Etude Computationnelle de la Stabilité Interlangue des Catégories Morphosyntaxiques

Rapport de Stage de L3

Matthieu Boyer

25 juin 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 Why ?

Martin Haspelmath sur la différence entre une catégorie linguistique descriptive dans un langage et une catégorie linguistique comparative dans le méta-langage :

> 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 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 en utilisant par exemple la distance cosinus. On trouve par exemple les résultats suivants :

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc
Median	0.44926	0.58137	0.44742	0.42992	0.51341	0.47854
Mean	0.50024	0.57176	0.48805	0.48151	0.53144	0.51641
NLow	91978	59432	65558	79386	33606	40618
NHigh	105318	138708	61716	71148	47751	54974
First Quartile	0.24595	0.31676	0.2513	0.24237	0.26514	0.27157
Third Quartile	0.76172	0.82656	0.71989	0.71855	0.78883	0.76127

TABLE 1 – Proximities for Case=Acc

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc
Median	0.43377	0.43776	0.47225	0.41885	0.44723	0.48964
Mean	0.48132	0.47951	0.49914	0.46633	0.4891	0.51874
NLow	66862	53262	54112	54008	28296	32922
NHigh	57633	45432	51652	40254	24589	41547
First Quartile	0.22958	0.24764	0.26381	0.23872	0.23842	0.27469
Third Quartile	0.7282	0.69449	0.71959	0.67239	0.7268	0.75314

TABLE 2 – Proximities for Case=Dat

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc
Median	0.43798	0.43274	0.41881	0.51997	0.46953	0.4496
Mean	0.49172	0.48325	0.47	0.53942	0.51311	0.49633
NLow	84864	73628	64464	68976	34294	43224
NHigh	84316	67125	50458	105768	39952	46619
First Quartile	0.23575	0.24233	0.23766	0.27291	0.24765	0.25358
Third Quartile	0.75452	0.72363	0.68244	0.81685	0.77729	0.73544

TABLE 3 – Proximities for Case=Gen

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc
Median	0.44588	0.46849	0.48794	0.46892	0.51927	0.57887
Mean	0.49841	0.51443	0.51995	0.50402	0.54114	0.58953
NLow	27886	21110	21518	21266	11000	18048
NHigh	29971	27910	28499	25073	16744	45864
First Quartile	0.23495	0.26545	0.27671	0.26711	0.26242	0.3383
Third Quartile	0.77133	0.76979	0.7566	0.72731	0.83677	0.87388

TABLE 4 – Proximities for Case=Loc

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc
Median	0.68656	0.45487	0.44484	0.44427	0.55354	0.48988
Mean	0.63322	0.5029	0.49385	0.49775	0.56313	0.528
NLow	48728	96048	86424	97554	36366	54370
NHigh	224224	112469	86575	103187	72176	77696
First Quartile	0.37983	0.24768	0.23348	0.23564	0.28434	0.25232
Third Quartile	0.88616	0.76515	0.75861	0.76997	0.85375	0.82257

TABLE 5 – Proximities for Case=Nom

Proximity with :	Case=Nom	Case=Acc	Case=Dat	Case=Gen	Case=Voc	Case=Loc
Median	0.53652	0.52529	0.49388	0.55171	0.88359	0.56814
Mean	0.55947	0.54856	0.53176	0.55858	0.76794	0.57436
NLow	10168	9358	10354	8638	2688	6416
NHigh	18219	14842	13198	14931	33512	12001
First Quartile	0.27545	0.27187	0.24879	0.28231	0.63158	0.27575
Third Quartile	0.87261	0.83938	0.81784	0.84647	0.97788	0.8884

TABLE 6 – Proximities for Case=Voc