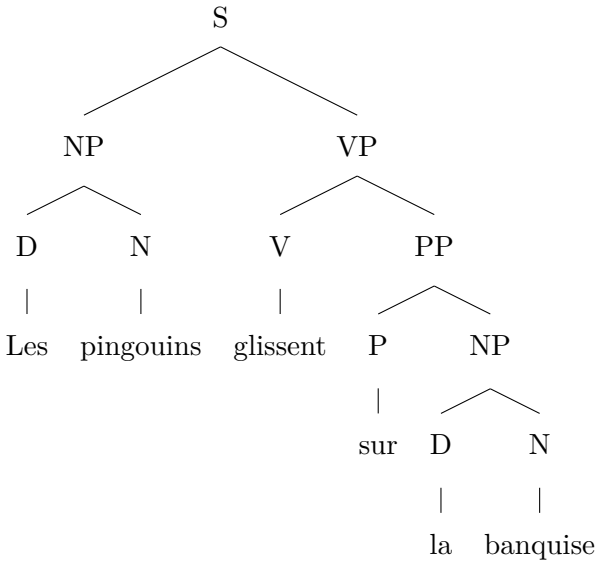


Figure 2.3: Arbre en constituants pour l'énoncé *les pingouins glissent sur la banquise*

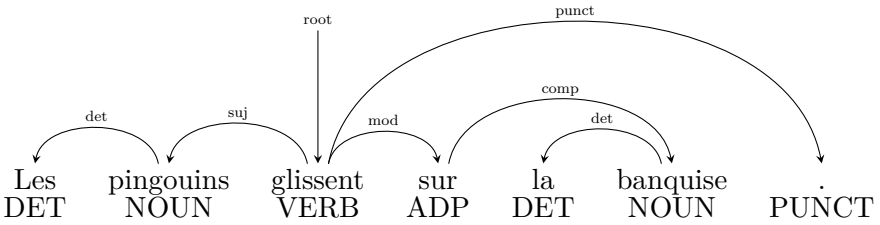


Figure 2.4: Arbre de dépendance syntaxique linéarisé, typé et avec ajouts d'étiquettes morpho-syntaxiques pour l'énoncé *les pingouins glissent sur la banquise*

Treebank Un treebank (ou corpus arboré) est un corpus enrichi d'une couche d'annotation qui vient expliciter la structure (syntaxique, sémantique, prosodique, discursive) des énoncés. Le choix du terme "treebank" peut s'expliquer par le fait que ces structures sont le plus souvent des arbres, mais d'autres types de structures (par exemples des graphes) sont également possibles. Par la suite, notre utilisation du terme treebank fera systématiquement référence aux treebanks annotés en syntaxe. Parmi les treebanks les plus connus, nous pouvons citer le Penn Treebank [Marcus et al., 1993] ou encore le Prague Dependency Treebank [Hajic et al., 2001], mais les premiers treebanks remonteraient au moins aux années 70 [Nivre, 2008].

Modèle de la syntaxe d'une langue Tout comme un modèle de langue est un modèle qui, à tout énoncé d'une langue, associe une probabilité (celle que cet énoncé soit réalisé parmi l'ensemble des énoncés possibles), un modèle de la syntaxe d'une langue associe à toute structure syntaxique d'une langue une probabilité. Ainsi, certaines structures auront des réalisations plus probables dans certaines langues que dans d'autres. En pratique, étudier les structures dans leur intégralité est complexe, puisque chaque structure a un nombre d'occurrences très faible. Pour pouvoir calculer des probabilités plus robustes, il est souvent nécessaire de découper chaque structure en configurations ou sous-structures afin d'étudier leur distribution. Nous dresserons dans la section 3.2.3.1 une typologie des structures partielles qu'il est possible d'extraire des treebanks.

Distance Le terme *distance* désigne à la fois une fonction de distance (ou métrique) utilisée pour mesurer la distance entre deux éléments appartenant à un même espace topologique, et la valeur obtenue par application de cette fonction aux deux éléments. Soit d une fonction de distance sur un ensemble X avec a, b et c des éléments de X, alors d présente les 3 propriétés suivantes :

- **transitivité** : $d(a, b) = d(b, a)$
- **inégalité triangulaire** : $d(a, c) \leq d(a, b) + d(b, c)$
- **identité des indiscernables** : $d(a, b) = 0 \iff a = b$. La distance entre deux objets n'est nulle que lorsque les propriétés des deux éléments à comparer ne sont pas discernables. Il est alors ontologiquement impossible de distinguer ces éléments l'un de l'autre.

Typologie La typologie est une branche de la linguistique qui s'intéresse à décrire l'amplitude des variations structurelles observables entre les langues naturelles, ainsi que les limites posées sur ces variations [Comrie, 1981]. Les typologues cherchent des régularités dans l'organisation interne des langues. Ces régularités permettent de dégager des types (possibles et attestés) dans lesquels classer les langues. Parmi les types possibles, certains semblent privilégiés puisqu'on observe plus fréquemment des exemples attestés de langues y appartenant. Cette amplitude de variations possible peut être représentée sous la forme d'un espace topologique, à l'intérieur duquel nous pourrions positionner les langues en fonction de certains critères. Par exemple dans un espace vectoriel multidimensionnel, nous pouvons assigner à chaque langue des coordonnées, et ces coordonnées permettraient d'observer de multiples proximités entre langues selon les axes considérés (où chaque axe représenterait un paramètre syntaxique). Ce qui est intéressant n'est donc pas nécessairement d'avoir une et une seule classification, mais de permettre de regarder comment se répartissent les langues dans cet espace (clusters (ou "amas") permettant de définir des types, répartition plus ou moins dense des langues, proximités inattendues etc...) en faisant varier les paramètres syntaxiques pris en compte.

2.2 Travaux précédents

Dans cette section, nous survolons rapidement des travaux qui se sont intéressés à des phénomènes syntaxiques dans le but de définir une typologie des langues. Nous verrons comment ces questionnements se sont renouvelés en intégrant des méthodes basées sur les corpus et sur des corpus annotés, notamment des treebanks. Enfin nous présenterons quelques exemples de travaux qui montrent que les interactions entre typologie et TAL peuvent être bénéfiques en proposant une validation empirique de théories typologiques, permettant d'accéder à de nouvelles connaissances typologiques basés sur des corpus (notamment annotés), ainsi qu'en facilitant l'extension

et le développement de systèmes de TAL moins naïfs, c'est-à-dire mieux informés quant à des propriétés linguistiques utiles pour les traitements qu'ils proposent, et plus opportunistes, car davantage à même d'utiliser des ressources et outils existants et de les adapter pour leurs besoins.

2.2.1 Vers une typologie quantitative

2.2.1.1 Classification en termes de proximité génétique

La typologie propose d'opérer une classification des langues en types, en fonction d'un certain nombre de propriétés présentées par ces langues. Une analogie souvent utilisée consiste à comparer la typologie à la classification scientifique des espèces.

Les premières classifications typologiques ont été proposées sur la base de similarités lexicales, ces similarités étant supposées être des traces d'une parenté commune. En lexico-statistique, des recherches ont porté sur l'étude des **cognats**, c'est-à-dire des mots qui sont phonétiquement (et/ou) orthographiquement proches dans plusieurs langues dû à une étymologie commune, et du degré de divergence entre langues basé sur la proportion de ces cognats partagés [Gudschinsky, 1956], afin de reconstruire une chronologie dans l'évolution des langues.

Ce type de typologie est davantage basé sur la reconnaissance d'une proximité génétique entre langues que sur le type de classification qui nous intéresse, c'est-à-dire une classification uniquement basée sur des similarités fonctionnelles.

2.2.1.2 Typologie moderne et similarité fonctionnelle

Universaux implicationnels La typologie moderne a été marquée par les travaux de Greenberg sur la définition d'universaux implicationnels [Greenberg, 1963] et de corrélations entre propriétés typologiques. Ces universaux sont formulés comme des propositions logiques : si une langue respecte la condition A alors elle respecte la condition B. Ainsi Greenberg propose l'universel suivant : "Les langues à ordre de mot dominant VSO sont toujours prépositionnelles." (nous traduisons) [1]

[1] "Languages with dominant VSO order are always prepositional."

Orientation statistique La plupart des corrélations que Greenberg décrit sont complexes et mettent en jeu plusieurs paramètres simultanément. En utilisant le terme de corrélations, Greenberg oriente dès le départ ses descriptions vers des analyses statistiques : il cherche à observer quelles propriétés corrèlent entre elles de manière significative sur des échantillons de langues données, censées être représentatives des langues en général, ce qui conférerait à ces proposition leur caractère "universel". En pratique, il s'avère assez difficile de prétendre décrire des universaux sur les langues car l'échantillonnage est particulièrement problématique. Les langues suffisamment décrites pour effectuer ce type de travaux ne représentent qu'une fraction des langues actuellement parlées (sans parler des langues qui ont été parlées par le passé, ou qui sont logiquement possibles). De plus, ces langues ont souvent en commun un certain nombre de caractéristiques qui biaisent les observations que l'on peut effectuer à leur sujet : les langues nationales sont surreprésentées, de même que certaines zones géographiques (en raison de conditions économiques, historiques et géo-politiques), et ces langues sont souvent parlées par un grand nombre de personnes, ce qui influe sur les types de structures qu'elles présentent [Raviv et al.,]. Dans ce contexte, il n'est pas raisonnable de considérer que l'échantillon peut être représentatif des "langues humainement possibles". De plus, même l'échantillon le plus aléatoire possible pourrait poser problème puisque les caractéristiques des langues sont en parties motivées par leur proximité génétique [Dryer, 1992] ainsi que par les contacts entre les locuteurs de ces langues, notamment lorsque les langues appartiennent à des zones géographiques proches [Dryer, 1989]. Il serait donc difficile d'établir si les caractéristiques que nous observons sont des caractéristiques des langues, où des groupes de langues auxquelles elles appartiennent.

Visée descriptive et non explicative Greenberg, cependant, ne cherche pas encore à proposer une explication cohérente sur les corrélations qu'il observe, son but premier est entièrement descriptif. Ce qui prime pour lui est la **validité empirique** plutôt que la **modélisation**, puisqu'il ne propose pas un système conceptuel cohérent qui viendrait expliquer les corrélations observées.

Hypothèses sur les tendances dans la structuration des langues

> "Le typologue constate souvent que dans un domaine particulier de la syntaxe, parmi des modes d'organisation a priori également envisageables, certains sont largement

> dominants à l'échelle des langues du monde, alors que d'autres ne sont attestés que de manière exceptionnelle, ou pas du tout."[Creissels, 2006]

À la suite de Greenberg, d'autres chercheurs ont également tenté de définir ce type d'universaux ou de tendances statistiques, mais en essayant cette fois-ci de les inscrire dans le cadre d'hypothèses plus générales sur la structuration des langues. C'est le cas de Vennemann notamment, pour qui la simple description des corrélations est une étape, qui doit permettre de décrire des principes d'organisation des langues

Opérateur-opérande [Vennemann, 1972] introduit la notion d'opérateur et d'opérande. En mathématiques, l'opérande est le terme sur lequel s'applique une opération, au moyen d'un ou plusieurs opérateurs. En linguistique, le terme d'opérande peut être utilisé afin de désigner un terme qui subit une transformation [2]. L'opérande serait donc l'unité qui est modifiée par l'opérateur, ce que nous pouvons reformuler en désignant l'opérande comme la tête, et l'opérateur comme un dépendant de cette tête. La figure 2.1 présente une partie de cette classification telle que la conçoit Vennemann :

Table 2.1: Tableau présentant une partie des répartitions opérateur-opérande proposées par Vennemann

catégorie de l'unité	opérateur	opérande
VP	verbe	auxiliaire
VP	objet	verbe
NP	adjectif	nom
NP	subordonnée relative	nom
NP	génitif	nom
PP	syntagme nominal	adposition

Principe de linéarisation naturelle Vennemann avance que les langues tendent soit à instancier les opérateurs avant leurs opérandes (tête-dépendant), soit à instancier les opérandes avant leurs opérateurs (dépendant-tête), ce qui correspond à son principe de linéarisation naturelle (Natural Serialization Principle). Les langues ne seraient pas toutes égales face au NSP,

[2] http://www.cnrtl.fr/definition/op%C3%A9rande

plus l'ordre relatif des opérateur-opérande à l'intérieur des constituants serait biaisé en faveur d'un type de linéarisation, plus la typologie de cette langue serait régulière ou harmonieuse.

Principe d'harmonie cross-catégorielle [Hawkins, 1983] défend une version modifié du principe précédent : le principe d'harmonie cross-catégorielle (CCH). Hawkins avance que les langues tendent à harmoniser la proportion de linéarisation tête-modifieur et modifieur-tête entre les catégories (NP, VP, AP, PP). Pour une langue deux catégories sont dites "harmoniques" l'une par rapport à l'autre lorsqu'on observe des proportions similaires entre modifieurs pré-posés et post-posés. Là encore, nous retrouvons l'idée de linéarisation "harmonieuse" ou régulière qui serait favorisée par les langues, c'est-à-dire que plus nous nous écarterions de cet "idéal" plus les langues se rarifieraient.

Variations selon les types de catégories Les principes que nous avons décrit ci-dessus sont des principes généraux, qui seraient plus ou moins respectés selon des paramètres divers. Hawkins souligne que les contraintes sur la linéarisation des modifieurs par rapport aux têtes sont plus fortes pour les langues qui ont strictement le verbe en position finale (comme le japonais par exemple).

Étudier la syntaxe et la typologie à partir des données

> "Si l'élaboration de théories formelles n'a de sens que dans la perspective d'une confrontation avec un éventail de langues aussi large que possible, inversement, le classement de données sur la diversité des structures syntaxiques n'a de sens que dans la mesure où il a comme perspective à plus ou moins long terme l'évaluation d'hypothèses sur les principes d'organisation communs à toutes les langues. Et il serait tout à fait illusoire de prétendre procéder à un pur classement de données qui n'impliquerait pas d'hypothèse préalable sur ce que peut être la structuration d'une langue." [Creissels, 2006]

Puisqu'il existe des régularités dans les types de structures observables dans les langues, il semble important de fonder la définition des unités, catégories et rôles syntaxiques en visant à permettre la description de la syntaxe d'une multitude de langues. Les projets d'annotations

multilingues comme Universal Dependencies sont un lieu intéressant pour développer ce type de travaux, puisqu'ils proposent des données empiriques et des schémas d'annotations communs qui facilitent les comparaisons multilingues, rapprochant ainsi typologie, syntaxe formelle et traitements automatiques basés sur les données.

2.2.1.3 Exploration quantitative d'hypothèses typologiques

Avec l'évolution des méthodes et le développement de la linguistique de corpus et du TAL, la typologie s'est également intéressée à baser ses analyses sur des corpus de données langagières attestées, puis dans un deuxième temps à enrichir ces corpus au moyen d'annotations pour effectuer des mesures quantitatives sur la fréquence et la distribution de certains types d'annotations. De plus en plus, ce corpus sont développés pour des langues multiples et dans un soucis d'harmonisation des schémas d'annotations qui permet de faciliter les études comparatives.

Ordre des mots Ainsi [Liu, 2010] reprend des hypothèses examinées entre autres par [Tesnière, 1959], [Greenberg, 1963], [Dryer, 1989] sur l'ordre des mots et teste la validité de leurs hypothèses empiriquement en mesurant la fréquence des dépendances orientées à droite (tête-initiale) et à gauche (tête-finale) pour décrire une langue en terme de la position linéaire des têtes par rapport à leurs dépendants dans l'énoncé. Cette utilisation des treebanks permet une caractérisation plus fine puisque ce n'est plus une catégorise discrète qui est attribué à chaque langue (tête-finale, tête-initiale) mais une catégorisation continue (par exemple 11% de dépendances à tête initiale et 89% à tête finale pour le Japonais).

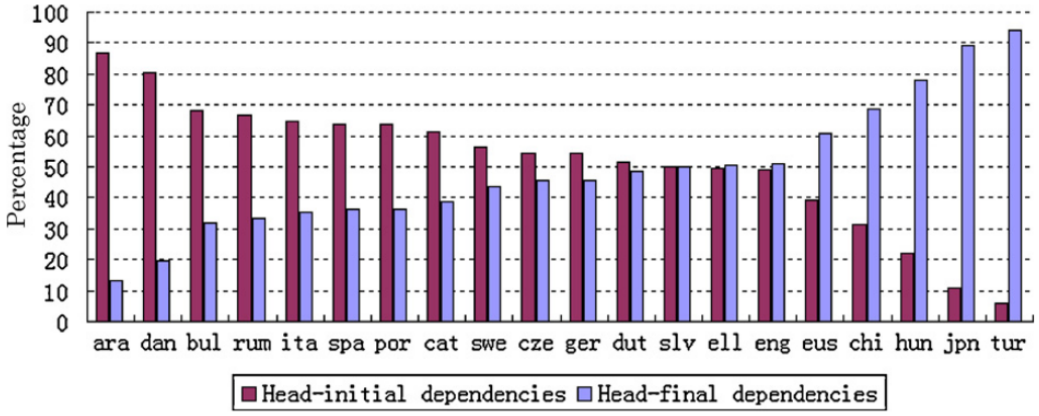


Fig. 4. Distribution of the dependency direction in 20 languages.

Figure 2.5: 20 langues placées sur un continuum tête-initiale tête-finale dans [Liu, 2010]

Cette méthode offre deux avantages considérables puisque d'une part elle ne présuppose pas qu'il y aurait un ordre de mot "basique" pour chaque langue, notion qui a été critiquée notamment par [Mithun, 1987], puisqu'elle permet d'observer pour chaque langue la fréquence de réalisation de chaque ordre de mot logiquement possible. D'autre part, en réutilisant les ressources (treebanks) et outils (étiqueteurs morpho-syntaxiques, analyseurs syntaxiques) développés en TAL, la typologie a accès à une énorme quantité d'annotations assez fines qui peuvent faciliter la découverte de nouvelles régularités en permettant la prise en compte de multiples critères (relation syntaxique et parties du discours par exemple). Ce dernier aspect est intéressant dans la mesure où dans certaines langues, l'ordre des mots n'est pas libre mais mixte [Tesnière, 1959], par exemple en Français nous trouvons un ordre SVO pour l'énoncé "J'achète le pain" mais un ordre SOV lorsque l'objet est pronominalisé "Je l'achète", ce qui signifie que la linéarisation de l'arbre de dépendance est bien conditionnée par des critères syntaxiques.

Typométrie et universaux configurationnels [Chen et al., 2018] proposent de faire émerger de nouveaux universaux en visualisant des classifications multi-dimensionnelles obtenues à partir de la distribution d'annotations syntaxiques extraits des corpus disponibles sur Universal Dependencies.

En particulier, ils observent l'existence de corrélations entre le pourcentage de dépendances avec une tête initiale, pour différentes combinaisons de relations syntaxiques. Dans la figure 2.6 nous observons la relation entre le pourcentage de sujet nominaux situés à droite du verbe, et le

pourcentage d'objet nominaux situés à droite du verbe.

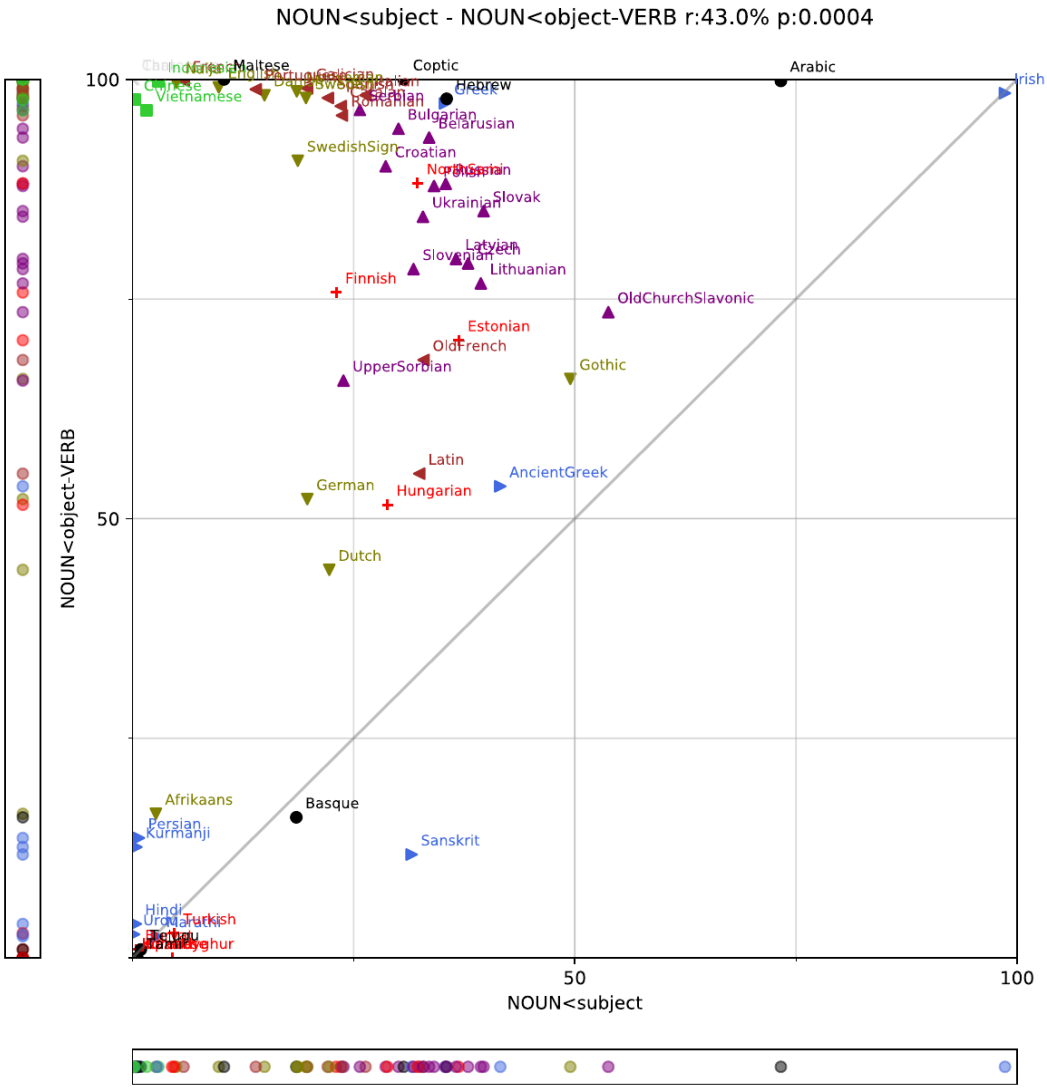


Figure 2.6: Sujets nominaux et objet nominaux dans [Chen et al., 2018]

Les chercheurs proposent également de considérer que ces diagrammes permettent de visualiser des configurations qui peuvent correspondre à des universaux quantitatifs. Ainsi, la figure 2.6 rentre dans un motif triangulaire qui indique qu'il y a une très forte tendances des langues à avoir plus souvent leur objet nominal à droite du verbe que leur sujet nominal. Ce type de diagramme dessine également une configuration avec des langues d'ordre libre ou mixte au centre du diagrammes, et les langues à ordre de mot rigide dans les coins.

Liberté d'ordre de mots Des travaux comme [Futrell et al., 2015] s'intéressent ainsi quant à eux à quantifier la liberté d'ordre de mot d'une langue [3], c'est-à-dire le degré de variabilité dans la linéarisation de l'arbre de dépendance sans modification du sens de l'énoncé. Pour y parvenir ils suivent les étapes suivantes :

1. segmentations des arbres de dépendance en deux types de sous-arbres pour calculer deux types d'entropie :
   - la première configuration prend en compte les parties du discours d'un gouverneur, d'un de ses dépendants et la relation syntaxique qui les lie, comme dans la figure 2.7.
   - la seconde configuration en figure 2.8 est plus complexe, puisqu'elle prend en compte tous les dépendants d'une même tête (parties du discours et relation syntaxique). De plus il y aura beaucoup plus de combinaisons possibles, ce qui limite l'efficacité statistiques des mesures. Les auteurs remarquent eux même qu'à l'heure actuelle ils disposent de trop peu de données pour pouvoir l'utiliser raisonnablement.
2. comptage sur les paires (X,C) avec X une séquence de mots et C le graphe de dépendance non-ordonné pour la séquence
3. estimation de l'entropie conditionnelle de la linéarisation X sachant le graphe C. Cette estimation est obtenue par maximisation de la vraisemblance d'obtenir la distribution mesurée en 2.

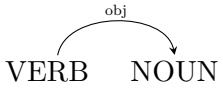


Figure 2.7: Entropie sur la direction des dépendances

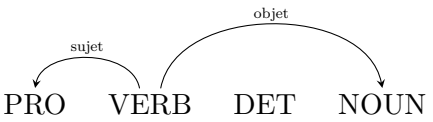


Figure 2.8: Entropie sur l'ordre des relations

[3]Plus précisément, et pour des raisons pratiques, les auteurs proposent d'estimer la borne supérieure de cette mesure de variabilité. Cela signifie que ces langues ne présentent pas davantage de variabilité d'ordre de mots, mais qu'il est possible que la variabilité réelle soit inférieure, notamment lorsque des critères de sélection lexicale sont pris en compte.

Pour rappel, l'entropie est une mesure qui vient du domaine de la théorie de l'information et a été introduite par Claude Shannon [Shannon, 1948]. Pour comprendre l'entropie il est préférable de commencer par définir le concept d'information tel qu'il est conçu en théorie de l'information. L'information est ce qui permet de diminuer notre incertitude. Dans le cas que nous étudions, si en rencontrant un graphe de dépendance non-ordonné nous sommes incertain quant à sa linéarisation, cela signifie que cette langue a un ordre de mot plutôt libre (plusieurs choix sont possibles).

Chaque événement X est porteur d'une certaine quantité d'information appelé *contenu informationnel* ou *surprise* (en anglais *self-information*, *information content* ou *surprisal*) représenté par $I(X)$, avec $I(X) = -log(p(xi))$, ce qui signifie que moins un événement est probable plus cet événement a un contenu informationnel fort.

L'entropie, qui représente notre incertitude quant au résultat de la linéarisation, se calcule ainsi :

$$H(X) = -\sum_{x \in X} p(x)\dot{l}ogp(x)$$

Ici ce qui nous intéresse plus particulièrement, c'est l'entropie conditionnelle de X sachant C, c'est-à-dire la quantité d'information apportée par la linéarisation X d'un graphe de dépendance C.

$$H(X|C) = \sum_{\substack{x \in X \\ c \in C}} p_c(c)p_{x|c}(x|c)logp_{x|c}(x|c)$$

L'entropie sur la direction des dépendances [4] est mesurée sur les configurations du type de la figure 2.7. Tandis que les configurations en 2.8 servent pour mesurer une entropie sur l'ordre des relations [5].

Utilisation de corpus parallèles Jusqu'ici nous n'avons abordé qu'un type de ressources : les treebanks annotés en syntaxe. D'autres corpus présentent des annotations qui peuvent être utilisées dans une optique de typologie quantitative. Ainsi [Östling, 2015] utilisent 1144 traductions alignées du Nouveau Testament (disponible pour 986 langues, ce qui en fait un

[4] Head Direction Entropy
[5] Relation Order Entropy

corpus très intéressant)[6] dans le but de mesurer la fréquences des différentes linéarisation pour les constructions (verbe, sujet, objet), (adjectif, nom), (nom, adposition). L'ordre le plus fréquent pour chaque langue est comparé aux description du WALS [Dryer and Haspelmath, 2013], et les auteurs observent un accord entre 86 et 96 % avec la ressource, selon le type de construction considérée. Ce type de méthode pourrait être utilisé pour vérifier ou compléter des bases de données existantes comme le WALS, tout en sachant que cette ressource impose de fournir un ordre dominant ou de choisir la valeur "ne présente pas d'ordre de mots dominant".

2.2.2 TAL et typologie : quels liens ?

D'après [Bender, 2009] le TAL a beaucoup à gagner de l'intégration de connaissances typologiques dans ses architectures. En effet, une certaine naïveté vis-à-vis de l'amplitude de variation des structures observables dans les langues, pourrait conduire les ingénieurs des systèmes de TAL à ne pas suffisamment interroger leurs biais, et à développer des systèmes qui encodent ces biais dans leur architecture. La chercheuse cite le cas des modèles n-grammes qui ont parfois été considérés comme n'encodant aucun savoir linguistique, et qui seraient donc en théorie adaptés à traiter toutes les langues. En réalité ces systèmes ont tendance à favoriser des langues dans lesquelles il y a un faible jeu de caractères, et à fonctionner de façon beaucoup moins satisfaisante sur des langues comme le mandarin dans lesquelles l'inventaire de caractères est plus grand, ce qui conduit à une explosion combinatoire du nombre de n-grammes rencontrés.

De plus en plus de travaux en parsing se concentrent sur l'apprentissage automatique d'analyseurs syntaxiques (ou parsers) multilingues [Ammar et al., 2016], le transfert d'annotation ou de modèles [Zeman and Resnik, 2008]. Dans ces contextes, plusieurs travaux ont essayé de cibler les apprentissages en choisissant des corpus d'apprentissages similaires aux corpus cibles.

[Zeman and Resnik, 2008] Entraîne ainsi un parser dé-lexicalisé (c'est-à-dire que les tokens sont remplacés par leur partie du discours) sur un treebank du danois, puis utilise le modèle sur un corpus dé-lexicalisé Suédois. Le parser obtient une précision de 66.4% sur le corpus Suédois, ce qui est encourageant étant donné qu'il faudrait 1546 phrases annotés pour entraîner un parser équivalent sur un treebank du Suédois.

[6] Cependant il s'agit tout de même d'un corpus assez particulier qui peut entraîner un certain nombre de biais, puisqu'il est constitué de traductions, qui plus est de textes sacrés et anciens, ce qui aura probablement une incidence sur les observations.

[Rosa and Zabokrtsky, 2015] avancent une extension à cette méthode, puisqu'ils proposent d'utiliser une mesure de divergence entre langues, pour sélectionner parmi un ensemble de corpus de langues sources ceux qui permettront d'apprendre le meilleur parser pour une langue cible. Pour calculer cette mesure de divergence, les auteurs observent la distribution de trigrammes de parties du discours dans les treebanks sources et le corpus cible qui est annoté en parties du discours. Ils utilisent la mesure de Kullback-Leibler qui permet de mesurer à quel point il est "coûteux" de passer d'une distribution de trigramme dans une des langues sources à la distribution de trigramme du corpus cible. Leur hypothèse consiste à soutenir que les corpus arborés dans lesquelles les distributions de trigrammes de catégories morpho-syntaxiques sont les plus proches se ressemblent en terme de comportements syntaxiques observables, puisque les n-grammes capturent une partie du comportement syntaxique, et que ce sont ces corpus arborés qui seront les plus utiles pour apprendre un analyseur syntaxique pour la langue cible. La contribution des parsers appris sur les corpus sources est ensuite pondérée en fonction de ces mesures de divergence, pour favoriser les parsers des langues "proches" en termes de distribution de trigrammes.

Ces deux travaux se limitent à utiliser des treebanks de langues "globalement similaires" pour apprendre un parser pour une nouvelle langue. Cependant, il n'y a pas toujours de corpus d'une langue proche de la langue cible qui soit disponible. Il est possible néanmoins d'utiliser le même type d'approche, mais de manière "locale", c'est-à-dire qu'au lieu d'identifier des corpus ou des langues proches, la tâche consiste à identifier à l'intérieur de ces corpus des **zones de compétences**. Deux langues peuvent présenter une syntaxe globalement assez différentes, mais se comporter similairement pour certaines configurations. Ainsi, [Naseem et al., 2012] les apprennent certaines configurations depuis les corpus des langues qui ont les mêmes propriétés typologiques (ainsi le rattachement des prépositions en anglais sera appris sur des langues qui ont elles aussi des prépositions plutôt que des postpositions).

[Søgaard, 2011] quant à lui propose de sélectionner à l'intérieur d'un corpus source les phrases dont les séquences de parties du discours sont les plus similaires aux phrases du corpus de langue cible que nous souhaitons parser. Pour cela il construit un modèle de langue basé sur les séquences de parties du discours dans la langue cible, puis sélectionne les phrases du corpus source qui ont les plus petites valeurs de perplexité par token. Cette méthode a l'avantage de permettre l'utilisation de corpus dont les langues ne sont pas nécessairement très proches

(l'auteur teste sur des corpus Bulgare, Arabe, Danois et Portugais).

2.3 Récapitulatif : que pouvons-nous dégager de ces travaux ?

Nous souhaitons retenir deux éléments principaux développés au cours de ce chapitre :

- Il peut être intéressant d'utiliser des treebanks pour comparer les propriétés syntaxiques de langues différentes dans le but de proposer des descriptions typologiques.
- Les systèmes de TAL, et notamment les analyseurs syntaxiques, peuvent intégrer des connaissances typologiques préalables (universaux, proximités syntaxiques entre langues...) ou des moyens d'induire ces similarités à l'intérieur de leur architecture pour adapter leurs traitements à de nouvelles langues.

Nous avons vu qu'il est possible de caractériser les langues en prenant en compte différents paramètres syntaxiques, parmi lesquels :

- l'ordre des mots, que nous reformulons en terme d'ordre des unités impliquées dans certaines relations syntaxiques (sujet, objet ...).
- la liberté d'ordre de mots, à savoir le degré de permissivité des linéarisations d'un même graphe de dépendance.
- la "cohérence" d'une langue dans la linéarisation des têtes et de leurs dépendants (est-ce que les têtes sont majoritairement initiales ou finales, et quelles variations observe-t-on selon le type de relations considérées).
- la structure des arbres de dépendances : profondeur des arbres, longueurs des dépendances.

Concernant les paramètres syntaxiques que nous souhaitons prendre en compte dans nos mesures de distances, nous constatons qu'il est possible de leur attribuer une valeur discrète (par exemple pour l'ordre des mots : SOV, SVO...) ou continue (un vecteur associant à chaque ordre des mots possible un poids proportionnelle à sa fréquence en corpus). Cette dernière option est rendue possible par l'utilisation de corpus qui permettent d'extraire des représentations plus

informatives, et ainsi de mieux rendre compte de l'existence de variations internes à chaque langue [7].

À partir de ces paramètres syntaxiques, et au moyen de mesures de distances, nous espérons observer comment les langues se ressemblent ou diffèrent dans le but d'observer si certains paramètres corrèlent entre eux.

[7] Ce qui est particulièrement important en TAL où on s'intéresse à automatiser des traitements sur tous les énoncés et pas seulement ceux qui ont une structure fréquente ou "canonique".

Chapitre 3

Méthodologie

3.1 Le corpus et son schéma d'annotation

3.1.1 Présentation du projet Universal Dependencies

Le projet Universal Dependencies est un projet collaboratif visant à proposer un schéma d'annotation syntaxique cohérent et cross-linguistiquement applicable afin de faciliter l'apprentissage d'analyseurs syntaxiques multilingues. Dans le cadre de ce projet, plus de 130 treebanks décrivant 73 langues ont été développés. Dans ce mémoire, nous utilisons la version 2.1 des treebanks UD [Nivre et al., 2017]. La table A.1 présente synthétiquement les 102 treebanks présents dans cette version.

Il est intéressant de noter qu'il existe une tension intrinsèque au projet que nous pouvons résumer ainsi : chaque langue doit être décrite de manière satisfaisante, ce qui peut impliquer un certain degré de précision afin d'opérer des distinctions entre des structures qui se comportent différemment à l'intérieur de la langue, mais parallèlement l'annotation doit rester cohérente au niveau "universel" afin de mettre en avant des similarités et régularités entre les langues. Il existe donc un compromis à faire afin de permettre la réconciliation de ces deux objectifs. Le projet UD a proposé une solution qui consiste à :

- mettre à disposition des inventaires universels fermés d'étiquettes morpho-syntaxiques et de relations syntaxiques.

- proposer des extensions spécifiques selon les langues pour les relations syntaxiques.
- autoriser l'introduction de nouveaux traits et valeurs pour la morphologie.
- encourager l'échange entre contributeurs dans le but d'harmoniser les annotations de phénomènes similaires (inter et intra-langues) notamment au travers d'un forum.

Nous souhaitons également relever le fait qu'il existe une certaine hétérogénéité dans la constitution de ces treebanks : certains ont été convertis depuis des schémas d'annotations préexistants, tandis que d'autres ont été pensés dès le départ dans ce schéma d'annotation, ce qui peut influer sur la qualité des annotations (par exemple certaines distinctions peuvent ne pas avoir été disponibles dans les treebanks d'origine). De plus, les annotations peuvent avoir été obtenues manuellement, automatiquement ou après corrections manuelles de sorties de systèmes automatiques. Ceci explique également le bruit que nous pouvons trouver dans les annotations, puisque celles-ci peuvent avoir été obtenues après de multiples conversions. Enfin, une autre différence importante réside dans l'utilisation que les contributeurs souhaitent faire de ces treebanks. Ainsi des personnes qui souhaitent utiliser les treebanks pour entraîner un parser dans le but de l'intégrer à une chaîne de traitement automatique n'attendront pas nécessairement le même niveau de granularité dans les annotations que des linguistes qui s'intéressent à décrire des phénomènes syntaxiques particuliers. Tous ces éléments permettent de mettre en évidence la complexité et fragilité de la cohérence et de l'universalité du schéma d'annotation.

3.1.2 Format des treebanks UD

Format ConLL-U Les treebanks sont disponibles au format CoNLL-U[1]. C'est un format tabulaire dans lequel les informations relatives à un token apparaissent sur une ligne qui comporte 10 colonnes :

1. identifiant du token
2. token
3. lemme

[1]Une version modifiée du format CoNLL-X [Buchholz and Marsi, 2006b], qui a été introduit lors de la 10ème itération de la CoNLL (Conference on Computational Natural Language Learning).

4. partie du discours universelle (upos)

5. partie du discours spécifique (xpos)[2]

6. traits morphologiques : chaque trait est annoté en suivant la convention *trait* = *valeur*, et les traits sont séparés par le symbole | .

7. identifiant du gouverneur

8. relation syntaxique par rapport au gouverneur

9. relations appartenant à la représentation UD "augmentée" (qui fournit des informations sur le graphe de syntaxe profonde)

10. informations supplémentaires que les personnes ayant constitué le corpus souhaitent ajouter : gloses, alignements temporels, commentaires etc...

```
# newdoc id = n02033
# sent_id = n02033063
# text = クーンは首を振ることしかできない。
# english_text = Kühn can only shake his head.
1  クーン  クーン  PROPN  NNP  _  5  nsubj  _  GHEAD=5|GHEADH=5|GID=1|Match=Yes|SpaceAfter=No
2  は  は  ADP  DP  _  1  case  _  GDPRL=compound:prt|GHEAD=1|GHEADH=1|GID=2|GUPOS=PART|Match=Yes|SpaceAfter=No
3  首  首  NOUN  NN  _  5  obj  _  GHEAD=5|GHEADH=5|GID=3|Match=Yes|SpaceAfter=No
4  を  を  ADP  CM  Case=Acc  3  case  _  GDPRL=compound:prt|GHEAD=3|GHEADH=3|GID=4|GUPOS=PART|Match=Yes|SpaceAfter=No
5  振る  振る  VERB  VV  Form=Adn  8  acl  _  GDPRL=acl:relcl|GHEAD=6|GHEADH=6|GID=5|Match=Yes|SpaceAfter=No
6  こと  こと  NOUN  NNB  _  5  aux  _  GDPRL=obj|GHEAD=8|GHEADH=8|GID=6|Match=Yes|SpaceAfter=No
7  しか  しか  ADP  DP  _  5  case  _  GDPRL=compound:prt|GHEAD=6|GHEADH=6|GID=7|GUPOS=PART|Match=Yes|SpaceAfter=No
8  でき  できる  VERB  VV  Form=Irr  0  root  _  GHEAD=0|GID=8|Match=Yes|SpaceAfter=No
9  ない  ない  AUX  VXP  Polarity=Neg|VerbForm=Fin  8  aux  _  GDPRL=advmod|GHEAD=8|GHEADH=8|GID=9|Match=Yes|SpaceAfter=No
10  。  。  PUNCT  .  _  8  punct  _  Match=Yes|GID=10|GHEAD=8|GHEADH=8
```

Figure 3.1: Bloc conllu pour une phrase du corpus UD_Japanese-PUD

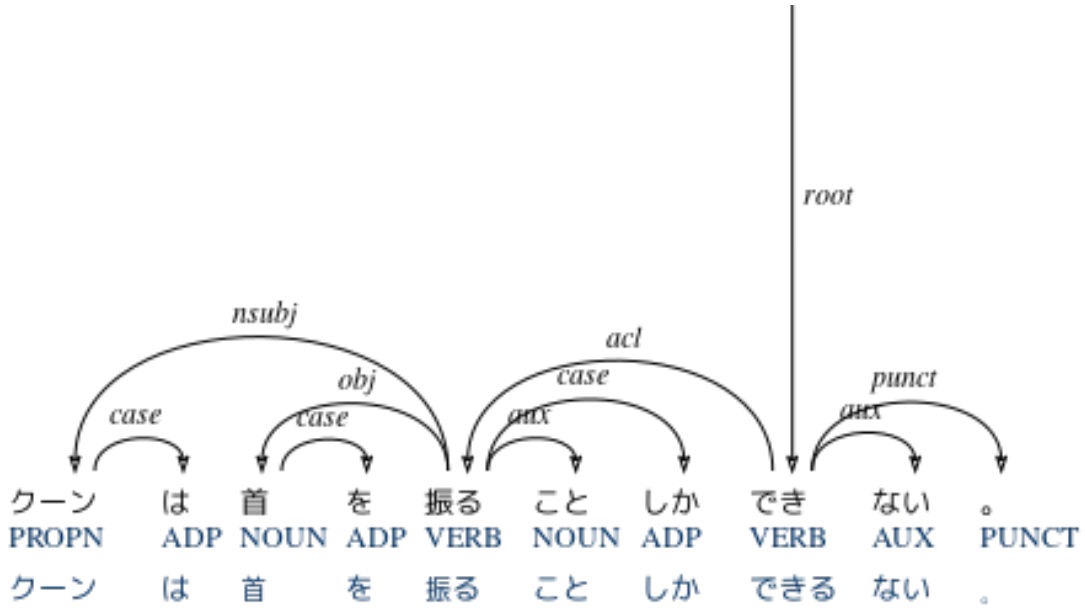


Figure 3.2: Arbre de dépendance associé à l'énoncé précédent dans le corpus UD_Japanese-PUD

[2]Elle permet notamment de conserver les parties du discours provenant du corpus original lorsque le corpus est issu d'une conversion, ce qui peut s'avérer utile puisque ces corpus utilisent parfois des jeux d'étiquettes d'une granularité plus fine

Sous-spécification Les seules informations requises par le schéma d'annotation concernent le token, son gouverneur et la relation le rattachant à son gouverneur. Les autres informations peuvent être sous-spécifiées si elles ne sont pas pertinentes ou lorsqu'elles ne sont pas disponibles, dans ce cas le symbole _ figure dans la colonne en lieu de la valeur.

Méta-données Chaque phrase est précédée d'un certain nombre de méta-données [3]. Celles-ci sont identifiables dans le fichier texte par le premier caractère #. Les seules méta-données obligatoires sont l'identifiant de la phrase ainsi qu'une transcription, mais il est également possible de rajouter des informations concernant le document depuis lequel a été extrait l'énoncé, le genre de texte, l'époque ou encore de proposer une traduction.

3.1.3 Schéma d'annotation : un inventaire de catégories et de relations syntaxiques "universelles"

Comme nous l'avons vu précédemment, à la base de ce projet existe l'idée qu'il est possible de proposer un inventaire de catégories et relations syntaxiques permettant de mettre en évidence des régularités dans le comportement syntaxique entre les langues. Cette idée correspond à ce que [Croft, 2001] désigne comme une version plutôt souple de la théorie de la Grammaire Universelle : qu'il existe un inventaire de catégories et relations syntaxiques partagées par toutes les langues, mais qu'individuellement chaque langue n'a pas nécessairement l'usage de toutes ces catégories et relations.

Pour définir les catégories morpho-syntaxiques, deux types d'approches sont possibles : une approche top-down et une approche bottom-up.

[Haegeman, 1994] défend l'approche top-down qui consiste à présupposer l'existence de certaines catégories morpho-syntaxiques primitives. C'est l'appartenance à cette catégorie qui déterminerait les contextes dans lesquels l'unité apparaît :

> "Les mots appartiennent à différentes catégories syntaxiques, comme la catégorie des noms ou des verbes etc... et la catégorie syntaxique à laquelle appartient un mot détermine sa distribution, c'est-à-dire les contextes dans lequel nous pouvons le

[3] La nature et le niveau de détail de ces méta-données varient selon les corpus.

rencontrer." [Haegeman, 1994] (nous traduisons)[4]

Cependant, de nombreux exemples indiquent que les mots appartiennent à différentes catégories selon les contextes qu'ils occupent. Ainsi dans la phrase "Their singing was exceptional" *singing* n'est pas un verbe mais un nom, puisque c'est avec ce type d'unités que le mot peut commuter, et parce que ce dernier présente les propriétés d'un nom, à savoir qu'il peut être modifié par un adjectif ou par une relative. Cette seconde approche, qui consiste à fonder les catégories morpho-syntaxiques sur les contextes dans lesquels les mots apparaissent, est appelée *approche distributionnelle*.

UD opte pour une approche majoritairement distributionnelle. Dans la majorité des cas, l'appartenance à une catégorie morpho-syntaxique est déterminée en contexte, selon un certain nombres de critères universaux ou spécifiques à la langue. En revanche pour des cas particuliers, le schéma d'annotation incite à fixer d'avance une catégorie morpho-syntaxique et à l'utiliser lorsque le comportement de l'unité est jugé "exceptionnel", par exemple lorsqu'un mot est mentionné, comme "precede" dans la phrase "He pronounced 'precede' in a funny way". Dans cet énoncé "precede" est utilisé comme un pronom, comme nous pouvons le constater par commutation dans "He pronounced it in a funny way". Cependant ce type de comportement est très inhabituel pour "precede" qui est habituellement un verbe, ce qui justifierait le choix d'annotation qui permet une généralisation plus facile dans le cas d'un apprentissage lexicalisé, bien qu'il aille à l'encontre des critères distributionnels.

Le schéma d'annotation développé par UD fixe l'inventaire de catégories morpho-syntaxiques suivant :

- ADJ: adjectif
- ADP: adposition
- ADV: adverbe
- AUX: auxiliaire
- CCONJ: conjonction de coordination

[4] "Words belong to different syntactic categories, such as nouns, verbs etc., and the syntactic category to which a word belongs determines its distribution, that is, in what contexts it can occur"

- DET: déterminant
- INTJ: interjection
- NOUN: nom
- NUM: numéral
- PART: particule
- PRON: pronom
- PROPN: nom propre
- PUNCT: ponctuation
- SCONJ: conjonction de subordination
- SYM: symbole
- VERB: verbe
- X: autre

Cet inventaire doit respecter simultanément plusieurs contraintes : il doit posséder suffisamment de catégories pour ne pas créer de fausses similarités entre des unités qui appartiendraient à des paradigmes différents, tout en contraignant suffisamment le nombre de catégories afin de mettre en évidence des régularités de comportement. Un autre critère important pour la constitution de l'inventaire est l'apprenabilité. Une grande partie de la communauté UD s'intéresse à ces treebanks pour entraîner des analyseurs syntaxiques, que ce soit pour proposer des améliorations sur la tâche de parsing en elle-même, ou afin d'utiliser ces analyseurs pour obtenir les structures d'énoncés afin de faciliter d'autres tâches de TAL (analyse de sentiment, extraction d'information…). L'inventaire doit donc se révéler suffisamment informatif pour être utile aux tâches en aval, en évitant une explosion du nombre de catégories qui rendrait plus complexe l'apprentissage des analyseurs par manque de généralisation.

Le projet fournit également l'inventaire de relations syntaxiques suivant :

- acl: modifieur clausal d'un syntagme nominal

- advcl: modifieur clausal adverbial d'un verbe ou prédicat
- advmod: modifieur adverbial
- amod: modifieur adjectival
- appos: apposition
- aux: auxiliaire
- case: marqueur de cas, adposition
- cc: conjonction de coordination
- ccomp: complément clausal
- clf: classifieur
- compound: compound
- conj: conjoint
- cop: copule
- csubj: sujet clausal
- dep: dépendance sous-spécifiée
- det: déterminant
- discourse: marqueur de discours
- dislocated: élément disloqué
- expl: explétif
- fixed: expression poly-lexicale
- flat: construction sans tête syntaxique
- goeswith: rattachement de tokens qui n'auraient pas dû être segmentés
- iobj: objet indirect
- list: liste

- mark: marqueur
- nmod: modifieur nominal
- nsubj: sujet nominal
- nummod: modifieur numéral
- obj: objet
- obl: oblique
- orphan: gapping et ellipses
- parataxis: parataxe
- punct: ponctuation
- reparandum: disfluence
- root: racine
- vocative: vocatif
- xcomp: complément clausal contrôlé

Cet inventaire de relations syntaxiques est mieux expliqué en présentant les distinctions qu'il opère et n'opère pas. Tout d'abord, il introduit une distinction entre les dépendants clausaux et nominaux, comme nous pouvons le constater avec la séparation entre les relations csubj/nsubj, ccomp/obj. Pour les modifieurs, cette séparation est ternaire avec les nominaux, d'une part, les clauses d'autre part, et les unités qui ne sont ni nominales ni clausales, pour lesquelles la relation est typée en fonction de la catégorie morpho-syntaxique du dépendant, ce qui nous donne obl pour les nominaux, advcl pour les clauses et advmod|nummod|amod pour les autres modifieurs. Nous relevons également une distinction entre deux types de compléments clausaux : ceux dont tous les arguments sont instanciés (ccomp), et les compléments pour lesquels au moins un des arguments est hérité, sans être instancié, comme dans les phénomènes de contrôle (xcomp).

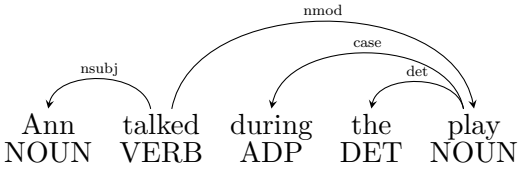
Dans le même temps, certaines distinctions ne sont pas opérées. Ainsi la relation obl est utilisée à la fois pour des arguments comme entre "victimes" et "Révolution" dans "Elles furent victimes

de la Révolution française", et pour les modifieurs comme dans l'exemple "Par la suite, le château fut démantelé", où "suite" est annoté comme un dépendant oblique de "démantelé"[5].

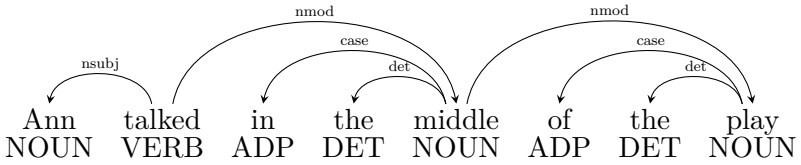
Choix de têtes lexicales Dans le schéma d'annotation UD, les gouverneurs sont principalement des têtes lexicales plutôt que des têtes fonctionnelles, alors que c'est plutôt l'inverse qui est généralement adopté en syntaxe de dépendance. Ce choix est particulièrement visible dans les configurations auxiliaire-verbe et adposition-complément.

Une expérience réalisée dans [Osborne and Niu, 2017] remarque que pour la combinaison auxiliaire-verbe, les participants interrogés préfèrent un chunking – c'est-à-dire un découpage de l'énoncé en séquences contigües et non-récursives d'unités lexicales liées à une unique tête forte [Abney, 1991] – qui est cohérent avec une analyse de l'auxiliaire comme gouverneur du verbe, plutôt que l'analyse favorisée par UD.

De plus, comme [Gerdes and Kahane, 2016] le remarquent, le fait de traiter différemment les unités porteuses de contenu lexical ("content word") et les unités fonctionnelles, occulte certaines similarités, par exemple entre des prépositions simples et des expressions prépositionnelles [6] comme dans la figure 3.3. Dans l'exemple, bien que "during" et "in the middle of" jouent le même rôle, le premier est dépendant du nom, tandis que le second, considéré comme porteur d'un contenu lexical accède au statut de gouverneur du nom.



(a)



(b)

Figure 3.3: Comparaison des annotations pour les prépositions et les expressions prépositionnelles

[5] Les deux exemples proviennent du corpus UD_French-Original.

[6] Plus particulièrement, dans le cas où ces unités sont compositionnelles, puisque les expressions prépositionnelles non-compositionnelles sont souvent annotées en fixed

Dans une de leurs expériences [Schwartz et al., 2012] montrent que, pour l'anglais, le choix des adpositions comme gouverneurs dans ce type de configuration est plutôt justifié puisqu'il retourne un meilleur score d'"apprenabilité", c'est-à-dire que les annotations sont potentiellement plus faciles à reproduire pour un analyseur syntaxique automatique.

En revanche [Wisniewski and Lacroix, 2017] testent cette hypothèse sur des treebanks de 38 langues appartenant à la version v1.3 du projet UD et constant que dans 65.25% des cas apprendre sur un treebank ayant suivi une conversion visant à remonter la préposition dans l'arbre de dépendance entraîne de moins bonnes performances des analyseurs syntaxiques que l'apprentissage sur le treebank initial, avec les adpositions annotés comme dépendant de leur complément. Le choix pourrait donc rester cohérent avec certains des objectifs d'UD.

3.2 Que mesurons-nous réellement lorsque nous mesurons des distances syntaxiques ?

3.2.1 Pourquoi se poser la question de savoir si deux langues ont des syntaxes similaires ?

> "Il faudrait peut-être commencer par observer que les langues sont en perpétuel changement, ce qui autorise à penser que, si un type d'organisation est peu ou pas du tout attesté, c'est probablement parce qu'il n'est pas l'aboutissement d'un type fréquent de changement linguistique, alors que les types d'organisation particulièrement bien attestés doivent constituer l'aboutissement de changement qui tendent à se produire fréquemment dans l'histoire des langues."[7]

Derrière cette question de proximité entre langues, il y a un objectif théorique : mieux comprendre comment les langues évoluent dans le temps et l'espace. À savoir, comment les variations en terme de structures syntaxiques observables permettent-elles de mieux comprendre les processus d'évolutions linguistiques (évolution vers un système plus économique, créolisation, contact entre langues...) ? Notons que certains types de corpus en particulier (corpus en diachronie, corpus de créoles ou pidgins, corpus de variétés d'une même langue, corpus de langues apparentées..) se prêtent particulièrement à des études de ce genre.

[7] Creissels (page 6)

Mieux comprendre ces proximités entre langues peut également avoir une application pédagogique, puisque cela permet de contextualiser l'apprentissage d'une langue pour un apprenant, en pointant les différences et similarités avec les langues qu'il connaît déjà.

Nous souhaitons maintenant revenir sur la notion de distances syntaxiques entre langues. Intuitivement, cette expression fait sens, puisque nos expériences (apprentissage de langues étrangères, étude linguistique de langues inconnues) nous ont appris que parmi les langues, certaines se ressemblent plus que d'autres (en termes de systèmes syntaxiques, phonologiques etc...). Pourtant, il n'est pas évident de déterminer des critères qui permettent de motiver cette caractérisation. Nous souhaitons dans cette section présenter tout d'abord des critères qui permettraient de dire que deux langues sont similaires [8], puis regarder plus concrètement les types d'éléments de structures que nous pouvons extraire et quantifier depuis les treebanks, qui nous servirons afin de construire des représentations, c'est-à-dire des versions simplifié de l'objet (ici la syntaxe d'une langue) qui permettent d'étudier son comportement.

3.2.2 Critères de proximité entre langues

Les langues présentent-elles le même type de constructions ? Deux langues peuvent être décrites comme similaires lorsqu'elles partagent certaines constructions. Ainsi, les corpus qui n'utilisent pas la relation syntaxique *det* partagent une caractéristique commune : l'absence de déterminants dans la langue, ce qui est un indicateur – parmi tous les indicateurs possible – d'une certaine proximité entre les langues de ces corpus. De la même façon, nous pouvons dire que l'anglais et le nigerian pidgin (ou naija) sont similaires d'une certaine façon, car dans ces deux langues il est possible de disloquer aussi bien un sujet qu'un objet, c'est-à-dire de le réaliser deux fois : sous forme complète, et pronominalisée comme dans les figures 3.4 et 3.5.

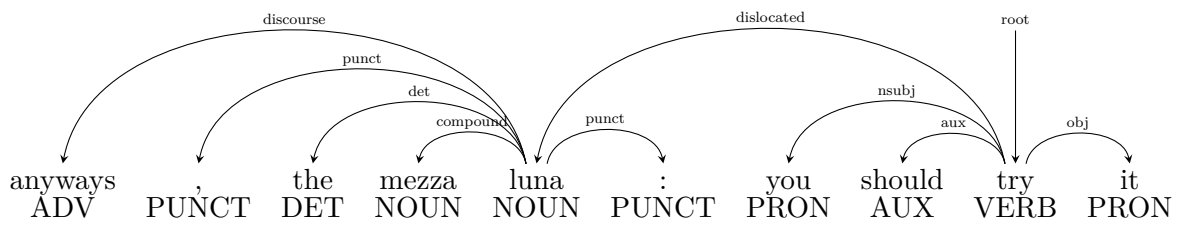


Figure 3.4: Exemple d'une phrase avec dislocation de l'objet dans le corpus d'Anglais UD_English-Original

[8] Sachant que cette similarité est toujours basé sur des multiples axes d'analyse, et que nous pouvons difficilement tous les prendre en compte simultanément.

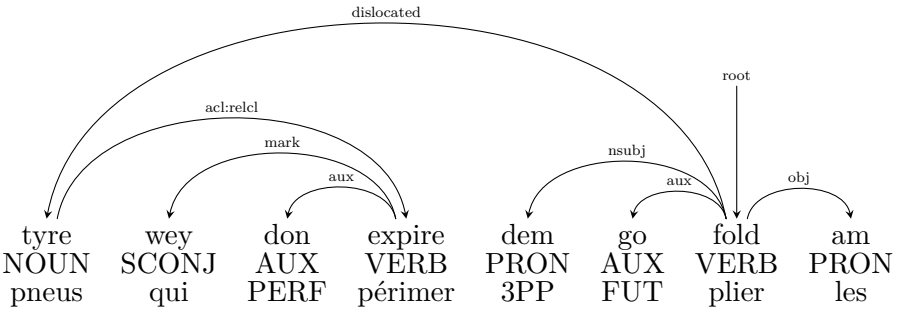


Figure 3.5: Exemple d'une phrase avec dislocation de l'objet dans le corpus d'Anglais UD_Naija-NSC. Traduction : "Des pneus qui ne sont plus en état d'être vendus, ils vont les plier."

9

Constructions similaires avec des annotations différentes En revanche, ce critère n'est pas sans ses limites, tout particulièrement dans notre cas puisque nous utilisons des annotations pour identifier ces constructions. Or malgré le schéma d'annotation commun, certaines constructions peuvent être annotées différemment selon les corpus. Ainsi, la relation *clf* peut être utilisé pour les classifieurs dans les langues comme le mandarin qui possèdent ce type d'unités. Cependant, l'absence de cette relation dans un treebank ne permet pas nécessairement d'exclure la présence de classifieurs dans la langue, comme nous pouvons l'observer dans les corpus du japonais qui proposent une annotation différente des classifieurs avec la relation *mark* comme en figure 3.6.

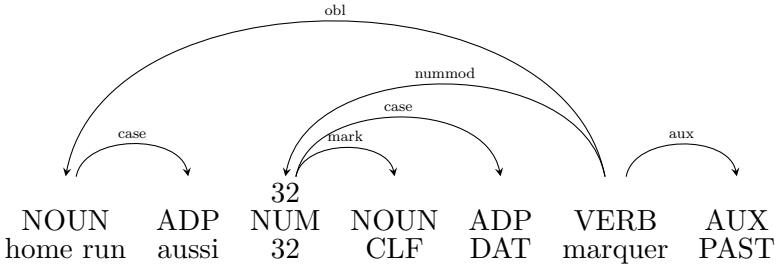


Figure 3.6: Exemple d'annotation d'un classifieur dans le corpus UD_Japanese-Original. Traduction : "Il est aussi parvenu à marquer 32 home run."

Fréquence des constructions D'autre part, même si les langues présentent des structures communes, cela ne signifie pas qu'elles sont utilisées de la même manière. La construction peut être très fréquente dans un corpus, et réservée à certain cas beaucoup plus contraints (ou simplement moins fréquents) dans un autre corpus. Si nous revenons sur notre exemple

des dislocations en anglais et naija, nous observons des fréquences avec une différence marquée entre les corpus de ces langues (1.67% des relations de dépendance[10] sont annotées *dislocated* dans le corpus du naija, contre 0.35% dans l'ensemble des corpus de l'anglais). Il nous paraît en conséquent pertinent de prendre en compte non seulement l'existence de certaines constructions, mais également leur fréquence pour observer des divergences entre les langues à comparer. En revanche, il faut également conserver à l'esprit que si certaines de ces variations sont réellement dues à des variations entre les langues, d'autres variations peuvent provenir de propriétés des corpus, comme le genre de texte (presse, textes religieux, blogs...), sa modalité (corpus oral transcrit, corpus écrit, corpus signé transcrit [11]), ou encore la variété de langue. Ainsi la sur-représentation des relations disloquées dans le corpus du naija (oral transcrit) par rapport aux corpus de l'anglais (en grande partie des textes écrits) pourrait n'être qu'en partie liée aux langues de ces corpus, et en partie liée au fait que les dislocations sont plus courantes à l'oral qu'à l'écrit.

Devons-nous plutôt considérer les phénomènes fréquents dans la langue ou les phénomènes rares par rapport à l'échantillon ? Il nous appartient également de déterminer les caractéristiques qui font qu'un phénomène syntaxique est pertinent pour caractériser le comportement syntaxique d'une langue. Nous pouvons décider qu'une langue est mieux décrite par les phénomènes qui sont fréquents dans cette langue, puisque ce sont ces types de structures que nous rencontrons le plus souvent. Une seconde approche en revanche, consiste à regarder les propriétés qui sont les plus spécifiques de cette langue, comparées aux propriétés de l'échantillon, c'est-à-dire les propriétés qui font que cette langue se démarque par rapport aux autres, puisque si deux langues partagent une propriété qui les distinguent réellement des autres langues, la saillance de cette propriété nous incitera à considérer ces langues comme proches, même si le phénomène n'est pas très fréquent dans la langue.

Structures équivalentes Qu'en est-il de l'isomorphisme ? Deux structures sont dites isomorphes l'une de l'autre lorsqu'elles sont équivalentes, c'est-à-dire qu'il y a une transformation possible qui permet de passer de la structure A à la structure A' et de repasser de la structure A' à la structure A par le moyen d'une seconde transformation. Dans ce mémoire, nous avons

[10] Les liens de ponctuations n'ont pas été pris en compte dans le décompte total, car ils sont très nombreux dans le corpus du naija, où ils sont utilisés pour annoter des informations prosodiques.

[11] Comme le corpus UD_Swedish_Sign_Language-SSLC .

toujours étudié les structures à un isomorphisme prêt, puisque nous n'avons pas pris en compte le fait que deux structures peuvent avoir l'air différentes alors qu'elle sont équivalentes.

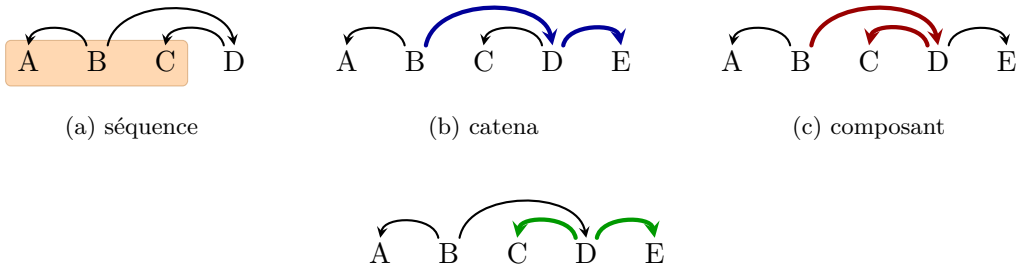
3.2.3 Que pouvons-nous mesurer sur des treebanks ?

3.2.3.1 Fréquence d'éléments de structures

Nous cherchons à comparer la syntaxe des langues, en utilisant des méthodes quantitatives qui nous permettront d'exploiter des treebanks annotés en syntaxe. Suivant la description de [Osborne and Niu, 2017], nous définissons dans la table 3.1 quatre types d'unités syntaxiques dont nous pouvons observer le comportement.

type d'unité	description
séquence	mot ou combinaison de mots qui sont linéairement ordonnés
catena	mot ou combinaison de mots qui sont structurellement ordonnés
composant	mot ou combinaison de mots qui sont linéairement et structurellement ordonnés
constituant	un composant qui est complet (tous les dépendants de la tête sont sélectionnés)

Table 3.1: Typologie des unités syntaxiques



(a) séquence (b) catena (c) composant

(d) constituant

Figure 3.7: Typologie visuelle des types d'unités syntaxiques

3.2.3.2 Propriétés des sous-structures

Pour tous les types d'unités syntaxiques décrites en 3.1, il est possible d'étudier les propriétés, ou de spécifier des contraintes sur les éléments suivants :

- les catégories associées aux nœuds.

- les types de relations syntaxiques associées aux arcs.
- la précédence linéaire ou structurale entre des nœuds.
- le distance linéaire entre des nœuds.
- la distance structurale entre des nœuds (ou profondeur du graphe).

Chapitre 4

Expérimentations

Ce chapitre est consacré à mise en place des expérimentations sur les treebanks. Dans la section 4.1, nous présentons l'outil qui nous a permis d'extraire les motifs, et les motifs syntaxiques que nous avons sélectionnés. La section 4.2 décrit les représentations sur lesquelles viendront s'appliquer les mesures introduites en section 4.3, et à partir desquelles nous pourrons essayer de faire émerger des groupes de langues de comportements similaires à l'aides des méthodes de clustering décrites dans la section 4.4.

4.1 Choix des structures à observer

4.1.1 Recherche de chemins dans des arbres

Afin d'extraire les configurations dont nous avons choisi d'étudier la distribution, nous avons utilisé 2 méthodes : des scripts écrits en Python et le logiciel de réécriture de graphes Grew [1] [Guillaume et al., 2012].

Nous avons commencé par utiliser des scripts en python pour transformer les fichiers conlls en arbres. Cependant, plus les structures que nous souhaitons extraire sont complexes (contraintes multiples sur les relations, l'ordre structural, l'ordre linéaire, les catégories des nœuds...) plus il devient intéressant d'utiliser un programme dédié à la recherche de chemins dans des graphes, puisque ces recherches peuvent être computationnellement exigeantes.

[1] https://gitlab.inria.fr/grew/grew

Nous avons utilisé le logiciel Grew qui permet de requêter les arbres de dépendance de façon assez fine, puisqu'il permet de spécifier des contraintes sur le placement relatif des nœuds dans l'arbre (précède, précède directement ..), de ne pas déclencher la règle de reconnaissance si certains motifs négatifs sont présents, ainsi que d'interroger les relations de dépendances, leur type, la catégorie, et les traits morphologiques des nœuds de l'arbre. Un script python permet ensuite d'automatiser le repérage et le comptage des motifs sur tous les corpus UD. Au préalable, nous avons lancé un pré-traitement sur l'arborescence des treebanks, visant à enlever les sous-types de relation [2].

Format des règles Dans les règles grew, les contraintes sur les nœuds ou sur les arcs sont indiquées entre crochets. Des variables sont utilisées afin de référer aux nœuds. La précédence linéaire est indiquée au moyen des symboles < (précédence directe) et « (précédence directe ou indirecte).

Exemple de règle spécifiant le motif SOV En langage naturel cette règle signifie : retourne-moi tous les verbes qui ont un dépendant sujet sur leur gauche et un dépendant objet sur leur gauche, avec le sujet précédant l'objet. Les caractères V, S et O sont des noms de variables associés à des nœuds de l'arbre. Les crochets permettent de poser des contraintes, par exemple cat impose une contrainte sur la catégorie morpho-syntaxique. Les relations de dépendances sont indiquées par des flèches, sur lesquelles peuvent s'ajouter des contraintes, ici la relation de dépendance entre le gouverneur V et le dépendant S doit être un sujet clausal (*csubj*) ou nominal (*nsubj*). Nous avons ici spécifié le motif SOV.

```
1 pattern {
2   V [cat=VERB];
3   V -[nsubj|csubj]-> S;
4   V -[obj|iobj|xcomp|ccomp]-> O;
5   S << V;
6   O << V;
7   S << O;
8 }
```

[2] Nous avons découvert par la suite que grew est capable de spécifier les noms de relations sous forme d'expression régulière, ce qui permet d'éviter ce genre de traitements.

4.1.2 Langues et treebanks : de multiples sources de variations

Jusqu'ici nous n'avons pas réellement abordé un point important de ces expérimentations. Nous disposons de treebanks, mais souhaitons mesurer des distances entre langues, or certaines langues ont plusieurs treebanks censés permettre de généraliser des comportements syntaxiques. Comment pouvons-nous nous accommoder de ces multiples treebanks par langue lorsqu'ils existent ?

Une première solution consiste à tout simplement à fusionner les différents treebanks d'une langue, et à traiter le treebank global résultant comme "le treebank de la langue". Cette solution a l'avantage de nous donner des mesures plus significatives, puisqu'elles seront basées sur davantage de données. D'un autre côté, dans cette solution la variabilité présente entre les treebanks d'une même langue n'est pas exploitée, au contraire elle est même occultée. Pourtant cette division en treebanks peut nous apporter des informations précieuses. En effet les variations entre les treebanks d'une même langue ne sont pas aléatoires, elles dépendent de propriétés comme le genre de texte, sa modalité ou encore les choix d'annotations qui ont été effectués. Nous pouvons donc également choisir de conserver cette séparation entre treebanks, ce qui nous permettrait d'évaluer à quel point les paramètres que nous avons choisis sont liés à d'autres propriétés qu'à la langue.

Enfin, une troisième alternative est envisageable, tout particulièrement lorsque nous souhaitons étudier une langue pour laquelle chaque sous-corpus est assez homogène. Nous pourrions alors dans un premier temps construire une représentation pour chaque sous-corpus, puis à partir de ces différentes représentations, proposer une représentation "moyenne" qui viserait à limiter l'influence des contributions spécifiques à chaque corpus. Pour une représentation vectorielle, cela reviendrait par exemple à calculer le barycentre des différents vecteurs. Cela permettrait que chaque représentation contribue équitablement (les grands corpus n'affecteraient pas plus la représentation finale que les corpus plus réduits).

Notons que les méthodes 1 et 3 adoptent des points de vue opposés : la méthode 1 considère que les variations entre les treebanks ne sont pas pertinentes, et qu'il n'y a qu'une vraie source de variation par rapport à l'objet d'étude, qui serait la langue des corpus; la méthode 3 en revanche considère qu'il y a plusieurs sources de variations, et que la meilleure façon de comparer les langues en minimisant les autres types de variations nécessite de calculer une représentation qui prenne en compte ces effets multiples et tente de les neutraliser.

4.1.3 Choix des configurations syntaxiques à étudier

Nous avons choisi d'étudier 3 types de phénomènes syntaxiques, ce qui nous a conduit à préciser le type de motifs syntaxiques à chercher. Le repérage de ces motifs nous permettra de construire des représentations qui captureront chacune une partie du comportement syntaxique de chaque langue. Ces trois paramètres ont quelque chose en commun : ils visent à étudier l'ordonnement d'unités vis-à-vis d'autres unités.

4.1.3.1 Ordre du verbe et de ses dépendants sujet et objet

Nous avons vu dans le chapitre 2 que l'ordre de ces unités a longtemps été un sujet d'intérêt pour les typologues qui y voyaient là un paramètre pertinent dans le but de dresser une typologie des langues. Si nous reprenons la typologie dans la table 3.1, caractériser l'ordre des mots en fonctions des unités du verbe et de ses dépendants objet et sujet revient à extraire des catenas, puis à quantifier la présences des 6 linéarisations possibles : OVS, OSV, SOV, SVO, VOS et VSO.

Le schéma d'annotation UD distingue deux types de sujets : les sujets nominaux (nsubj) et les sujets clausaux (csubj). Pour cette étude, nous fusionnons les deux relations dans nos règles d'extraction. En ce qui concerne les objets, nous fusionnons les relations des objets (obj et iobj) et les compléments clausaux (ccomp et xcomp).

Il serait également possibles de caractériser plus finement les linéarisations en prenant en compte le type de réalisation des sujets et objets, par exemple en introduisant à chaque fois une subdivision en (pronominal, nominal, clausal) puisque ce type d'information peut être pertinent pour certaines langues (par exemple en français une pronominalisation de l'objet peut changer son ordre par rapport au verbe).

C'est en partie la solution que [Chen et al., 2018] ont adopté puisque les chercheurs ont cherché à observer la relation entre objets nominaux rattachés à droite du verbe et objets pronominaux rattachés à droite du verbe. En revanche, pour une configuration prenant en compte trois unités, introduire ce niveau de granularité nécessiterait de subdiviser chaque type original 9 sous-types en fonction des 3 catégories possibles pour le sujet comme pour l'objet, ce qui résulterait en 54 types au total. Nous n'introduisons pas cette distinction pour le moment, car certains treebanks

ne sont pas très grand et étudier des configurations trop complexes risquerait de nous donner des comptages trop peu nombreux pour en tirer de réelles conclusions.

Pour ce motif en particulier, nous faisons le choix de n'autoriser qu'un comptage par verbe, c'est-à-dire que nous filtrons les résultats et ne gardons qu'une occurrence de chaque combinaison d'un identifiant de phrase et de l'identifiant du verbe. Cela signifie que si le verbe a plusieurs objets, la linéarisation d'un seul d'entre eux est étudiée ce qui n'est pas idéal, mais autoriser plusieurs comptages par verbe dédoublerait les comptes pour les sujets dans des constructions avec plusieurs objets, ce qui nous paraît être un biais encore moins satisfaisant.

4.1.3.2 Proportions de dépendances orientées à gauche pour certaines configurations

Nous souhaitons ici étudier la direction des liens de dépendance pour quatre types de couples : verbe-objet, auxiliaire-verbe, adposition-nom et nom-adjectif. Pour cela, nous extrayons ces motifs, et mesurons la quantité de liens orientés vers la gauche (ce qui correspond à une linéarisation dépendant-gouverneur) sur le nombre d'occurrence total du motif, ce qui revient à mesurer la proportion de linéarisation dépendant-gouverneur.

Nous avons choisi ce paramètre pour étudier l'hypothèse de Hawkins mentionnée en 2.2.1.2 selon laquelle à l'intérieur d'une langue, ce ratio serait équilibré pour ces quatre configurations.

4.1.4 Trigrammes d'étiquettes morpho-syntaxiques

Nous étudions ici des séquences de catégories morpho-syntaxiques. Comparé aux deux autres paramètres, celui-ci ne requière pas nécessairement pas un corpus arboré, un simple corpus étiqueté pourrait convenir. En revanche, ce paramètre a aussi l'avantage de permettre l'étude de comportements syntaxiques locaux qui ne sont pas captés par les deux autres motifs, puisqu'il ne nécessite pas que les unités soient en relation structurale. Ce paramètre permet donc d'étudier la linéarisation sous un autre angle de vue : comment des unités qui ne sont pas en relation syntaxique en arrivent-elles à se retrouver côte à côte dans l'énoncé ?

4.2 Représentations

Deux types de représentations ont été étudiées : une représentation vectorielle, et une représentation sous forme de distribution.

Représentation vectorielle Chaque représentation sous forme de vecteur désigne en fait les coordonnées d'un point dans un espace vectoriel de n dimensions. Les n dimensions de ce vecteurs correspondent aux types possibles pour un motif (dans le cas de l'ordre du sujet, verbe et objet), ou aux unités pour lesquelles nous étudions une proportion (verbe-objet, auxiliaire-verbe, adposition-nom et adjectif-nom) pour le deuxième paramètre.

Deux langues sont situées à des endroits proches dans l'espace vectoriel, lorsque leur représentation vectorielles sont proches, c'est-à-dire lorsqu'elles se comportent de façon similaires par rapport aux paramètres choisis pour construire les représentations. En revanche, lorsque les langues sont très différentes, nous pouvons nous attendre à ce qu'elles soient distantes dans l'espace vectoriel. Nous nous appuyons ici sur une notion intuitive de ce que désigne une distance. En fait, ce sont les mesures de distances qui vont déterminer ce qui fait que deux langues seront désignées comme similaires ou plus différentes. Certaines distances donnent plus de poids aux similarités, tandis que pour d'autres mesures, un écart dans une seule des dimensions peut suffire à obtenir une valeur élevée. Chaque mesure apporte donc un éclairage différent sur les relations entre les langues étant donné certains paramètres.

Distribution Pour le troisième paramètre, nous adoptons une représentation un peu différente, sous la forme d'une distribution. Cette distribution est une représentation de la répartition des séquences de trigrammes entre toutes les séquences de trigrammes possibles ("ADJ_ADJ_ADJ", "ADJ_ADJ_ADP"...). Les fréquences obtenues sommeront donc nécessairement à 1.

4.3 Mesures de distance

Nous motivons dans cette section le choix des mesures de distances qui permettent, à partir d'une représentation (vectorielle ou sous forme de distribution), d'obtenir une matrice de distance de

m^2 dimensions avec m le nombre de langues.

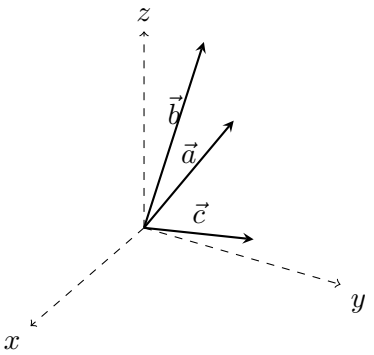
4.3.1 Distances vectorielles

Pour chacun des motifs décrit en section 4.1.3, nous pouvons créer un vecteur avec une dimension par occurrence du motif dans la langue (ou le treebank).

Ainsi, pour le paramètre de l'ordre des unités sujet, verbe et objet, nous obtenons une représentation vectorielle avec 6 dimensions (1 pour chacun des types observables). Les corpus étant de tailles différentes, nous exprimons les valeurs en terme de fréquence plutôt que de nombre d'occurrences. En effet, un nombre d'occurrences plus élevés pourrait simplement être dû à un corpus plus long. La représentation sera donc plus lisible en visualisation puisque les observations seront directement comparables (par exemple sur un histogramme).

Pour le paramètre qui concerne la proportion de liens de dépendances orientés à gauche (c'est-à-dire les linéarisations dépendant-gouverneur), la représentation est un peu différente puisque chaque dimension concerne un des motifs et encode une proportion comprise entre 0 et 1. En revanche le vecteur ne somme pas nécessairement à 1.

Nous utilisons la mesure de similarité cosinus, qui est sensible à la direction du vecteur mais pas à sa magnitude (c'est-à-dire que notre choix d'assigner des fréquences plutôt que des nombres d'occurrences pour le paramètre de linéarisation des unités sujet, verbe et objet n'aura pas d'incidence sur la valeur obtenue).



Plus cette mesure est faible (comme entre $\vec{a}$ et $\vec{b}$ plutôt qu'entre $\vec{b}$ et $\vec{c}$ dans la figure ci-dessus), plus les syntaxes de ces langues sont proches étant donné les paramètres pris en compte. La similarité cosinus est toujours comprise entre 0 et 1.

Nous rappelons ci-dessous le calcul de la similarité cosinus :

$$\cos\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Il est intéressant de noter que cette mesure est davantage sensible aux fortes différences pour un facteur qu'à de multiples petites différences. Ainsi si nous définissons 3 vecteurs :

$$a = [0.1, 0.3, 0.4, 0.3]$$

$$b = [0.2, 0.4, 0.6, 0.3]$$

$$c = [0.1, 0.3, 0.4, 0.7]$$

Si nous additionnons les différences entre les éléments de b et a d'une part, et les éléments de c et a d'autre part, nous obtenons la même différence positive de 0.4. Pourtant la similarité cosinus entre a et b est de 0.01, et celle de a et c est de 0.08.

Cette propriété est liée à la mesure, et le choix d'une mesure différente pourrait mettre en évidence d'autres relations entre les langues. Ce travail de réflexion autour des mesures est donc important, puisqu'il aura des implications sur les résultats obtenus, en mettant en avant différentes propriétés des données et en proposant une lecture parmi toutes les lectures possibles.

4.3.2 La distance comme divergence entre deux distributions

Nous avons introduit dans le chapitre 2 la théorie de l'information et et de certaines mesures qu'elle propose. Ainsi la mesure de divergence de Kullback-Leibler en particulier permet de calculer à quel point deux distribution semblent diverger, en mesurant la perte d'information attendue si nous approximons la distribution A par une distribution B. Cependant, cette mesure ne valide pas le critère de symétrie :

$$KL(A \parallel B) = x \implies KL(B \parallel A) = x$$

Cela signifie que si nous utilisons cette mesure les paires de langues n'auraient pas nécessairement la même "distance" entre elles (il ne s'agit donc pas d'une réelle mesure de distance). En

revanche, il existe une mesure basée sur cette mesure de Kullback-Leibler qui respecte cette propriété de symétrie. Il s'agit de la distance de Jensen-Shannon dont le résultat est compris entre 0 et 1 et qui s'obtient en moyennant le "coût" de passage de la distribution A à la distribution mixte de A et B, et le coût de passage de la distribution B à la distribution mixte de A et B :

$$JSD(A \parallel B) = \frac{1}{2}KL(A \parallel \frac{1}{2}(A+B)) + \frac{1}{2}KL(B \parallel \frac{1}{2}(A+B))$$

Cette mesure de distance a aussi ses particularités. Avec une similarité cosinus, deux langues qui sont distantes avec les mêmes langues sont proches entre elles. Il n'en est pas de même pour cette mesure de distance. Prenons un exemple : faisons l'hypothèse que la distribution du français est éloignée de celle du vietnamien, et que la distribution du russe est elle aussi éloignée du vietnamien. Cela ne signifie pas pour autant que français et russe auront des distributions similaires, et donc ces deux langues peuvent obtenir une distance élevée entre elles.

4.4 Des distances entre paires de langues aux groupes de langues similaires : clustering

Le clustering est une tâche de classification non-supervisée. Elle consiste à faire émerger des "paquets" (ou clusters) de données, en se basant sur des critères de proximités et de distances. Un bon clustering est censé faire émerger des paquets qui sont homogènes (les éléments à l'intérieur du paquet sont dans l'ensemble plutôt similaires) et bien distincts entre eux (les paquets sont facilement identifiables car bien séparés entre eux).

Clustering et hiérarchie Les méthodes de clustering peuvent être hiérarchiques ou non-hiérarchiques. Lorsque le clustering est non-hiérarchique, il n'y a qu'une étape de partitionnage, et chaque donnée est associée à un et un seul cluster (ce qui rappelle notamment la tâche d'étiquetage en classification supervisée). L'algorithme de clustering est dit hiérarchique lorsqu'il fait apparaître une structure qui précise des clusters à différents niveaux de granularité.

Fusion et division des clusters dans les méthodes hiérarchiques Les méthodes hiérarchiques peuvent être **agglomératives** (bottom-up) ou **divisives** (top-down). Dans le premier

cas, à l'initialisation chaque donnée appartient à son propre cluster, et à chaque itération, un ou plusieurs clusters sont fusionnés entre eux pour former un cluster plus grand, jusqu'à obtenir un unique amas qui rassemble toutes les données. Nous obtenons ainsi une structure qui explicite les différentes étapes de fusion (merging) des paquets. L'approche top-down quant à elle est initialisée avec un grand cluster qui contient toutes les données, et les nouveaux clusters s'obtiennent par séparation du cluster initial.

Clustering et contraintes sur le nombre de clusters Certaines méthodes de clustering nécessitent de prédéfinir le nombre de clusters que nous souhaitons obtenir. C'est le cas par exemple de la méthode des "k plus proches voisins" ou *k-means*, pour laquelle le paramètre k qui précise le nombre de clusters à obtenir doit être défini. En revanche, pour d'autres méthodes le nombre de clusters finaux "optimal" fait partie des paramètres à estimer pendant l'apprentissage.

4.4.1 Évaluation des clusters obtenus

La tâche à laquelle nous sommes confrontés est une tâche de clustering non-supervisée, c'est-à-dire que nous ne disposons pas d'étiquettes de référence qui nous permettraient d'observer à quel point les clusters obtenus sont cohérents avec les classes des éléments. Il est donc beaucoup plus complexe d'évaluer les résultats.

Une possibilité consiste à regarder la silhouette des clusters, afin de voir si ils sont plutôt homogènes ou chaotiques. Une mesure comme le coefficient de silhouette [Rousseeuw, 1987], qui calcule le ratio entre la moyenne des distances intra-cluster et la moyenne des distances entre chaque élément et le cluster le plus proche peut être utilisée dans cette optique. Dans ce cas, le score obtenu est compris entre -1 et 1, avec -1 pour un mauvais clustering et 1 pour un clustering dans lequel les clusters sont très denses, c'est-à-dire que la séparation entre les clusters est bonne et qu'il n'y a pas de chevauchements.

Chapitre 5

Résultats

Reproductibilité Le format de ce document ne nous permet pas de présenter systématiquement l'ensemble des résultats sur tous les treebanks et les langues (tant en terme de tableaux que de visualisations). Les résultats complets (données, graphiques et scripts) sont mis en ligne à l'adresse https://github.com/marinecourtin/syntactic_distances.

5.1 Ordre de mot

5.1.1 Linéarisation des catenas <s,v,o>

Nous commençons par extraire les motifs pour chaque langue. Puis, nous calculons pour toutes les combinaisons de langues la distance cosinus, entre les représentations vectorielles normalisées des ordres de mot. Nous obtenons donc une matrice de distances entre les paires de langues. Afin de visualiser cette matrice, nous utilisons une représentation graphique de type "heatmap", c'est-à-dire une carte qui fait varier la couleur des cellules selon l'intensité de la valeur. Dans notre exemple, plus la couleur de la cellule est foncée plus la distance entre les représentations vectorielles des deux langues est importante. Au contraire, deux langues qui ont des représentations vectorielles similaires se verront attribuer une cellule blanche ou bleu claire. La diagonale correspond à des distances nulle puisque les deux représentations vectorielles comparées représentent la même langue. Nous reproduisons également les mêmes étapes pour mesurer cette fois-ci les distances entre treebanks, dont nous trouvons une illustration dans la figure 5.1. Les graphiques ainsi créés sont très riches en informations et permettent de relever des régularités

en regardant l'ensemble des langues. Par exemple les "barres" violacées autour des treebanks du japonais, coréens et kazakh indiquent que ces langues sont généralement distantes des autres. Au contraire de grandes zones blanches ou bleu clair indiquent des groupes de treebanks assez similaires vis-à-vis du paramètre étudié.

Par soucis d'espace disponible, nous focalisons notre attention sur l'étude d'un groupe de langues : les langues romanes. Cet échantillon contient 2 langues plus anciennes, le latin et l'ancien français, et 7 langues modernes qui sont le français, l'italien, l'espagnol, le catalan, le galicien, le roumain et le portugais.

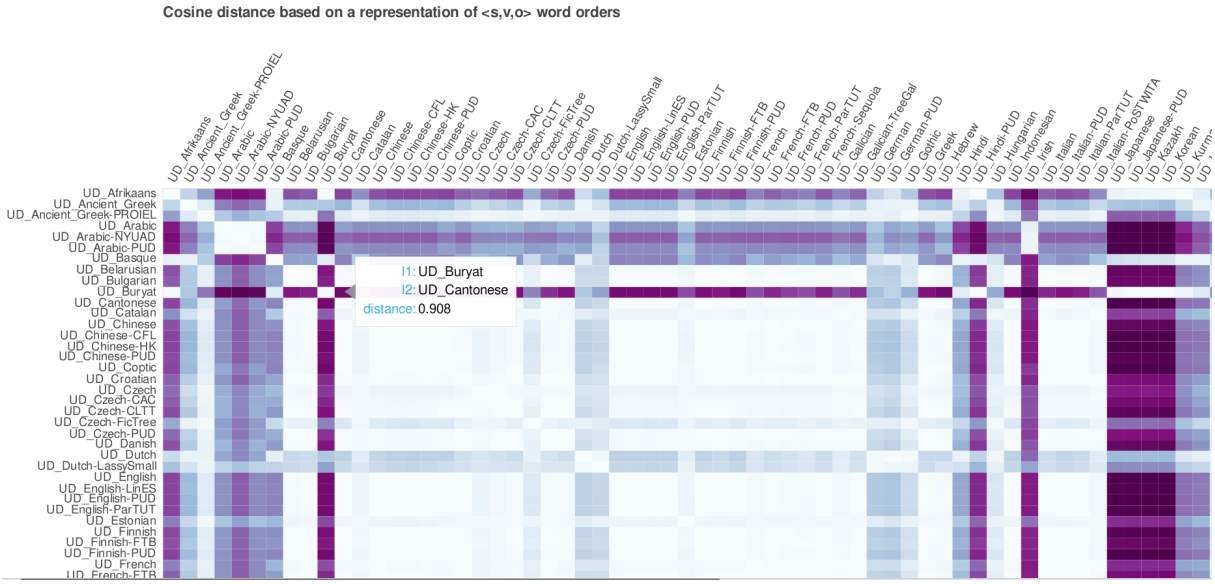


Figure 5.1: Capture d'écran montrant un aperçu de la carte interactive avec séparation en treebanks

Dans la figure 5.2, nous mesurons la similarité cosinus entre les langues romanes, à partir des représentations vectorielles construites sur les fréquences des motifs d'ordre de mot de la configuration sujet verbe et objet. Les deux paires les plus proches sont l'espagnol et le catalan d'une part (1.07e-4), et l'italien et le roumain d'autre part (6.85e-4). Nous constatons également que certains motifs plus larges se dessinent. Le français, galicien et italien, portugais et roumain forment un groupe avec cinq langues qui sont proches entre elles. De l'autre côté du spectre, le latin et l'ancien français[1] semblent présenter systématiquement des distances assez importantes avec les autres langues (les lignes et colonnes qui les concernent sont particulièrement saillantes). Latin et ancien français entretiennent davantage de proximité entre eux (2.39e-2) qu'avec les

[1]Le treebank UD_Old_French-SRCMF est constitué à partir de textes qui vont du 9ème au 13ème siècle.

autres langues. Contrairement à ce à quoi nous pourrions nous attendre, l'ancien français est plus éloigné des langues romanes modernes que ne l'est le latin. Et parmi les langues romanes modernes, le français n'est que la quatrième langue la plus proche de l'ancien français derrière l'espangol, le catalan et le galicien.

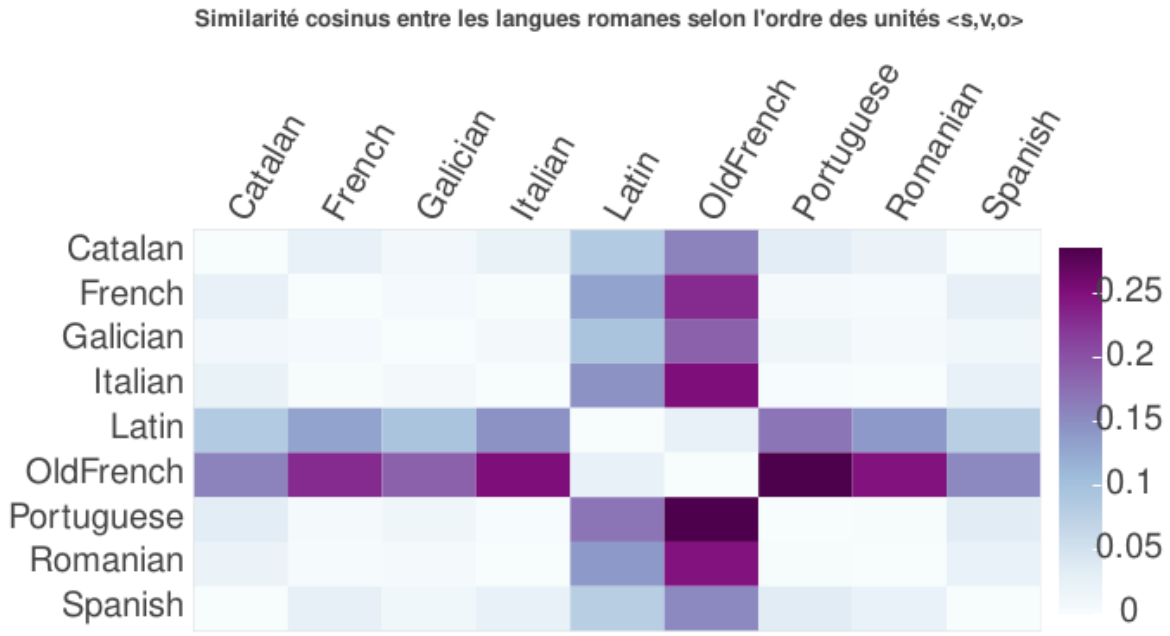


Figure 5.2: Distance cosinus entre langues romanes selon une représentation vectorielle de l'ordre des unités <s,v,o>

Ce type de visualisation est intéressant pour nous donner une vision globale de l'échantillon, mais il n'est pas adapté pour une analyse fine. La figure 5.3 nous donne un meilleur aperçu des dimensions qui contribuent à ces distances. Par rapport à l'échantillon, nous retrouvons bien les particularités de l'ancien français qui expliquent ses distances généralement élevées avec les autres langues. En effet, c'est la langue pour lequel le motif dominant svo a la proportion la plus faible (37.12%) puisqu'il n'atteint pas même la moitié des motifs rencontrés, ce qui est également le cas du latin (42.18%). La distance réduite entre le latin et l'ancien français s'explique surtout par leur partage de cette caractéristique : les deux langues ont une faible proportion de motifs svo. Cette différence s'accompagne d'un plus fort taux de motifs sov, avec tout de même une différence de 9.61 points entre les deux. C'est probablement sur ce motif que se joue le rapprochement de l'ancien français avec le galicien, l'espagnol et le catalan. Le portugais est la langue la plus fortement svo avec 84.13% de motif de ce type. Italien et français ont des proportions très semblables. De la même façon, les différences entre catalan et espagnol sont à peine perceptibles.

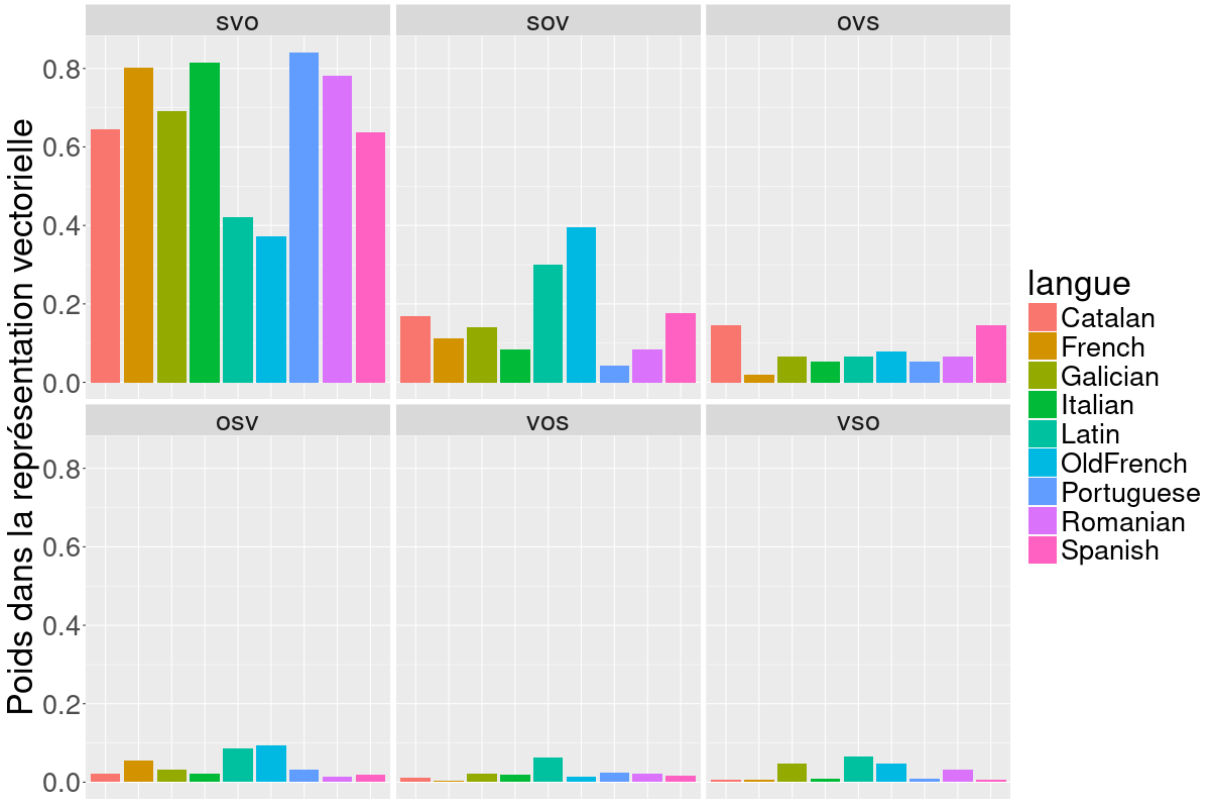


Figure 5.3: Poids des ordres de mots <s,v,o> dans les représentations vectorielles des langues romanes

Réduction de dimensionnalité Il existe des méthodes de positionnement multidimensionnel (MDS) qui permettent de représenter des données multi-dimensionnelles dans un nombre de dimensions réduit. La compréhension et l'utilisation de ces techniques permet de parvenir à des visualisations qui mettent en lumière des relations entre les données qui seraient autrement difficile d'accès pour les humains, puisque notre capacité à conceptualiser simultanément de multiples dimensions est limitée. Ces méthodes appliquent une transformation sur l'espace vectoriel, visant à minimiser la déformation des distances entre les données, c'est-à-dire à résumer de la façon la plus juste possible les données initiales. L'analyse en composante principale (ACP) est un sous-type de MDS qui utilise uniquement des transformations linéaires [2] et cherche "un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes et ces plans permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-à-dire avoir des images les moins déformées possibles." [3]

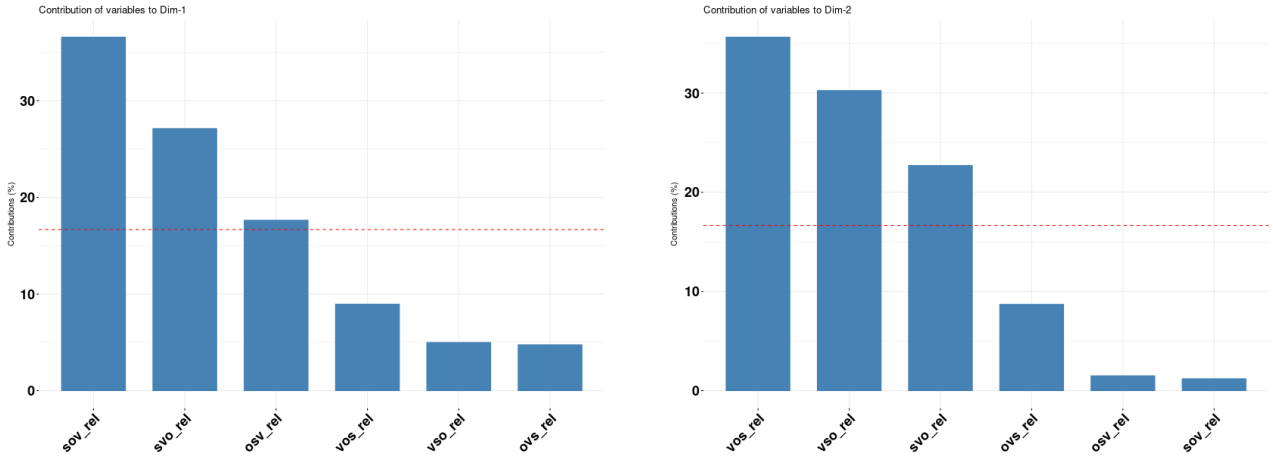
[2] D'autres méthodes de MDS utilisent des transformations non-linéaires.

[3] Extrait d'un support de cours rédigé par Camille Duby et Séphane Robin : http://www2.agroparistech.fr/IMG/pdf/AnalyseComposantesPrincipales-AgroParisTech.pdf

Il y a en fait deux analyses possibles : la première prend en entrée la matrice de n langues sur m variables, c'est-à-dire que chaque langue sera positionnée par rapport à son comportement vis-à-vis de ces variables qui représentent les fréquences relatives d'ordre de mots. Cependant, ce nombre de variable est limité, donc il peut-être intéressant d'observer une autre projection, avec en entrée la matrice de distances cosinus de dimensions n sur n. Dans ce deuxième cas, c'est le comportement d'une langue vis-à-vis de toutes les autres langues de l'échantillon qui sera caractérisé. Cette deuxième méthode est plus adaptée pour découvrir des clusters, puisque ces derniers émergent mieux lorsqu'il y a une certaine homogénéité au sein du groupe.

Pour ces analyses, nous utilisons les bibliothèques R *factoextra* [Kassambara and Fabian,] et *FactoMineR* [Husson et al.,], ainsi que la bibliothèque python [Pedregosa et al., 2011] [Pedregosa et al., 2011] qui implémentent des fonctions de réductions de dimensionnalité afin de produire des représentations visuelles en deux et trois dimensions des matrices de distance. Les scripts manipulant ces bibliothèques sont mis en ligne sur https://github.com/marinecourtin/syntactic_distances.

Nous commençons par la première analyse. Puisque le nombre de variables est limité, nous pouvons observer leur contribution respective dans les deux dimensions sélectionnées (aussi appelées composantes principales) comme sur la figure 5.4. Nous pouvons voir que le positionnement des langues sur l'axe des abscisses va surtout se jouer sur la présence des motifs sov et svo, tandis que le positionnement sur l'axe des ordonnées dépend de la fréquence des motifs vos, vso et également svo.



(a) contribution des variables au PC1

(b) contribution des variables au PC2

Figure 5.4: Contribution respective des variables aux deux dimensions sélectionnées

Une fois les représentations des ordres de mot projetées sur le nouvel espace vectoriel, nous obtenons la répartition de la figure 5.5, avec la direction des vecteurs représentant les variables en noir. Ainsi plus une dimension est fortement activée pour une langue, plus le point la représentant s'éloignera de l'origine dans la direction du vecteur représentant cette dimension.

Pour cette transformation, le programme mesure également la quantité de variation exprimée dans les deux dimensions choisies. Plus la quantité de variation exprimée est élevée, moins la réduction de dimensionnalité à entraîné une déformation des données initiales. La transformation entraîne une forte déformation puisque seulement 63.6 % des variations sont conservées. La variable *cos*2 décrit la qualité de représentation d'une langue sur le graphique en deux dimensions, sachant qu'une langue mieux représentée aura une couleur orangée tandis qu'une langue pour laquelle la transformation entraîne une plus grande déformation sera associée à une couleur bleutée. Nous pensons que cela est dû au trop faible nombre de variable, ainsi qu'aux corrélations positives et négatives fortes entre les variables (voir figure 5.6), qui entraîne une polarisation (par exemple entre sov et svo) qui rend difficile la visualisation des variations d'ordre de motà l'intérieur d'une langue. En effet la direction presque opposée des vecteurs représentant les variables sov et svo est dûe à une corrélation négative entre les deux motifs, qui sont souvent dominants dans les langues de cet échantillon. Cela entraîne un positionnement des langues principalement sur cet axes, avec peu de prise en compte des autres motifs, qui pourraient être pertinents pour rapprocher les langues similaires.

Nous passons donc à la seconde analyse, où nous essayons de résumer les données en étudiant cette fois-ci le comportement de chaque langue vis-à-vis des autres langues. Pour cela nous donnons en entrée à l'algorithme la matrice de distances basées sur la similarité cosinus.

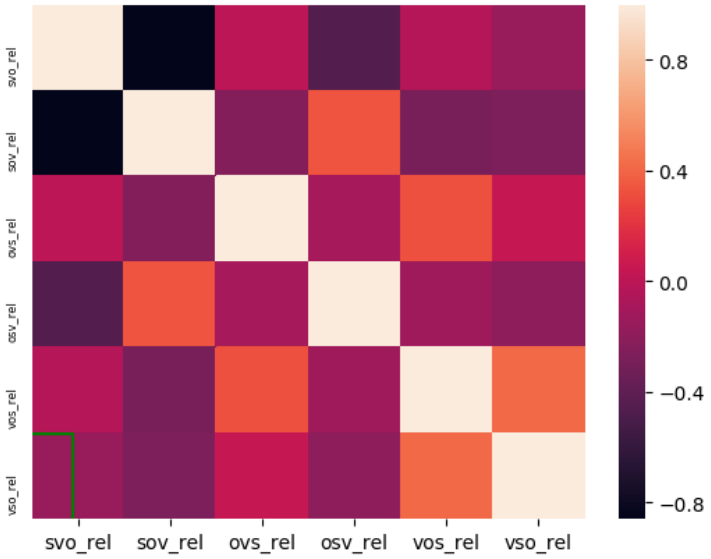


Figure 5.6: Corrélations positives et négatives entre la présence des différents ordres de mots

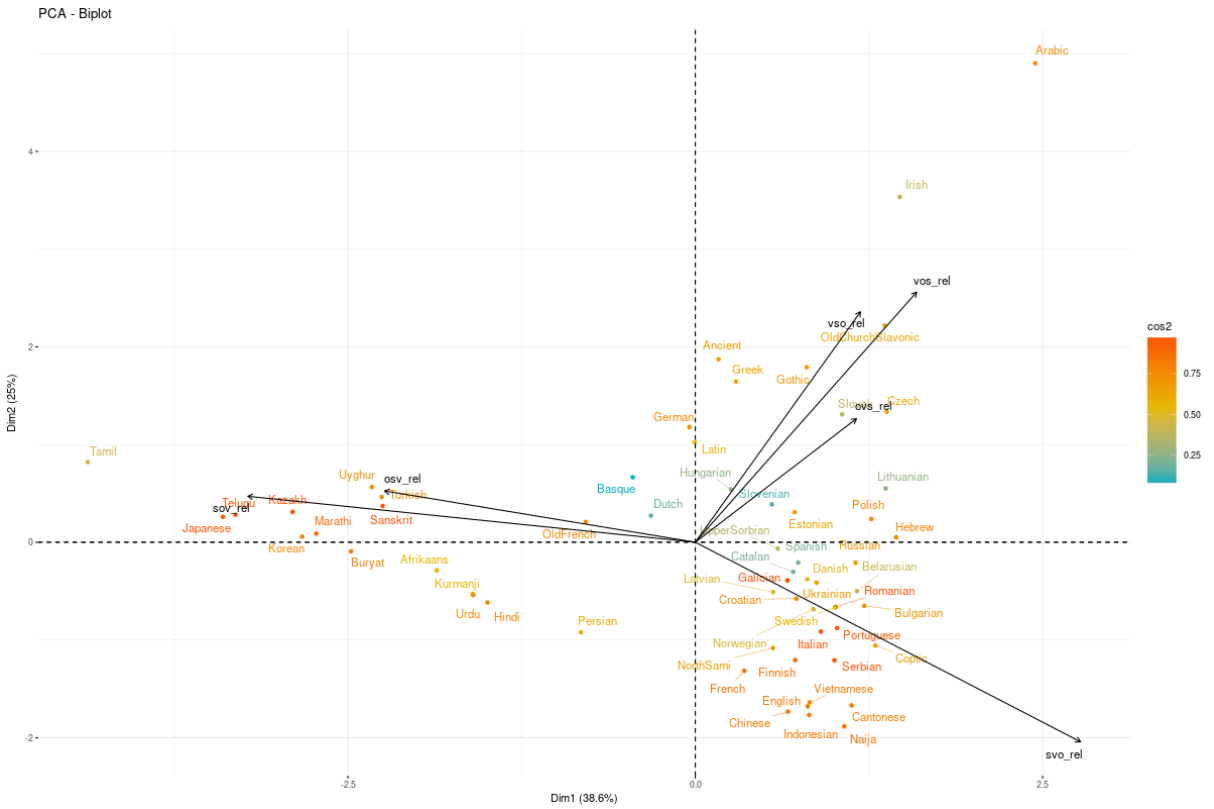


Figure 5.5: ACP des langues depuis l'observation des ordres de mot

La figure 5.7 permet de résumer les observations que nous avons faites sur les langues romanes, en utilisant une réduction de dimensionnalité qui prend en entrée la matrice de distance des langues romanes. Ici, nous obtenons une quantité de variations représentant 99.79% de la variations dans les données originales, ce qui signifie que les distances sont beaucoup moins déformées que dans

le cas précédent où nous avions positionné les langues selon leur comportement par rapport aux motifs. Nous remarquons que lorsque les langues romanes sont positionnées en fonction de leur distance avec toutes les autres langues, le regroupement du portugais et du galicien avec l'italien, le français et le roumain n'est plus aussi certain.

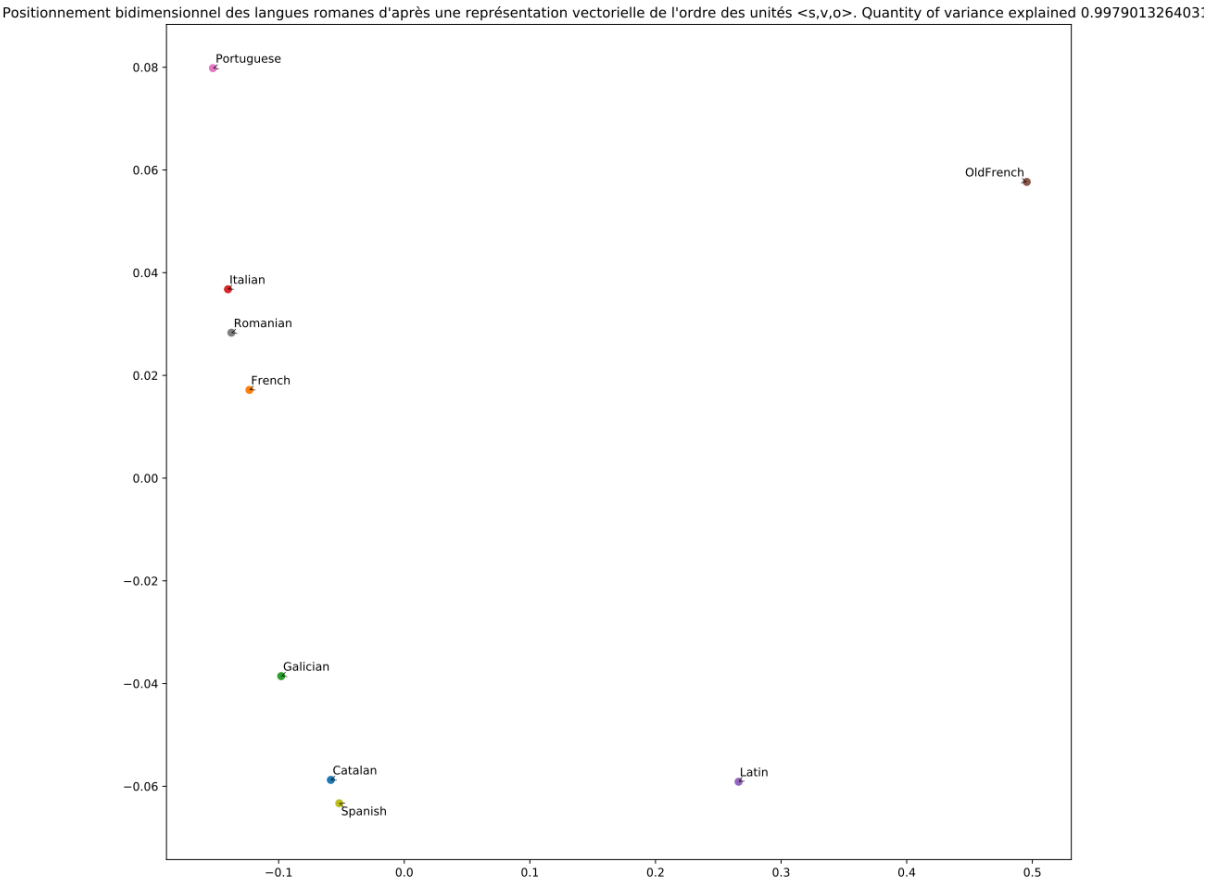


Figure 5.7: Positionnement bidimensionnel des langues romanes d'après une représentation vectorielle de l'ordre des unités <s,v,o>

Lorsque nous réitérons ces étapes pour un échantillon de langues plus variées nous obtenons la figure suivante. Pour améliorer la lisibilité, nous utilisons la bibliothèque adjustText [Flyamer, 2018] afin d'optimiser le placement des étiquettes, et reprenons les conventions graphiques de [Chen et al., 2018] : brun : langue romane, violet : balte et slave, olive : germanique, bleu : autres langues indo-européenne, vert : langues sinitiques et austronésiennes, rouge : langues agglutinantes, noir : langues sémitiques ou non-classées. Nous commençons plus nettement à voir des groupes se dessiner, avec quelques amas dont certains ressemblent à des amas de langues apparentées: indo-iraniennes : kurmanji, urdu; grec ancien, latin et grec; altaïques (sauf le telugu) : japonais, coréen, kazakh, telugu et ouïghour, et quelques langues plus isolées comme le tamil, l'arabe ou l'irlandais. Le français est dans une position intéressante,

puisqu'il se trouve entre l'ancien français et un amas qui réunis des langues indo-européennes (germanique : anglais; créole à base d'anglais : naija; uralique : same du Nord, finnois; sémitique : copte, hébreu; sino-tibétain : chinois, mandarin, cantonais). Pour cet échantillon, l'ACP avec sélection des deux meilleures dimensions permet d'exprimer 98.84% des variations présentes dans l'échantillon de départ.

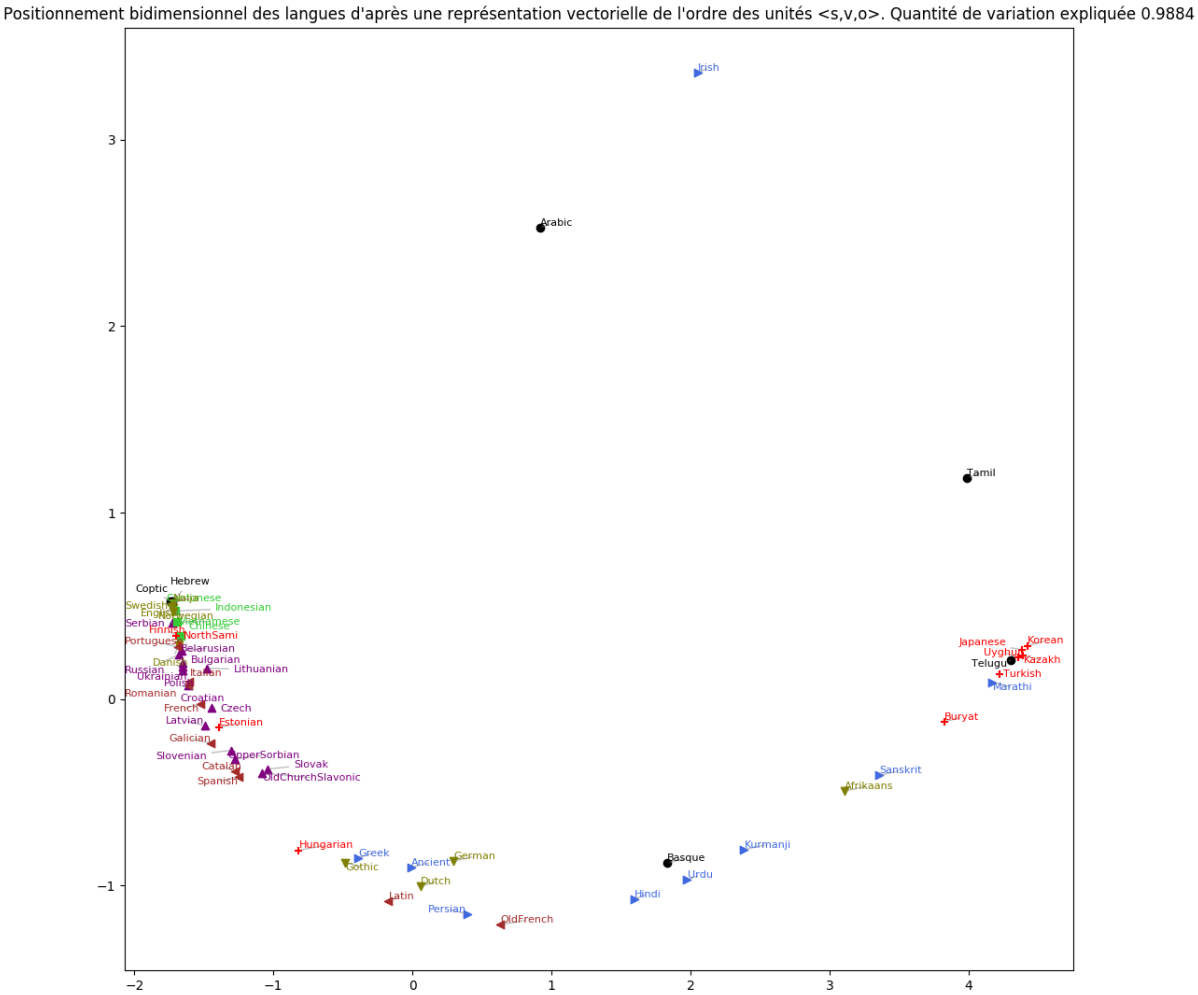


Figure 5.8: Positionnement bidimensionnel des langues d'après une représentation vectorielle de l'ordre des unités <s,v,o>

Évaluation et clustering Jusqu'ici nous avons mesuré des distances en nous basant sur un treebank "global" construit à partir de la fusion de tous les treebanks disponibles pour une même langue. Nous n'avons donc pas pris en compte le fait qu'il y a parfois plusieurs treebanks pour une même langue, et que la comparaison des résultats pour ces treebanks constitue un bon indicateur qui permet d'une part l'évaluation des mesures (deux treebanks d'une même langue devraient être proches), et d'autre part l'évaluation de l'influence d'autres sources de variations,

notamment du genre de texte ou de sa modalité.

Nous proposons donc de réitérer ces mesures en gardant les treebanks séparés, et d'observer les résultats d'un clustering non-supervisé hiérarchique agglomératif. Nous choisissons un seuil de 0.2 et découpons le dendrogramme obtenus en plusieurs dendrogrammes comme celui de la figure 5.9. Comme nous pouvons le constater, l'algorithme de clustering n'a pas entièrement réussi à rapprocher les treebanks du japonais ou ceux du turc. La distance indiquée en ordonnée correspond à la distance maximale entre les items de chaque cluster[4]. La fusion retardée entre les treebanks du japonais – même si ceux-ci sont très proches avec une distance à $4.64e{-}3$ – s'explique par la plus faible distance entre le treebank du kazakh et le treebank japonais-PUD $1.53e{-}3$. Il semblerait donc que le kazakh et le japonais soient très similaires en termes d'ordre de mot, puisque tous les deux seraient très fortement sov avec quelques occurrences du type osv. Pour le turc, la distance entre les deux treebanks est également faible, mais dix fois plus grande que pour les treebanks du japonais ($1.11e{-}2$), tandis que le treebank turc est proche de l'ouïghour avec $1.51e{-}3$, et turc-PUD très proche du treebank coréen ($3.89e{-}4$).

[4]Méthode "complete" dans R. Beaucoup d'autres méthodes sont disponibles ("centroid", "ward" etc...), et les résultats varient considérablement selon la méthode choisie. Toutes les méthodes ne se valent pas, et certaines sont plus appropriées selon l'usage que nous souhaitons en faire.

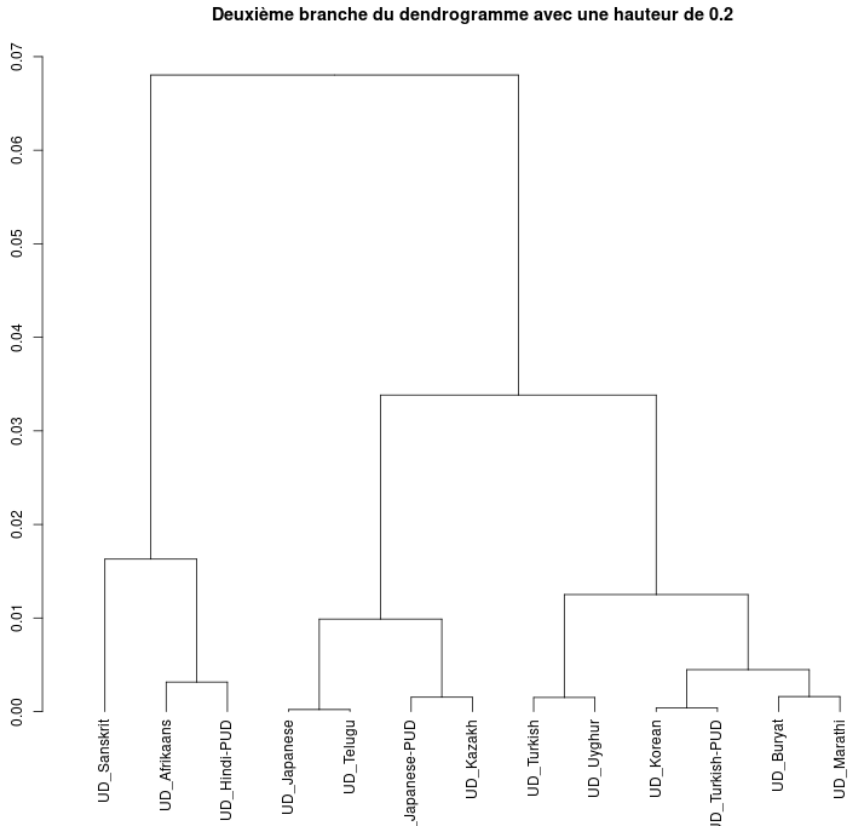


Figure 5.9: Dendrogramme résultant d'un clustering basé sur des similarités dans l'ordre des mots

L'algorithme réussi parfois à rassembler les treebanks d'une même langue, avec des scores de confiance très forts pour l'arabe (ce qui n'est pas étonnant vu sa position isolée sur le graphique sur les figures 5.5 et 5.8.

Distance moyenne entre les treebanks d'une même langue Nous mesurons également la distance moyenne entre les treebanks d'une même langue. Pour chaque langue pour laquelle sont disponibles de multiples treebanks, nous calculons la similarité cosinus, entre toutes les combinaisons de ces treebanks et divisons la somme de ces distances par le nombre de combinaisons. Les résultats de ce calcul sont présentés dans la table 5.1.

Ces mesures, lorsqu'elles sont élevées, peuvent être indicatives de plusieurs choses : il peut y avoir une source de variation par aux ordres de mots rencontrés dans les textes utilisées pour constituer les treebanks, ou alors certaines incohérences peuvent subsister entre l'annotation intra-langue des corpus arborés. Nous remarquons notamment que le portugais, l'ancien grec et le finnois sont particulièrement affectées. Pour l'ancien grec cela peut se comprendre puisque

langues	distance-moyenne
Ancient_Greek	0.132593
Arabic	0.011048
Chinese	0.058559
Czech	0.092939
Dutch	0.071684
English	0.024268
Portuguese	0.260037
Romanian	0.081124
Russian	0.036550
Slovenian	0.063450
Spanish	0.081049
Swedish	0.016202
Turkish	0.010187
French	0.073217
Finnish	0.110363
Galician	0.067484
German	0.024218
Hindi	0.002137
Italian	0.076877
Japanese	0.004641
Latin	0.034842
Norwegian	0.013531

Table 5.1: Table des distances moyennes entre treebanks d'une même langue

ce corpus intègre de la diachronie, et provient de textes littéraires, ce qui pourrait contribuer à expliquer de telles variations dans l'ordre des mots. En revanche le latin pour lequel ces éléments sont également vrais est beaucoup moins affecté, donc peut-être existe-il une autre explication.

Il est intéressant de noter que l'arabe et l'allemand ne sont pas les langues pour lesquelles les distance entre les treebanks sont les plus faibles (hindi, japonais, turc). Si l'algorithme a réussi à regrouper les multiples treebanks pour ces langues, ce n'est donc pas uniquement parce qu'il y a peu de variations inter-treebanks, mais également parce que leurs distances avec les autres langues sont plus élevées que pour les autres langues qui ont aussi de faibles distances entre treebanks.

Conclusion de l'expérience Nous avons étudié des distances entre langues en nous basant sur l'extraction d'unités syntaxiques appartenant au type des *catenas*. Les catenas visées renseignaient des informations sur la linéarisation des unités verbe, sujet et objet, et les représentations vectorielles nous ont permis d'observer des distances à la fois macro (entre des langues très variées) et plus micro (entre des langues apparentées). L'évaluation de ces mesures en com-

parant les treebanks d'une même langue entre eux montre que d'autres types de variation (genre de texte, période, choix d'annotations) peuvent affecter les résultats et rendre deux treebanks d'une même langue plus éloignés que deux treebanks de langues différentes mais proches.

5.1.2 Proportion des linéarisations dépendant-gouverneur

Nous avons ensuite voulu tester l'hypothèse d'Hawkins mentionnée dans le paragraphe 2.2.1.2, sur une harmonisation des proportions de linéarisation dépendant-gouverneur entre les constituants principaux. Pour cela nous avons regardé 4 types de motifs : les verbe-auxiliaire, objet-verbe, adjectif-nom et nom-adposition. Les résultats, présentés dans la table 5.3 ne semblent pas soutenir cette hypothèse, puisque dans la plupart des cas la proportion de linéarisations dépendant-gouverneur n'est pas cohérentes pour les 4 types de motif. Les seuls exemples qui ont des proportions régulières sont les exemples qui sont quasi systématiquement linéarisés avec le gouverneur à droite du dépendant (japonais, kazakh, tamil, ouïghour) et l'arabe qui opte pour la solution contraire. Nous ne poursuivons donc pas cette hypothèse, en revanche nous allons étudier les distances entre langues selon les préférences de linéarisations de ces configurations.

Une difficulté se pose puisque nous ne pouvons comparer que les langues pour lesquelles le motif existe avec au moins une linéarisation, puisque lorsque le motif n'est pas présent, il n'y a aucun sens à chercher à calculer la proportion des linéarisations. Cela exclut 9 langues pour lesquelles au moins 1 des motif est absent des treebanks : copte (adj-nom)[5], indonésien (verb-aux)[6], irlandais (verb-aux, même commentaire que pour l'indonésien), coréen (verbe-aux), letton (verbe-aux) [7], same du Nord (verb-aux) [8], sanskrit (verbe-aux, même commentaire que pour l'indonésien et l'irlandais), telugu (verbe-aux) et vietnamien (verbe-aux, même commentaire que pour le same du Nord).

La matrice de distances cosinus, nous permet surtout d'observer les langues qui adoptent des stratégies similaires comme le tchèque et le biélorusse ($8.42e{-}3$) ou opposées l'hébreu et le cantonais (0.99) (voir figure 5.10), ou encore le naija et kurmanji (0.96).

[5]Le treebank du copte possède bien quelques occurences de la relation des modifieurs adjectivaux, mais les unités sont étiquetées comme des noms

[6]il y a bien des auxiliaires dans le treebank de l'indonésien, mais la relation choisie est cop

[7]Le treebank a des instances de la relation aux mais les unités sont étiquetées comme des verbes

[8]Le corpus possède à la fois des unités étiquetées comme auxiliaires mais en relation de copule avec le verbe, et des verbes dans une relation d'auxiliaire vis-à-vis d'un autre verbe

Langue	verbe-auxiliaire	adjectif-nom	objet-verbe	nom-adposition
Afrikaans	0.439983	0.986607	0.76831	0
AncientGreek	0.367647	0.459446	0.451799	0.0284063
Arabic	0.00818554	0.0110467	0.0122237	0.00174468
Basque	0.913062	0.163778	0.807126	1
Belarusian	0.075	1	0.197393	0.00155763
Bulgarian	0.0169119	0.968989	0.160684	0
Buryat	1	0.984991	0.958204	1
Cantonese	0.538462	1	0	0.333333
Catalan	0.00647361	0.211655	0.288591	0.00524371
Chinese	0.11498	0.99799	0.0642263	0.120858
Croatian	0.322311	0.99037	0.189662	0.000211134
Czech	0.161973	0.930792	0.275077	0.000114026
Danish	0.0100104	0.971064	0.112419	0.00629774
Dutch	0.0911558	0.993207	0.487006	0.00363742
English	0.00280939	0.979375	0.0332171	0.00120374
Estonian	0.0342988	0.994564	0.294126	0.74929
Finnish	0.00231029	0.995909	0.231899	0.85014
French	0.000345781	0.29355	0.163218	8.06697e-05
Galician	0	0.209965	0.161309	0.00695696
German	0.3436	0.997506	0.456497	0.00704676
Gothic	0.980198	0.505848	0.362322	0
Greek	0.367647	0.709886	0.429222	0.0229834
Hebrew	0	0	0.0324119	0.000808407
Hindi	1	0.99924	0.795714	0.999035
Hungarian	0.655172	0.999522	0.407625	1
Italian	0.000185684	0.320405	0.130617	0
Japanese	1	1	1	1
Kazakh	1	1	1	0.991379
Kurmanji	1	0.0348432	0.702991	0.314086
Latin	0.816035	0.477027	0.459062	0.000489716
Lithuanian	0	0.992126	0.274376	0
Marathi	0.940678	1	0.921569	1
Naija	0	1	0.00567644	0
Norwegian	0	0.944952	0.0538331	0.00254473
OldChurchSlavonic	0.973727	0.213405	0.26703	0.00800291
OldFrench	0.0935252	0.736772	0.545826	0.000375799
Persian	0.00187617	0.0648787	0.573349	0.000200615
Polish	0.795591	0.652367	0.147739	0.00283956
Portuguese	0.000709975	0.295399	0.107509	0.000566342
Romanian	0.00799526	0.114978	0.184464	0.00095673
Russian	0.0310782	0.984422	0.172978	0.00164364
Serbian	0.362291	0.998313	0.0865424	0
Slovak	0.43712	0.964633	0.329675	0.000150083
Slovenian	0.0650699	0.980304	0.333958	0.000339703
Spanish	0.00754212	0.262716	0.272165	0.00373317
Swedish	0.000422476	0.992574	0.0405826	0.00526184
Tamil	0.996248	0.998198	0.998597	1
Turkish	1	0.993597	0.962213	0.991223
Ukrainian	0.298343	0.95886	0.198445	0.000539084
UpperSorbian	0	0.994485	0.316916	0
Urdu	1	0.994526	0.792678	0.987978
Uyghur	0.995434	1	0.99897	1

Table 5.3: Proportion des linéarisations dépendant-gouverneur pour 4 types de catenas

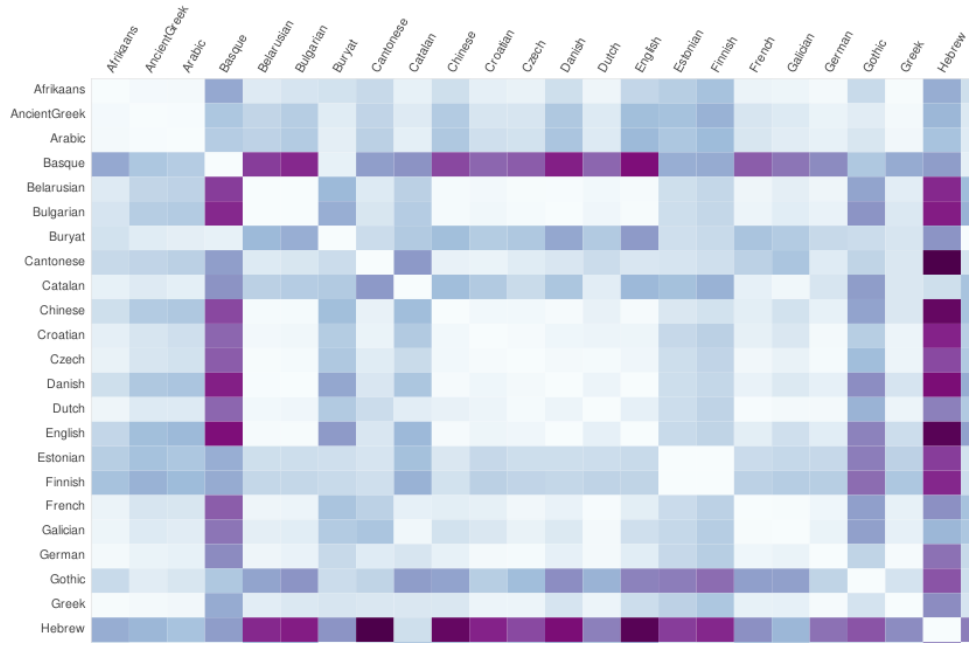


Figure 5.10: Visualisation des distances entre un échantillon de langues à partir de la proportion de linéarisation dépendant-gouverneur dans les relations verbe-auxiliaire, objet-verbe, adjectif-nom et nom-adposition

Comparé au paramètre précédent, celui-ci relève beaucoup moins de fortes similarités ou de fortes différences, une grande partie des distances sont aux alentours de 0.4 (en bleu clair). Le paramètre est donc moins polarisant, il y a probablement plus de petites différences réparties sur plusieurs dimensions qu'une différence importante dans une dimension, puisque c'est ce dernier critère qui donne de grande distance avec la similarité cosinus comme nous l'avons souligné dans le chapitre précédent.

Notons que ce ne sont pas exactement les mêmes groupes qui ressortent avec ce paramètre qu'avec le précédent. Le latin adopte une stratégie assez différente des autres langues, et n'est pas suivi par l'ancien français qui rejoint le groupe du français. De plus, cette fois-ci, le roumain est groupé avec l'espagnol et le catalan plutôt qu'avec le français, le galicien et l'italien, et le portugais comme auparavant.

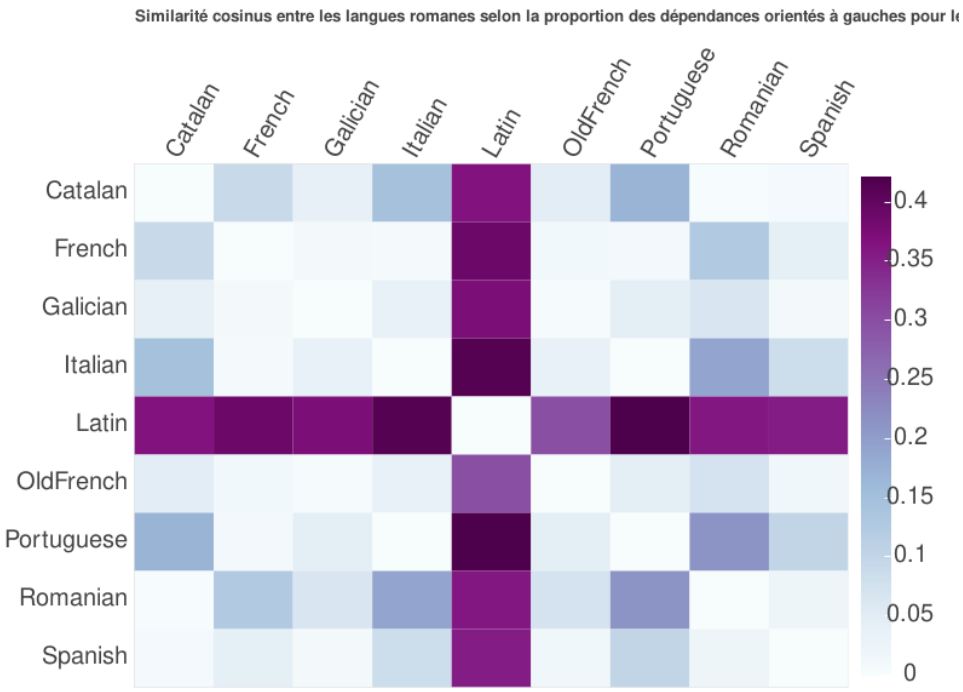


Figure 5.11: Visualisation des distances entre les langues romanes à partir de la proportion de linéarisation dépendant-gouverneur dans les relations verbe-auxiliaire, objet-verbe, adjectif-nom et nom-adposition

Caractéristiques du latin Cette différence entre le latin et le reste des langues s'explique assez facilement puisque cette langue présente un haut taux linéarisation verbe-auxiliaire c'est-à-dire dépendant-gouverneur), avec respectivement 81.60%, contrairement aux autres langues romanes dans lesquelles ces linéarisations sont quasiment inexistantes (voir figure).

En revanche, l'ancien français se remarque par son haut taux de linéarisations objet-verbe (54.58%), même si cette construction existe dans les langues romanes modernes avec des fréquences allant de 10.75% pour le portugais à 28.86% pour le catalan (et 45.91 % pour le latin). C'est également le cas pour la linéarisation adjectif-nom qui est plus fréquente en ancien français (73.68%) et en latin (47.70%) que dans les autres langues, avec un maximum à 32.04% pour l'italien.

Il est possible que cette diminution de la fréquence des linéarisations auxiliaire-verbe, objet-verbe et adjectif-nom soit corrélée avec l'évolution dans le temps des langues romanes. Il est intéressant de noter que l'ancien français a pour deux de ces unités un comportement plus extrême que le latin, et que les langues romanes modernes sont relativement proches avec tout de même une séparation entre catalan, espagnol et roumain d'une part, et français, galicien, italien d'autre

part.

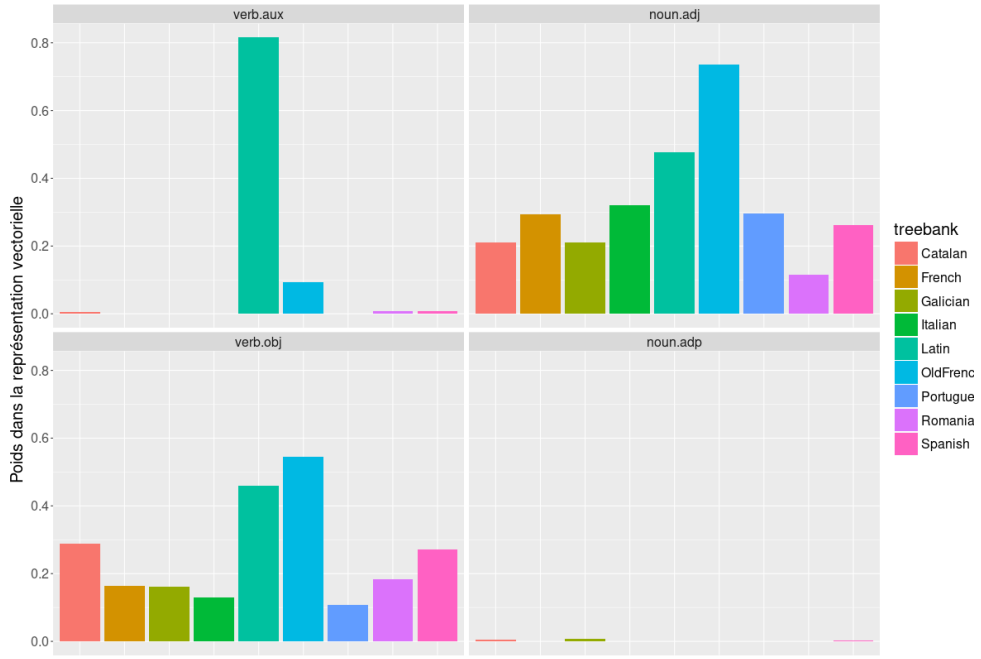


Figure 5.12: Proportions des linéarisations dépendant-gouverneur dans les relations verbe-auxiliaire, objet-verbe, adjectif-nom et nom-adposition des langues romanes.

Corrélation entre les variables Une matrice mesurant la corrélation de Pearson – une forme de mesure d'association entre deux variables, qui permet d'indiquer si elles semblent liées de façon linéaire – pour les combinaisons de ces 4 variables indique que les paires verbe-auxiliaire/objet-verbe et nom-adposition/objet-verbe sont les plus corrélées, avec des coefficients respectifs de 0.76 et 0.75, suivi par verbe-auxiliaire/nom-adposition à 0.69. En comparaison, adjectif-nom a une corrélation systématiquement faible entre 0.12 et 0.3. Cela pourrait constituer une piste pour dire que le paramètre adjectif-nom n'interagit pas tellement avec les autres variables, à l'inverse des paramètres verbe-auxiliaire, objet-verbe et nom-adposition pour lesquels nous pouvons faire l'hypothèse que ces paramètres ont des comportement liés dans les processus d'évolution de l'ordre des mots.

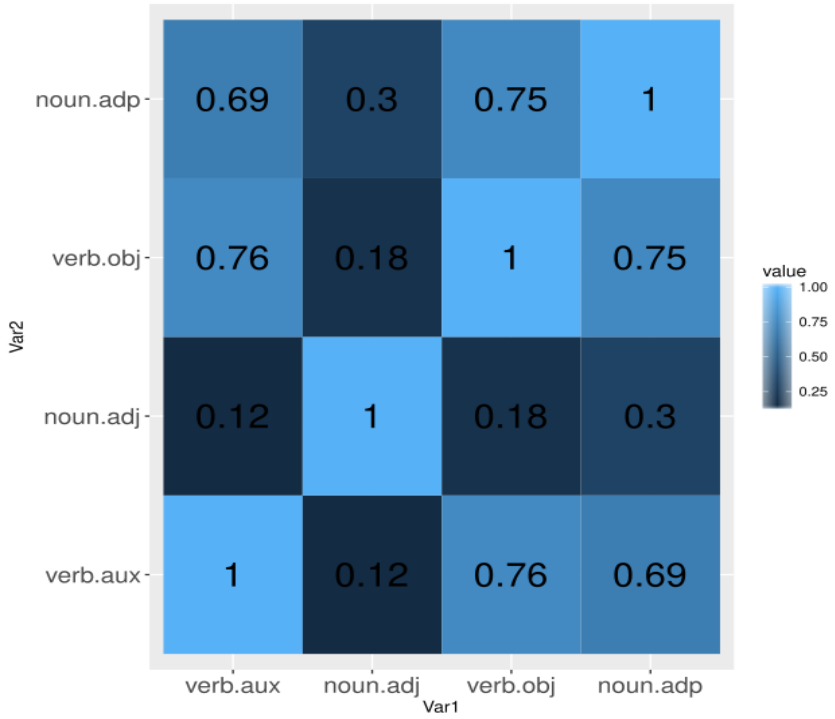


Figure 5.13: Corrélations de Pearson entre les variables représentant la proportion de linéarisations dépendant-gouverneur pour 4 types d'unités syntaxiques

Comparaison inter-treebanks Si nous regardons les distances en fonction des treebanks (figure 5.5), dans la plupart des cas, les variations pour une même langues sont très faibles, ce qui pourrait indiquer que ces paramètres sont généralement assez peu sensibles au genre des textes, ou aux variations entre écrit et oral. Nous notons cependant des variations extrêmement importantes entre les deux treebanks de l'arabe qui contiennent les 4 motifs : 62.5% de linéarisations verbe-aux dans UD_Arabic contre 0.1% dans UD_NYUAD. La comparaison entre les deux treebanks est assez difficile puisque UD_NUYAD est distribué en version dé-lexicalisé (c'est-à-dire avec uniquement les annotations, mais il faut une licence pour avoir accès aux mots), nous ne pouvons donc pas vraiment savoir à quel point cette divergence est due à des choix d'annotations différents. Il faut également relever qu'il existe des cas où les 4 motifs existent pour certains treebanks d'une langue mais pas pour tous (par exemple UD_French-PUD ne contient pas le motif verbe-aux), ce qui les exclue de ce calcul de distance à la moyenne. Les raisons tiennent en général à des choix d'annotations, notamment du rattachement des auxiliaires par le lien *cop* plutôt qu'*aux*.

langue	distance-moyenne
Arabic	0.934286
Chinese	0.097363
Czech	0.017547
Dutch	0.005408
English	0.000163
Portuguese	0.000019
Romanian	0.091346
Russian	0.002537
Slovenian	0.004201
Spanish	0.007486
Swedish	0.000206
French	0.008974
Finnish	0.000294
German	0.004034
Hindi	0.001933
Italian	0.021248
Japanese	0.000000
Latin	0.002693
Norwegian	0.004547

Table 5.5: Distance cosinus moyenne entre les treebanks d'une même langue concernant la proportion de linéarisation dépendant-gouverneur dans les couples verbe-objet, nom-adjectif, adposition-nom et auxiliaire-verbe

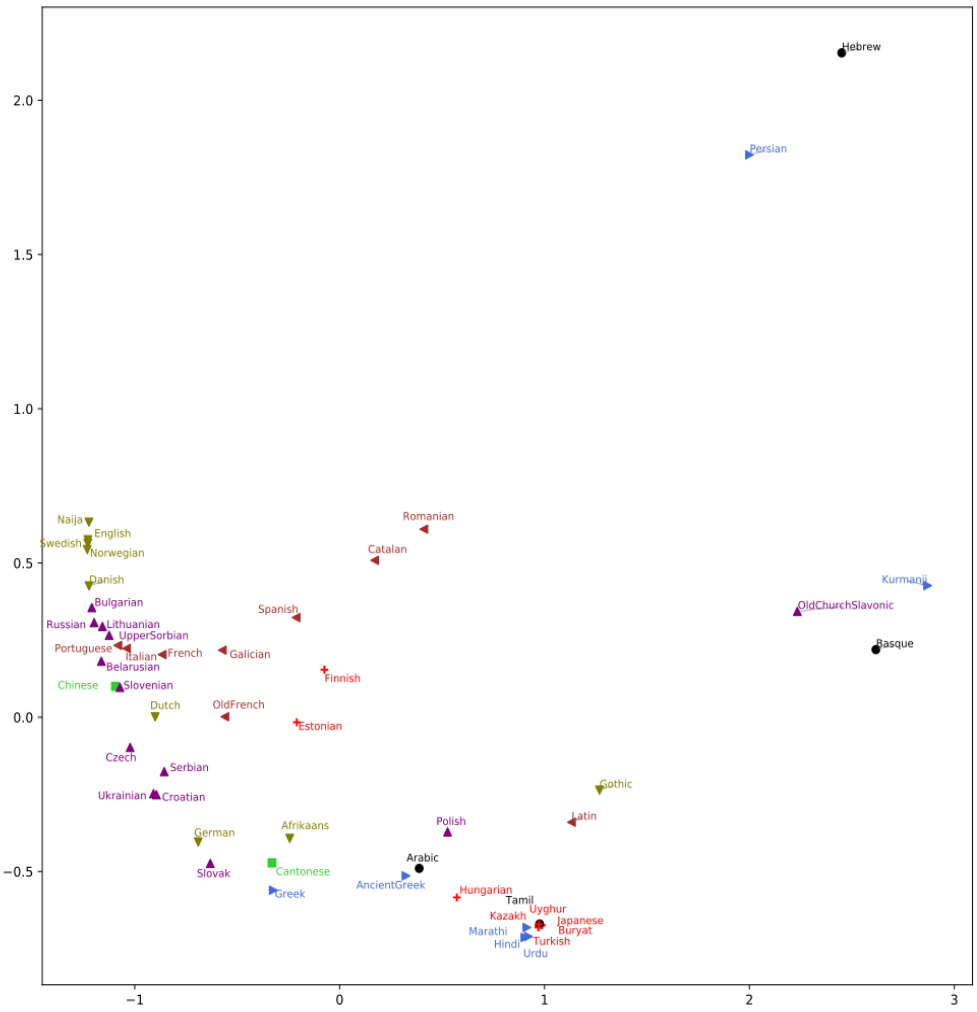


Figure 5.14: Positionnement bidimensionnel des langues selon la proportion des linéarisations dépendant-gouverneur

Dans la figure 5.14 nous pouvons voir ce que l'échantillon donne une fois placé sur un espace en deux dimensions (la transformation a conservé 85.97% des variations originales). Un petit groupe de langue se forme autour des langues altaïques (coréen, buryat, turc, japonais et kazakh) et indiennes (marathi, urdu, hindi). Dans langues comme l'hébreu ou le persan semble avoir adopté des stratégies de linéarisations assez différentes des autres langues, puisqu'elles sont seules dans le coin en haut à droite du graphe. Le naija quant à lui est groupé avec certaines langues germaniques comme l'anglais, le suédois et le norvégien. Certaines familles de langues semblent former des lignes ou des courbes, à quelques exception près, ainsi on retrouve une courbe associées aux langues romanes, de laquelle s'échappe l'ancien français et le latin qui se rapproche du gothique et du groupes des langues altaïques et indiennes.

Conclusion de l'expérience L'étude de ces paramètres semble généralement moins sensible aux variations de genre ou de modalité entre les treebanks d'une même langue. Pour s'en assurer, il faudrait investiguer davantage les distances importantes entre les treebanks de l'arabe, et dans une moindre mesure du tchèque et du chinois, afin de s'assurer qu'il n'existe pas dans ces langues des variations régulières qui viendraient expliquer des proportions de linéarisations dépendant-gouverneur différentes entre les treebanks de ces langues. Nous avons également vu que ces paramètres sont moins polarisants que les paramètres précédents, dans le sens où ils n'entraînent d'aussi grandes distances ou similarités entre langues. Nous faisons l'hypothèse que cela vient du fait que les différences entre langues sont distribuées sur les paramètres plutôt que de concerner une ou deux dimensions majoritairement.

5.1.3 Distribution de trigrammes d'étiquettes morpho-syntaxiques

Pour notre troisième expérience, nous étudions la distribution de séquences d'étiquettes morpho-syntaxiques. Plus particulièrement nous étudions des séquences de trois catégories, c'est-à-dire des trigrammes. Pour chaque combinaison possible nous initialisation 1 occurrence afin de nous permettre de mesurer des fréquences sans être bloqués par l'absence d'une séquence pour une langue.

Pour ce paramètre, nous étudions une distance adaptée aux divergences entre distributions : la distance de Jensen-Shannon. Cette distance a également une différence par rapport à la

similarité cosinus : le fait que deux langues a et b soient distantes des mêmes langues n'implique pas a et b seront proches.

Comparé aux mesures de distances jusqu'à présent, les distances sont généralement plus élevées, ce qui tendrait à indiquer que ce type de paramètre capture une partie importante du comportement syntaxique et que des langues qui par d'autres aspects pourraient être similaires, ont tout de même leurs particularités. Ce comportement est visible une fois la mesure de distance appliquée aux langues romanes, comme sur la figure 5.15. Les similarités qui étaient très fortes jusque là sont mitigées. Le latin, l'ancien français et le roumain se distinguent particulièrement des autres langues.

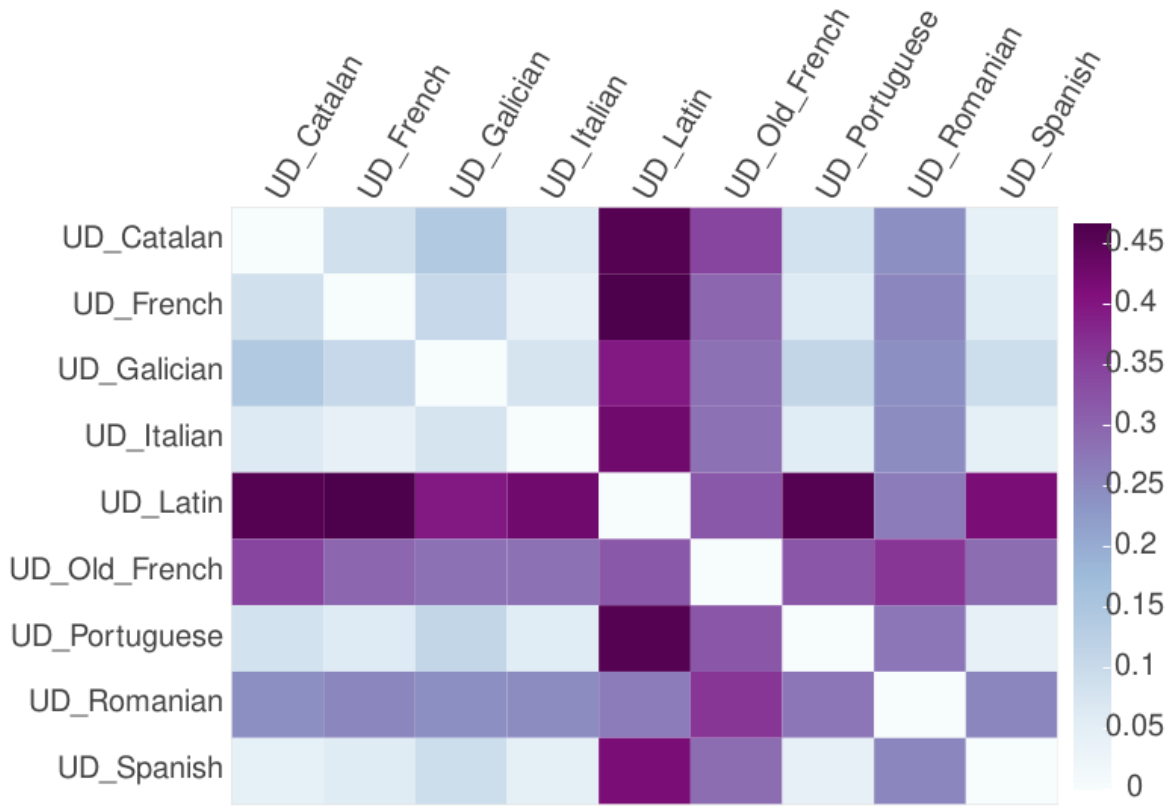


Figure 5.15: Distance de Jensen-Shannon entre les distributions de trigrammes de catégories morpho-syntaxiques des langues romanes

Nous ne pouvons pas particulièrement attribuer cela à des variations autres que la langue, du point pas davantage que pour les autres mesures. La comparaison entre treebanks retourne des distances moyennes qui sont tout à fait similaires des distances moyennes jusqu'à présent :

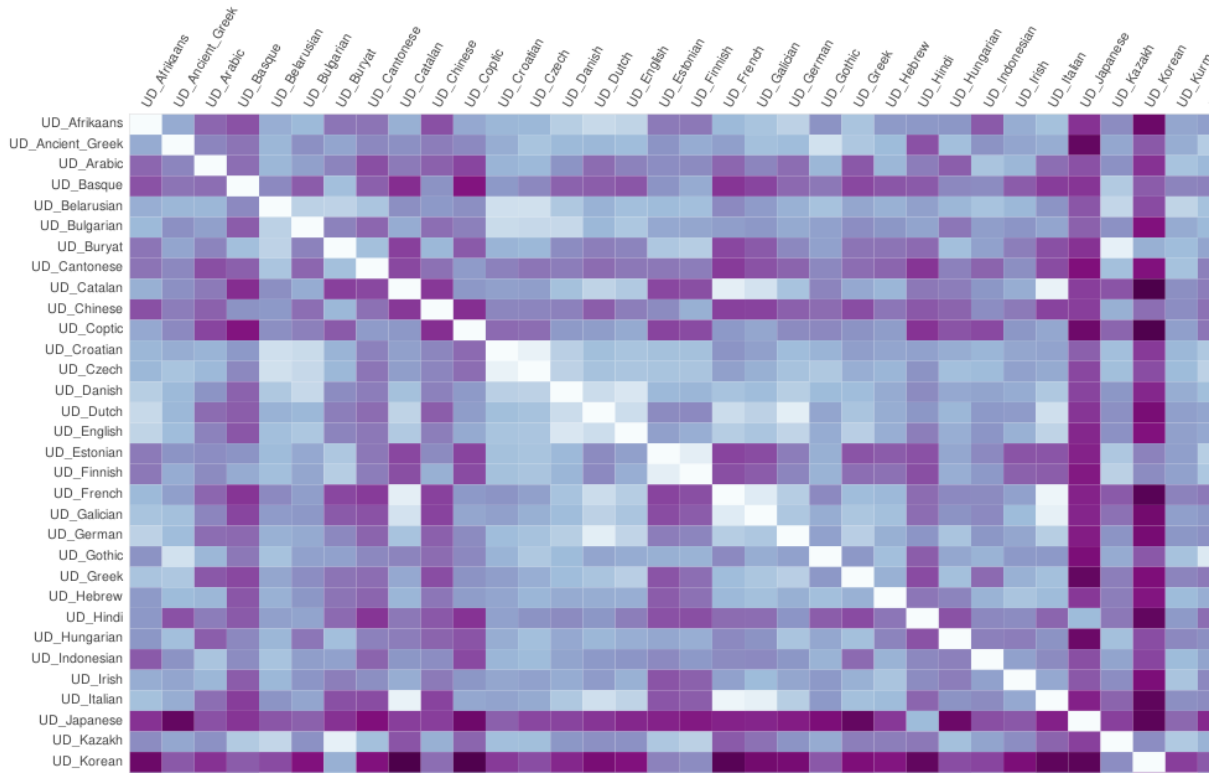


Figure 5.16: Distance de Jensen-Shannon entre les distributions de trigrammes de catégories morpho-syntaxiques pour un échantillon de langues

Il y a assez peu d'extrêmes (langues très similaires ou très éloignées). Sur la figure 5.16 coréen et le japonais ont généralement des distances élevées avec les autres langues, mais également entre elles, ce qui est plus inattendu étant donné que ces langues sont généralement considérées comme étant proches. Cela peut s'expliquer notamment par la segmentation des corpus de japonais en unités de granularité plus fine que dans les treebanks du coréen (par exemple il y a beaucoup plus d'auxiliaires qui sont considérés comme des unités à part dans les corpus du japonais, alors qu'ils font partie de l'unité verbale dans le corpus du coréen).

Les proximités lorsqu'elles existent se font souvent entre langues d'un même famille : italien/-catalan (romane), buryat/kazakh (altaïque), (uralique) finnois/same du nord. Il est également possibles que ces proximités entre langues apparentés soient aussi le reflet d'interactions entre les contributeurs travaillant sur ces langues qui cherchent souvent à harmoniser à l'intérieur d'une famille de langues, et s'appuient sur des exemples de langues proches pour construire les guides d'annotations.

Cependant, les langues qui sont proches pour ce paramètre le sont aussi pour les autres puisque cette distance semble plus exigeante envers les langues. Des trois paramètres étudiés, celui-

ci retourne les distances les moins saillantes, il y a davantage de proximités et distances en demi-mesure plutôt que des comportements très tranchés.

Conclusion de l'expérience Cette troisième expérience nous a permis de mieux nous rendre compte de comment le choix d'une mesure de distance et d'un paramètre syntaxique particulier influe sur les résultats obtenus. En particulier cette mesure retourne des distances moyennes pour la plupart des paires de langues, à l'exception de certaines paires apparentées. Une comparaison entre les treebanks de la même langue nous montre à quel point obtenir une similarité est difficile puisque la distribution varie énormément entre les différents corpus. Ce paramètre est également beaucoup plus difficile à interpréter puisqu'il y a 2744 trigrammes possibles, et que chacun d'eux ne représente qu'une très petite partie du total des tri-grammes. Cette mesure n'est donc peut-être pas la plus adaptée pour observer des comportements typologiques et apprendre quelque chose sur les proximités entre langues. En revanche, l'extraction de ces tri-grammes pourrait permettre d'observer si certains sont plus spécifiques à certaines langues, afin de repérer des motifs intéressants.

Chapitre 6

Conclusion

6.1 Résumé

Typologie et corpus arborés Nous avons montré que les descriptions typologiques peuvent désormais s'appuyer sur des corpus annotés en syntaxe. La disponibilité de grands corpus de langues variées partageant le même schéma d'annotation est une véritable avancée, qui permet de comparer les langues en extrayant des unités simples ou des configurations plus complexes, comme nous l'avons proposé dans ce mémoire. L'extraction de tels motifs nous pousse à envisager des représentations multidimensionnelles, plus à même de capturer les variations entre les langues, et qui permettent de représenter des distances et des similarités en variant les paramètres étudiés. Cela nous permet d'observer des corrélations entre certains paramètres, ce qui s'inscrit dans les buts poursuivis par les fondateurs de la typologie moderne comme Greenberg. Dans une optique de didactique, ce type de méthode pourrait être utilisé afin de mettre en avant des différences ou des similarités entre la langue d'un apprenant et une langue qu'il souhaite apprendre. Enfin, les systèmes de TAL peuvent bénéficier de ce type de connaissances pour faciliter le transfert de ressources. Il est déjà courant de transférer des ressources entre langues proches, mais nous pensons que des connaissances sur les différences qui existent entre certaines langues pourraient également être exploités, et servir à échantillonner des langues pour évaluer les systèmes sur des langues typologiquement variées.

Visualisation de données multidimensionnelles La question de la visualisation de ces distances et similarités est également une question théorique intéressant, puisqu'elle doit se faire de concert avec une réflexion sur la lecture et l'interprétation des résultats. Dans le cas de données multidimensionnelles, il est crucial de réfléchir à ce que les visualisations mettent en avant ou occultent. Les techniques de réduction de dimensionnalité peuvent nous aider à proposer des représentations plus simples et lisibles, tout en conservant au mieux les variations présentes dans les données. Nous pensons qu'il est intéressant de combiner visualisations mettant en avant des régularités macroscopiques (sur un très grand ensemble de langues) ainsi que des visualisations permettant une étude plus fine du comportement d'un échantillon de langue vis-à-vis de chaque paramètre.

Distances et similarités Nous avons volontairement limité le nombre de mesures de distances et de similarités utilisées dans ce travail, afin de focaliser notre attention sur les langues et l'étude des configurations syntaxiques choisies. Néanmoins, il existe de nombreuses possibilités à explorer, et chacune de ces mesures peut proposer une clé de lecture différente, mettant en avant certaines propriétés des données plutôt que d'autres. Il paraît également important de développer une intuition quant à ce que ces mesures représentent une fois appliquées à l'objet étudié afin de mieux exploiter tout leur potentiel.

6.2 Perspectives

Étudier d'autres types de variations Ce travail s'ouvre sur un certain nombre de questions et de pistes à explorer. Tout d'abord, une direction qui nous semblerait intéressante à poursuivre concerne l'application de cette méthode à d'autres types de variations dont la modalité, le genre ou le registre. En s'appuyant sur la littérature qui décrit des marqueurs syntaxique de ces variations, nous pourrions essayer observer la répartition de certaines configurations à l'intérieur d'une langue (entre un corpus de français écrit et un corpus d'oral spontané par exemple), afin de déterminer des spécificités propres aux corpus présentant ces propriétés. Ce travail pourrait également s'effectuer en intégrant plusieurs langues (est-ce que l'italien écrit et le français écrit se ressemblent plus que l'italien oral et le français oral ?) éventuellement dans une perspective analogique (est-ce que le français oral est au français écrit ce que l'italien oral est à l'italien écrit;

ou plus généralement, il y a-t-il des trajectoires typiques pour passer de l'oral à l'écrit, ou d'un genre à un autre ?).

Régularités entre paramètres syntaxiques Une autre possibilité pourrait être d'étudier plus particulièrement les regroupements entre langues, selon les paramètres syntaxiques choisis. Nous avons fais quelques essais de clustering, mais un travail plus axé dans cette direction, qui viserait à identifier des similarités dans les amas de langues pour des paramètres syntaxiques différents, pourrait permettre de faire émerger des régularités inattendues.

Appendix A

Information sur les treebanks

Table A.1: Table présentant les treebanks de la version UD 2.1 (extraite depuis le site du projet)

treebank	language	genres	major family
af_afribooms	Afrikaans	government text	Indo-european (IE)
grc_perseus	AncientGreek	fiction	IE
grc_proiel	AncientGreek	bible nonfiction	IE
ar_nyuad	Arabic	news	Afro-asiatic
ar	Arabic	news	Afro-asiatic
ar_pud	Arabic	news wiki	Afro-asiatic
eu	Basque	news	_
be	Belarusian	news	IE
bg_btb	Bulgarian	fiction legal misc news	IE
bxr	Buryat	fiction grammar-examples news	mongolic
yue	Cantonese	spoken	Sino-tibetan
ca_ancora	Catalan	news	IE
zh	Chinese	wiki	Sino-tibetan
zh_pud	Chinese	news wiki	Sino-tibetan
zh_cfl	Chinese	learner-essays	Sino-tibetan
zh_hk	Chinese	subtitle	Sino-tibetan
cop_scriptorium	Coptic	bible fiction nonfiction	Afro-asiatic
hr	Croatian	news web wiki	IE

cs	Czech	news	IE
cs_cac	Czech	legal medical news nonfiction reviews	IE
cs_fictree	Czech	fiction	IE
cs_cltt	Czech	legal	IE
cs_pud	Czech	news wiki	IE
cu_proiel	OldChurchSlavonic	bible	IE
da	Danish	fiction news nonfiction spoken	IE
nl	Dutch	news	IE
nl_lassysmall	Dutch	wiki	IE
en_lines	English	fiction nonfiction spoken	IE
en	English	academic fiction news nonfiction spoken web wiki	IE
en_partut	English	legal news wiki	IE
en_pud	English	news wiki	IE
et	Estonian	fiction news nonfiction	Uralic
fi	Finnish	blog fiction grammar-examples legal news wiki	Uralic
fi_ftb	Finnish	grammar-examples	Uralic
fi_pud	Finnish	news wiki	Uralic
fr_ftb	French	newswire	IE
fr	French	blog news reviews wiki	IE
fr_sequoia	French	medical new nonfiction wiki	IE
fr_partut	French	legal news wiki	IE
fr_pud	French	news wiki	IE
gl	Galician	legal medical news nonfiction	IE
gl_treegal	Galician	news	IE
de	German	news reviews wiki	IE
de_pud	German	news wiki	IE
got	Gothic	bible	IE
el	Greek	news spoken wiki	IE
he	Hebrew	news	Afro-asiatic
hi	Hindi	news	IE

hi_pud	Hindi	news wiki	IE
hu	Hungarian	news	Uralic
id	Indonesian	blog news	Austronesian
ga	Irish	fiction legal media news web	IE
it	Italian	legal news wiki	IE
it_postwita	Italian	social	IE
it_partut	Italian	legal news wiki	IE
it_pud	Italian	news wiki	IE
ja_ktc	Japonais	news	_
ja	Japonais	blog news	_
ja_pud	Japonais	news wiki	_
kk	Kazakh	fiction news wiki	Turkic
ko	_	blog news	_
kmr	Kurmanji	fiction wiki	IE
la_ittb	Latin	nonfiction	IE
la_proiel	Latin	bible nonfiction	IE
la	Latin	bible fiction nonfiction	IE
lv	Latvian	academic fiction legal misc news spoken	IE
lt	Lithuanian	news nonfiction	IE
mr	Marathi	fiction wiki	IE
no_nynorsk	Norwegian	blog news nonfiction	IE
no_nynorsklia	Norwegian	spoken	IE
fa	Persian	fiction legal medical news nonfiction social	IE
pl_sz	Polish	fiction news nonfiction	IE
pt_br	Portuguese	blog news	IE
pt	Portuguese	blog news	IE
pt_pud	Portuguese	news wiki	IE
ro	Romanian	fiction legal medical news nonfiction science	IE
ro_nonstandard	Romanian	bible folklore	IE
ru_syntagrus	Russian	fiction news nonfiction	IE
ru	Russian	wiki	IE

ru_pud	Russian	news wiki	IE
sa	Sanskrit	fiction	IE
sr	Serbian	news wiki	IE
sk_snk	Slovak	fiction news nonfiction	IE
sl	Slovenian	fiction news nonfiction	IE
sl_sst	Slovenian	spoken	IE
es_ancora	Spanish	news	IE
es	Spanish	blog news reviews wiki	IE
es_pud	Spanish	news wiki	IE
sv	Swedish	news nonfiction	IE
sv_lines	Swedish	fiction nonfiction spoken	IE
sv_pud	Swedish	news wiki	IE
ta	Tamil	news	Dravidian
te	Telugu	grammar-examples	Dravidian
tr	Turkish	news nonfiction	Turkic
tr_pud	Turkish	news wiki	Turkic
uk	Ukrainian	email fiction legal news social web wiki	IE
ur	Urdu	news	IE
ug	Uyghur	fiction	Turkic
vi	Vietnamese	news	Austro-asiatic
part of v2.2			
fro	OldFrench	literary, religious, historical, juridic, didactic	IE
pcm	Naija	conversations, interviews	Creole

Appendix B

Règles d'extraction

B.1 SOV

```
pattern {
  V [cat=VERB];
  V -[nsubj|csubj]-> S;
  V -[obj|iobj|xcomp|ccomp]-> O;
  S << V;
  O << V;
  S << O;
}
```

B.2 SOV

```
pattern {
  V [cat=VERB];
  V -[nsubj|csubj]-> S;
  V -[obj|iobj|xcomp|ccomp]-> O;
  S << V;
  O << V;
  S << O;
}
```

B.3 OVS

```
pattern {
V [cat=VERB];
V -[nsubj|csubj]-> S;
V -[obj|xcomp|ccomp|iobj]-> O;
O << V;
V << S;
O << S;
}
```

B.4 OSV

```
pattern {
  V [cat=VERB];
  V -[nsubj|csubj]-> S;
  V -[obj|xcomp|ccomp|iobj]-> O;
  O << V;
  S << V;
  O << S;
}
```

B.5 VSO

```
pattern {
V [cat=VERB];
V -[nsubj|csubj]-> S;
V -[obj|xcomp|ccomp|iobj]-> O;
  V << O;
  V << S;
  S << O;
}
```

B.6 VOS

```
pattern {
V [cat=VERB];
V -[nsubj|csubj]-> S;
V -[obj|xcomp|ccomp|iobj]-> O;
  V << O;
```

```
  V << S;
  O << S;
}
```

B.7 nom-adposition

```
pattern {
  operateur [cat=NOUN];
  operande [cat=ADP];
  operateur -[case]-> operande;
}
```

B.8 nom-adjectif (modifieur)

```
pattern {
  operande [cat=NOUN];
  operateur [cat=ADJ];
  operande -[amod]-> operateur;
}
```

B.9 verbe-objet

```
pattern {
  operande [cat=VERB];
  operande -[obj|iobj|xcomp|ccomp]-> operateur;
}
```

B.10 verbe-auxiliaire

```
pattern {
  operateur [cat=VERB];
  operande [cat=AUX];
  operateur -[aux]-> operande;
}
```

Bibliographie

[Abney, 1991] Abney, S. P. (1991). Parsing by chunks. In *Principle-based parsing*, pages 257–278. Springer.

[Ammar et al., 2016] Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. (2016). Many languages, one parser. *Transactions of the Association of Computational Linguistics*, 4(1):431–444.

[Bender, 2009] 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. Association for Computational Linguistics.

[Buchholz and Marsi, 2006a] Buchholz, S. and Marsi, E. (2006a). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

[Buchholz and Marsi, 2006b] Buchholz, S. and Marsi, E. (2006b). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

[Chen et al., 2018] Chen, X., Gerdes, K., and Kahane, S. (2018). Typometrics : from implicational to quantitative universals in word order typology. (en préparation).

[Comrie, 1981] Comrie, B. (1981). *Language universals and linguistic typology: Syntax and morphology*. Basil Blackwell, Oxford.

[Creissels, 2006] Creissels, D. (2006). *Syntaxe générale: une introduction typologique*. Hermes Science.

[Croft, 2001] Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

[Dryer, 1989] Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 13(2):257–292.

[Dryer, 1992] Dryer, M. S. (1992). The greenbergian word order correlations. *Language*, pages 81–138.

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

[Flyamer, 2018] Flyamer, I. (2016–2018). adjustText: A small library for automatically adjusting text position in matplotlib plots to minimize overlaps. https://github.com/Phlya/adjustText.

[Futrell et al., 2015] 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.

[Gerdes and Kahane, 2016] Gerdes, K. and Kahane, S. (2016). Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140.

[Greenberg, 1963] Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

[Gudschinsky, 1956] Gudschinsky, S. C. (1956). The abc's of lexicostatistics (glottochronology). *Word*, 12(2):175–210.

[Guillaume et al., 2012] Guillaume, B., Bonfante, G., Masson, P., Morey, M., and Perrier, G. (2012). Grew : un outil de réécriture de graphes pour le TAL. In Georges Antoniadis, Hervé Blanchon, G. S., editor, *12ième Conférence annuelle sur le Traitement Automatique des Langues (TALN'12)*, pages 1–2, Grenoble, France. ATALA.

[Haegeman, 1994] Haegeman, L. (1994). *Introduction to Government and Binding Theory*. Blackwell, Oxford, 2nd ed. edition.

[Hajic et al., 2001] Hajic, J., Vidová-Hladká, B., and Pajas, P. (2001). The prague dependency treebank: Annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 105–114.

[Hawkins, 1983] Hawkins, J. A. (1983). Word order universals.

[Husson et al.,] Husson, F., Josse, J., Le, S., and Mazet, J. FactoMineR: Multivariate exploratory data analysis and data mining. https://cran.r-project.org/web/packages/FactoMineR/index.html.

[Kahane, 2001] Kahane, S. (2001). Grammaires de dépendance formelles et théorie sens-texte. *TALN 2001*.

[Kahane and Gerdes, 2018] Kahane, S. and Gerdes, K. (2018). Structures syntaxiques : Modéliser les langues. (en préparation).

[Kassambara and Fabian,] Kassambara, A. and Fabian, M. factoextra: Extract and visualize the results of multivariate data analyses. https://cran.r-project.org/web/packages/factoextra/index.html.

[Littell et al., 2017] 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*, volume 2, pages 8–14.

[Liu, 2010] Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

[Marcus et al., 1993] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

[Mel'čuk, 1988] Mel'čuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press.

[Mithun, 1987] Mithun, M. (1987). Is basic word order universal. In *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984*, volume 11, page 281. John Benjamins Publishing Company.

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

[Nivre, 2008] Nivre, J. (2008). Treebanks. In Lüdeling, A. and Kytö, M., editors, *Corpus linguistics: An international handbook*, chapter 13, pages 225–242. Walter de Gruyter GmbH.

[Nivre et al., 2017] Nivre, J., Agić, Ž., and Ahrenberg, L. e. a. (2017). Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[Nivre et al., 2007] 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)*.

[O'Horan et al., 2016] O'Horan, H., Berzak, Y., Vulic, 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.

[Osborne and Niu, 2017] Osborne, T. and Niu, R. (2017). The component unit. introducing a novel unit of syntactic analysis. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 165–175.

[Östling, 2015] 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)*, volume 2, pages 205–211.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Raviv et al.,] Raviv, L., Meyer, A., and Lev-ari, S. The role of community size in the emergence of linguistic structure.

[Rosa and Zabokrtsky, 2015] Rosa, R. and Zabokrtsky, Z. (2015). Klcpos3-a language similarity measure for delexicalized parser transfer. 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)*, volume 2, pages 243–249.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[Schwartz et al., 2012] Schwartz, R., Abend, O., and Rappoport, A. (2012). Learnability-based syntactic annotation design. *Proceedings of COLING 2012*, pages 2405–2422.

[Seddah et al., 2014] Seddah, D., Kübler, S., and Tsarfaty, R. (2014). Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, pages 379–423.

[Søgaard, 2011] Søgaard, A. (2011). Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 682–686. Association for Computational Linguistics.

[Tesnière, 1959] Tesnière, L. (1959). Eléments de syntaxe structurale. *Paris, Klincksieck*.

[Vennemann, 1972] Vennemann, T. (1972). Analogy in generative grammar: the origin of word order. In *Proceedings of the eleventh international congress of linguists*, volume 2, pages 79–83. Il Mulino Bologna.

[Wisniewski and Lacroix, 2017] Wisniewski, G. and Lacroix, O. (2017). A systematic comparison of syntactic representations of dependency parsing. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 146–152.

[Zeman and Resnik, 2008] Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.