Sur la Stabilité Interlangue des Catégories Morphosyntaxiques

Rapport de Stage de L3

Matthieu Boyer

Laboratoire Lattice

CNRS — ENS-PSL — Université Sorbonne Nouvelle

Sous la direction de Mathieu Dehouck

Table des matières

Résumé

Dans ce rapport, nous nous intéressons à la stabilité interlangue des catégories morphosyntaxiques. Nous nous sommes interrogés quant à la réalité de la comparaison interlangue de catégories descriptives du langage, et tout particulièrement aux cas grammaticaux. Nous avons observé les différences dans la forme syntaxique de certains cas, et remarqué les similarités entre d'autres plus sémantiquement distincts, en observant toutefois une forme géométrique générale. Nous avons déterminé des prototypes syntaxiques de cas et étudié la représentation des adpositions.

1 Introduction

1.1 Contextualisation

> There is a fundamental distinction between language-particular categories of languages (which descriptive linguists must describe by descriptive categories of their descriptions) and comparative concepts (which comparative linguists may use to compare languages).
>
> *Martin Haspelmath* [Has18]

Selon Haspelmath, il est possible que la manière de décrire les langues en linguistique soit basée sur des envies de comparaison, parfois mal placées. Dans ce rapport, nous allons donc nous intéresser à la notion fondamentale de catégorie morphosyntaxique, et comparer les descriptions dans différents langages de catégories linguistiques comparatives. Ceci permettrait de justifier la transposition de résultats d'une langue vers une autre.

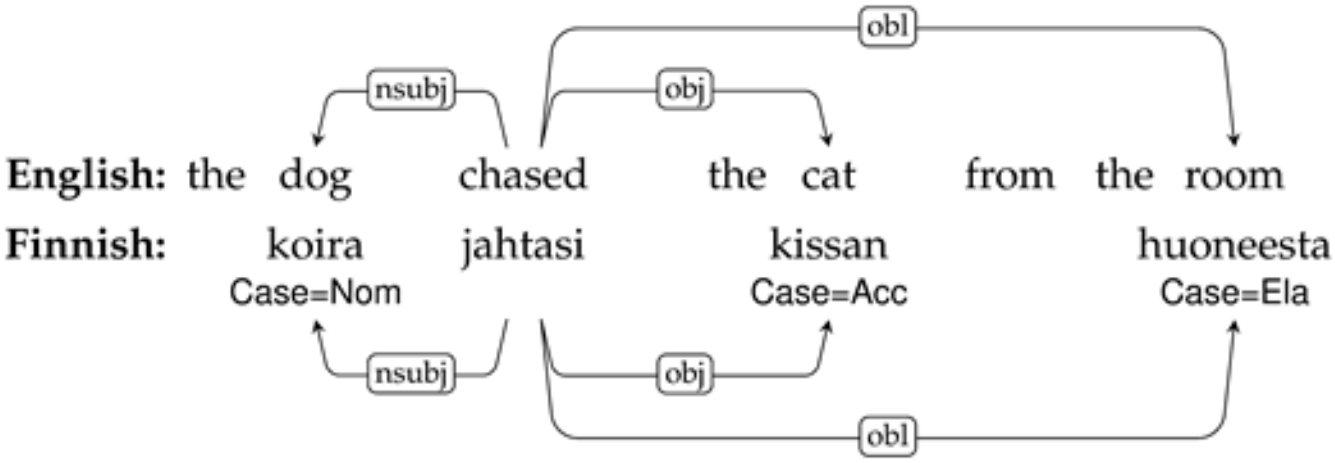

Figure 1 – Représentation d'une Phrase en Anglais et en Finnois et de ses Relations de Dépendances, source :[dMMNZ21]

En linguistique, la morphologie est l'étude des mots, de la manière dont ils sont formés et des relations entre eux au sein du langage. La syntaxe est l'étude de la manière dont se combinent les morphèmes (plus petites unités de son faisant sens dans un langage) et les mots pour former des structures plus grandes comme des phrases. La sémantique, enfin, est l'étude du sens linguistique, de comment les mots ont du sens, et de la manière dont le sens de parties d'une phrase influe sur le sens de celle-ci.
La morphosyntaxe est la combinaison des aspects morphologiques et syntaxiques du langage, et examine notamment comment les formes des mots et structures grammaticales interagissent pour transmettre du sens dans une phrase. Une catégorie morphosyntaxique est une propriété syntaxique, c'est à dire ayant des influences sur la structure grammaticale de la phrase, qui est marquée morphologiquement sur certains mots. En français par exemple, le pluriel est une catégorie morphosyntaxique : il est marqué à la fin des mots et donne des informations sur quel groupe gouverne un autre groupe. Il existe également des catégories morphosémantiques, qui ne donnent que des informations sur la signification d'une phrase et pas sur sa structure.
Les cas grammaticaux sont des exemples de catégories morphosyntaxiques et morphosémantiques.

1.2 Données et Universal Dependencies

Pour étudier la thèse d'Haspelmath, nous allons considérer que les relations de dépendances (*reldep*) décrites par les annotations de Universal Dependencies (UD, version 2 décrite dans [dMMNZ21]) sont une manière de représenter des catégories comparatives. Une relation de dépendance est une manière de décrire les relations syntaxiques dans une phrase (cf 1). Elles se déduisent de la construction par une grammaire de dépendance ou contextuelle de la phrase.

Il existe 37 relations de dépendances de base, mais les personnes annotant les corpus ont la possibilité d'en créer de nouvelles, sous la syntaxe `str1:str2:...` où `str1` doit être une relation de dépendance décrite comme relation basique dans la table 3 de [dMMNZ21].
Par ailleurs, les mots sont annotés avec les propriétés morphologiques qu'ils possèdent, par exemple leur temps ou leur aspect pour un verbe ou leur cas et leur genre pour un nom. Les propriétés morphologiques universelles utilisées par UD sont décrites table 2 dans [dMMNZ21].
Enfin, les mots sont annotés avec leur nature grammaticale (e.g. Nom, Verbe, Pronom...) comme décrit table 1 dans [dMMNZ21].

1.3 Méthode

Nous considérons tout d'abord que chaque *reldep* décrit une unique catégorie comparative et que plusieurs *reldep* ne peuvent instancier une même catégorie comparative. En comptant le nombre d'instances de chaque *reldep* pour un mot vérifiant une propriété grammaticale de la langue (i.e. une catégorie descriptive, que l'on représente par une propriété morphologique d'UD, typiquement les cas pour des langues en utilisant), on obtient une représentation des catégories descriptives et on peut donc mesurer la proximité de deux catégories descriptives dans deux langues différentes. Les corpus utilisés dans cette première partie sont ceux du projet Universal Dependencies, version 2.14 décrite dans [Z+24].
On n'utilise jamais la signification sémantique a priori d'un cas pour en induire un résultat. C'est pourquoi on ne donne la définition théorique d'aucun cas, sauf pour donner un exemple ou expliquer un résultat.

2 Approche Géométrique

Pour cette première approche, on considère la représentation obtenue comme une représentation vectorielle : on considère qu'on se place sur $\mathbb{R}^{|R|}$ où R est l'ensemble des *reldep* et dont une base est l'ensemble des relations de dépendance. Ceci nous permet d'étudier la structure syntaxique de chaque cas d'une manière géométrique. Dans la suite, on notera $\nu(T, C)$ la représentation du cas C dans le corpus T avec la convention $\nu(T, C) = 0_{\mathbb{R}^{|R|}}$ si T ne contient pas de mot au cas C. On définit également $\mathcal{E}(C)$ l'ensemble des corpus T pour lesquels $\nu(T, C)$ est non nul et $\mathcal{C}(T)$ l'ensemble des cas C tels que $\nu(T, C)$ est non nul.

2.1 Avec la distance Cosinus

On mesure d'abord la proximité de deux vecteurs en utilisant la distance [1] cosinus entre ceux-ci :

$$d_{\cos}(v_1, v_2) = \frac{\langle v_1 \mid v_2 \rangle}{\|v_1\| \, \|v_2\|} \tag{1}$$

Ici, on considère le produit scalaire canonique sur $\mathbb{R}^{|R|}$. On calcule alors, pour C_1, C_2 deux cas donnés, l'ensemble des $d_{\cos}\left(\nu\left(T_1, C_1\right), \nu\left(T_2, C_2\right)\right)$ où $T_1, T_2 \in \mathcal{E}(C_1) \times \mathcal{E}(C_2)$. Ceci nous donne des informations statistiques sur la proximité angulaire (donc la proximité des directions) de deux cas. On obtient alors les résultats décrits dans 1, 2 et 3.

On observe notamment le fait que le nominatif comme le vocatif ont des directions très particulières, et sont très différents de tous les autres cas. Pour le reste, les résultats sont assez flous, et il semble difficile de tirer des résultats généraux.

Néanmoins, cette méthode est très limitée. En effet, on ne considère ici que 9 des 45 cas définis dans au moins un corpus. De plus, les résultats donnés ici sont à pondérer par la présence de nombreux corpus/langages ne possédant pas au moins l'un des cas ci-dessus, ce qui amène à une représentation trop brouillée des informations.

2.2 Avec l'algorithme de Zassenhaus

On considère les espaces vectoriels engendrés par la représentation vectorielle du système de cas d'une langue, que l'on appellera *espaces de cas*. Ceux-ci sont d'une certaine dimension finie. On applique alors sur toute paire de système de cas l'algorithme de Zassenhaus (voir [LRW97]), permettant de générer une base de l'espace somme et de l'espace intersection.
Toutefois, la grande variance au niveau des coordonnées, et la trop faible dimension (au plus 45, mais souvent de

1. Ce n'est pas une distance au sens mathématique

Cas	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
Premier Quartile	0.037	0.020	0.022	0.032	0.056	0.027	0.026	0.000
Médiane	0.198	0.123	0.134	0.317	0.249	0.188	0.104	0.006
Troisième Quartile	0.416	0.302	0.341	0.823	0.449	0.400	0.225	0.047
Moyenne	0.259	0.196	0.214	0.421	0.282	0.243	0.159	0.058

TABLE 1 – Proximité Angulaire pour le *Génitif*

Cas	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
Premier Quartile	0.020	0.038	0.018	0.026	0.035	0.020	0.620	0.003
Médiane	0.067	0.137	0.072	0.104	0.113	0.078	0.815	0.026
Troisième Quartile	0.158	0.272	0.161	0.225	0.211	0.156	0.912	0.075
Moyenne	0.115	0.188	0.119	0.159	0.153	0.115	0.739	0.072

TABLE 2 – Proximité Angulaire pour le *Nominatif*

Cas	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
Premier Quartile	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.500
Médiane	0.004	0.007	0.006	0.006	0.012	0.007	0.026	0.885
Troisième Quartile	0.025	0.040	0.032	0.047	0.055	0.038	0.075	0.973
Moyenne	0.035	0.042	0.036	0.058	0.053	0.042	0.072	0.681

TABLE 3 – Proximité Angulaire pour le *Vocatif*

l'ordre de 5) dans un grand espace (dimension 228), rend l'intersection toujours nulle numériquement. De plus, cet algorithme est très lent à exécuter car il demande de nombreux appels mémoire pour obtenir la matrice de l'espace de cas de chaque paire de cas, et demande de trouver une base de l'espace de colonnes, ce qui n'est pas nécessairement la famille de colonnes de la matrice.

2.3 Angle entre Cas et Système de Cas

On considère à nouveau la distance cosinus, mais cette fois-ci, non pas entre deux vecteurs, mais entre un vecteur et un espace de cas. Ceci est fait en considérant le projeté orthogonal d'un vecteur sur un espace de cas et en mesurant l'angle entre les deux (ou la distance cosinus). En observant les données de plus près, on trouve une valeur inattendue : l'angle entre le vocatif du farsi et le système de cas arabe, deux langues syntaxiquement proches, est de l'ordre de $\arccos 10^{-16}$. En regardant de plus près les corpus en farsi, on observe que cela découle d'une idiosyncrasie [2] dans les annotations. En farsi, le lemme (unité morphologique abstraite : *fais* et *fait* sont deux graphies du même lemme *faire*, conjugué à deux personnes différentes) est décrit comme une interjection portant le vocatif et se reliant à un nom au cas absolu par la relation de transmission de cas (c'est à dire de marquer le cas pour un autre mot). Ce lemme agit donc en réalité plus comme une adposition [3]. Le vocatif n'apparaît que très peu en farsi, et majoritairement dans cette situation.
Ainsi, il semble que nous ne pouvons pas appliquer de propriétés apprises du farsi à une autre langue, du moins sur le système de cas. Il est toutefois bon de noter que plusieurs corpus adoptent cette convention d'annotation des adpositions (cf 3.2) et certains l'adoptent également sur certains adverbes.

2. Caractère propre à une langue
3. Mot-outil immédiatement associé à un élément subordonné qui en précise les relations syntaxiques et sémantiques avec le reste de la phrase

Cas	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
Premier Quartile	0.605	0.315	0.552	0.521	0.535	0.595	0.569	0.842
Médiane	0.781	0.476	0.725	0.715	0.695	0.764	0.722	0.992
Troisième Quartile	1.002	0.695	0.950	0.936	0.929	0.993	0.965	1.115
Moyenne	0.802	0.526	0.751	0.733	0.732	0.790	0.760	0.982

TABLE 4 – Proximité pour la distance euclidienne avec l'*Accusatif*

	Dat RU	Gen CZ	Gen RU	Dat CZ
`Total`	1711	2631	2070	277
`obl`	450	208	219	48
`iobj`	340	0	0	0
`amod`	243	736	475	54
`nmod`	300	1000	980	24
`conj`	112	225	84	21
`case`	0	340	1	87
`det`	34	80	79	3

TABLE 5 – Extraits des vecteurs de *Reldep* pour le Russe et le Tchèque

Par ailleurs, il n'est pas rare qu'au sein d'une même langue, deux corpus produisent des résultats assez différents. Ceci peut venir de la variance des phrases considérées, mais plus souvent de la présence ou non des reldeps `conj`, `case` et de la manière d'annoter le cas d'une adposition, comme décrit plus haut..

2.4 Distance Euclidienne

On considère cette fois la distance euclidienne entre tous deux vecteurs, qu'on aura au préalable normalisés pour qu'ils représentent des distributions de probabilités. On calcule alors, pour C_1, C_2 deux cas donnés, l'ensemble des $d\left(\nu\left(T_1, C_1\right), \nu\left(T_2, C_2\right)\right)$ où $T_1, T_2 \in \mathcal{E}(C_1) \times \mathcal{E}(C_2)$ et d est la distance euclidienne associée au produit scalaire canonique. On obtient le tableau 4

On utilise ces données pour déterminer, entre deux corpus (ici le `Czech-CLTT` et le `Russian-GSD`), quel cas de l'autre langage est le plus proche d'un cas du premier. Dans l'exemple de 5, le datif russe est plus proche du génitif tchèque que du datif tchèque.

On enlève ensuite `conj, det` puisque ces *reldep* démontrent l'accord vers la tête du groupe/de la proposition, et donc des doublons dans les données, ceci permet d'éviter de compter comme plusieurs instances d'un même cas un groupe nominal de la forme *Alice, le boulanger, la laitière et Bob*, qui remplit un usage sémantique et syntaxique unique dans la phrase. On enlève aussi `case`, qui est souvent utilisé (on peut notamment le voir dans l'exemple ci-dessous) pour marquer le cas avec une adposition, et ceci dépend très fortement de la personne qui a annoté le corpus, et de l'usage dans les grammaires du langage.

On observe les relations de proximités entre les cas de deux langues en considérant le graphe orienté des plus proches voisins. Celui-ci permet de découvrir quels groupes de cas ont les mêmes fonctions syntaxiques dans les deux langues, et pourraient être interchangés lors de l'analyse grammaticale (parsing), cf 2

Pour éviter encore plus d'avoir des redites, on décide de ne se concentrer que sur des mots de même nature. Ceci permet par exemple d'éviter qu'un groupe nominal ayant la même fonction sémantique (e.g. objet direct du verbe) apporte plusieurs instances d'un même cas (e.g. avec un adjectif et un nom à l'accusatif). On obtient alors

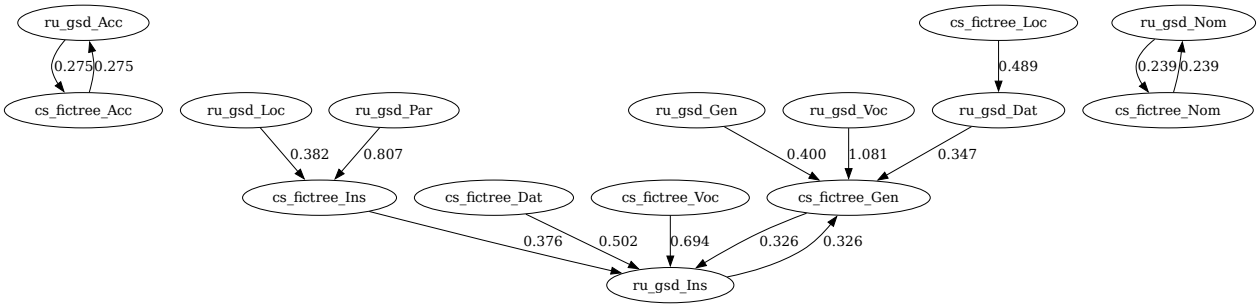

FIGURE 2 – Graphes des Plus Proches Voisins Russe-Tchèque.

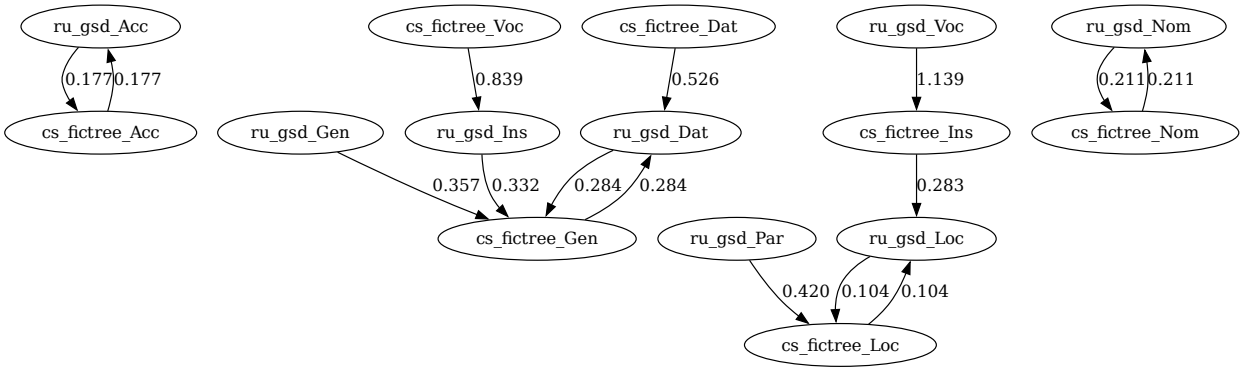

FIGURE 3 – Graphes des Plus Proches Voisins Russe-Tchèque pour les Noms uniquement.

le graphe 3 pour le tchèque et le russe en ne considérant que les noms :

On observe notamment que pour les noms, la structure du graphe reste la même. Le vocatif russe et le vocatif tchèque, peu utilisés et le partitif russe n'ayant pas d'équivalent en tchèque, ils sont bien plus éloignés des autres cas. On retrouve par ailleurs un bloc datif – génitif qui était déjà présent auparavant, à variance dans le corpus près. Par ailleurs, on observe également que les paires accusatif – accusatif et nominatif – nominatif restent stables et plus proches que toutes les autres paires de cas. En ne considérant que les pronoms, on obtient le graphe 4.

Cette fois ci, il y a une variance bien plus forte dans les distances, sans doute due à la variance dans les données. En effet, à part au nominatif et à l'accusatif les échantillons de données sont bien plus faibles (cf 6).

Finalement, il semble que considérer les pronoms fait perdre en information car ceux-ci sont bien moins usités en général. Toutefois, on remarque également qu'une structure générale de la langue semble transparaître de ces graphes. Il faut cependant noter que les graphes dépendent très fortement du corpus proposé comme le montre 5

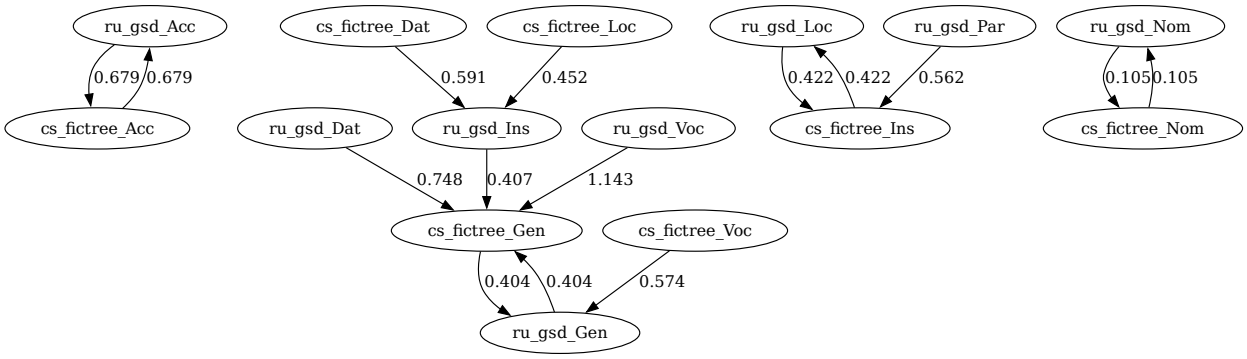

FIGURE 4 – Graphes des Plus Proches Voisins Russe-Tchèque pour les Pronoms uniquement.

Échantillons pour les noms :

Cas	Russe	Tchèque
Acc	2807	5960
Dat	1029	861
Gen	7616	4378
Ins	1642	2100
Loc	2809	2583
Nom	4571	5970
Par	1	0
Voc	1	203

Échantillons pour les pronoms :

Cas	Russe	Tchèque
Acc	206	5960
Dat	129	2743
Gen	241	448
Ins	151	400
Loc	128	221
Nom	631	1427
Par	1	0
Voc	1	8

TABLE 6 – Taille d'Échantillons sur les cas en Russe et en Tchèque.

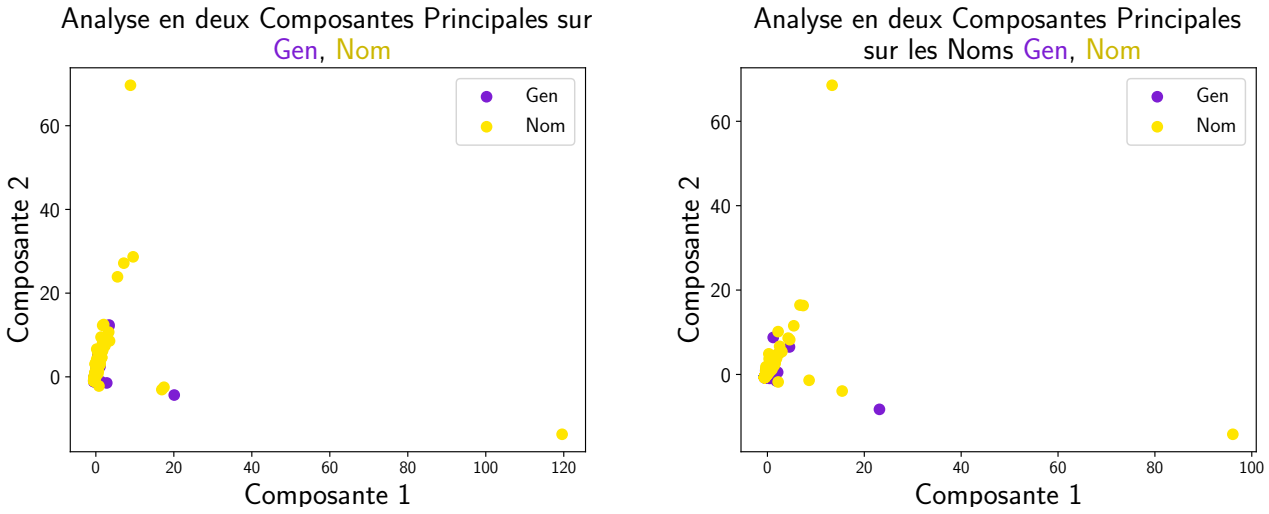

FIGURE 6 – Représentations de l'Analyse en deux Composantes Principales sur le Génitif et le Nominatif

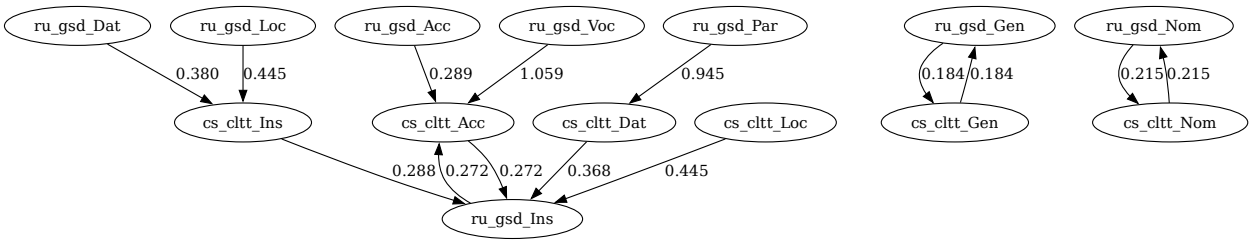

FIGURE 5 – Graphe des Plus Proches Voisins Russe-Tchèque

Ici, on ne considère plus le corpus `cs_fictree` mais le `cs_cltt`, bien plus petit (467 phrases contre 10160). Même si la forme du graphe ne semble que peu changer, la variance dans le corpus joue énormément.
Par ailleurs, cette méthode n'est que peu applicable, puisqu'elle nécessite d'étudier toutes les langues par paire.

2.5 Visualisation des Données

Les résultats précédemment obtenus ne permettent d'observer les données ou bien qu'à une échelle très importante, ce qui brouille les résultats, ou bien qu'à une échelle trop faible pour qu'on puisse généraliser un résultat. Dans la suite, on propose donc des méthodes pour regrouper les vecteurs de cas dans différentes langues, afin d'essayer de constater l'uniformité (ou non) de certains groupes de cas. L'objectif est ici de réussir à repérer, au sein des données, des patterns, afin de pouvoir mener les analyses présentées plus haut sur des groupes de corpus intéressants.

2.5.1 PCA

On commence par appliquer une analyse en deux composantes principales (PCA) aux vecteurs représentant deux cas différents. Cette analyse va procéder à une réduction linéaire de la dimension des données de départ afin de trouver deux vecteurs (les composantes) qui forment un système de coordonnées pour lequel les variations selon chaque axe sont les plus fortes. Pour ceci, on calcule la première composante comme combinaison linéaire des vecteurs d'entrée qui maximise la variance des projections des vecteurs d'entrée sur cette première composante. Puis on recommence pour la seconde composante en prenant un vecteur orthogonal à la première composante.

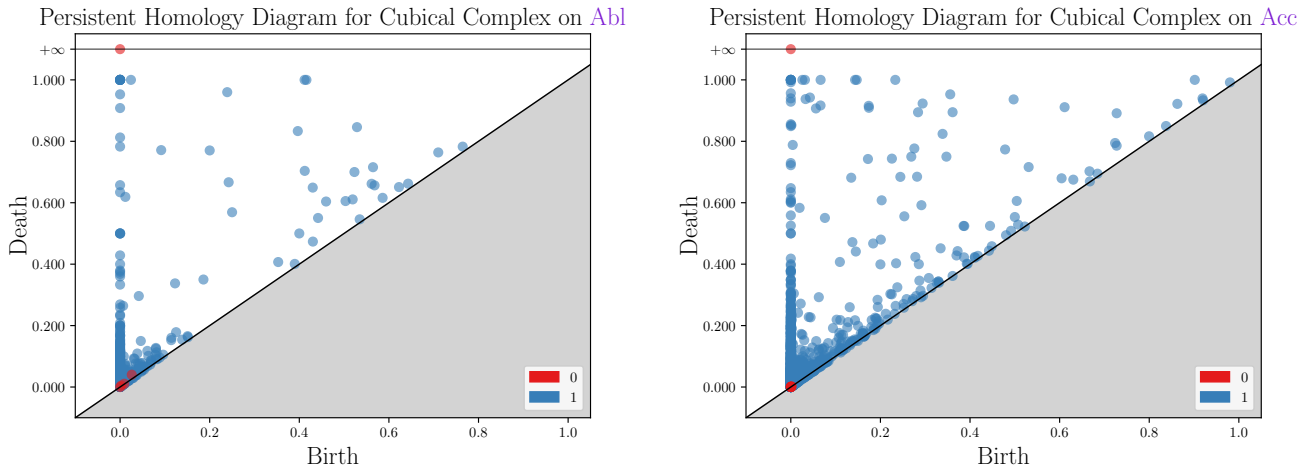

FIGURE 7 – Représentations de l'Homologie Persistente du Complexe Cubique sur {`Abl`, `Acc`}

En regardant les corpus qui sont éloignés du groupe principal (cf 6), on observe que les composantes déterminées par l'algorithme de PCA sont en fait basées sur deux langues, et donc les deux composantes n'ont pas de sens au regard de la syntaxe.

2.5.2 Analyse Topologique des Données

Dans ce paragraphe, on applique des méthodes topologiques dans l'espoir d'obtenir une manière de différencier (ou de regrouper) des ensembles de données (les $\nu(T, C)$ pour $T \in \mathcal{E}(C)$). En effet, on essaie de reconstruire topologiquement la variété dont les vecteurs d'entrée forment une triangulation [4], ce qui permet d'obtenir des invariants algébriques et de les comparer entre deux variétés.

On rappelle qu'un complexe simplicial est une sorte de généralisation à plusieurs dimensions des graphes. Le complexe simplicial k-dimensionnel d'un ensemble $X = \{x_0, \dots, x_k\}$ de points est l'enveloppe convexe de X. Les points de X sont appelés sommets et les simplexes [5] engendrés par les parties de X sont appelés faces du complexe. Le k-ème groupe d'homologie H_k est un espace vectoriel dont la dimension représente le nombre de « *trous* » k-dimensionnels. H_0 représente ainsi les composantes connexes, H_1 les boucles unidimensionnelles et H_2 les cavités 2-dimensionnelles. Les nombres de Betti sont les dimensions des groupes d'homologie et donnent donc le nombre de « *trous* » k-dimensionnels (voir [CM21] pour plus de détails).

Pour l'implémentation de ce qui suit, on utilise les bibliothèques Python `Gudhi`, `POT` et `Hera`, dont les implémentations sont décrites dans [MBGY14],[FC+21], [KMN17] qui permettent de calculer efficacement les complexes simpliciaux des données et leurs diagrammes d'homologie, la distance de Wasserstein (ou Earth Mover's Distance) entre deux distributions ou diagrammes d'homologie.

On utilise d'abord un complexe cubique plutôt qu'un complexe simplicial pour essayer de représenter plus efficacement les groupes d'homologie de la variété triangulée par les points de chaque cas. Un complexe cubique est un ensemble $\mathcal{K}$ de cubes [6] tel que la frontière de chaque cube soit dans $\mathcal{K}$ (voir [KMM04]). On utilise un complexe cubique puisqu'ici, il est compliqué de calculer directement un complexe simplicial de par le nombre de points et la dimension de l'espace.

On a également calculé les complexes sur les Nominatifs et Accusatifs. On remarque sur 7 qu'une forme générale se retrouve dans les diagrammes de persistence de chacun des cas. Pour vérifier l'intuition, on calcule la distance-1

4. i.e. la représentation de la variété par des espaces linéaires par morceaux, des simplexes (voir ci-dessous)
5. Généralisation à n-dimension du triangle. Le tétraèdre est un simplexe en dimension 3
6. produits d'intervalles

Cas	Abl	Acc	Dat	Loc	Gen
Abl	0.00	2.45	3.11	1.94	3.85
Acc	2.45	0.00	1.33	1.25	1.79
Dat	3.11	1.33	0.00	1.63	1.27
Loc	1.94	1.25	1.63	0.00	2.26
Gen	3.85	1.79	1.27	2.26	0.00

TABLE 7 – Distances de Wasserstein entre les Diagrammes de Persistence des Complexes Cubiques pour quelques Cas

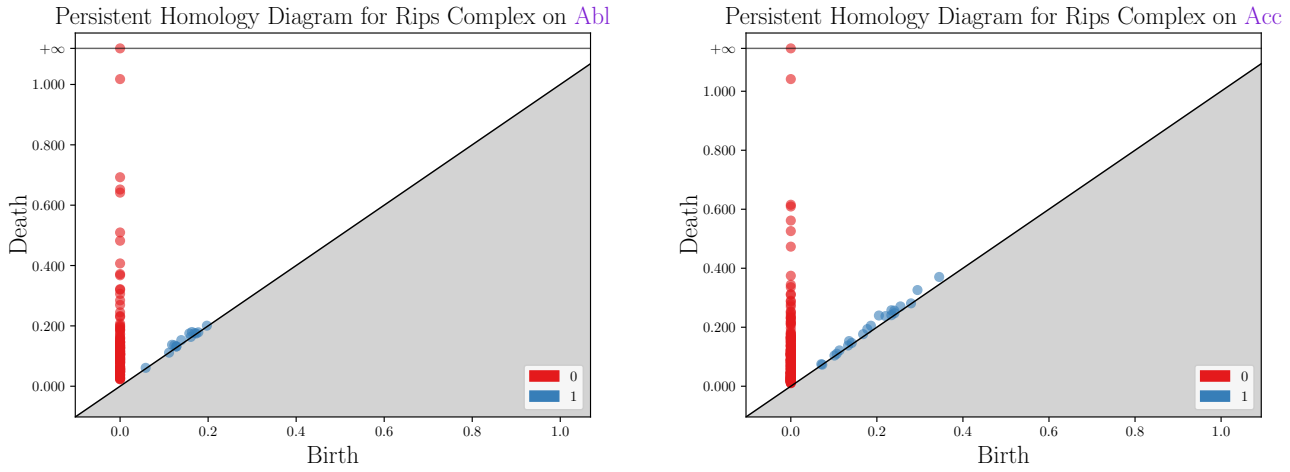

FIGURE 8 – Représentations de l'Homologie Persistente du Complexe de Rips sur {`Abl`, `Acc`}

de Wasserstein entre les distributions de points. On rappelle que la distance-p de Wasserstein pour $p \in [1, +\infty]$ entre deux mesures de probabilités μ, π dont les moments d'ordre p sont finis est :

$$W_p(\mu, \pi) = \inf_{\gamma \in \Gamma(\mu,\pi)} \sqrt[p]{\mathbb{E}_{(x,y)\sim\gamma} d(x,y)^p}$$

où $\Gamma(\mu, \pi)$ est l'ensemble des couplages [7] de μ, π.

Les distances présentées dans 7 étant assez faibles compte tenu le nombre de points (on n'a, contrairement à 3.1, pas des distributions de probabilité), on obtient bien le résultat suggéré par les figures, il semble y avoir une structure générale de la notion topologique de variété engendrée par un cas. Pour vérifier cette hypothèse, on teste de même en créant le complexe de Rips [8], avec une valeur de paramètre assez faible $\alpha = .5$. Toutefois, les diagrammes présentés dans 8 restant très brouillons, il est difficile de trouver une variété plus abstraite avec la même homologie.

On remarque à nouveau qu'une forme générale se retrouve dans les diagrammes de persistence de chacun des cas. Lorsqu'on calcule la distance de Wasserstein entre deux diagrammes pour quelques cas, on obtient le tableau suivant :

Cela signifie notamment que la représentation d'un cas comme variété topologique ne varie que peu d'un cas à l'autre, sans toutefois pouvoir tirer plus d'informations que cela. Intuitivement, cela signifie que si deux cas peuvent avoir des directions *générales* différentes, leurs représentations seront semblables, à une rotation près.

Cependant, d'un point de vue linguistique, il est difficile de tirer des conclusions de ces données. En effet, il faudrait pouvoir comparer les diagrammes pour plus de catégories morphosyntaxiques que simplement quelques

7. Loi de probabilité à deux variables dont les lois marginales sont μ, π
8. Généralisation des graphes à α-voisinage, c'est l'ensemble des simplexes dont tous les sommets sont à distance au plus une constante

Cas	Abl	Acc	Dat	Gen	Loc	Nom
Abl	0.00	0.89	1.27	0.82	1.09	0.94
Acc	0.89	0.00	1.01	0.42	1.03	0.91
Dat	1.27	1.01	0.00	0.87	1.47	0.76
Gen	0.82	0.42	0.87	0.00	0.84	0.87
Loc	1.09	1.03	1.47	0.84	0.00	1.48
Nom	0.94	0.91	0.76	0.87	1.48	0.00

TABLE 8 – Distances de Wasserstein entre les Diagrammes de Persistence des Complexes de Rips pour quelques Cas

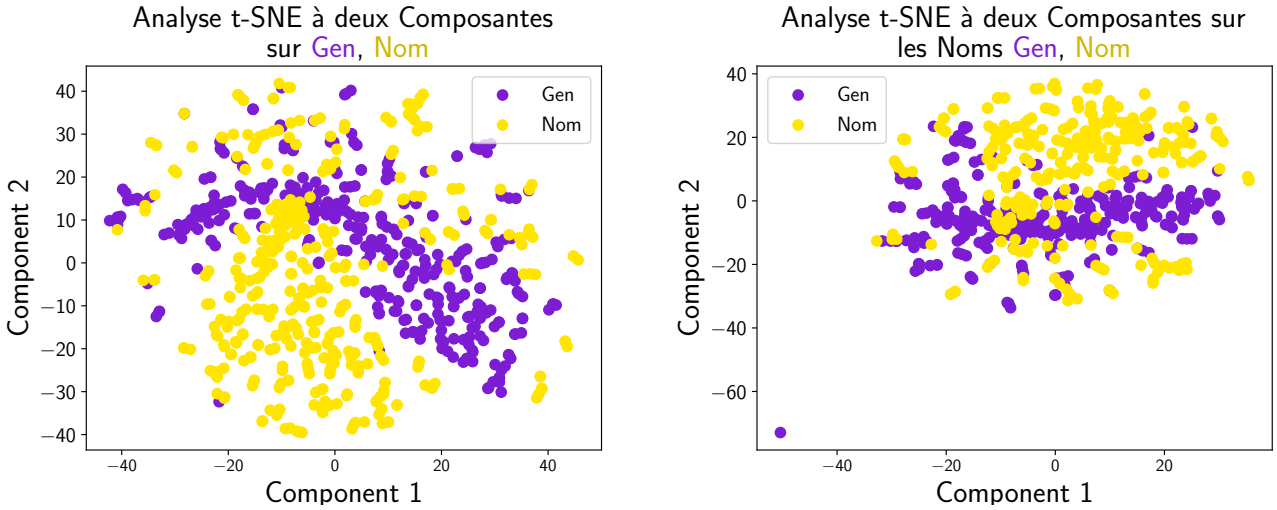

FIGURE 9 – Représentations de l'Analyse t-SNE à deux Composantes sur le Génitif et le Nominatif

cas. Ceci n'étant pas le coeur de notre étude et demandant un temps de calcul assez long, nous n'irons pas plus loin dans ce rapport.

2.5.3 t-SNE

Dans la suite, nous allons essayer différentes techniques de clustering, afin de vérifier si deux cas supposément distincts peuvent être aisément différenciés. On essaie d'abord d'appliquer une analyse t-SNE en 2D, décrite dans [vdMH08]. L'algorithme définit une distribution [9] D_1 de probabilités sur les paires de points de l'espace de départ (si x, y sont proches alors $D_1(x \mid y)$ est proche de 1) et construit alors une distribution D_2 sur les paires de points d'un espace à 2 dimensions minimisant la divergence de Kullback-Leibler $\sum_{x,y} D_1(x \mid y) \log \frac{D_1(x|y)}{D_2(x|y)}$ entre D_1 et D_2. Ici, on l'applique sur le Génitif et le Nominatif, voir 9

Il semble que deux clusters se dégagent, avec des frontières toutefois assez floues. Il semblerait donc que syntaxiquement, le génitif et le nominatif aient des représentations assez différentes, sans toutefois savoir à quel point.

2.5.4 Clustering avec ToMATo

On applique l'algorithme présenté dans [CGOS11] avec la bibliothèque `Gudhi` ([MBGY14]) sur des paires de cas, pour essayer de construire des clusters. Cet algorithme utilise les ensembles de sur-niveau associés à une fonction f (les $F^\alpha = f^{-1}([\alpha, +\infty[)$) pour construire le diagramme de persistence (voir [CM21] et [CGOS11]).

9. On le note plutôt comme des probabilités conditionnelles, mais c'est équivalent

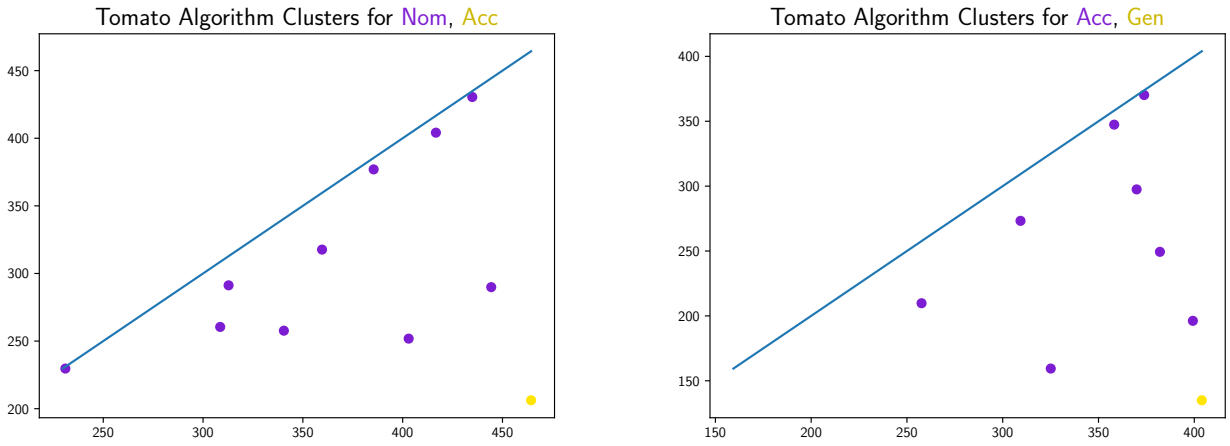

FIGURE 10 – Représentations des Clusters trouvés par l'Algorithme ToMATo sur les Paires de {Acc, Nom, Gen}

Acc	Gen	Loc	Nom
130	62	51	34
69	156	16	42
35	57	29	34
29	28	9	227

TABLE 9 – Heatmap de l'Algorithme KNN avec $k = 11$ sur `Acc, Gen, Loc, Nom`

Dans 10, aucun des cas n'est clairement regroupé au sein de certains clusters puisque, comme vu dans 2.5.2, les variétés sont très similaires au niveau de la persistence et de l'homologie. L'algorithme ne produit aucun résultat véritablement utilisable d'un point de vue linguistique.

2.5.5 Clustering avec KNN

Pour vérifier l'intuition qui apparaît avec le clustering proposé par l'algorithme t-SNE, on essaie d'appliquer à des listes de cas l'algorithme KNN. On obtient la matrice de confusion de 9 en appliquant l'algorithme pour $k = 11$ aux noms qui sont à l'Accusatif, au Génitif, au Locatif ou au Nominatif.

On observe notamment que le *Locatif* est plus difficile à reconnaître que les autres cas. Ceci peut venir du faible nombre de langues possédant un locatif en comparaison aux trois autres cas.

2.6 Conclusion

Finalement, la comparaison des cas entre eux montre notamment qu'il est impossible d'entièrement différencier tous deux cas. De plus, du point de vue géométrique, l'ensemble des instances [10] de chaque cas est similaire, indépendamment du cas. Toutefois, il est clair que plusieurs types de cas se dégagent, et que chacun de ces types est fortement différenciable des autres sémantiquement, mais que cela ne suffit pas à former plusieurs types.

3 Approche Probabiliste

On considère désormais qu'à un cas donné on associe une variable aléatoire sur les distributions de relations de dépendances dont on connaît certaines réalisations (les représentations du cas dans les différents corpus). C'est à dire

10. des vecteurs représentation

Cas	Prototype	iobj	nmod	nsubj	obj	obl
Abs	Uniforme	0.001	0.033	0.272	0.367	0.224
	Wasserstein	0.000	0.016	0.286	0.522	0.112
Erg	Uniforme	0.000	0.007	0.924	0.005	0.059
	Wasserstein	0.000	0.005	0.976	0.014	0.003
Nom	Uniforme	0.001	0.080	0.556	0.074	0.050
	Wasserstein	0.000	0.049	0.654	0.093	0.038
Acc	Uniforme	0.006	0.078	0.038	0.625	0.205
	Wasserstein	0.000	0.072	0.019	0.576	0.259
Gen	Uniforme	0.009	0.674	0.039	0.056	0.149
	Wasserstein	0.000	0.729	0.031	0.045	0.179
Dat	Uniforme	0.144	0.149	0.019	0.000	0.572
	Wasserstein	0.190	0.164	0.005	0.000	0.605
Loc	Uniforme	0.000	0.166	0.009	0.017	0.696
	Wasserstein	0.000	0.188	0.000	0.000	0.762
Ins	Uniforme	0.000	0.172	0.014	0.000	0.660
	Wasserstein	0.000	0.213	0.000	0.000	0.738
Abl	Uniforme	0.000	0.165	0.013	0.001	0.700
	Wasserstein	0.000	0.172	0.000	0.000	0.785

Table 10 – Représentation des Principales *reldep* des Prototypes pour quelques Cas sur les Noms

une variable aléatoire qui a un langage va associer une distribution de probabilité sur les relations de dépendances des mots à ce cas.

3.1 Barycentrisation

On a cherché jusque-là à comparer les cas. On va désormais essayer de les prototyper [11] et de mesurer l'écart au prototype, afin de tirer une définition des cas. Celle-ci devrait coller à la description théorique proposée ci-dessous :

Nom Agent (sujet) d'un verbe

Acc Patient (objet direct) d'un verbe

Erg Agent (et donc patient) d'un verbe intransitif (sans objet), ou patient d'un verbe transitif

Abs Agent d'un verbe transitif

Gen Complément du nom

Dat Objet indirect d'un verbe

Pour calculer les prototypes, on procède comme suit : On calcule l'espérance de la variable aléatoire de chaque cas (en considérant les mesures équiprobables), mais également le barycentre de ses réalisations pour la distance-1 de Wasserstein (ou Earth Mover's Distance, voir [FC+21] pour plus de détails). On rappelle par ailleurs que le barycentre de n points pour une distance d est le point P qui minimise l'énergie E associée :

$$P = \arg\min_x E\left(x, (x_i)_{i\in[\![1,n]\!]}\right) \text{ avec } E\left(x, (x_i)_{i\in[\![1,n]\!]}\right) = \frac{1}{n}\sum_{i=1}^{n} d(x, x_i)$$

Ceci nous donne une forme de vecteur prototypique pour la distribution des relations de dépendance du cas. On obtient la distribution 10, en ne considérant à nouveau que les noms, pour quelques *reldep* :

11. Trouver un vecteur *typique* du cas

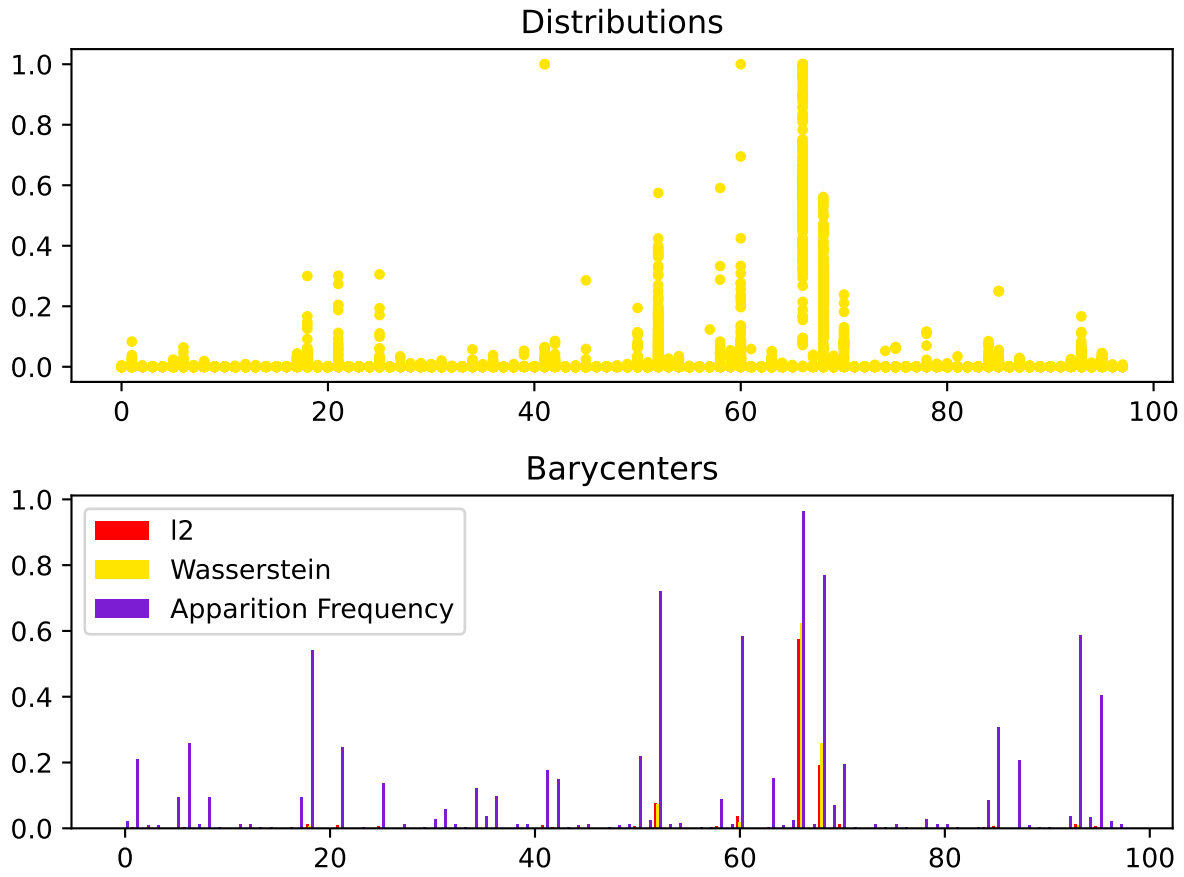

FIGURE 11 – Représentation des Données et des Prototypes proposés pour les Noms à l'Accusatif

On retrouve bien la représentation attendue, ce qui découle notamment de la définition *théorique* des cas. Il est important de noter que les vecteurs ci-dessus ne permettent pas nécessairement de différencier tous deux cas, mais seulement de différencier leurs usages syntaxiques. Prenons un exemple : l'instrumental est le cas qui marque l'outil utilisé pour une action/un objet : Dans une traduction de *J'ai mangé* avec *une fourchette*, on utiliserait l'instrumental pour *fourchette*. Le Locatif est le cas qui marque le lieu d'une action, ou d'un objet : *J'ai péleriné jusqu'à Saclay* ou *J'ai couru à côté de l'école* se traduiraient avec des locatifs [12]. Le locatif et l'instrumental sont principalement des cas morphosémantiques, au sens où ils ne modifient pas la structure de la phrase mais seulement son sens. Ainsi, il est difficile de les différencier syntaxiquement : ils agissent principalement sur un verbe en précisant son action (`obl`), mais également sur les noms en précisant leur fonction ou leur position (`nmod`). À l'inverse, le nominatif (sujet du verbe), l'accusatif (objet du verbe), le datif (objet indirect du verbe), l'ergatif (sujet du verbe transitif) et l'absolutif (sujet du verbe intransitif et objet du verbe transitif), sont des cas plus fortement marqués syntaxiquement et sont donc plus facilement reconnaissables.

Par ailleurs, l'énergie associée au prototype obtenu en prenant la moyenne des réalisations est de l'ordre de celle du barycentre pour la distance-1 de Wasserstein. Comme on pouvait s'y attendre au vu des données, le barycentre pour la distance de Wasserstein présente un profil similaire à la moyenne des instances.

Sur la figure 11, l'axe des abscisses représente les différentes *reldep* apparaissant pour des noms à l'accusatif.

12. certaines langues comme le finnois possèdent plusieurs *locatifs* appelés inessif, élatif, illatif, adessif, ablatif, allatif... Ceux-ci servent à différencier différent types de lieux liés à une action : la destination (à Saclay), la direction (vers le Sud), un lieu à proximité (près du Panthéon)...

Adposition	advcl	iobj	nmod	nsubj	obj	obl
À	0.01668		0.17343	0.00048	0.00381	0.63373
Dans	0.00466		0.13780		0.00196	0.78694
Par	0.00264		0.13715	0.00107	0.00178	0.74632
Pour	0.29543		0.15867		0.00024	0.41168
En	0.08128		0.17115		0.00358	0.54076
Vers	0.00262		0.35741			0.62160
Avec	0.00613		0.32369			0.62606
De	0.02096		0.67966	0.00138	0.01312	0.14296
Sans	0.24451		0.21142		0.00781	0.43802
Sous	0.00217		0.22898	0.00020	0.72797	
Sur	0.00477		0.36267		0.00096	0.59385
Sauf	0.10714		0.22619			0.38095

Table 11 – Représentation de quelques adpositions en français

On représente sur le graphe du haut en jaune la fréquence associée à une reldep pour chaque corpus. Le graphe du bas représente, en rouge, la moyenne uniforme des distributions, en jaune, le barycentre des distributions pour la distance de Wasserstein, et violet la proportion des corpus pour lesquels la *reldep* apparaît. On vérifie bien notamment que pour la moyenne uniforme, certaines relations sont légèrement représentées car très présentes dans quelques langues, ce qui n'est pas le cas pour le barycentre associé à la distance de Wasserstein.

3.2 Représentation des adpositions dans Universal Dependencies

Dans UD, certains corpus dénotent, lorsqu'un groupe est combiné avec une adposition (e.g. les prépositions *à, dans, par* en français, *to, into, above* en anglais, ou la postposition *ile* en turc), l'adposition comme gouvernée par la relation `case` et donnent un cas à l'adposition (cf. l'interjection farsie présentée plus haut 2.3). Ceci découle du postulat linguistique selon lequel tous les langages humains sont également expressifs, ce qui implique qu'on peut traduire les cas par des adpositions (en français, tous exceptés le nominatif et l'accusatif). Dans [KSGK+17], Kirov, Sylak-Flassman, Knowles et Cotterell décrivent une manière d'annoter des corpus en anglais (langue morphologiquement pauvre) selon les caractéristiques morphologiques du tchèque (langue morphologiquement riche).
Nous avons donc cherché à déterminer le cas syntaxiquement équivalent à certaines adpositions. A priori, il n'a aucune raison d'être également le cas sémantiquement équivalent à une adposition, même si l'on considère différents représentants d'une même classe syntaxique de cas (inessif, allatif, essif sont syntaxiquement équivalents parmi les locatifs par exemple). Pour cela, on compte les relations de dépendances des cibles des relations de dépendance depuis chaque adposition. On obtient une distribution des usages syntaxiques des adpositions, ce qui permet de comparer une adposition à un marqueur de cas. On ne donne que la moyenne uniforme des distributions uniformes pour une même langue, le barycentre pour la distance de Wasserstein étant trop volatile pour si peu de distributions (on n'en n'a que 9 pour le français). Dans la table 11 on retrouve le fait que les prépositions françaises s'utilisent dans des constructions similaires, et ne se différencient presque que sémantiquement. Plus spécifiquement, il semble que *de* ait un caractère semblable au génitif et que *à, dans, par, en, vers, avec, sans, sous, sur* ont plutôt un caractère semblable à l'ablatif, au locatif et autres cas syntaxiquement similaires. Les prépositions *sauf* et *pour* quant à elle ont des usages plus particuliers, puisqu'elles servent souvent à introduire des clauses adverbiales (e.g. *il mange* pour *vivre*).

4 Conclusion

Nous avons pu montrer que si les cas ont une structure syntaxique générale assez similaire géométriquement, il existe plusieurs classes syntaxiques assez distinctes au sein desquelles les cas ne diffèrent que sémantiquement. Il faut également noter que dans le formalisme de Universal Dependencies, avoir nombre d'annotateurs différents cause de fortes disparités dans les manières d'annoter, notamment sur l'annotation de cas sur les adverbes et les adpositions selon la langue. Ainsi, comme le supposait Martin Haspelmath, il est justifiable d'utiliser des catégories descriptives d'un langage pour en décrire un autre, mais il faut rester vigilant et ne pas calquer d'idiosyncrasies d'une méta-langue de comparaison lors de la description d'une autre. Pour les cas cela ne pose pas de problème du point de vue syntaxique a priori, mais il est nécessaire de revérifier pour une paire de langues avant de traduire les propriétés syntaxiques d'une langue sur une autre.
Par ailleurs, il n'est pas illogique de donner des cas aux adpositions, de par leurs fonctions sémantiques, mais il apparaît que syntaxiquement elles ne sont que rarement distinguées, et pas toujours de la manière dont on l'attendrait du cas sémantiquement équivalent.

Pour complètement finaliser les résultats de cet étude et justifier leur applicabilité, il reste à vérifier :

- si un parser syntaxique [13] entraîné sur une langue dont on aurait fusionné des catégories morphosyntaxiques ayant des significations syntaxiques (des vecteurs de *reldep*) proches est indistingable probabilistiquement d'un parser syntaxique entraîné sur cette même langue sans fusions ;
- si un parser entraîné sur une première langue peut, sur une langue ayant un système de cas proches (à fusion de cas près), obtenir des résultats dignes d'un parser entraîné sur cette seconde langue.

Ceci a été étudié par Mathieu Dehouck [14] durant mon stage. La vérification n'est pas encore statistiquement robuste, mais les premiers essais indiquent que les deux questions ont des réponses positives.

Références

[CGOS11] Frédéric Chazal, Leonidas Guibas, Steve Oudot et Primoz Skraba : Persistence-based clustering in riemannian manifolds. *Journal of the ACM*, 60, 06 2011.

[CM21] Frédéric Chazal et Bertrand Michel : An introduction to topological data analysis : Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4, 2021.

[dMMNZ21] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre et Daniel Zeman : Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021.

[FC+21] Rémi Flamary, Nicolas Courty *et al.* : Pot : Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[Has18] Martin Haspelmath : *How comparative concepts and descriptive linguistic categories are different*, pages 83–114. 09 2018.

[KMM04] T. Kaczynski, K. Mischaikow et M. Mrozek : *Computational Homology*. Applied Mathematical Sciences. Springer New York, 2004.

[KMN17] Michael Kerber, Dmitriy Morozov et Arnur Nigmetov : Geometry helps to compare persistence diagrams. *ACM J. Exp. Algorithmics*, 22, sep 2017.

[KSGK+17] Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell et Matt Post : A rich morphological tagger for english : Exploring the cross-linguistic tradeoff between morphology and syntax. pages 112–117, 01 2017.

13. Annotant les mots avec leurs *reldeps*
14. qui se trouve être mon encadrant

[LRW97] Eugene M. Luks, Ferenc Rákóczi et Charles R.B. Wright : Some algorithms for nilpotent permutation groups. *Journal of Symbolic Computation*, 23(4):335–354, 1997.

[MBGY14] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse et Mariette Yvinec : The gudhi library : Simplicial complexes and persistent homology. *In* Hoon Hong et Chee Yap, éditeurs : *Mathematical Software – ICMS 2014*, pages 167–174, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[vdMH08] Laurens van der Maaten et Geoffrey Hinton : Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[Z+24] Daniel Zeman *et al.* : Universal dependencies 2.14, 2024. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.