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

Plan

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]

Données et Universal Dependencies

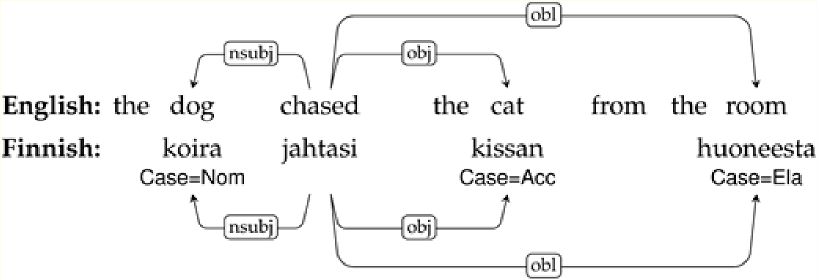


Figure: Représentation d'une Phrase en Anglais et en Finnois et de ses Relations de Dépendances, source:[dMMNZ21], [Z^+24]

Plan

Méthode Géométrique

On pose:

- $\mathbb{R}^{|R|}$ où $R = \{\textit{reldep}\}$
- $\mathcal{T}$ l'ensemble des corpus et $\mathscr{C}$ l'ensemble des cas présents dans au moins un corpus
- $\nu(T, C)$ la représentation du cas C dans le corpus T
- $\mathcal{E}(C) = \{T | \nu(T, C) \neq 0\}$
- $\mathcal{C}(T) = \{C | \nu(T, C) \neq 0\}$

Avec la distance Cosinus

Pour:

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

On calcule:

$$d_{\cos}\left(\nu\left(T_1, C_1\right), \nu\left(T_2, C_2\right)\right) \text{ où } T_1, T_2 \in \mathcal{E}(C_1) \times \mathcal{E}(C_2)$$

Avec la distance Cosinus

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: 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: Proximité Angulaire pour le *Nominatif*

Système de Cas

On définit le Système de Cas d'une langue:

$\mathcal{S}(T) = \mathrm{Vect}\left((\nu(C,T))_{C\in\mathcal{C}(T)}\right)$

- Avec l'algorithme de Zassenhaus, on calcule:

$$\left\{\dim(\mathcal{S}(T_1)\cap\mathcal{S}(T_2)) \mid (T_1,T_2)\in\mathcal{T}^2\right\}$$

$$\left\{\dim(\mathcal{S}(T_1)+\mathcal{S}(T_2)) \mid (T_1,T_2)\in\mathcal{T}^2\right\}$$

Système de Cas

On définit le Système de Cas d'une langue:
$\mathcal{S}(T) = \text{Vect}\left((\nu(C,T))_{C\in\mathcal{C}(T)}\right)$

- Avec l'algorithme de Zassenhaus, on calcule:

$$\left\{\dim(\mathcal{S}(T_1) \cap \mathcal{S}(T_2)) \mid (T_1, T_2) \in \mathcal{T}^2\right\}$$

$$\left\{\dim(\mathcal{S}(T_1) + \mathcal{S}(T_2)) \mid (T_1, T_2) \in \mathcal{T}^2\right\}$$

- Avec la distance cosinus, on calcule:

$$\left\{d_{\cos}(\nu(T_1, C_1), p_{\mathcal{S}(T_2)}(\nu(T_1, C_1))) \mid T_1, T_2 \in \mathcal{T}^2, C_1 \in \mathcal{C}(T_1)\right\}$$

Distance Euclidienne

On calcule alors, pour C_1, C_2 deux cas donnés,

$$\{d\left(\nu\left(T_1, C_1\right), \nu\left(T_2, C_2\right)\right) \mid T_1, T_2 \in \mathcal{E}(C_1) \times \mathcal{E}(C_2)\}$$

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: Proximité pour la distance euclidienne avec l'*Accusatif*

Graphes des Voisins

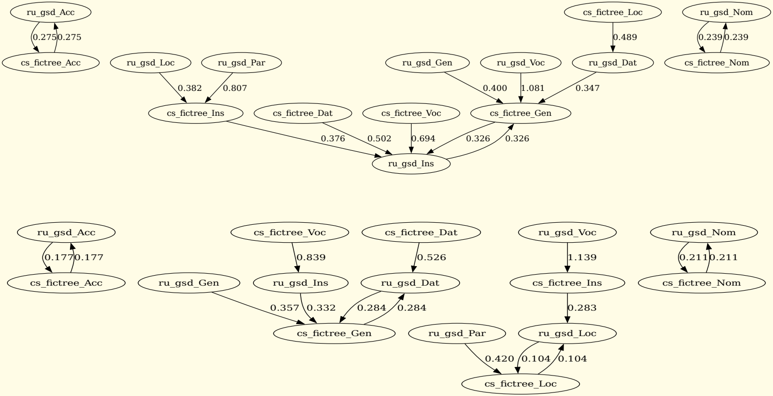


Figure: Graphes des Plus Proches Voisins Russe-Tchèque

Plan

PCA

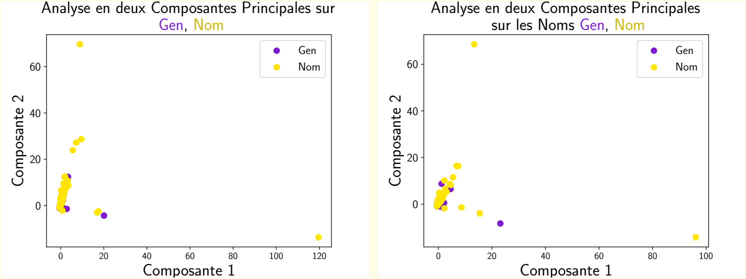


Figure: Représentations de l'Analyse en deux Composantes Principales sur le Génitif et le Nominatif

Analyse Topologique des Données – 1

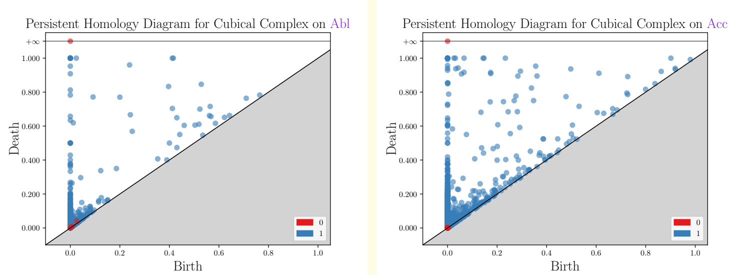


Figure: Représentations de l'Homologie Persistente du Complexe Cubique sur $\{\texttt{Abl}, \texttt{Acc}\}$

Analyse Topologique des Données – 2

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:

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

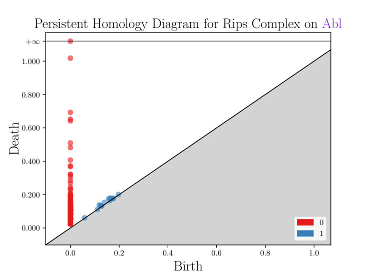
où $\Gamma\left(\mu, \pi\right)$ est l'ensemble des couplages de μ, π.

Analyse Topologique des Données – 2

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: Distances de Wasserstein entre les Diagrammes de Persistence des Complexes Cubiques pour quelques Cas

Analyse Topologique des Données – 3



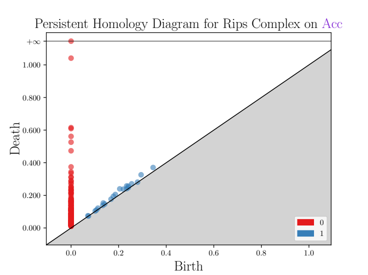


Figure: Représentations de l'Homologie Persistente du Complexe de Rips sur $\{\texttt{Abl}, \texttt{Acc}\}$

Analyse Topologique des Données – 3

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: Distances de Wasserstein entre les Diagrammes de Persistence des Complexes de Rips pour quelques Cas

t-SNE

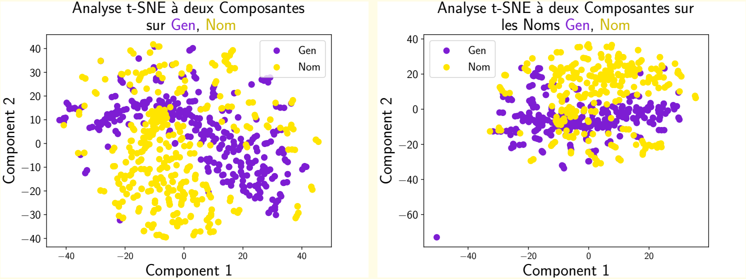


Figure: Représentations de l'Analyse t-SNE (cf. [vdMH08]) à deux Composantes sur le Génitif et le Nominatif

Clustering avec ToMATo

Cet algorithme utilise les ensembles de sur-niveau associés à une fonction f (les $F^\alpha = f^{-1}\left([\alpha, +\infty[\right)$) pour construire le diagramme de persistence (voir [CM21] et [CGOS11]).

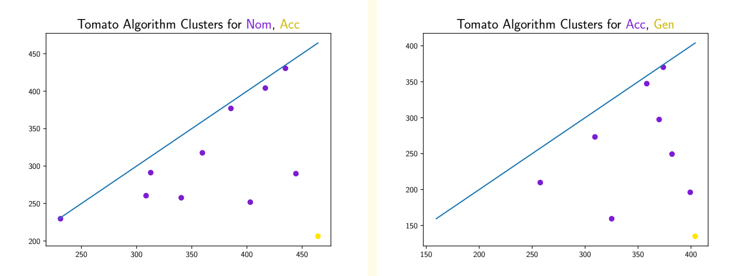


Figure: Représentations des Clusters trouvés par l'Algorithme ToMATo sur les Paires de $\{\texttt{Acc}, \texttt{Nom}, \texttt{Gen}\}$

Clustering avec KNN

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

Table: Heatmap de l'Algorithme KNN avec $k = 11$ sur `Acc`, `Gen`, `Loc`, `Nom`

Plan

Barycentrisation – Théorie

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

Barycentrisation – Empirique

On associe à un cas une variable aléatoire $\mathcal{R}(C)$. On peut alors calculer $\mathbb{E}(\mathcal{R}(C))$ mais également le barycentre de ses réalisations pour la distance-1 de Wasserstein. On rappelle par ailleurs que le barycentre de n points pour une distance d est le point P qui minimise l'énergie E associée:

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

Barycentrisation – Empirique

Cas	Prototype	iobj	nmod	nsubj	obj	obl
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

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

Comparaison aux Données

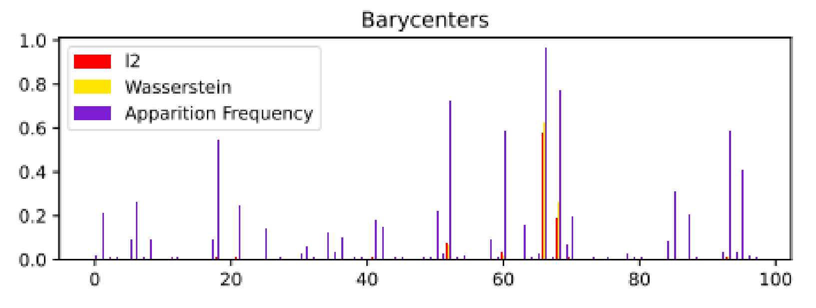


Figure: Représentation des Données et des Prototypes proposés pour les Noms à l'Accusatif

Représentation des adpositions dans UD

Adposition	`advcl`	`iobj`	`nmod`	`obj`	`obl`
À	0.01668		0.17343	0.00381	0.63373
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
De	0.02096		0.67966	0.01312	0.14296
Sous	0.00217		0.22898	0.72797	
Sauf	0.10714		0.22619		0.38095

Table: Quelques adpositions françaises, suivant [KSGK+17]

Plan

Conclusion

Nous avons trouvé:

- Une structure géométrique générale de la syntaxe des cas
- Des classes d'équivalence syntaxique de cas
- Des disparités dans les manières d'annoter les cas
- Des preuves appuyant la thèse de Martin Haspelmath
- Une manière d'annoter des adpositions

Il reste à étudier:

- la viabilité d'un parser syntaxique basé sur des fusions de cas;
- l'applicabilité d'un parser d'une langue sur une autre langue proche en système de cas.

Bibliographie I

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

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

Bibliographie II

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

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

Bibliographie III

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.

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

Bibliographie IV

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.