



MODAL D'INFORMATIQUE

INF473G – Graphe Global Géant

DÉVELOPPEMENT ET UTILISATION D'OUTILS NUMÉRIQUES POUR
L'ÉTUDE COMPARATIVE DE LA LINGUISTIQUE

Tristan FRANÇOIS et Christian KOTAIT

Motivation du projet

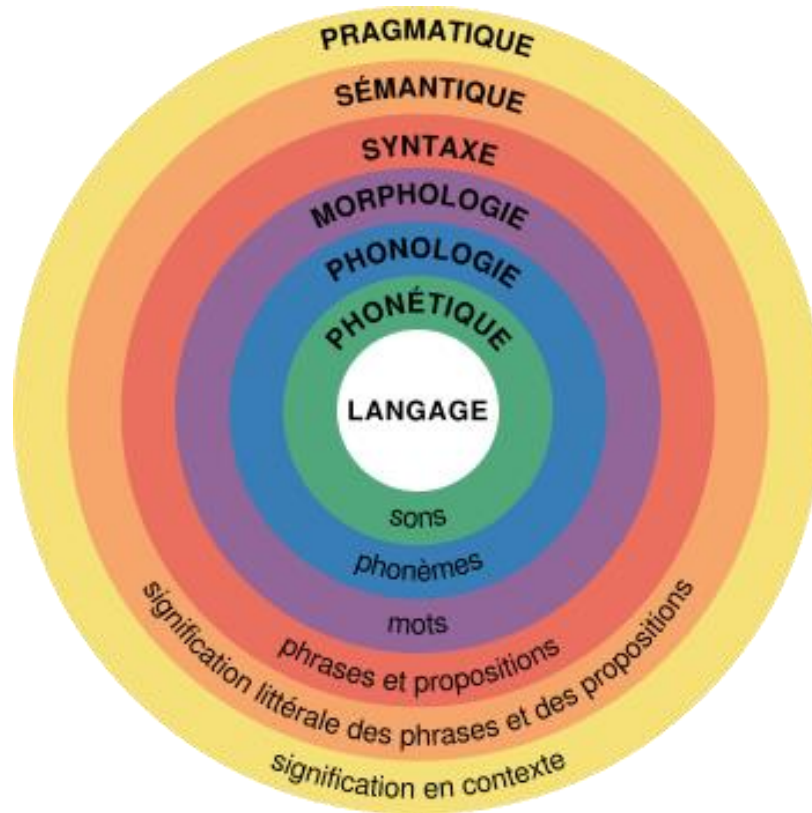


Figure 1 : Principaux domaines de la linguistique

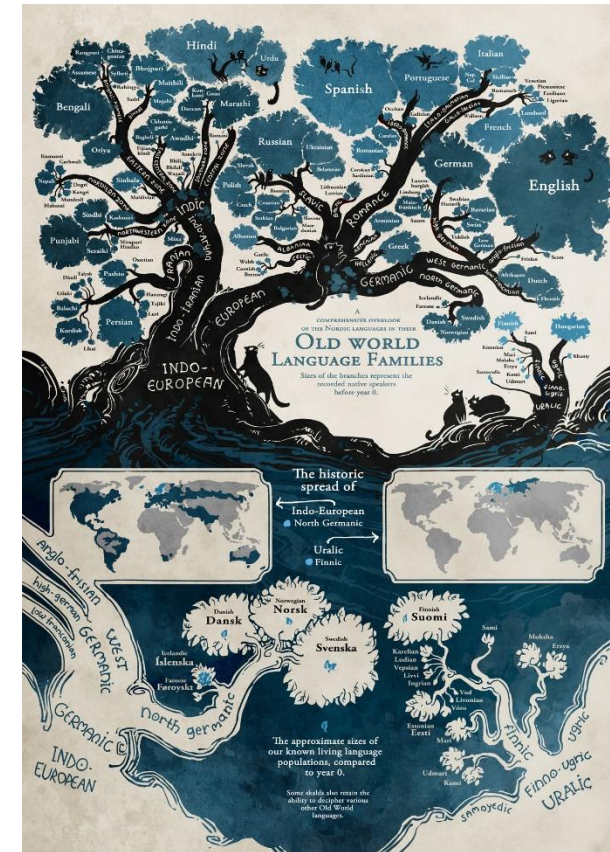


Figure 2 : Infographie des familles de langues

Problématiques et objectifs

Une métrique difficile à définir entre les langues

- Une métrique existante
- Définition et validation de notre propre métrique

Extraction, fusion, requêtes et visualisation

- Abondance de bases de données propres à chaque domaine de la linguistique
- Extraction et fusion des données
- Confrontation des domaines de la linguistique et de l'économie

**Mots-clefs : Python, R, Gephi, Neo4J, WALS, PHOIBLE,
Glottolog**

Déroulé

1. Recherche, définition et validation de la métrique
2. Confrontation des domaines de la linguistique sur Neo4J
3. Etude de la proximité linguistique et des échanges économiques

Déroulé

1. Recherche, définition et validation de la métrique
2. Confrontation des domaines de la linguistique sur Neo4J
3. Etude de la proximité linguistique et des échanges économiques

Lang2vec : une métrique existante

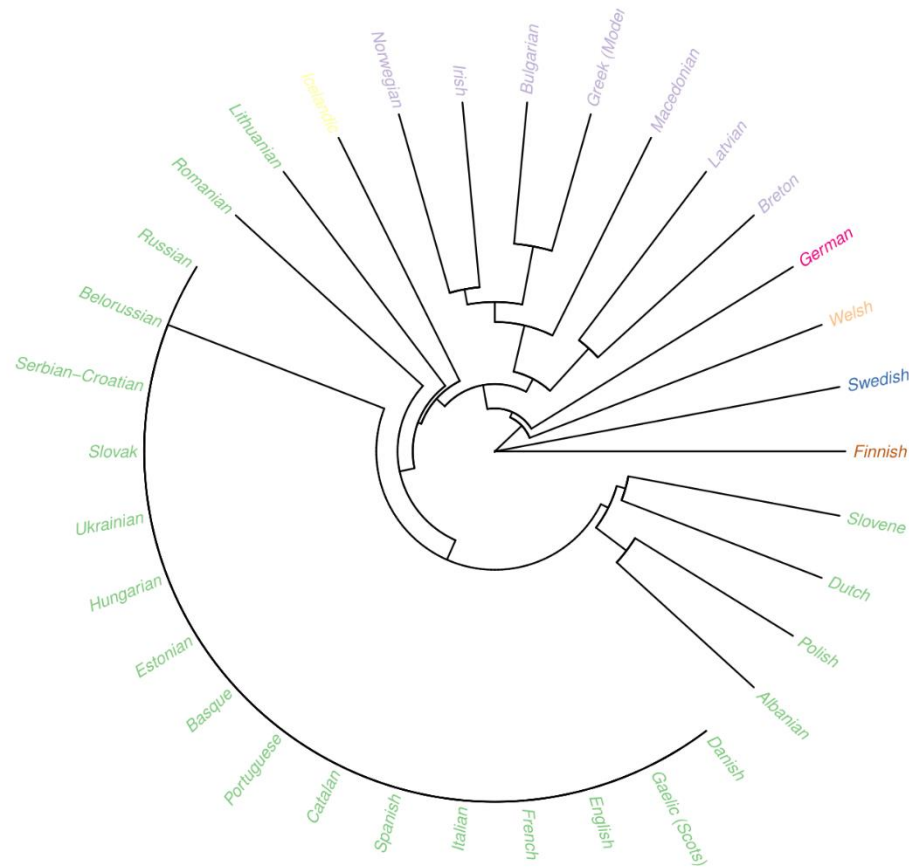


Figure 3 : Dendrogramme des langues d'Europe avec *lang2vec*

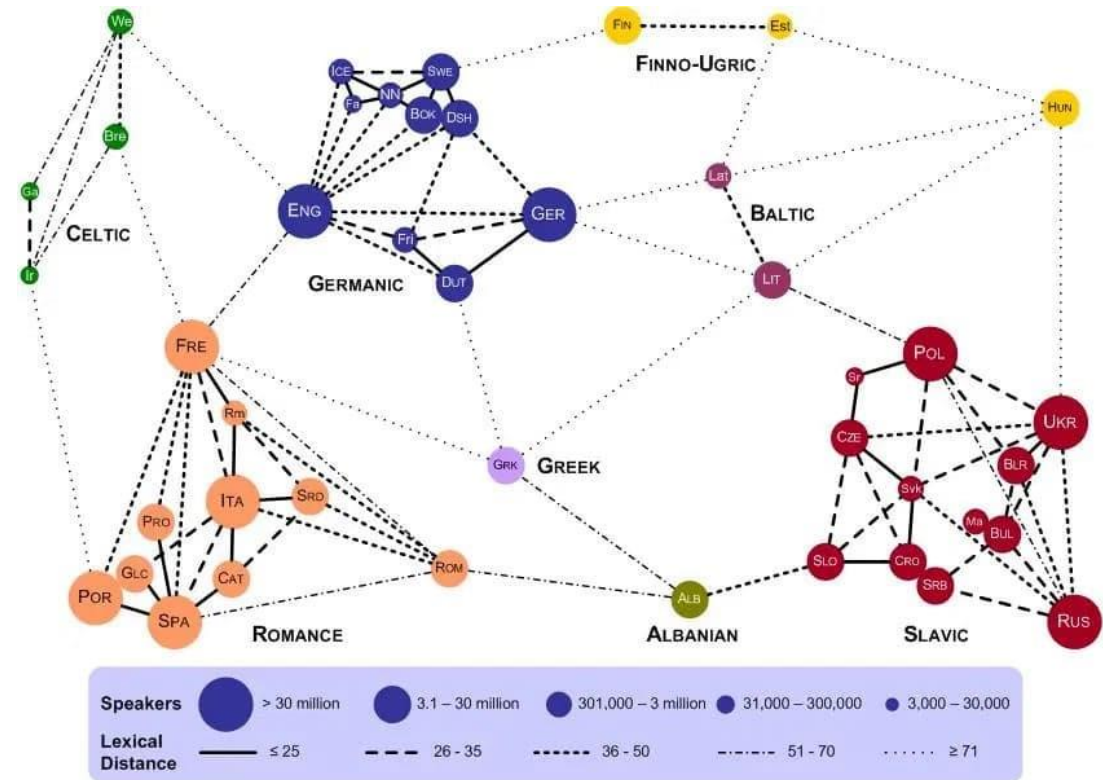


Figure 4 : Clusters attendus pour les langues d'Europe

WALS, PHOIBLE, Glottolog

PHOIBLE

Répositoire en ligne de phonèmes et de données linguistiques sur plus de 2500 langues

Glottolog

Catalogue des langues, dialects, languoides et familles de langues dans le monde

WALS

Base de données de propriétés structurelles, grammaticales et lexicales

Définition de notre métrique

- Métrique basée sur l'ensemble P propriétés structurelles de WALs
- Chaque langue L est vectorisée vers $V_L = (v_i)$

$$\forall i \in P, v_i = \begin{cases} 0 & \text{if } i \notin P \\ n \in \mathbb{N} & \text{if } i \in P \end{cases}$$

$$\forall (L_1, L_2) \in L | L_1 = (v_i)_{i \in P}, L_2 = (w_i)_{i \in P}, d(L_1, L_2) = \frac{\sum_{i \in P} \delta_{v_i, w_i}}{\sum_{i \in P} 1_{|v_i w_i| > 0}(i)}$$

Validation de la métrique

- Codé sur R
- 34 langues d'Europe
- $k = 7$ clusters
- Average clustering

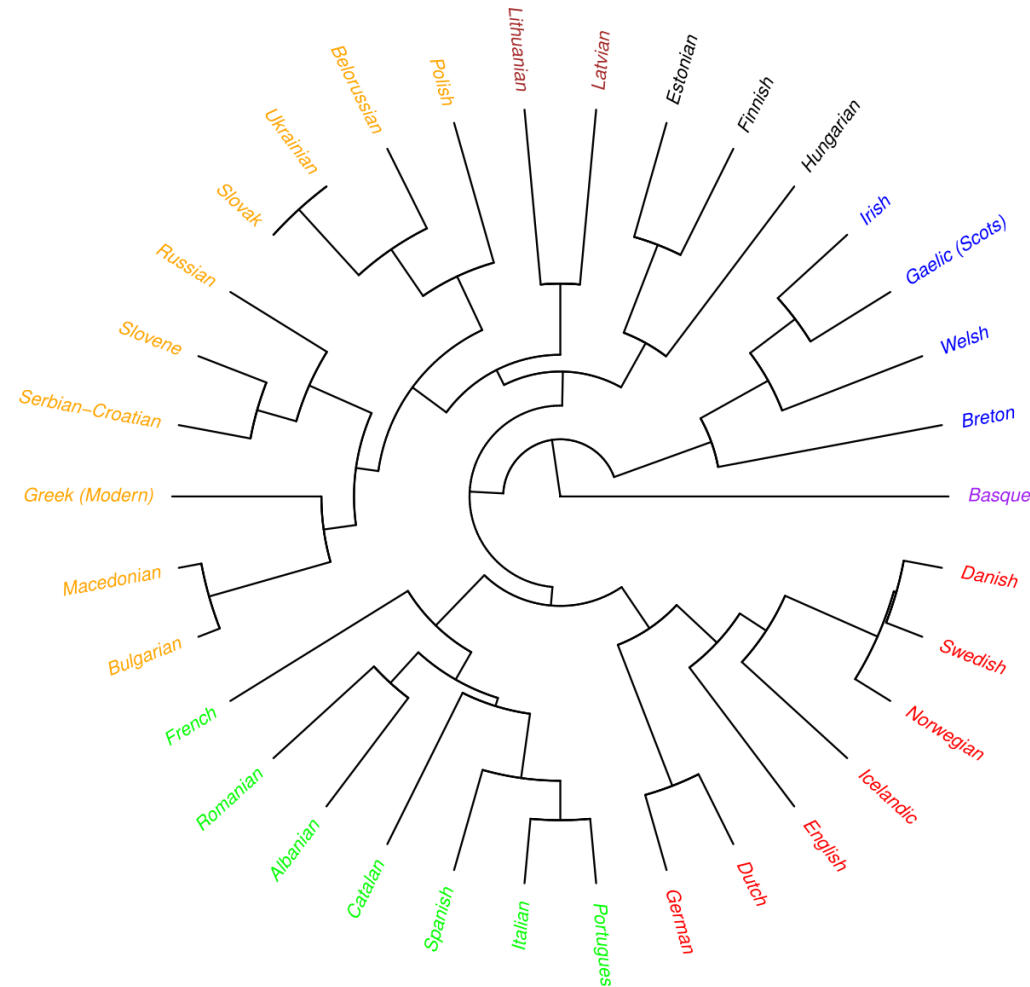


Figure 5 : Dendrogramme des langues d'Europe avec notre métrique

Validation de la métrique

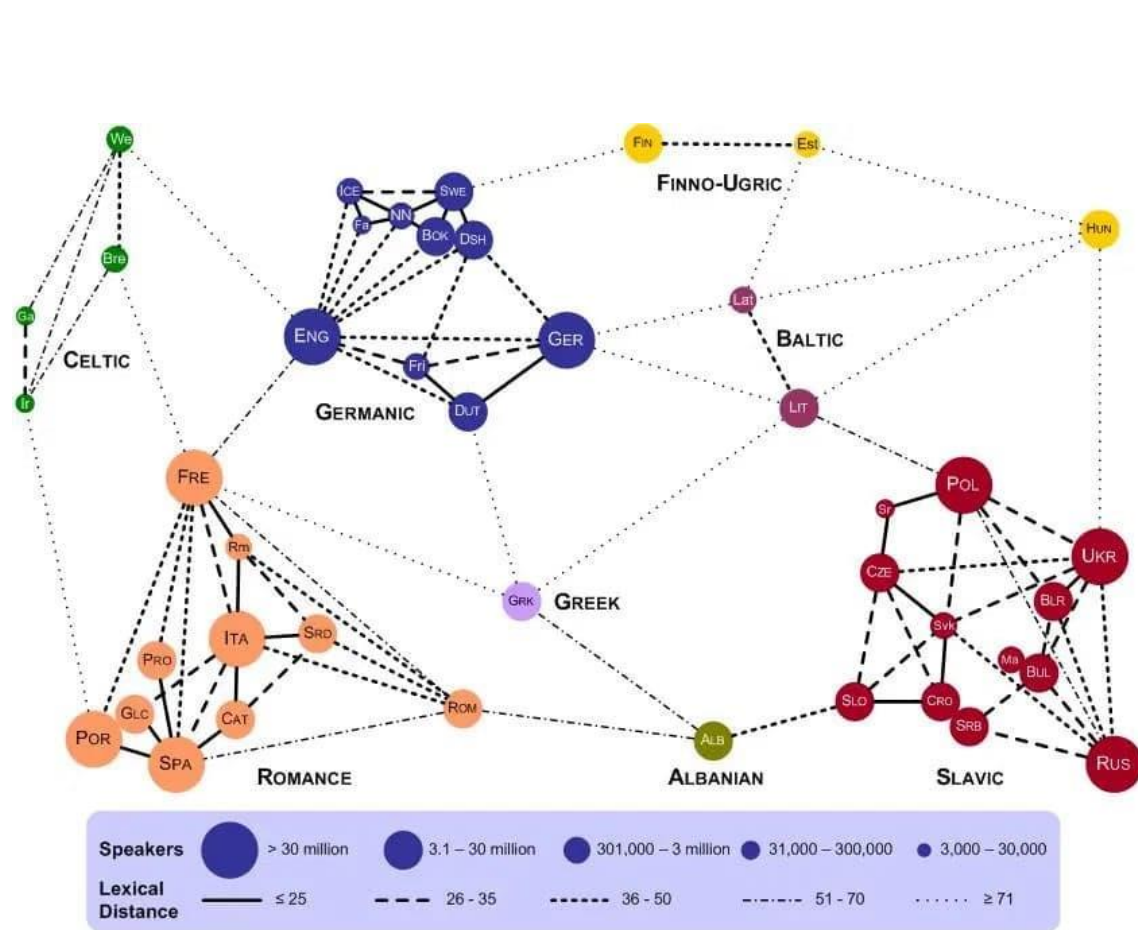


Figure 4 : Clusters attendus pour les langues d'Europe



Validé !

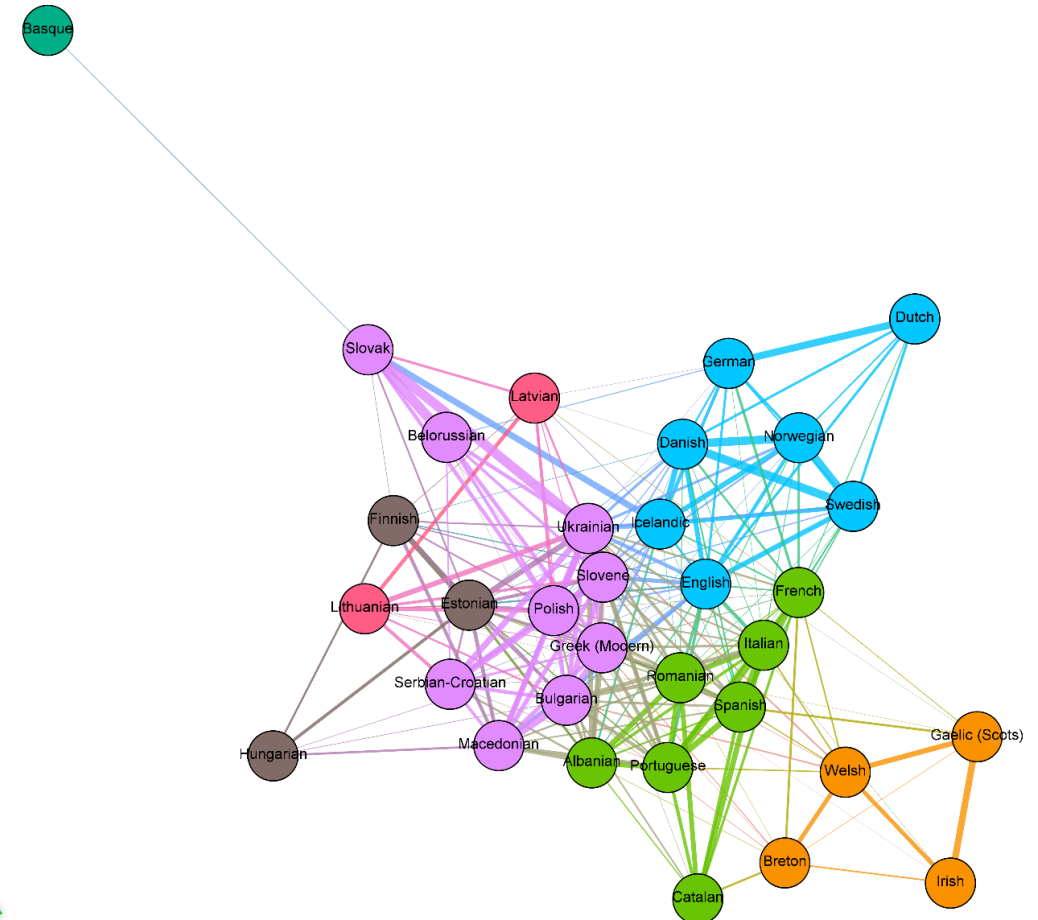
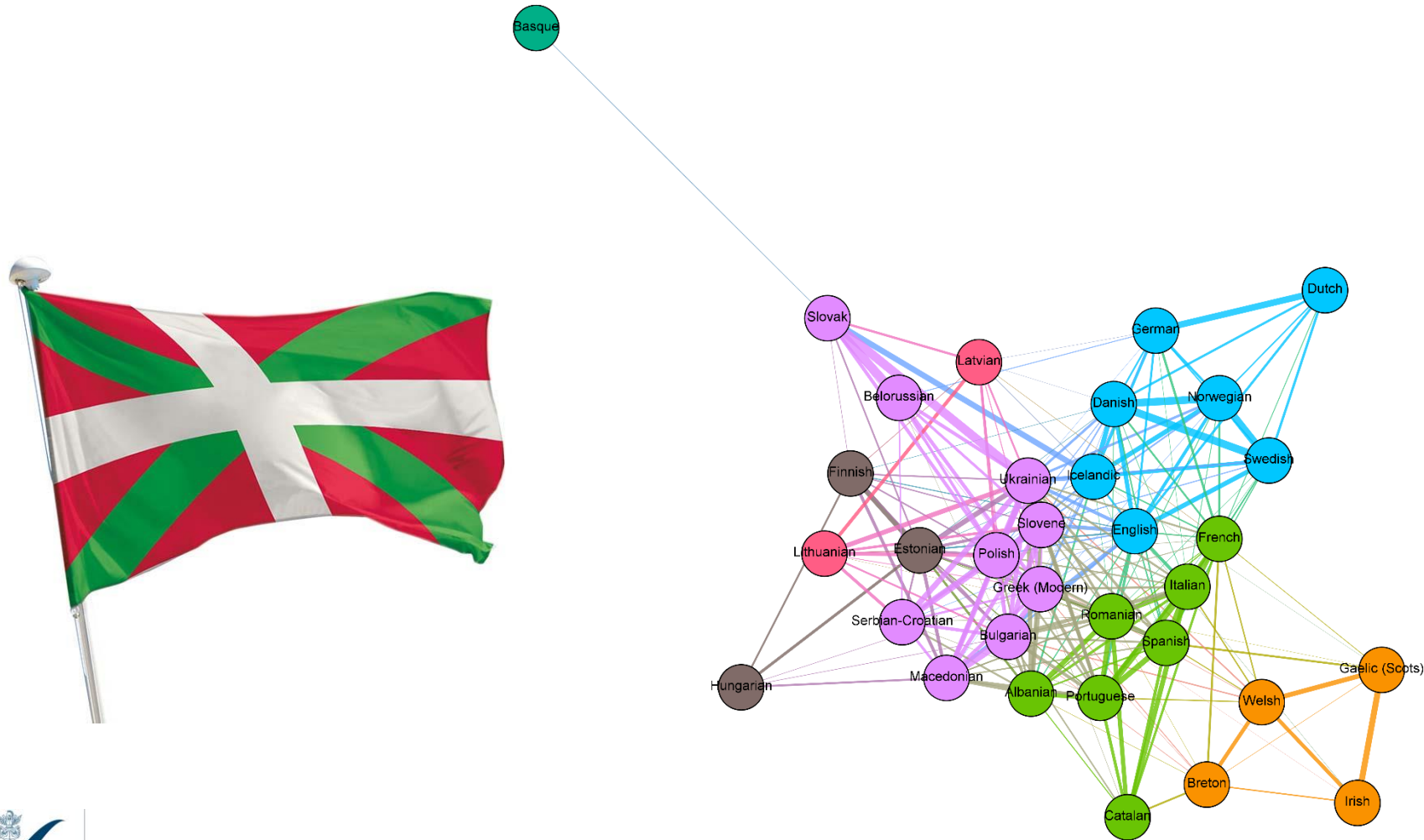


Figure 6 : Clusters des langues d'Europe sur Gephi

Validation de la métrique

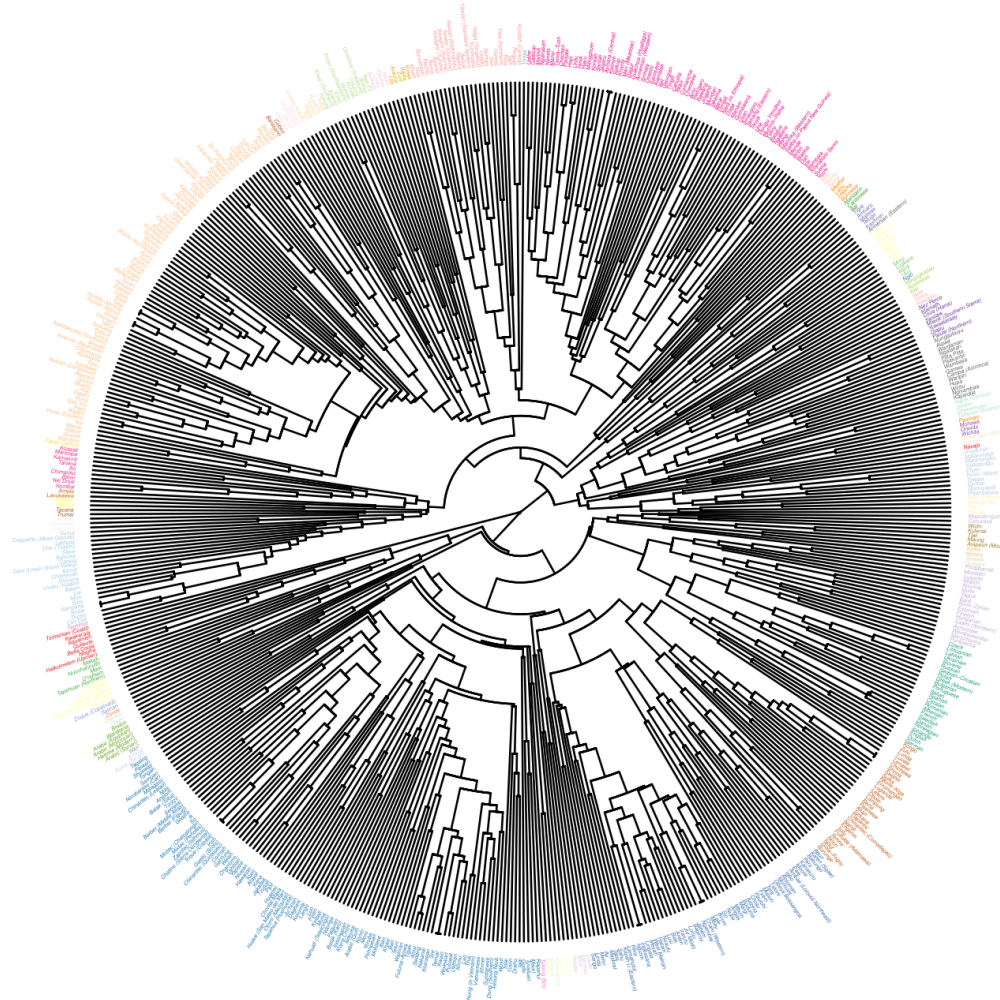


Validé !

Figure 6 : Clusters des langues d'Europe sur Gephi

Les clusters des langues avec le plus d'attributs WALs

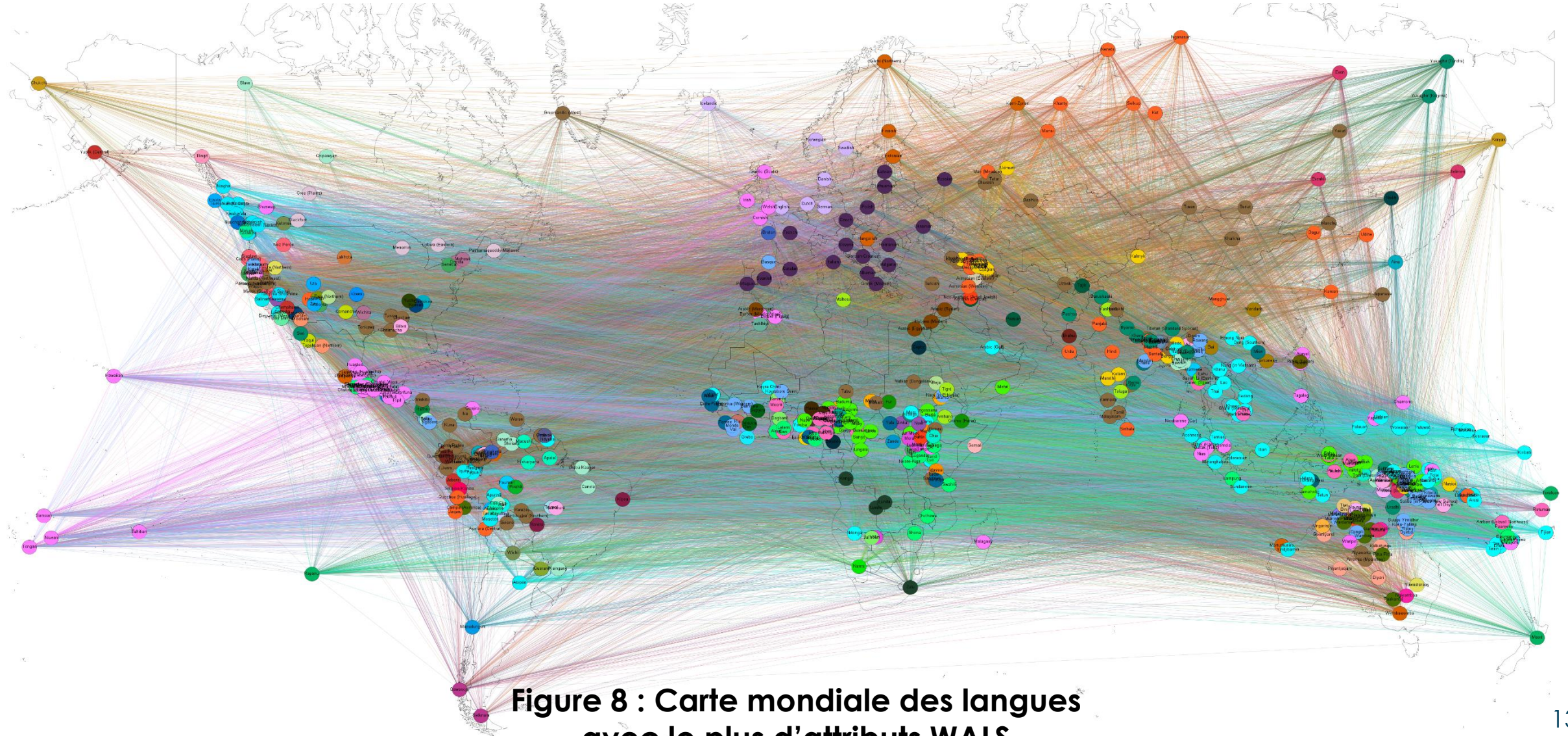
- Nombres de locuteurs récupérés sur Wikipedia
- Codé sur R
- 611 langues
- $k = 60$ clusters
- Average clustering



Gephi
➔

Figure 7 : Dendrogramme des langues du monde avec notre métrique

Gephi : les clusters des langues avec le plus d'attributs WALs



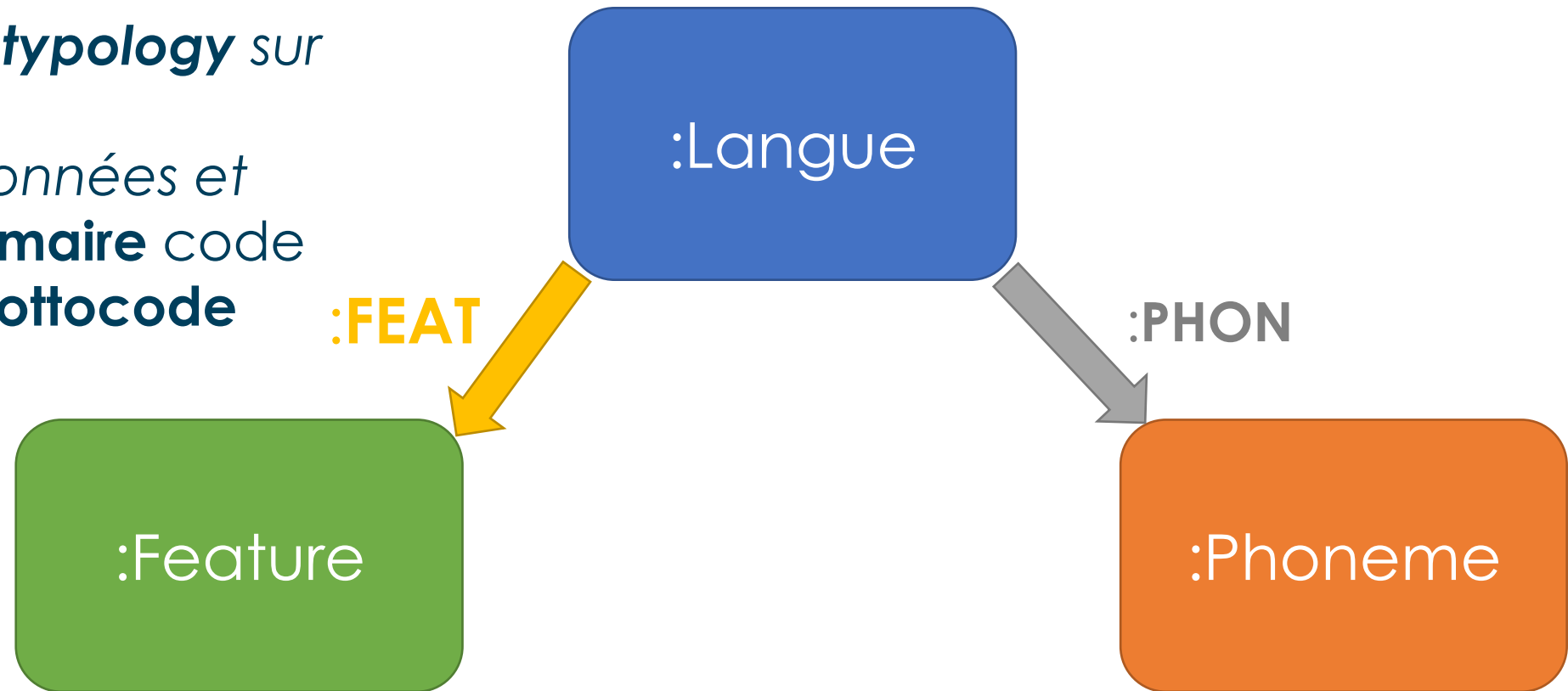
Déroulé

1. Recherche, définition et validation de la métrique
- 2. Confrontation des domaines de la linguistique sur Neo4J**
3. Etude de la proximité linguistique et des échanges économiques

Création du graphe Neo4J

Fusion de **WALS** et **PHOIBLE**

- Bibliothèque *lingtypology* sur *R*
- Extraction des données et fusion sur **clef primaire** code ISO 639-1 puis **Glottocode**



Création du graphe Neo4J

```
01 | LOAD CSV WITH HEADERS FROM 'file:///nodes_features.csv' AS
    line
02 | CREATE (F:Feature {id: line.id, name: line.name});
03 |
04 | LOAD CSV WITH HEADERS FROM 'file:///nodes_lang.csv' AS line
05 | CREATE (L:Language {name: line.Name, glottocode: line.
    glottocode, genus :line.Genus, family : line.Family,
    latitude :toFloat(line.Latitude), longitude :toFloat(line.
    Longitude}));
06 |
07 | LOAD CSV WITH HEADERS FROM 'file:///phonemes_nodes.csv' AS
    line
08 | CREATE (P:Phoneme {description: line.description, equivalent:
    line.equivalence_class, id :line.id, name : line.name,
    type : line.segment_class});
```

```
10 | LOAD CSV WITH HEADERS FROM "file:///edges_lang_features.csv"
    AS row
11 | MATCH (L:Language), (F:Feature)
12 | WHERE L.glottocode = row.Source AND F.id = row.Target
13 | CREATE (L)-[edgewals:FEAT]->(F)
14 | SET edgewals=row, edgewals.Value = toInteger(row.Value);
15 |
16 | LOAD CSV WITH HEADERS FROM "file:///edges_lang_phonemes.csv"
    AS row
17 | MATCH (L:Language), (P:Phoneme)
18 | WHERE L.glottocode = row.Glottocode AND P.name = row.Phoneme
19 | CREATE (L)-[edgephoible:PHON]->(P)
20 | SET edgephoible=row;
```

Relation entre les langues et les phonèmes

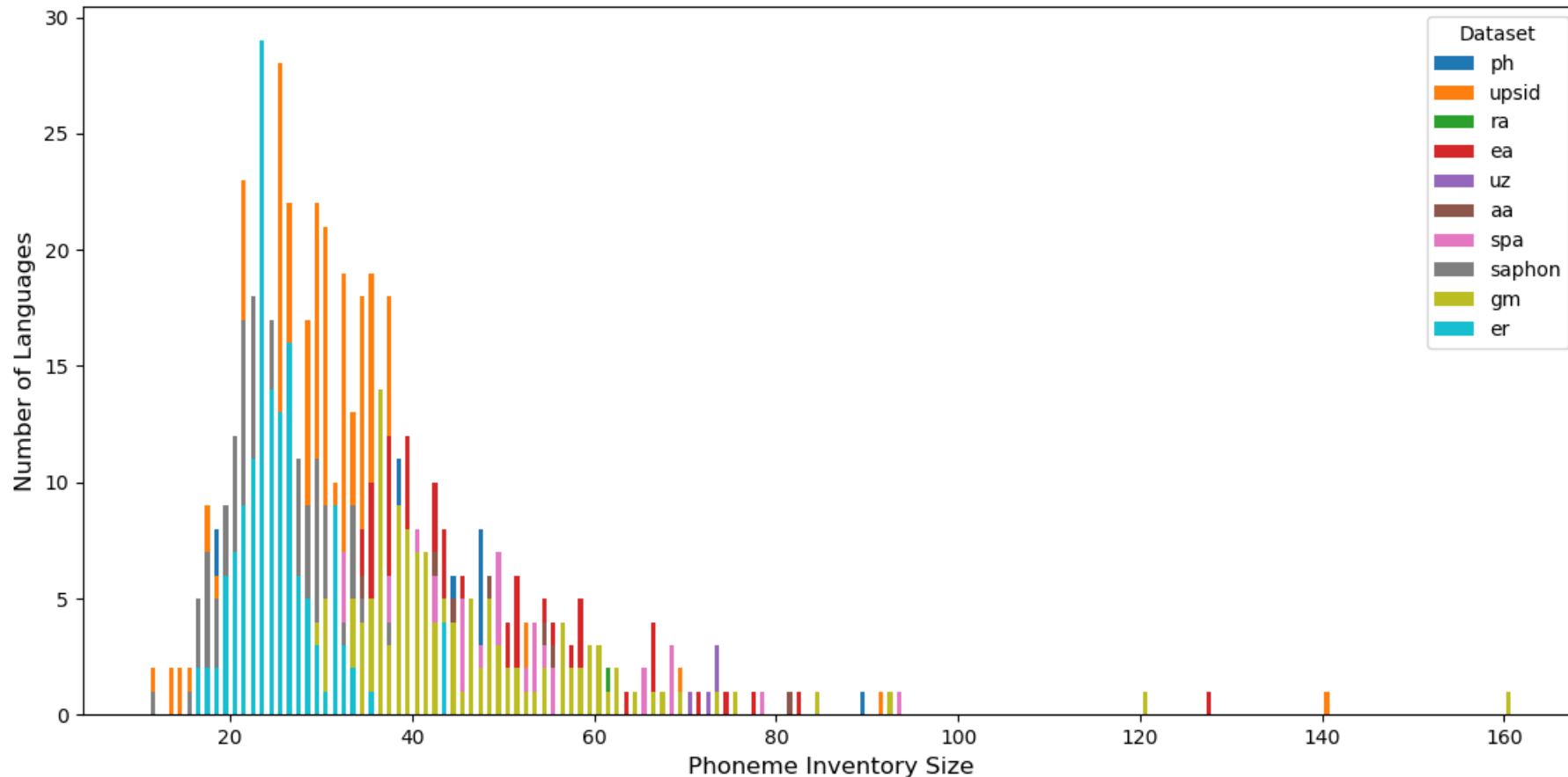


Figure 9 :
Relation
entre les
langues et le
nombre de
phonèmes

```
01 | MATCH (L)-[r:PHON]->()  
02 | WITH r.Dataset AS ds, r.Dialect AS dial, L, COUNT(r) AS c  
03 | RETURN L.glottocode, L.name AS name, ds, AVG(c)
```

Les voyelles et les consonnes dans les langues

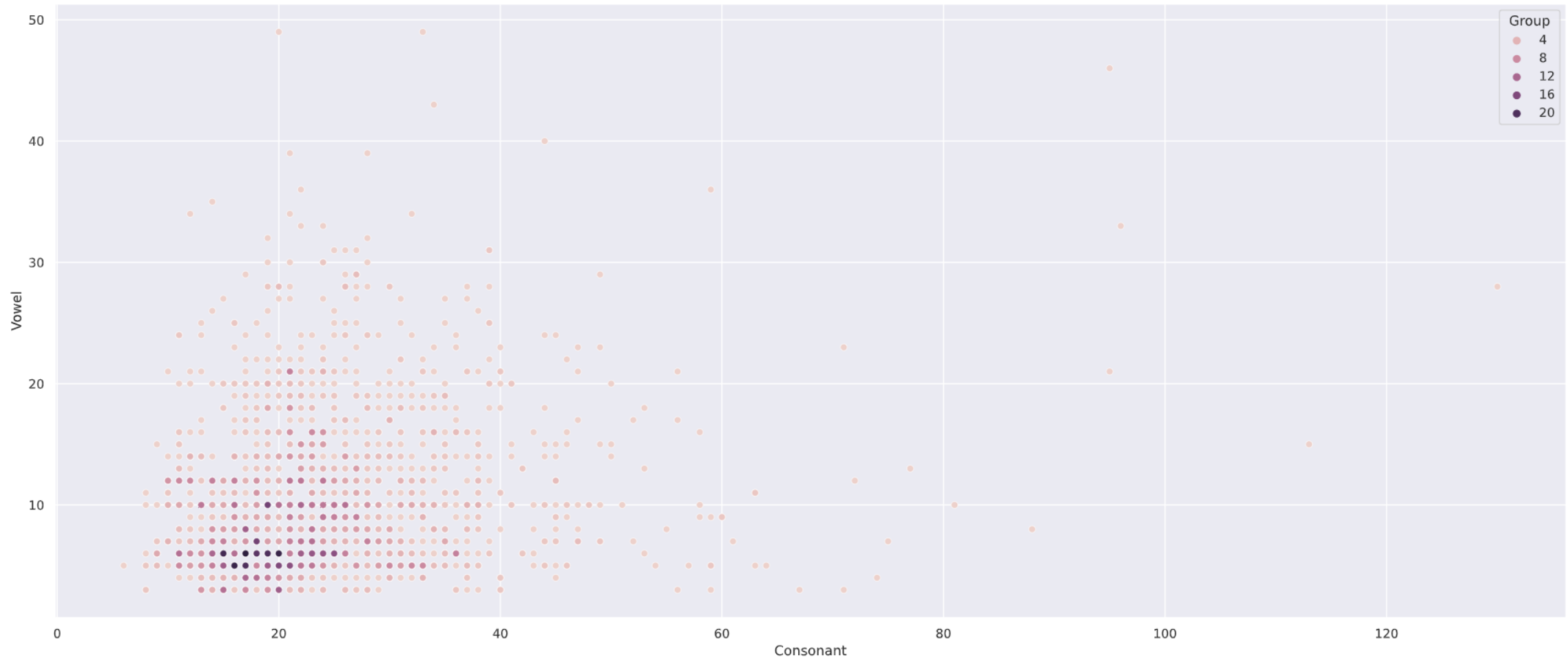


Figure 10 : Relation entre les nombres de voyelles et consonnes

Les voyelles et les consonnes dans les langues

- Les points de couleur plus foncée représentent les langues avec les mêmes nombres de consonnes et de voyelles
- Python : pour chaque augmentation d'environ 13 ou 14 consonnes, il y a augmentation d'une voyelle.

```
01 | MATCH (L)-[r:PHON]->(P)
02 | WHERE P.type = "consonant"
03 | WITH r.Dataset AS ds, r.Dialect AS dial, L, COUNT(P) AS
      consonant
04 | MATCH (L)-[r:PHON]->(P)
05 | WHERE P.type = "vowel" AND r.Dataset = ds AND r.Dialect =
      dial
06 | RETURN L.glottocode, L.name AS name, ds AS dataset, dial AS
      dialect, consonant, COUNT(P) as vowel
```

Les langues bien documentées et segmentées

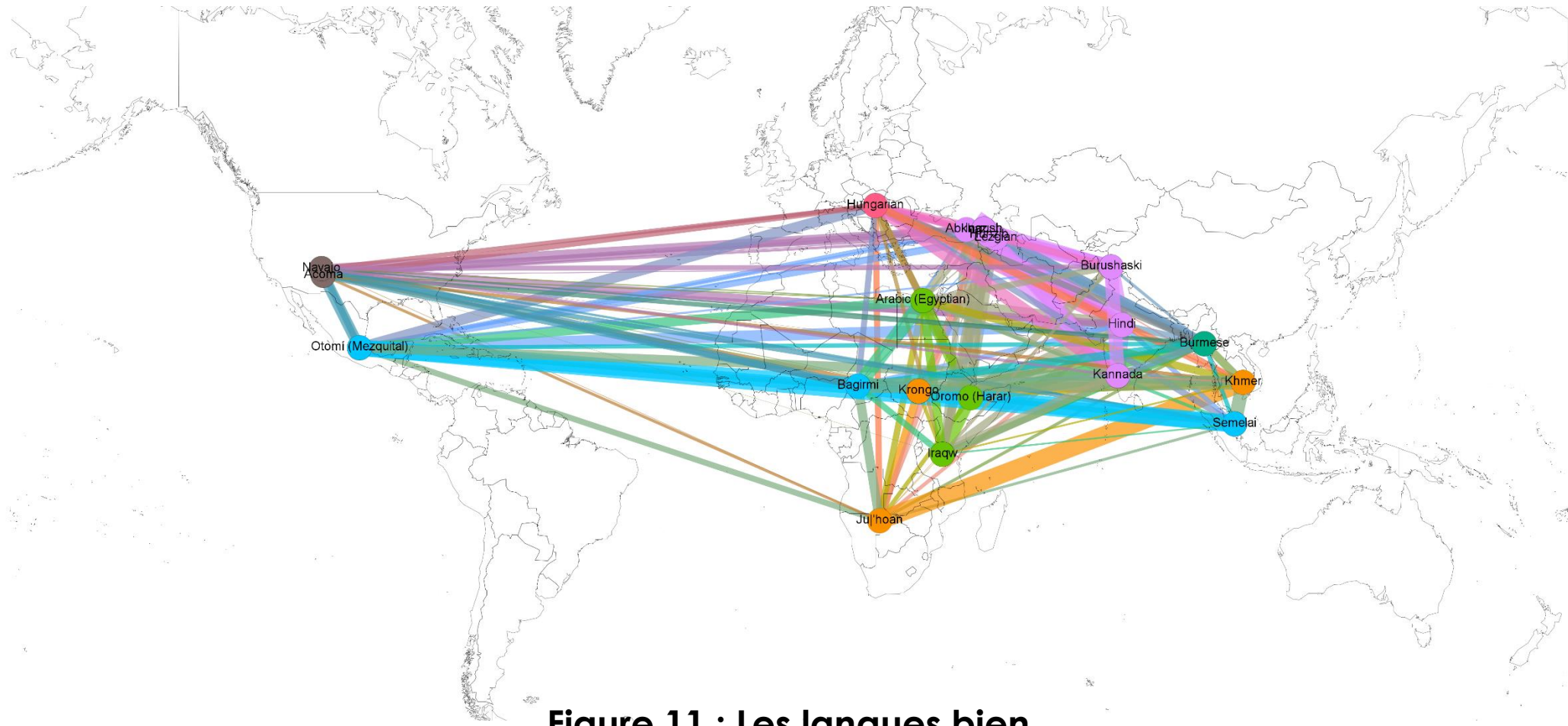


Figure 11 : Les langues bien documentées et segmentées

Déroulé

1. Recherche, définition et validation de la métrique
2. Confrontation des domaines de la linguistique sur Neo4J
- 3. Etude de la proximité linguistique et des échanges économiques**

Extraction des données

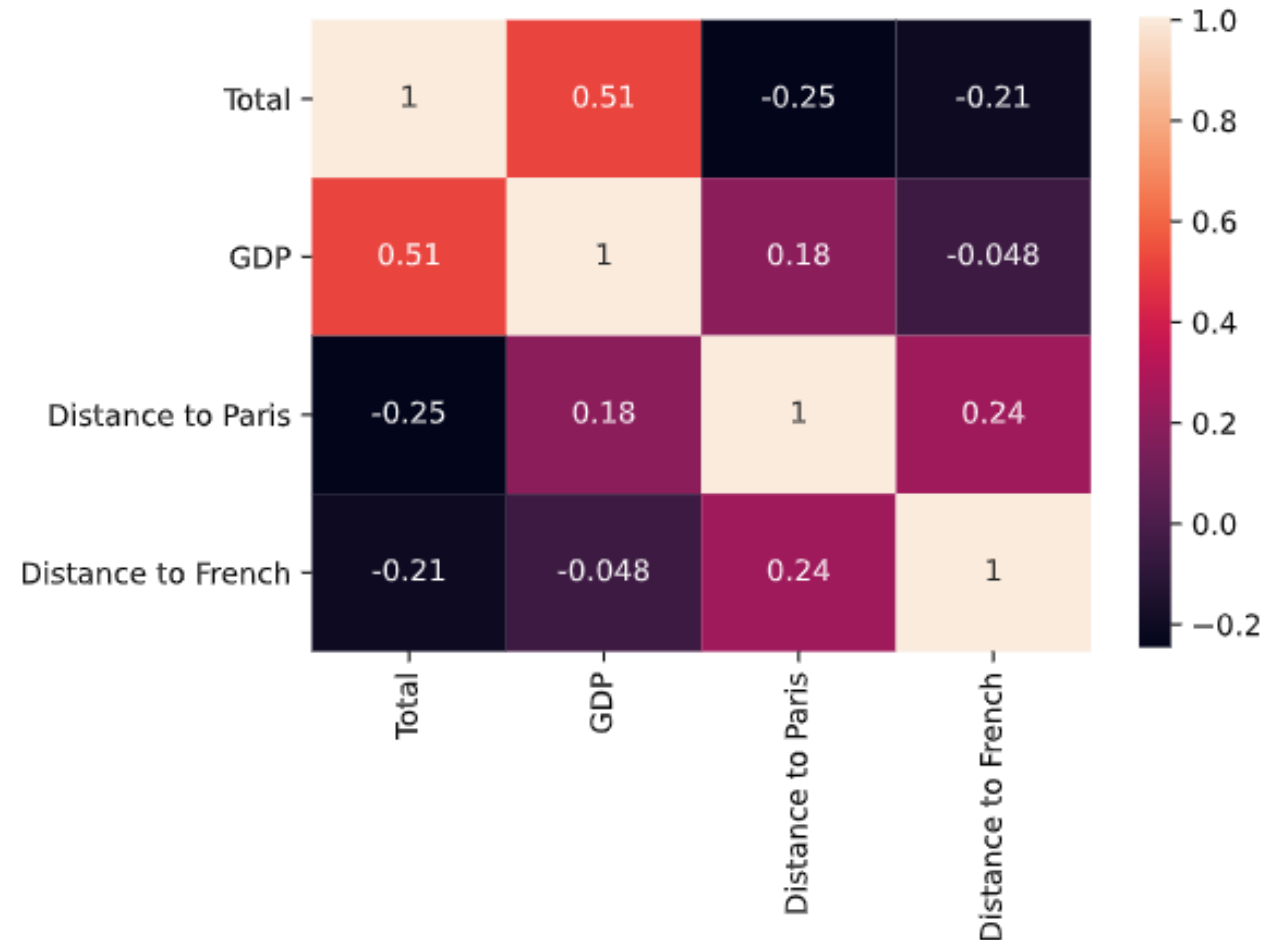
Données des **50 premiers partenaires** commerciaux de la France.

- **Le Produit Intérieur Brut (PIB) de chaque pays.**
- **La liste des langues officielles nationales** de chaque pays.
- **Les imports et exports vers la France.**
- **La liste des distances géographique.**
- **La liste des distances linguistiques.**

Corrélogramme

- Corrélation négative entre le Total et la distance linguistique.
- Coefficient de corrélation similaire à celui de la distance géographique

Figure 12 : Corrélogramme entre GDP, distance linguistique et géographique



Cartographie : échanges économiques et proximité linguistique

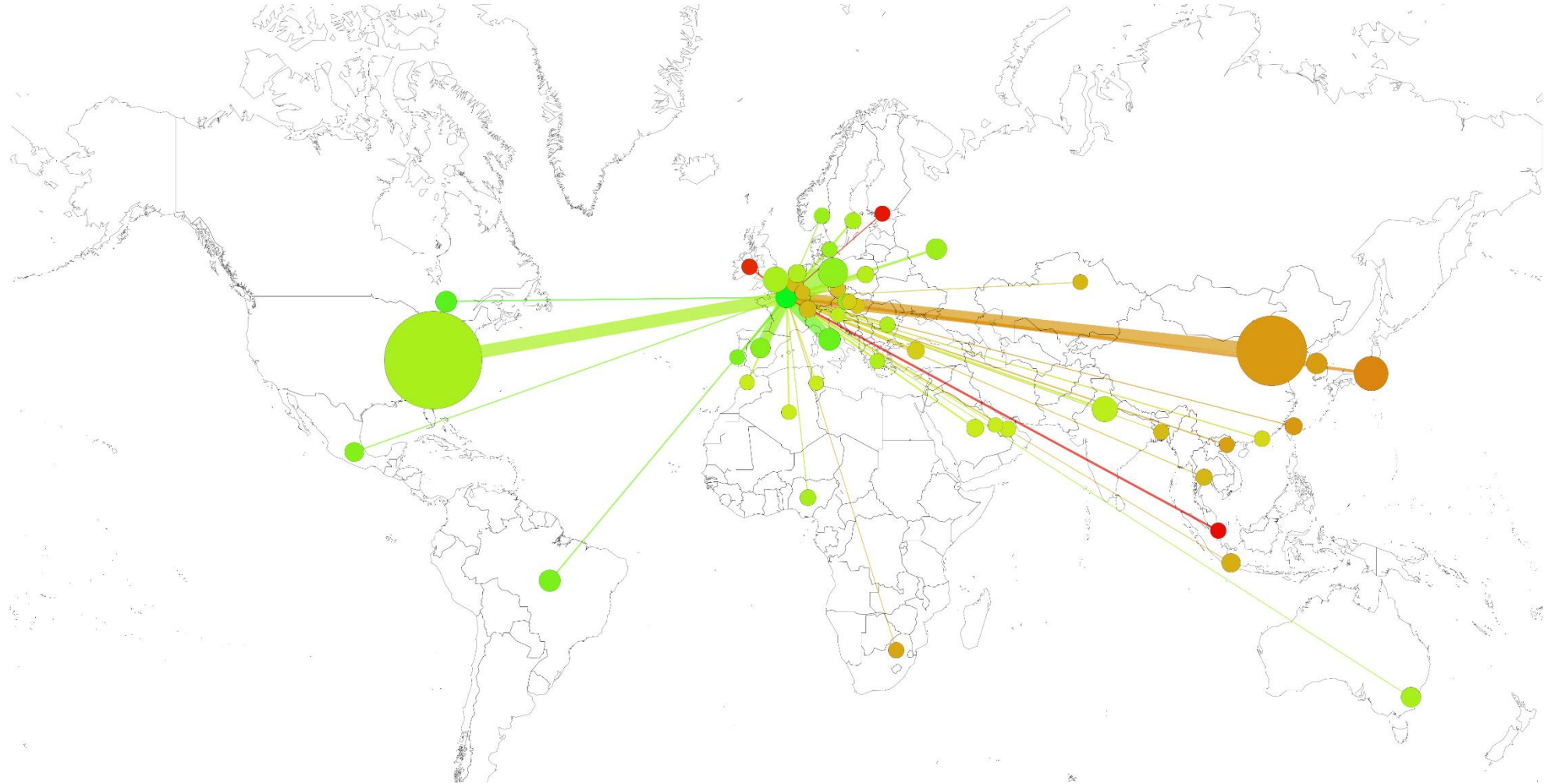


Figure 13 : Cartographie : échanges économiques et proximité linguistique

Conclusion et ouverture

- Etendre l'analyse faite à la France sur plusieurs pays.
- Fusionner d'autres bases de données : **Ethnologue, WOLD...**
- D'autres requêtes sur le graphe Neo4J :
 - Nombre de phonèmes en fonction du nombre de locuteurs.
 - Etude de la corrélation entre proximité géographique et partage de phonèmes.
- Etude de la proximité des phonèmes selon leurs attributs.