Etude Computationnelle de la Stabilité Interlangue des Catégories Morphosyntaxiques

Rapport de Stage de L3

Matthieu Boyer

9 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).

Martin Haspelmath, How comparative concepts and descriptive linguistic categories are different 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.
Pour ce faire, nous allons considérer que les relations de dépendances (*reldep*) décrites par les annotations de Universal Dependencies (UD) sont une manière de représenter des catégories comparatives.

2 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, accessibles en ligne.

2.1 Avec la distance Cosinus

On mesure en utilisant la distance Cosinus entre deux vecteurs [1], la proximité entre ceux ci. On obtient alors les résultats suivants pour quelques cas :

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.06954	0.19691	0.06857	0.06515	0.00631	0.04309	0.02606
NLow	79473	29324	44917	59485	28892	28821	14785
NHigh	1917	31291	3466	1736	159	2688	1429
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.06013	0.32862	0.00437	0.0206	0.0	0.0	0.0

Table 1 – Proximities for Case=Acc

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.03328	0.05376	0.09264	0.05258	0.00444	0.05714	0.0343
NLow	57901	38008	28140	34565	21502	16842	8188
NHigh	412	2446	9658	2324	12	7240	4837
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2 – Proximities for Case=Dat

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.05607	0.06261	0.06254	0.13129	0.00793	0.04625	0.03021
NLow	68940	54020	39559	41116	25695	25608	14532
NHigh	2093	1564	2948	22214	133	2639	1576
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.02968	0.01221	0.0	0.05743	0.0	0.0	0.0

Table 3 – Proximities for Case=Gen

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

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.0149	0.02145	0.04055	0.02964	0.00232	0.0558	0.02732
NLow	25261	17671	11662	14847	9525	9900	3493
NHigh	220	935	4763	1495	68	10558	5737
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 4 – Proximities for Case=Loc

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.26295	0.07799	0.04112	0.06237	0.01181	0.02566	0.0162
NLow	18404	80944	75574	79415	35362	48157	28185
NHigh	54192	2105	506	2145	254	211	107
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.54978	0.0799	0.01559	0.05144	0.0	0.0	0.0

TABLE 5 – Proximities for Case=Nom

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.00425	0.00289	0.00255	0.00417	0.03415	0.0022	0.00047
NLow	13024	12083	10996	10622	1402	7273	1828
NHigh	88	55	15	74	8299	51	2
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 6 – Proximities for Case=Voc

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc	Case=Abl
Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean	0.00599	0.00569	0.0167	0.0087	0.00033	0.01607	0.02558
NLow	7840	6501	3013	6616	1545	2368	2056
NHigh	49	453	2916	457	0	3704	6022
First Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Third Quartile	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE 7 – Proximities for Case=Abl

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.

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, 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.

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 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 [2], 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).

2.4 Distance Euclidienne

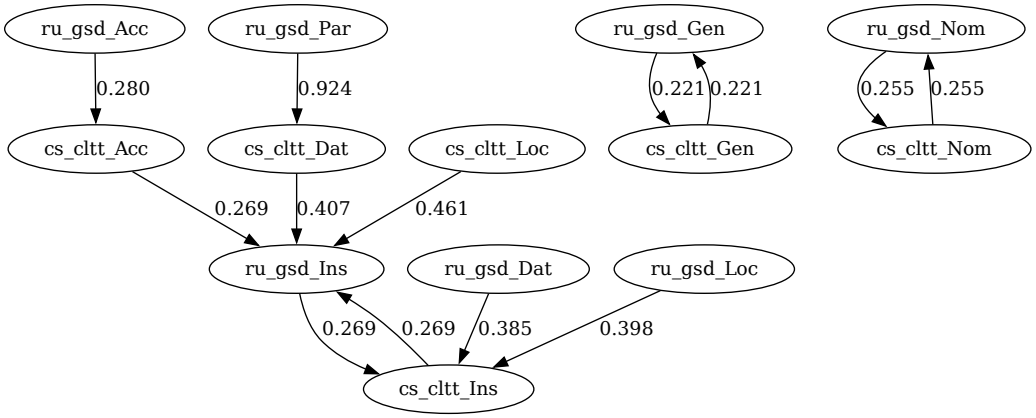
On considère cette fois la distance euclidienne entre tous deux vecteurs, qu'on aura au préalable normalisés pour qu'ils représentent des distributions de probabilité. On utilise ces données pour déterminer, entre deux corpus (ici le Czech-CLTT et le Russian-GSD), quel cas de l'autre langage est le plus proche d'un cas du premier. On obtient qu'ici, le datif russe est plus proche du génitif tchèque que du datif tchèque.

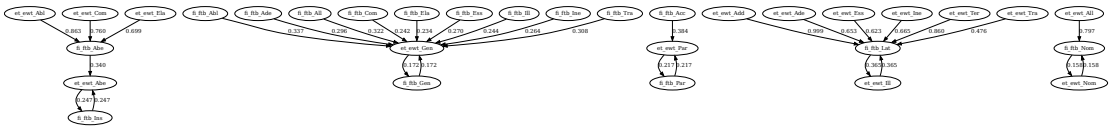
	Dat RU	Gen CZ	Gen RU	Dat CZ
`Total`	1711	2631	2070	277
`obl`	450	208	219	48
`iobj`	340	0	0	0
`amod`	243	736	475	54
`nmod`	300	1000	980	24
`conj`	112	225	84	21
`case`	0	340	1	87
`det`	34	80	79	3

TABLE 8 – Extraits des Vecteurs de Reldep pour le Russe et le Tchèque

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. On obtient alors le graphe suivant pour le tchèque et le russe :

2. et non pas les dindes.





On enlève ensuite `conj, det` puisque ces reldep démontrent l'accord vers la tête, et donc des doublons dans les données et on se restreint aux mots de nature *Nom*. Ceci permet d'éviter la variance liée aux adjectifs, pronoms et dans de rares cas (ceux-ci restant majoritairement au cas absolu) aux adverbes. 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.
On obtient alors le graphe suivant pour le tchèque et le russe :