# Etude Computationnelle de la Stabilité Interlangue des Catégories Morphosyntaxiques

Rapport de Stage de L3

Matthieu Boyer

11 juillet 2024


## Table des matières



**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.

# 1 Pourquoi ?

Cette citation de Martin Haspelmath sur la différence entre une catégorie linguistique descriptive dans un langage et une catégorie linguistique comparative dans le méta-langage est le point de départ de notre étude.

> 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).

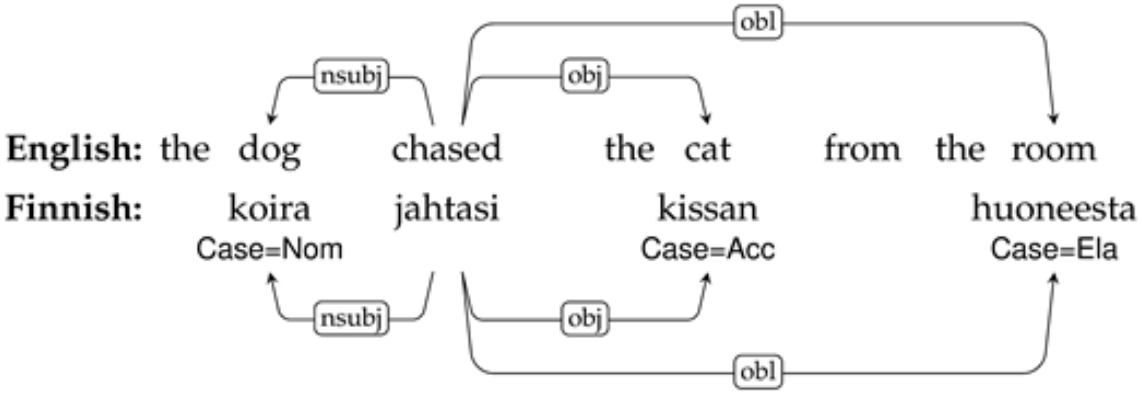


FIGURE 1 – Représentation d'une phrase et de ses relations de dépendances en anglais, source :[dMMNZ21]

*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.

## 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. 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 décrits table 1 dans [dMMNZ21].

## 3 Première Approche

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 vectorielle 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].

# 4 Avec la distance Cosinus

On mesure en utilisant la distance Cosinus entre deux vecteurs la proximité entre ceux ci :

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

On obtient alors les résultats suivants pour quelques cas :

| Proximity with : | Abl | Acc | Dat | Gen | Ins | Loc | Nom | Voc |
|---|---|---|---|---|---|---|---|---|
| First Quartile | 0.300 | 0.028 | 0.070 | 0.037 | 0.408 | 0.199 | 0.020 | 0.000 |
| Median | 0.840 | 0.202 | 0.340 | 0.198 | 0.631 | 0.622 | 0.067 | 0.004 |
| Third Quartile | 0.942 | 0.393 | 0.690 | 0.416 | 0.870 | 0.915 | 0.158 | 0.025 |
| Mean | 0.644 | 0.242 | 0.395 | 0.259 | 0.595 | 0.570 | 0.115 | 0.035 |

Table 1 – Proximities for Case=Abl

| Proximity with : | Abl | Acc | Dat | Gen | Ins | Loc | Nom | Voc |
|---|---|---|---|---|---|---|---|---|
| First Quartile | 0.028 | 0.394 | 0.030 | 0.020 | 0.042 | 0.014 | 0.038 | 0.000 |
| Median | 0.202 | 0.711 | 0.181 | 0.123 | 0.236 | 0.176 | 0.137 | 0.007 |
| Third Quartile | 0.393 | 0.860 | 0.379 | 0.302 | 0.408 | 0.381 | 0.272 | 0.040 |
| Mean | 0.242 | 0.616 | 0.236 | 0.196 | 0.255 | 0.230 | 0.188 | 0.042 |

Table 2 – Proximities for Case=Acc

| Proximity with : | Abl | Acc | Dat | Gen | Ins | Loc | Nom | Voc |
|---|---|---|---|---|---|---|---|---|
| First Quartile | 0.070 | 0.030 | 0.085 | 0.022 | 0.128 | 0.041 | 0.018 | 0.000 |
| Median | 0.340 | 0.181 | 0.320 | 0.134 | 0.403 | 0.303 | 0.072 | 0.006 |
| Third Quartile | 0.690 | 0.379 | 0.632 | 0.341 | 0.637 | 0.620 | 0.161 | 0.032 |
| Mean | 0.395 | 0.236 | 0.372 | 0.214 | 0.407 | 0.360 | 0.119 | 0.036 |

Table 3 – Proximities for Case=Dat

| 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 4 – Proximities for Case=Gen

| Proximity with : | Abl | Acc | Dat | Gen | Ins | Loc | Nom | Voc |
|---|---|---|---|---|---|---|---|---|
| First Quartile | 0.408 | 0.042 | 0.128 | 0.056 | 0.443 | 0.334 | 0.035 | 0.000 |
| Median | 0.631 | 0.236 | 0.403 | 0.249 | 0.609 | 0.577 | 0.113 | 0.012 |
| Third Quartile | 0.870 | 0.408 | 0.637 | 0.449 | 0.800 | 0.829 | 0.211 | 0.055 |
| Mean | 0.595 | 0.255 | 0.407 | 0.282 | 0.596 | 0.549 | 0.153 | 0.053 |

Table 5 – Proximities for Case=Ins

| Proximity with : | Abl | Acc | Dat | Gen | Ins | Loc | Nom | Voc |
|---|---|---|---|---|---|---|---|---|
| First Quartile | 0.199 | 0.014 | 0.041 | 0.027 | 0.334 | 0.131 | 0.020 | 0.000 |
| Median | 0.622 | 0.176 | 0.303 | 0.188 | 0.577 | 0.556 | 0.078 | 0.007 |
| Third Quartile | 0.915 | 0.381 | 0.620 | 0.400 | 0.829 | 0.904 | 0.156 | 0.038 |
| Mean | 0.570 | 0.230 | 0.360 | 0.243 | 0.549 | 0.526 | 0.115 | 0.042 |

TABLE 6 – Proximities for Case=Loc

| 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 7 – 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 8 – Proximities for Case=Voc

On obtient 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

Toutefois, cette méthode est très limitée. En effet, on ne considère ici que 9 des 45 cas définis dans au moins un corpus. Par ailleurs, 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.

## 5 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, 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 l'odrdre de 5) dans un grand espace (dimension 228), rend l'intersection 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-trivialement la matrice.

## 6 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 anomalie : l'angle entre le vocatif du farsi et le système de cas arabe est de l'ordre de $10^{-16}$. En regardant de plus près les corpus farsis [1], on observe que cela découle d'une idiosyncracie 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 apposition. 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 semble toutefois bon de noter qu'il y a sans doute de nombreuses autres anomalies du style dans les corpus.

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 apposition (cf. supra).

# 7 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 9 – 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. On obtient qu'ici, le datif russe est plus proche du génitif tchèque que du datif tchèque.

| | 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 10 – 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, 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 uniforme 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

1. et non pas les dindes.

avoir de $n$-cycle pour $n \geq 3$ et décrit des relations de proximité minimale. On obtient alors le graphe suivant pour le tchèque et le russe :

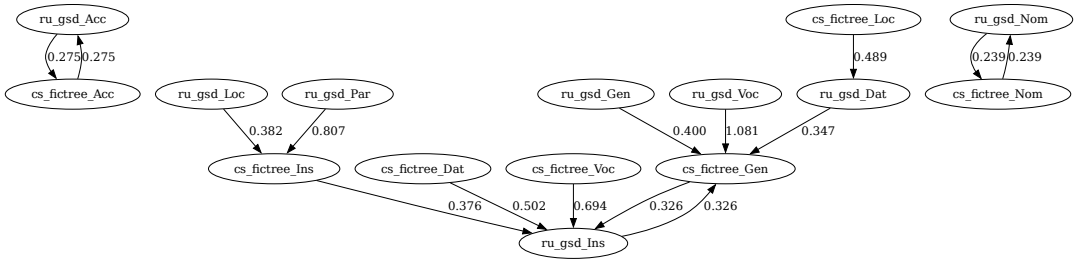


FIGURE 2 – Graphes des Plus Proches Voisins Russe-Tchèque.

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 :

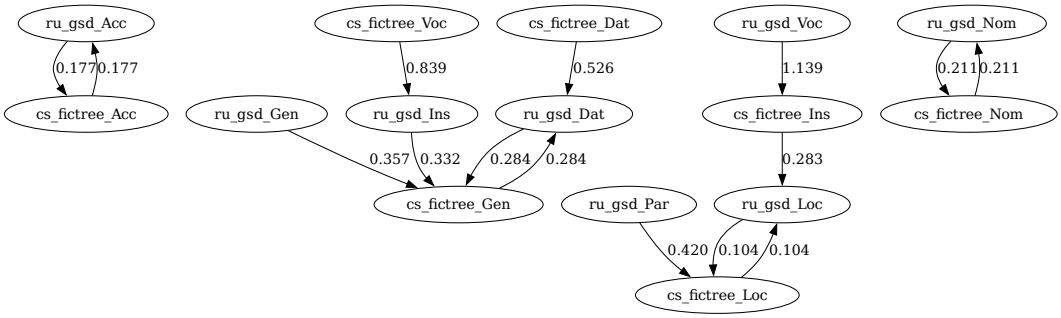


FIGURE 3 – Graphes des Plus Proches Voisins Russe-Tchèque pour les Noms uniquement.

On observe notamment que pour les noms, la structure du graphe reste la même. Le vocatif russe et le vocatif tchèque, peu utilisées 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 :

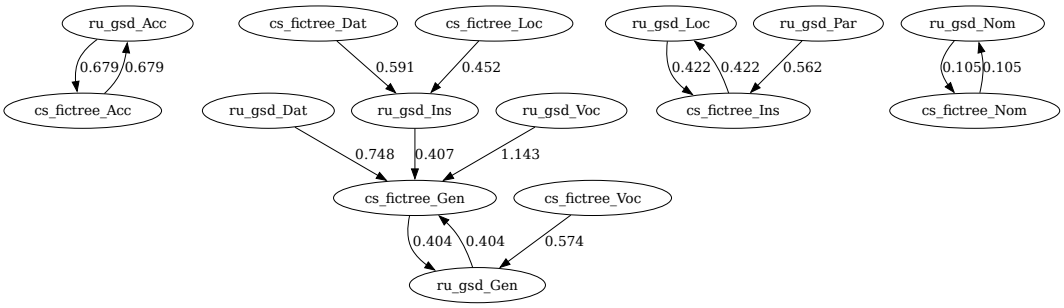


FIGURE 4 – Graphes des Plus Proches Voisins Russe-Tchèque pour les Pronoms uniquement.

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.

*Échantillons pour les noms :*

| Cas | Russe | Tchèque |
|---|---|---|
| Acc | 2807 | 5960 |
| Dat | 1029 | 0861 |
| 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 11 – 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 transparaitre de ces graphes. Il faut cependant noter que les graphes dépendent très fortement du corpus proposé :

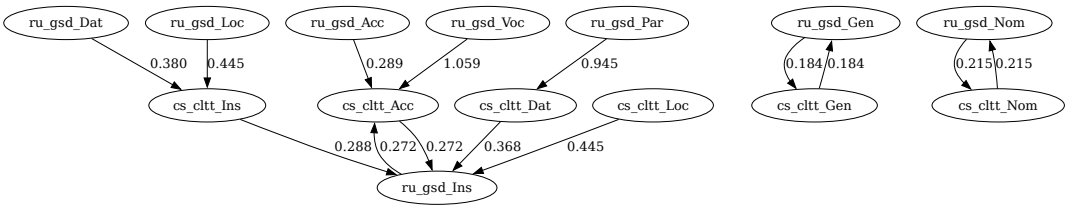


FIGURE 5 – Graphe des Plus Proches Voisins Russe-Tchèque

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.

# 8 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.

## 8.1 PCA

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

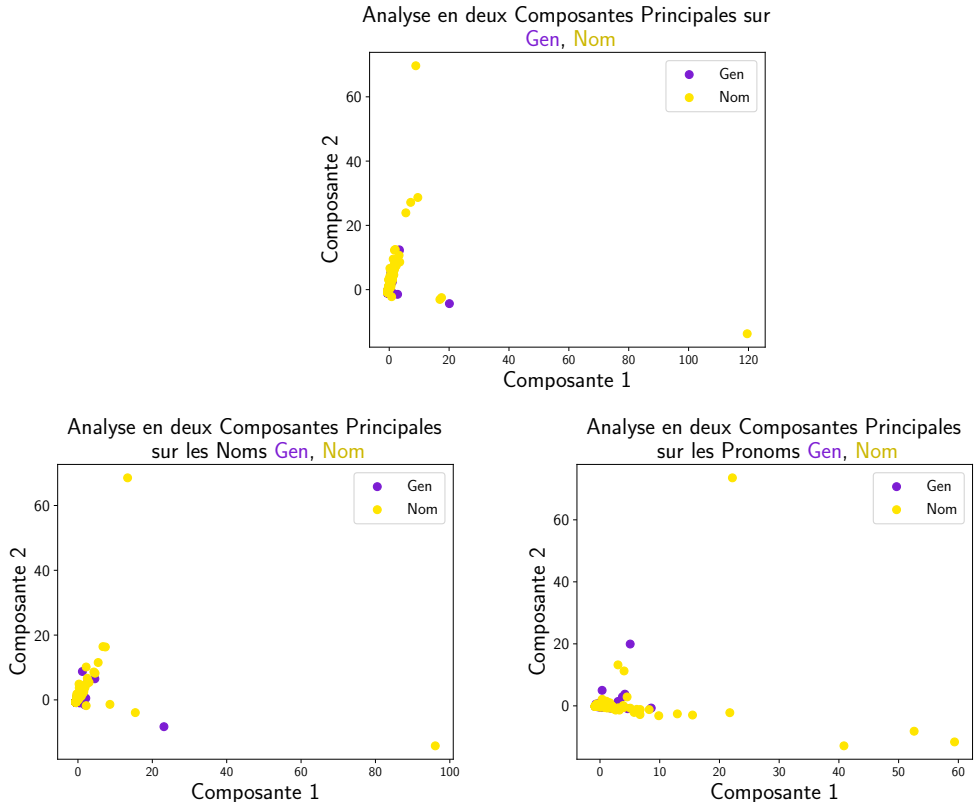


FIGURE 6 – Représentations de l'Analyse en deux Composantes Principales sur le Génitif et le Nominatif

## 8.2 t-SNE

On essaie ensuite d'appliquer une analyse t-SNE en 2D, décrite dans [vdMH08]

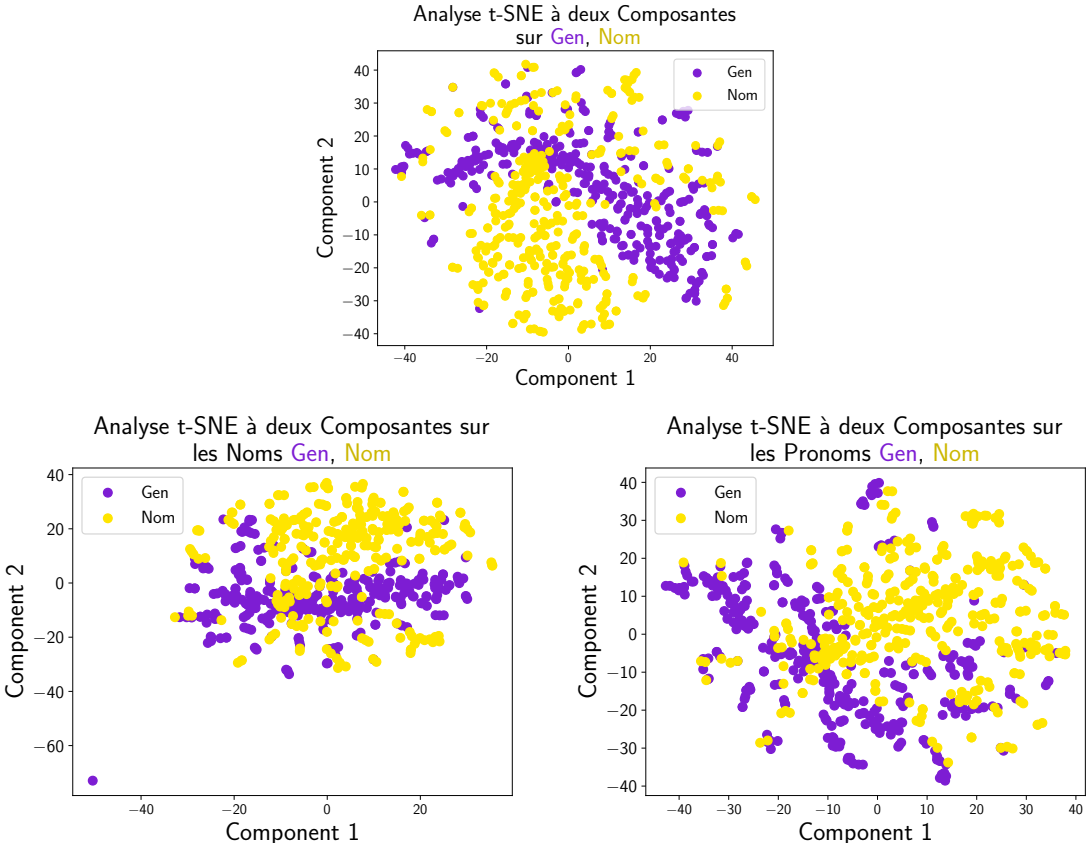


FIGURE 7 – Représentations de l'Analyse t-SNE à deux Composantes sur le Génitif et le Nominatif

# Références


[dMMNZ21] Marie-Catherine de MARNEFFE, Christopher D. MANNING, Joakim NIVRE et Daniel ZEMAN : Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021.

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

[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.