

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

1	Introduction	1
1.1	Contextualisation	1
1.2	Définitions Linguistiques	1
1.3	Données et <i>Universal Dependencies</i>	2
1.4	Méthode	2
2	Approche Géométrique	3
2.1	Avec la distance Cosinus	3
2.2	Avec l'algorithme de ZASSENHAUS	4
2.3	Angle entre Cas et Système de Cas	4
2.4	Distance Euclidienne	4
2.5	Visualisation des Données	7
2.5.1	PCA	7
2.5.2	Analyse Topologique des Données	8
2.5.3	t-SNE	10
2.5.4	Clustering avec ToMATo	10
2.5.5	Clustering avec KNN	11
3	Approche Probabiliste	12
3.1	Barycentrisation	12
3.2	Représentation des adpositions dans UD	14
4	Conclusion	14

Résumé

Dans ce rapport, nous nous intéressons à la stabilité interlangue des catégories morphosyntaxiques. Nous avons quantifié la manière dont différentes catégories descriptives d'un langage ont différentes significations dans différents langages, et particulièrement la manière dont un concept est matérialisé dans différents langages. Nous nous sommes tout particulièrement intéressés à la notion de cas grammatical, aux différences entre cas et à la définition fondamentale d'un cas. Nous avons également étudié la représentation des adpositions en parallèle à la notion de cas.

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.

1.2 Définitions Linguistiques

En linguistique, la morphologie est l'étude des mots, de la manière dont ils sont formés et des relations entre eux au sein d'un 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 influent sur le sens de celle-ci.

La morphosyntaxe est la combinaison des aspects morphologiques et syntaxiques du langage, et examinent 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.

1.3 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. Elles se déduisent de la construction par une grammaire de dépendance ou contextuelle de la phrase.

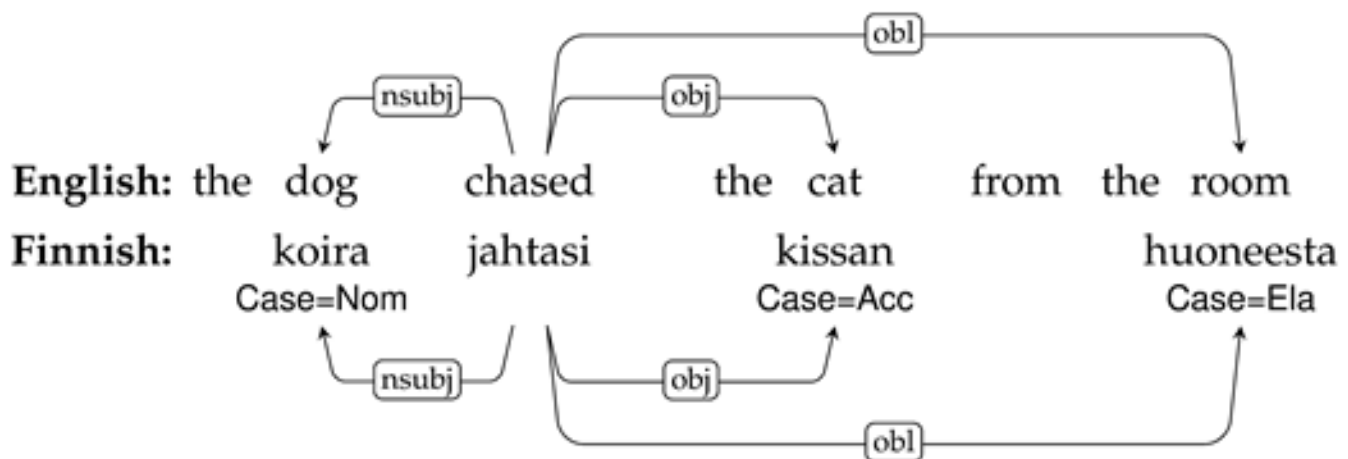


FIGURE 1 – Représentation d'une phrase et de ses relations de dépendances en anglais, source : [dMMNZ21]

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 décrits table 1 dans [dMMNZ21].

1.4 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 *feature* 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].

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.

2.1 Avec la distance Cosinus

On mesure d'abord la proximité de deux vecteurs en utilisant la distance¹ Cosinus entre ceux-ci :

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

On obtient alors les résultats suivants pour quelques cas :

Proximity with :	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
First Quartile	0.037	0.020	0.022	0.032	0.056	0.027	0.026	0.000
Median	0.198	0.123	0.134	0.317	0.249	0.188	0.104	0.006
Third Quartile	0.416	0.302	0.341	0.823	0.449	0.400	0.225	0.047
Mean	0.259	0.196	0.214	0.421	0.282	0.243	0.159	0.058

TABLE 1 – Proximities for Case=Gen

Proximity with :	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
First Quartile	0.020	0.038	0.018	0.026	0.035	0.020	0.620	0.003
Median	0.067	0.137	0.072	0.104	0.113	0.078	0.815	0.026
Third Quartile	0.158	0.272	0.161	0.225	0.211	0.156	0.912	0.075
Mean	0.115	0.188	0.119	0.159	0.153	0.115	0.739	0.072

TABLE 2 – Proximities for Case=Nom

Proximity with :	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
First Quartile	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.500
Median	0.004	0.007	0.006	0.006	0.012	0.007	0.026	0.885
Third Quartile	0.025	0.040	0.032	0.047	0.055	0.038	0.075	0.973
Mean	0.035	0.042	0.036	0.058	0.053	0.042	0.072	0.681

TABLE 3 – Proximities for Case=Voc

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.

Par ailleurs, 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.

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

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 intersubsection. Toutefois, la grande variance au niveau des coordonnées, et la trop faible dimension (au plus 45, mais souvent de l’ordre de 5) dans un grand espace (dimension 228), rend l’intersubsection toujours nulle numériquement. Par ailleurs, 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 colonne, ce qui est non-nécessairement 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 farsis², on observe que cela découle d’une idiosyncrasie 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. Le vocatif n’apparaît que très peu en farsi, et majoritairement dans cette situation. Ainsi, il semble que nous ne pouvons pas tirer d’enseignements du farsi vers une autre langue, du moins sur le système de cas. Il est toutefois bon de noter que plusieurs corpus adopte cette convention d’annotation des adpositions (cf 3.2) et certains l’adoptent également sur les adverbes.

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 (cf. supra).

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 probabilité.

Proximity with :	Abl	Acc	Dat	Gen	Ins	Loc	Nom	Voc
First Quartile	0.605	0.315	0.552	0.521	0.535	0.595	0.569	0.842
Median	0.781	0.476	0.725	0.715	0.695	0.764	0.722	0.992
Third Quartile	1.002	0.695	0.950	0.936	0.929	0.993	0.965	1.115
Mean	0.802	0.526	0.751	0.733	0.732	0.790	0.760	0.982

TABLE 4 – Proximities for Case=Acc

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 notre cas par exemple, le datif russe est plus proche du génitif tchèque que du datif tchèque.

2. et non pas les dindes.

	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

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 souvent (notamment visible dans l'exemple ci-dessous), est utilisé pour marquer le cas avec une apposition (sur, sous), et ceci dépend très fortement de la personne qui a annoté le corpus, et de l'usage dans les grammaires du langage.

En considérant le graphe orienté des plus proches voisins, on remarque que celui-ci ne peut pas avoir de n -cycle pour $n \geq 3$ et décrit des relations de proximité minimale. Pour le tchèque et le russe, le graphe est le suivant :

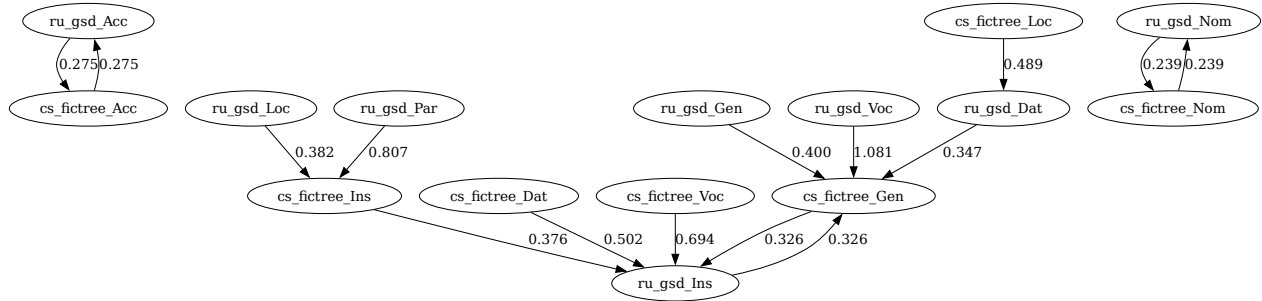


FIGURE 2 – Graphes des Plus Proches Voisins Russe-Tchèque.

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

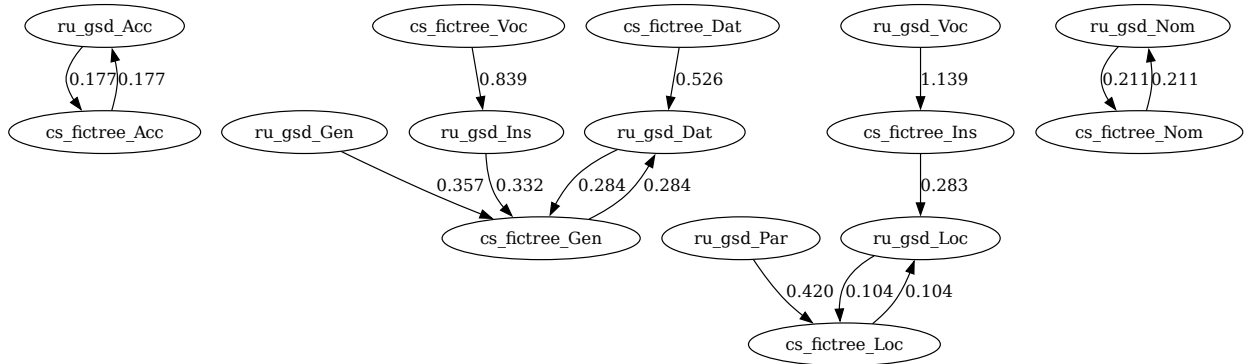


FIGURE 3 – Graphes des Plus Proches Voisins Russe-Tchèque pour les Noms uniquement.

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 suivant :

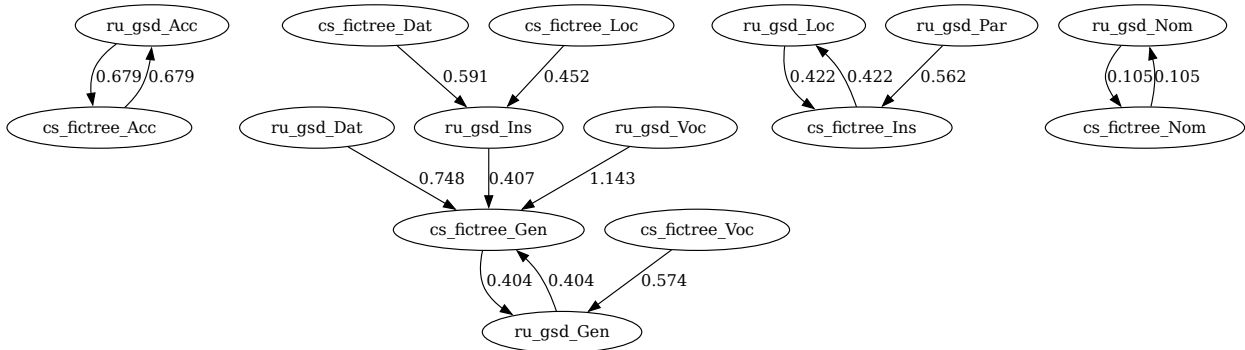


FIGURE 4 – Graphes des Plus Proches Voisins Russe-Tchèque pour les Pronoms uniquement.

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.

<i>Échantillons pour les noms :</i>			<i>Échantillons pour les pronoms :</i>		
Cas	Russe	Tchèque	Cas	Russe	Tchèque
Acc	2807	5960	Acc	206	5960
Dat	1029	861	Dat	129	2743
Gen	7616	4378	Gen	241	448
Ins	1642	2100	Ins	151	400
Loc	2809	2583	Loc	128	221
Nom	4571	5970	Nom	631	1427
Par	1	0	Par	1	0
Voc	1	203	Voc	1	8

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

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 :

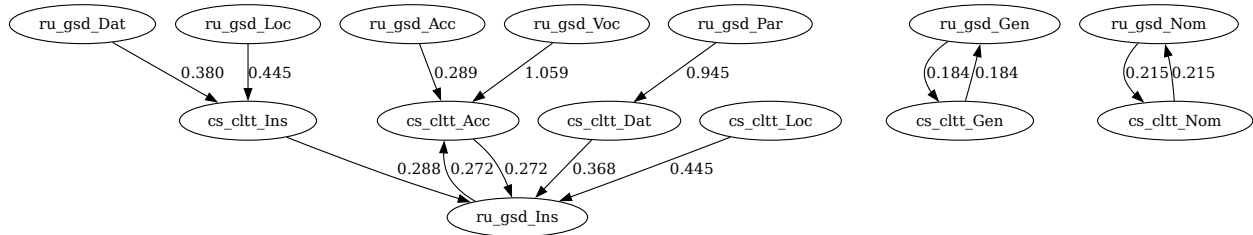


FIGURE 5 – Graphe des Plus Proches Voisins Russe-Tchèque

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 à une échelle très importante, ce qui brouille les résultats, ou bien 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.

2.5.1 PCA

On commence par appliquer une analyse en deux composantes principales (PCA) aux vecteurs représentant deux cas différents.

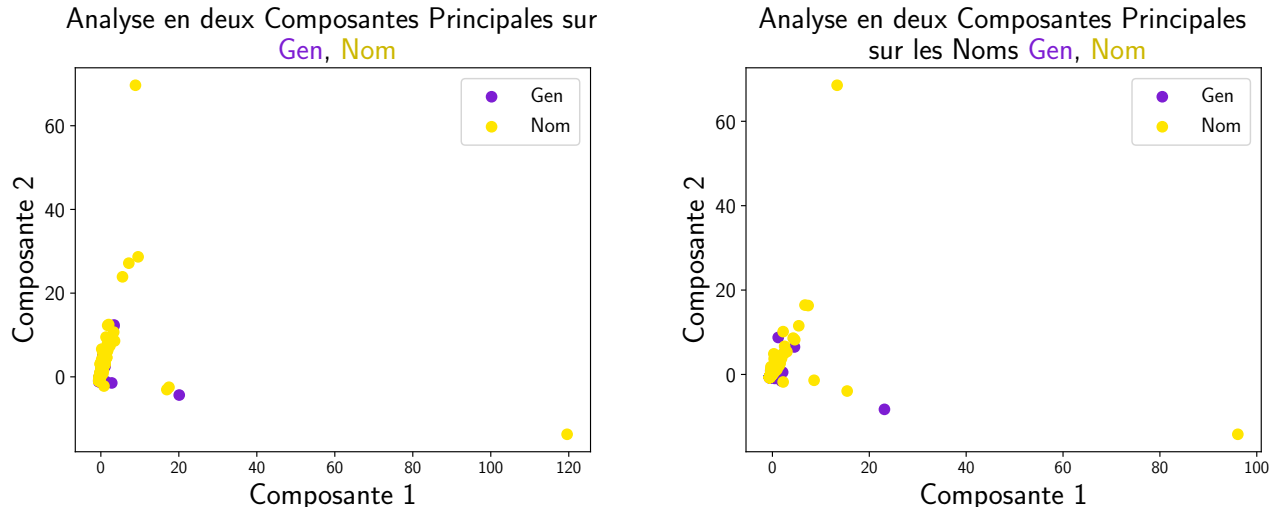


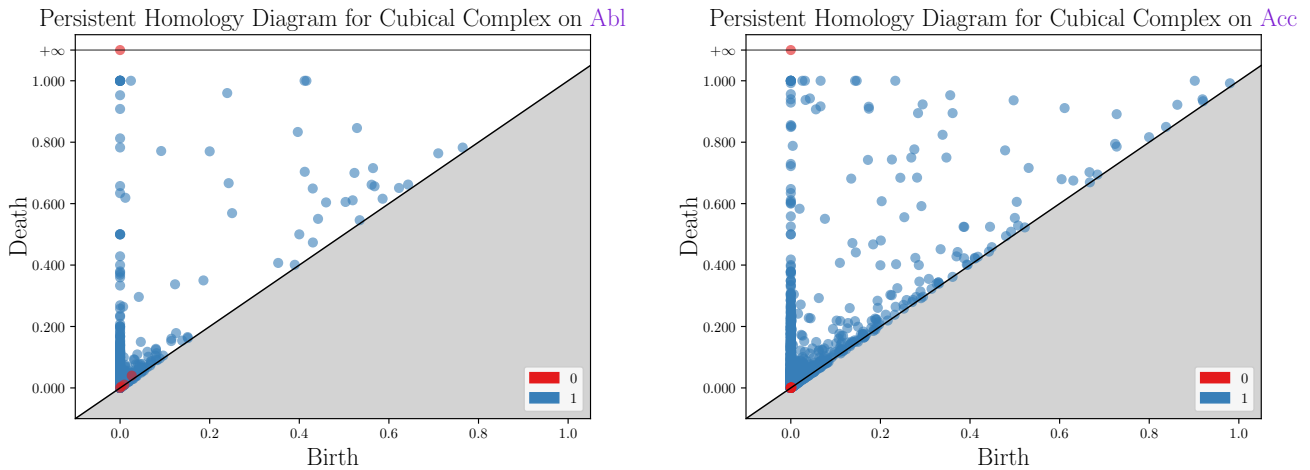
FIGURE 6 – Représentations de l'Analyse en deux Composantes Principales sur le Génitif et le Nominatif

En regardant les corpus qui sont éloignés du groupe principal, 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

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]

On utilise un complexe cubique pour essayer de représenter au mieux les groupes d'homologie (cf. [CM21] pour les détails) de la variété triangulée par les points de chaque cas. En effet, ici, il est compliqué de calculer directement un complexe moins régulier de part le nombre de points et la dimension de l'espace.

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

On a également calculé les complexes sur les Nominatifs et Accusatifs. On remarque qu'une forme générale se retrouve dans les diagrammes de persistences de chacun des cas. Lorsqu'on calcule la distance de Wasserstein entre deux diagrammes, on obtient le tableau suivant :

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

Les distances 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. Toutefois, les diagrammes restant très brouillons, il est difficile de trouver une variété plus abstraite avec la même homologie.

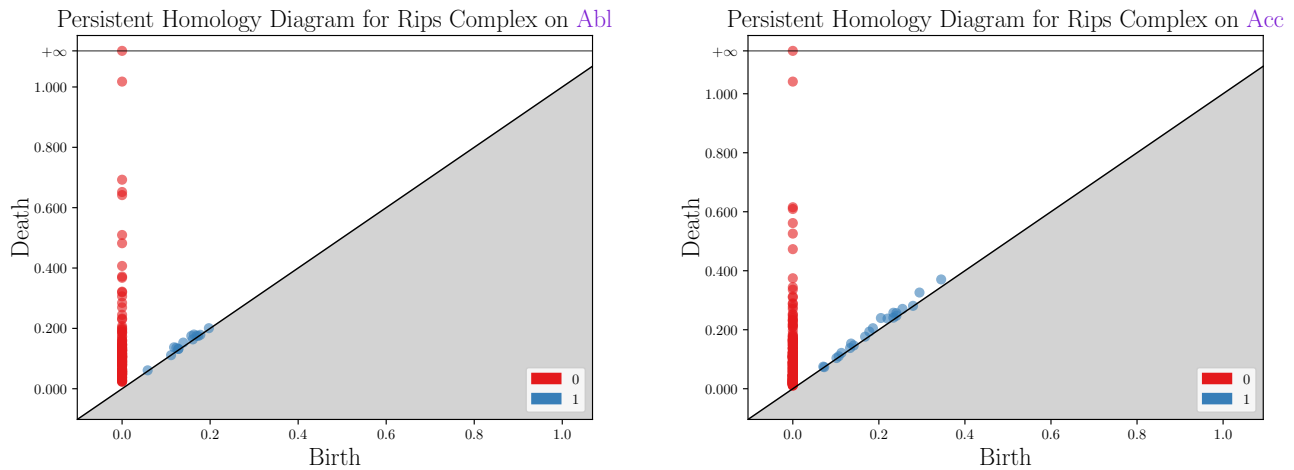


FIGURE 8 – Représentations de l'Homologie Persistente du Complexe de Rips sur $\{\text{Abl}, \text{Acc}\}$

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

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

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.

2.5.3 t-SNE

On essaie ensuite d'appliquer une analyse t-SNE en 2D, décrite dans [vdMH08]. Ici, on l'applique sur le Génitif et le Nominatif :

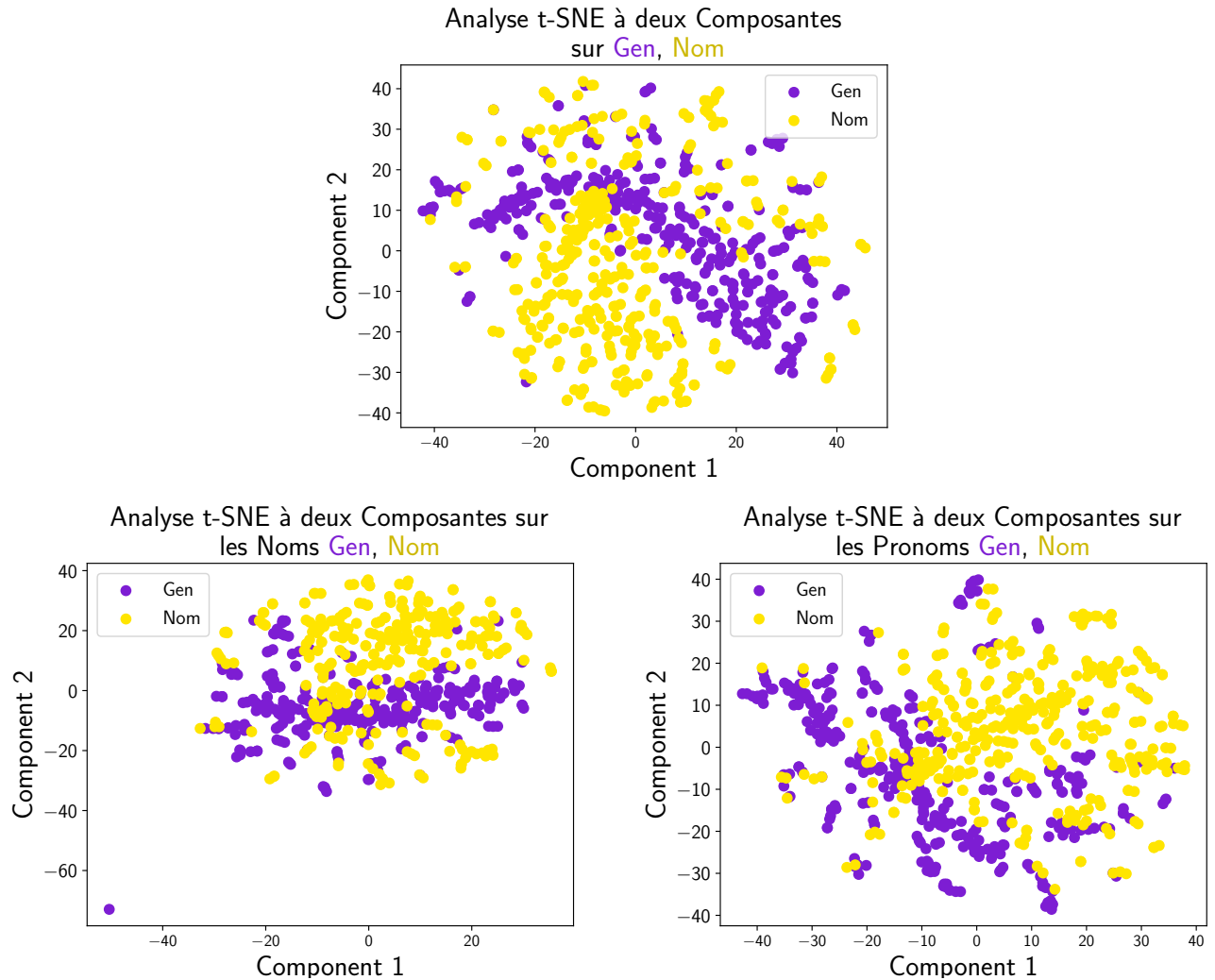
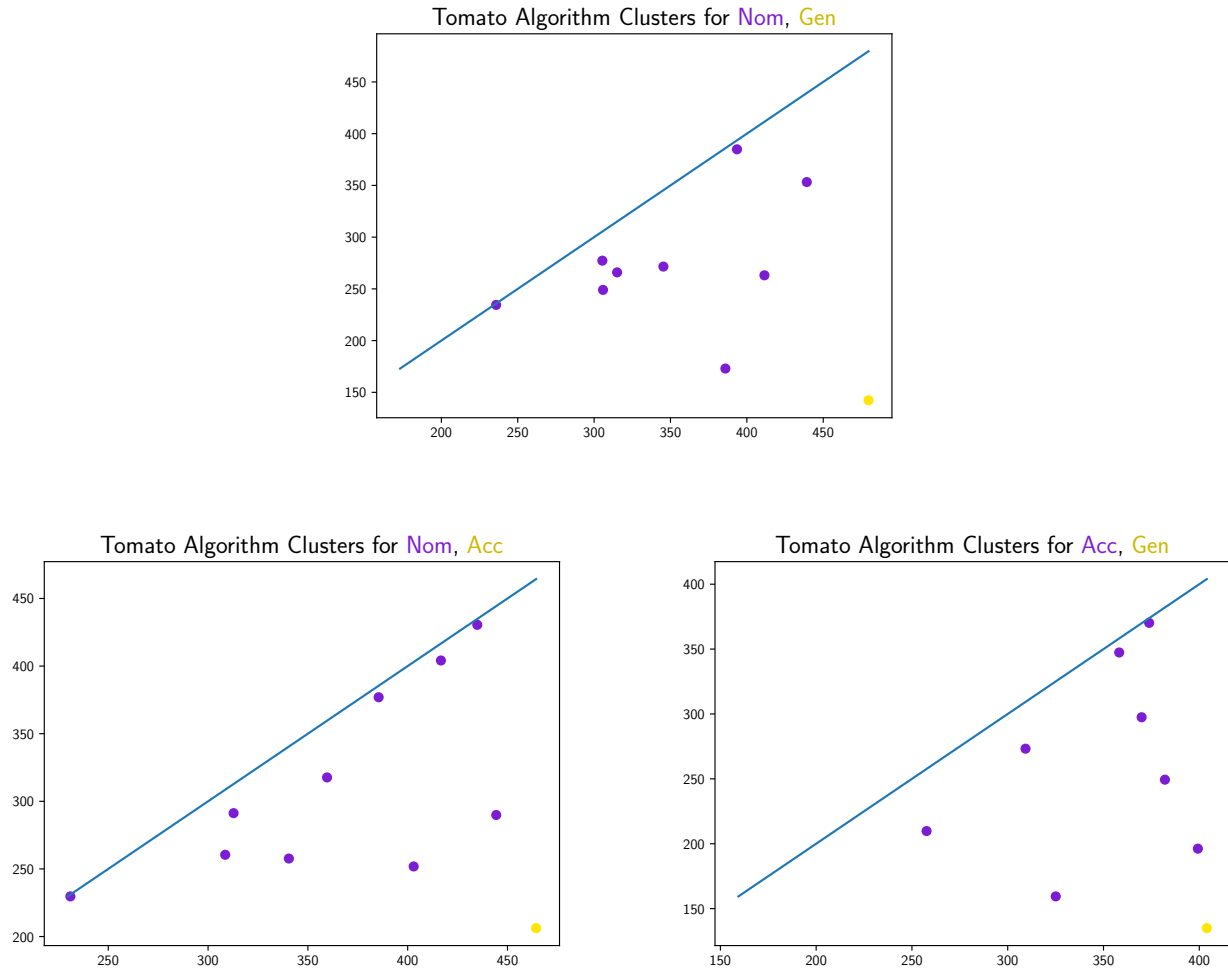


FIGURE 9 – Représentations de l'Analyse t-SNE à deux Composantes sur le Génitif et le Nominatif

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 représentation 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.

FIGURE 10 – Représentations des clusters trouvés par l’algorithme ToMATo sur les paires de $\{\text{Acc}, \text{Nom}, \text{Gen}\}$

Ici cependant, aucun des cas n’est clairement regroupé au sein de certains clusters, et l’algorithme ne produit aucun résultat utilisable.

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 suivante en appliquant l’algorithme pour $k = 11$ aux noms qui sont à l’Accusatif, au Génitif, au Locatif ou au Nominatif :

Case	Acc	Gen	Loc	Nom
Acc	130.000	62.000	51.000	34.000
Gen	69.000	156.000	16.000	42.000
Loc	35.000	57.000	29.000	34.000
Nom	29.000	28.000	9.000	227.000

TABLE 9 – KNN Confusion Matrix for $k = 11$ on **Acc**, **Gen**, **Loc**, **Nom**

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.

3 Approche Probabiliste

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

3.1 Barycentrisation

On a cherché jusque-là à comparer les cas. On va désormais essayer de les prototyper 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 nsubj d'un verbe

Abs nsubj d'un verbe avec obj

Acc obj d'un verbe

Gen nmod

Erg nsubj d'un verbe sans obj, ou obj d'un verbe

Dat iobj

Pour calculer les prototypes, on procède comme suit. On peut ainsi calculer son espérance (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 qui minimise l'énergie associée :

$$P = \arg \min_x \frac{1}{n} \sum_{i=1}^n d(x, x_i) = \arg \min_x E \left(x, (x_i)_{i \in \llbracket 1, n \rrbracket} \right)$$

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

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

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. Par exemple, l'Instrumental (qui marque l'outil utilisé pour

une action/un objet, e.g. dans une traduction de *J'ai mangé avec une fourchette*, on utiliserait l'instrumental pour *fourchette*) et le Locatif (qui marque le lieu d'une action, ou d'un objet, e.g. *J'ai pérégriné jusqu'à Saclay* ou *J'ai couru à côté de l'école* se traduiraient avec des locatifs³) 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, ce qui justifie ce que l'on pouvait inférer du tableau ci-dessus, le barycentre pour la distance de Wasserstein présente un profil similaire à la moyenne des instances. Par exemple :

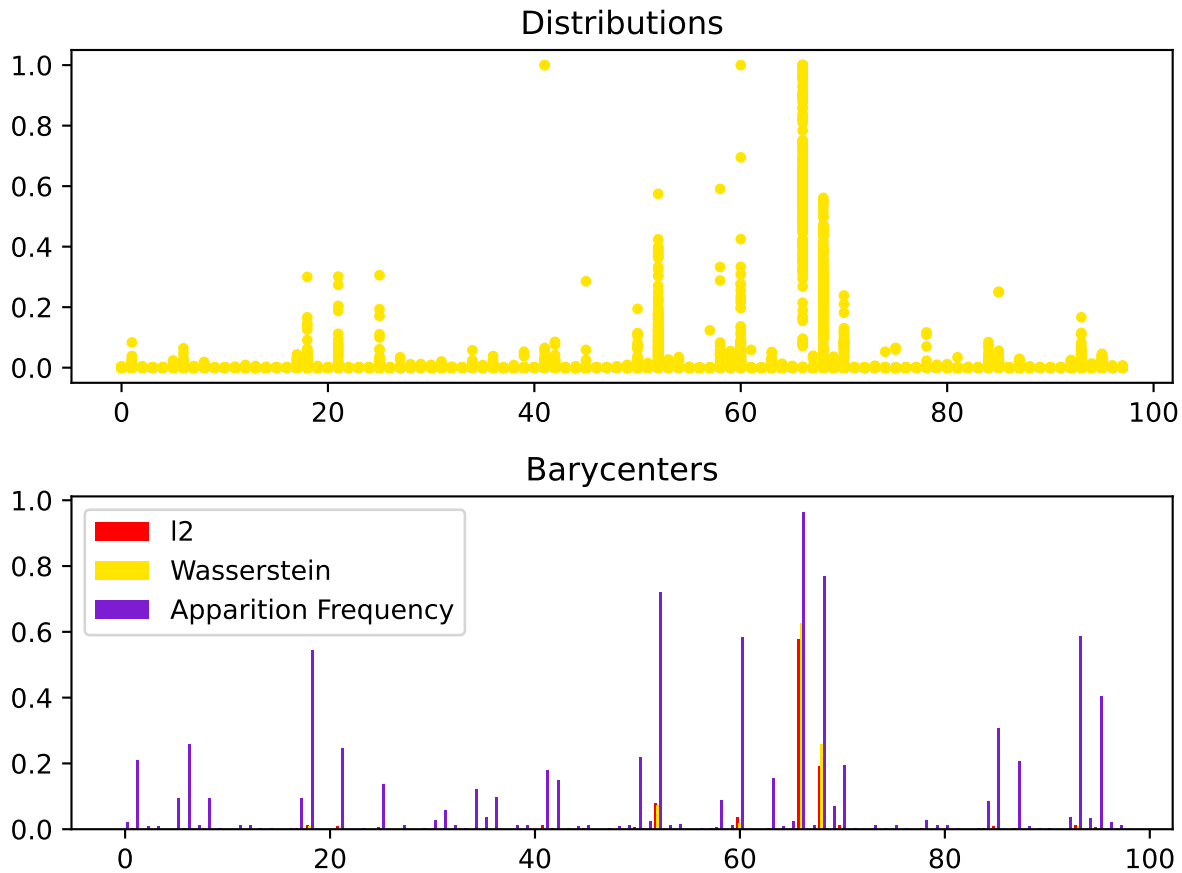


FIGURE 11 – Représentation des Données et des Prototypes proposés pour les noms à l'Accusatif

Sur la figure ci-dessus, l'axe des abscisses représente les différentes *reldep* apparaissant pour des noms à l'accusatif. 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. certaines langues comme le finnois possèdent plusieurs *locatifs* appelés inessif, élatif, illatif, adessif, ablatif, allatif... Ceux-ci servent à différencier, la destination, la direction, le lieu à côté...

3.2 Représentation des adpositions dans UD

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 anglais selon les caractéristiques morphologiques du tchèque, langage morphologiquement riche.

Nous avons donc cherché à déterminer le cas de certaines adpositions. 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 a que 9 pour le français).

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

Ici, 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, il existe plusieurs familles assez distinctes au sein desquelles les cas ne diffèrent que sémantiquement. Par ailleurs, il faut noter que le formalisme de UNIVERSAL DEPENDENCIES et la différence dans les annotateurs cause de fortes disparités dans les manières d'annoter, notamment sur l'annotation de cas sur les adverbes et les adpositions. Toutefois, il n'est pas illogique d'annoter un cas sur les adpositions, de par leurs fonctions sémantiques. Cependant, il apparaît que syntaxiquement elles ne sont que rarement distinguées.

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.
- [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.
- [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 : Visualizing 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.

Liste des tableaux

1	Proximities for Case=Gen	3
2	Proximities for Case=Nom	3
3	Proximities for Case=Voc	3
4	Proximities for Case=Acc	4
5	Extraits des Vecteurs de Reldep pour le Russe et le Tchèque	5
6	Taille d’Échantillons sur les cas en Russe et en Tchèque.	7
7	Distances de Wasserstein entre les Diagrammes de Persistence des Complexes Cubiques pour quelques Cas	9
8	Distances de Wasserstein entre les Diagrammes de Persistence des Complexes de Rips pour quelques Cas	9
9	KNN Confusion Matrix for $k = 11$ on Acc, Gen, Loc, Nom	11
10	Représentation des Principales <i>reldep</i> des Prototypes pour quelques Cas sur les noms	12
11	Représentation de quelques adpositions en français	14

Table des figures

1	Représentation d’une phrase et de ses relations de dépendances en anglais, source :[dMMNZ21] . . .	2
2	Graphes des Plus Proches Voisins Russe-Tchèque.	5
3	Graphes des Plus Proches Voisins Russe-Tchèque pour les Noms uniquement.	6
4	Graphes des Plus Proches Voisins Russe-Tchèque pour les Pronoms uniquement.	6
5	Graphe des Plus Proches Voisins Russe-Tchèque	7

6	Représentations de l'Analyse en deux Composantes Principales sur le Génitif et le Nominatif	8
7	Représentations de l'Homologie Persistente du Complexe Cubique sur $\{\mathbf{Abl}, \mathbf{Acc}\}$	8
8	Représentations de l'Homologie Persistente du Complexe de Rips sur $\{\mathbf{Abl}, \mathbf{Acc}\}$	9
9	Représentations de l'Analyse t-SNE à deux Composantes sur le Génitif et le Nominatif	10
10	Représentations des clusters trouvés par l'algorithme ToMATo sur les paires de $\{\mathbf{Acc}, \mathbf{Nom}, \mathbf{Gen}\}$. . .	11
11	Représentation des Données et des Prototypes proposés pour les noms à l'Accusatif	13