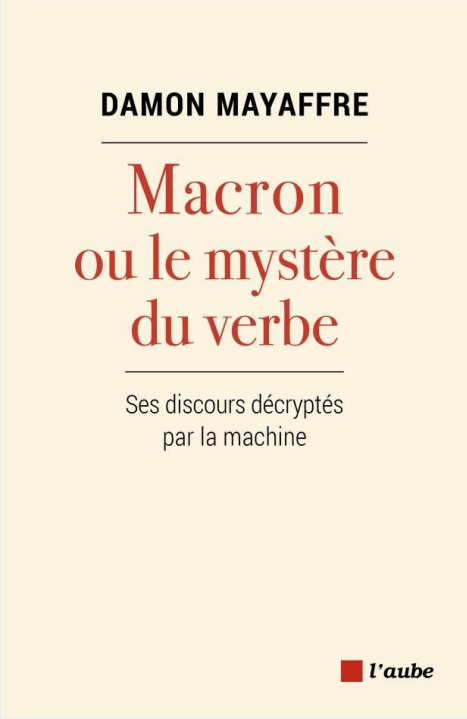
Grenoble, 11 mars 2022

Des unités pertinentes des textes
Prolégomènes à la sémantique de corpus



Damon Mayaffre

UMR 7320, Bases, Corpus, Langage

CNRS – Université Côte d'Azur

http://bcl.cnrs.fr/

✓ Le postulat

Le texte est un artefact et l'Intelligence artefactuelle pourrait être indiquée pour en témoigner…

- ✓ Le texte n'est pas un objet naturel (Adam, Rastier…) et la question qui nous poursuit est « qu'est-ce qui fait texte » ?
- ✓ L'IA donne une ***représentation*** du texte, très artificielle, qui semble pouvoir mettre au jour des « artefactures textuelles » (Bachimont).

Définir/découvrir les « unités » constituantes/constitutives du texte ?

(« énigme insoluble » Legallois 2006)

Le syntacticien a la phrase… le phonologue a le phonème… le morphologue a le morphème… mais que serait un « textème » ?

Sur quoi s'appuie l'IA pour traiter du texte ?

« grandeurs textuelles », « marqueurs », « unités textuelles », « patterns », « n-grams », « expressions régulières », « segments textuels », « unités phraséologiques », « lexies simples », « séquences », « paragraphes », « cooccurrences », « passages », « isotopies », « routines », « formules » « collocations », « texèmes », « herménèmes », « zones de localité », etc.

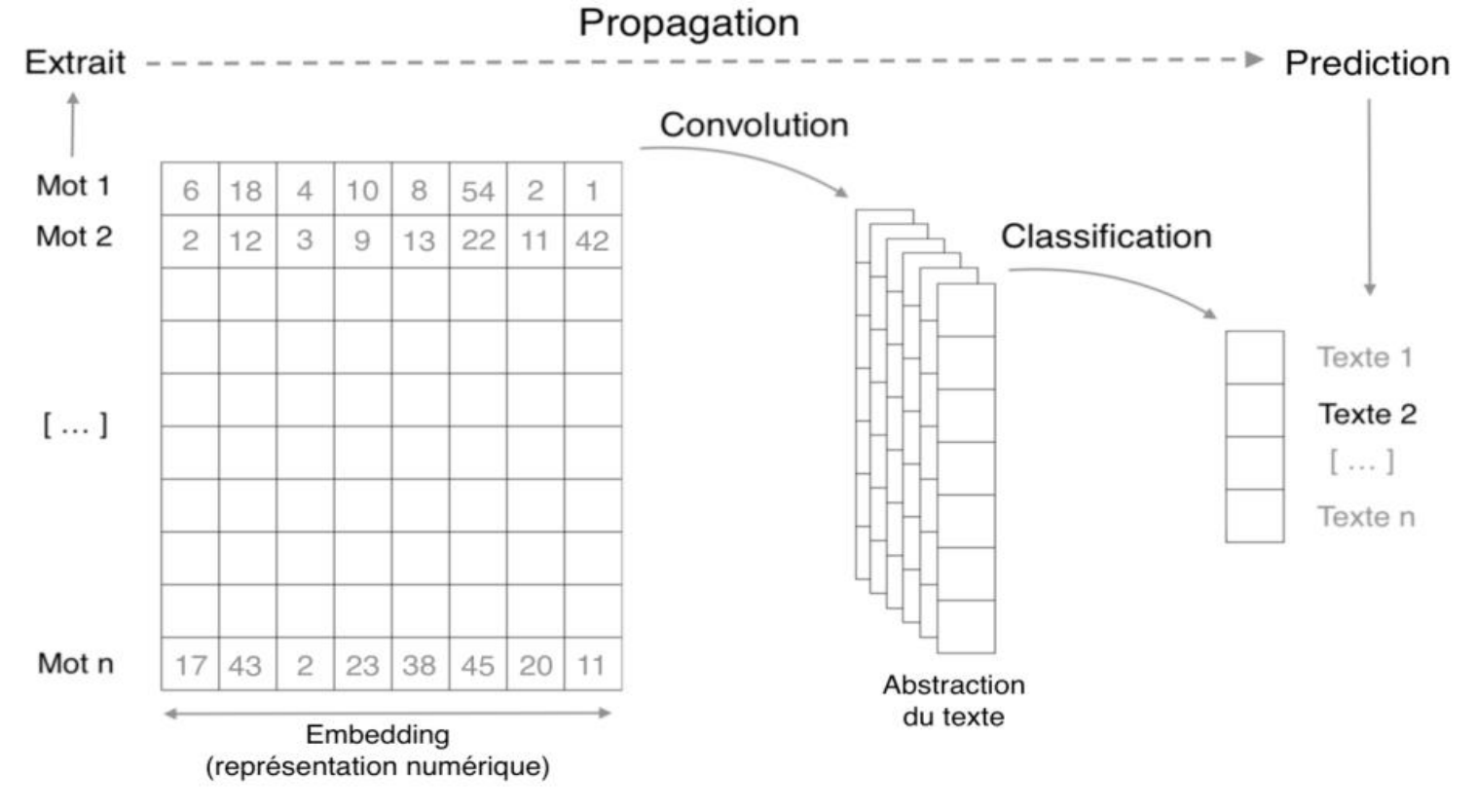
✓ La posture

Non pas néo-positiviste mais herméneutique…

- Le texte/le sens n’est pas un *donné* mais un *interprété* (Rastier)
- Le numérique est lui-même une représentation/interprétation du monde et des données (Doueihi)
- Une sortie-machine n’est pas une preuve mais un interprétable ; n’est pas la fin de l’analyse mais son commencement.

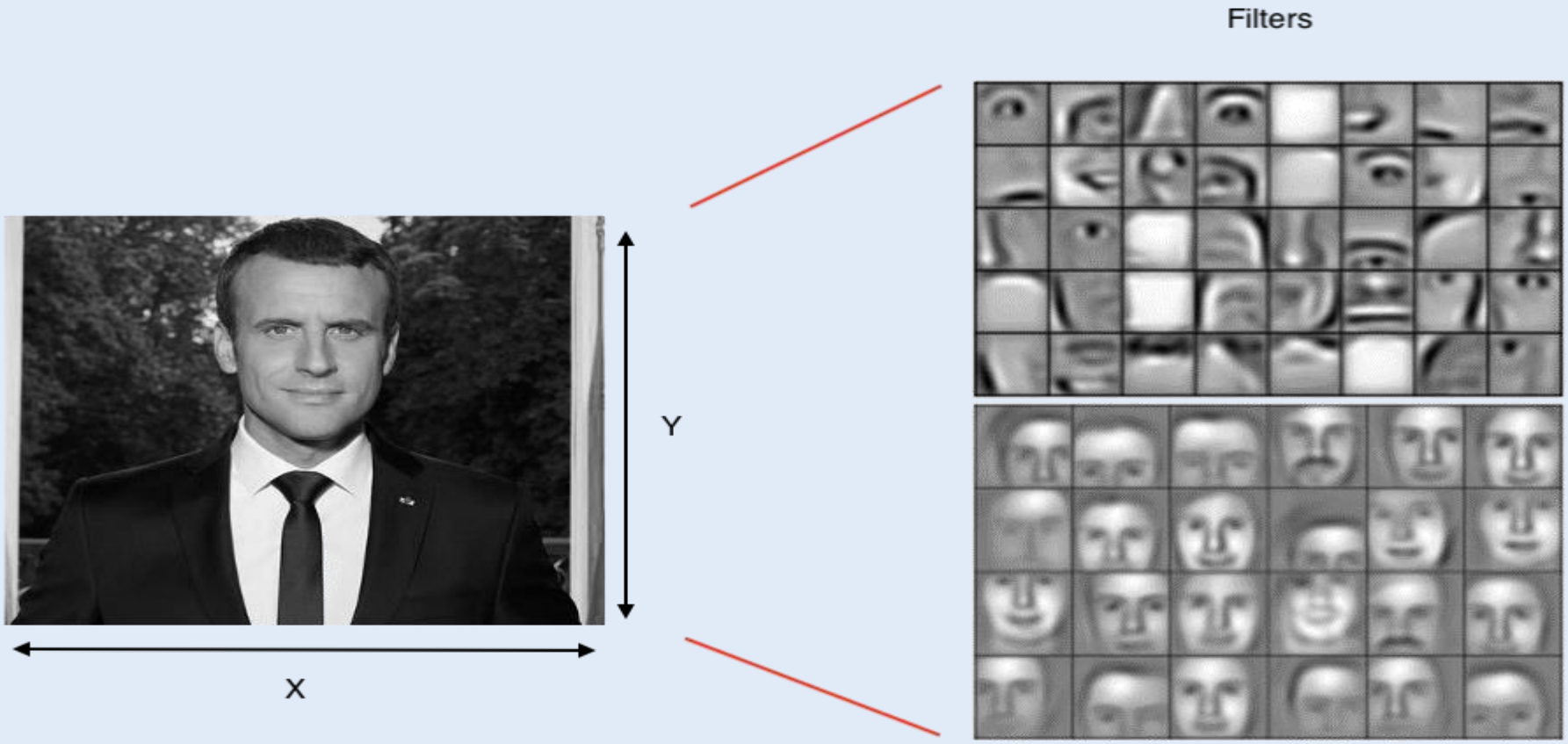
✓ Comment on s'y prend ?

Principes généraux du *deep texte*

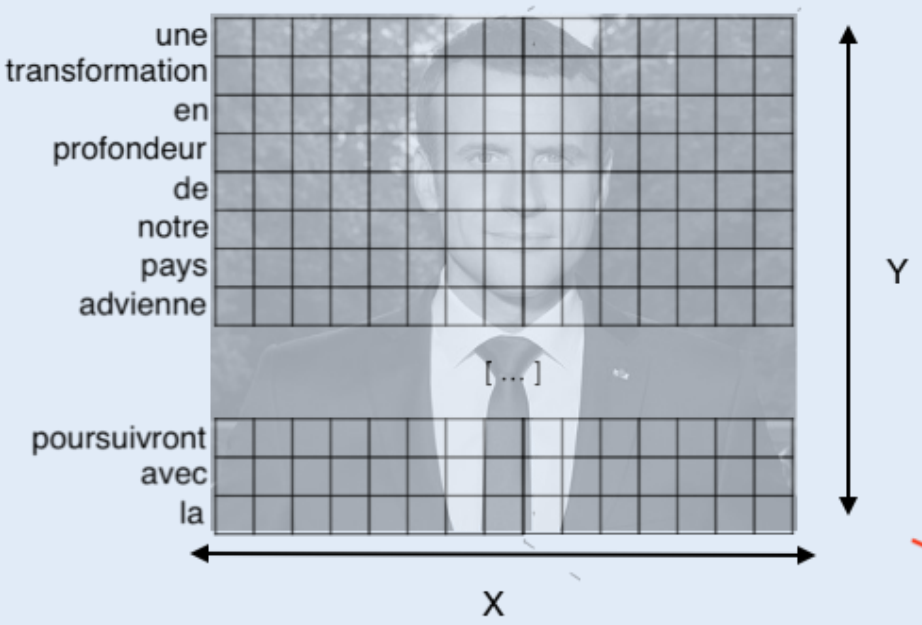


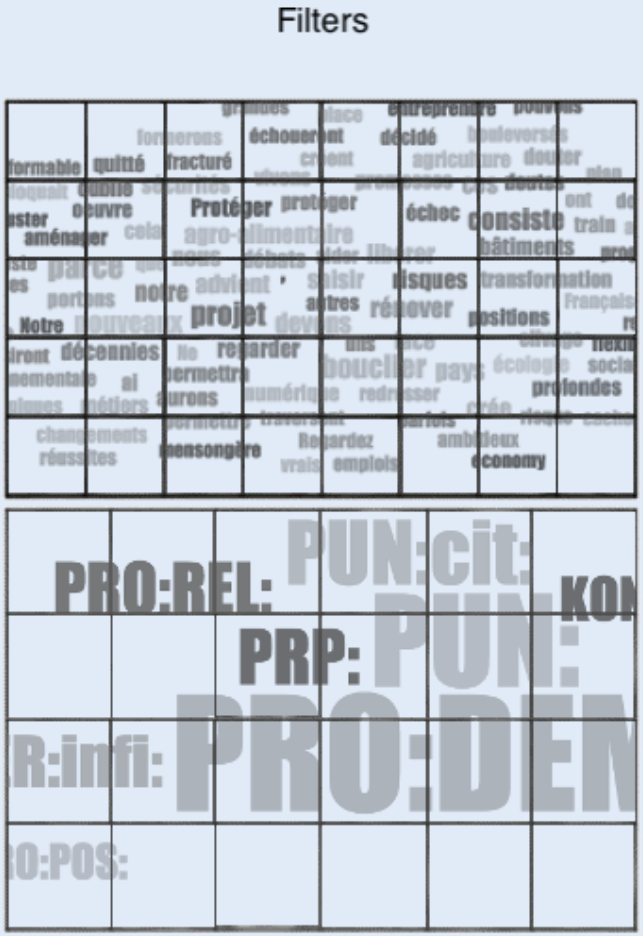
Le verrou : Prédiction (classification) => description
La richesse : le séquentiel et le fréquentiel

Deep learning : reconnaissance d’images sur la base des combinaisons de pixels et de filtres successifs



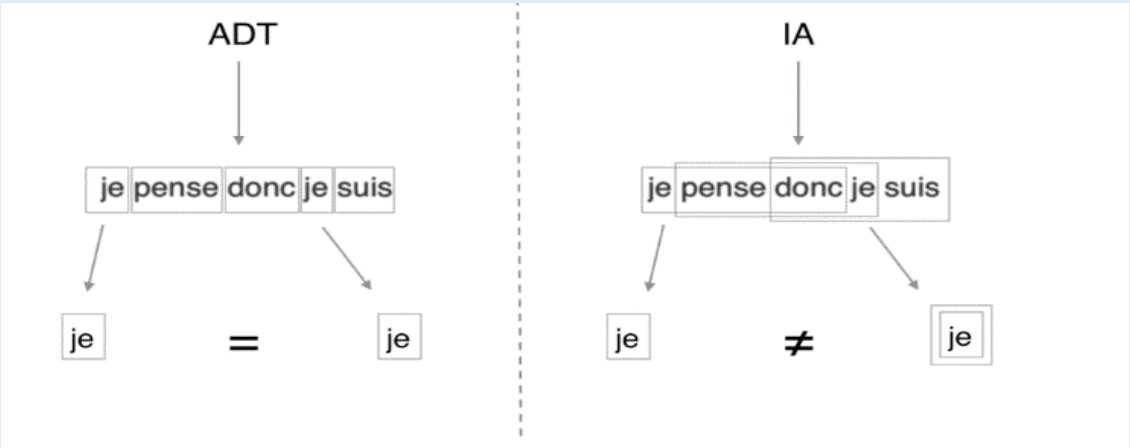
Deep learning : reconnaissance des textes sur la base de combinaisons de mots/lemmes/catégories morphosyntaxiques et des filtres successifs





- Micro : co(n)textualisation dans une fenêtre de 3, 6… mots.

= le mot à une représentation unique selon son contexte d'utilisation



Tokenisation Convolution

- Séquentiel complémentaire de fréquentiel

- Méso : co(n)textualisation dans le paragraphe

Attention : limite technique du programme et programme de recherche, pour des raisons de comparaison le paragraphe de (Adam) est ramené à une fenêtre fixe de X mots.

- Macro : co(n)textualisation dans le corpus. Le corpus réflexif ou ici le corpus d'apprentissage, puis de travail étant « la forme maximale du contexte » (Rastier, Mayaffre).

Résultats : le texte macronien ou le repérage de « motifs » (Mellet) qui « font texte »

<<… **permettre** aux **acteur européen** d'émerger dans un marché loyal et qui VER:FUTUR aussi de compenser les *profondes désorganisations* sur l'économie traditionnelle que DET:DEM *transformation* parfois **créer**. Les *grandes plateformes numériques*, la protection des *données* sont au cœur de **notre** *souveraineté* à cet égard. Et…>>

Un passage-clef de Macron – les items typographiquement marqués sont les zones du texte activées par le réseau qui ont permis à la machine de reconnaitre la prose du président (en vert les codes, en rouge les lemmes, en bleu les formes)

« Passage » (Rastier) : une notion à travailler

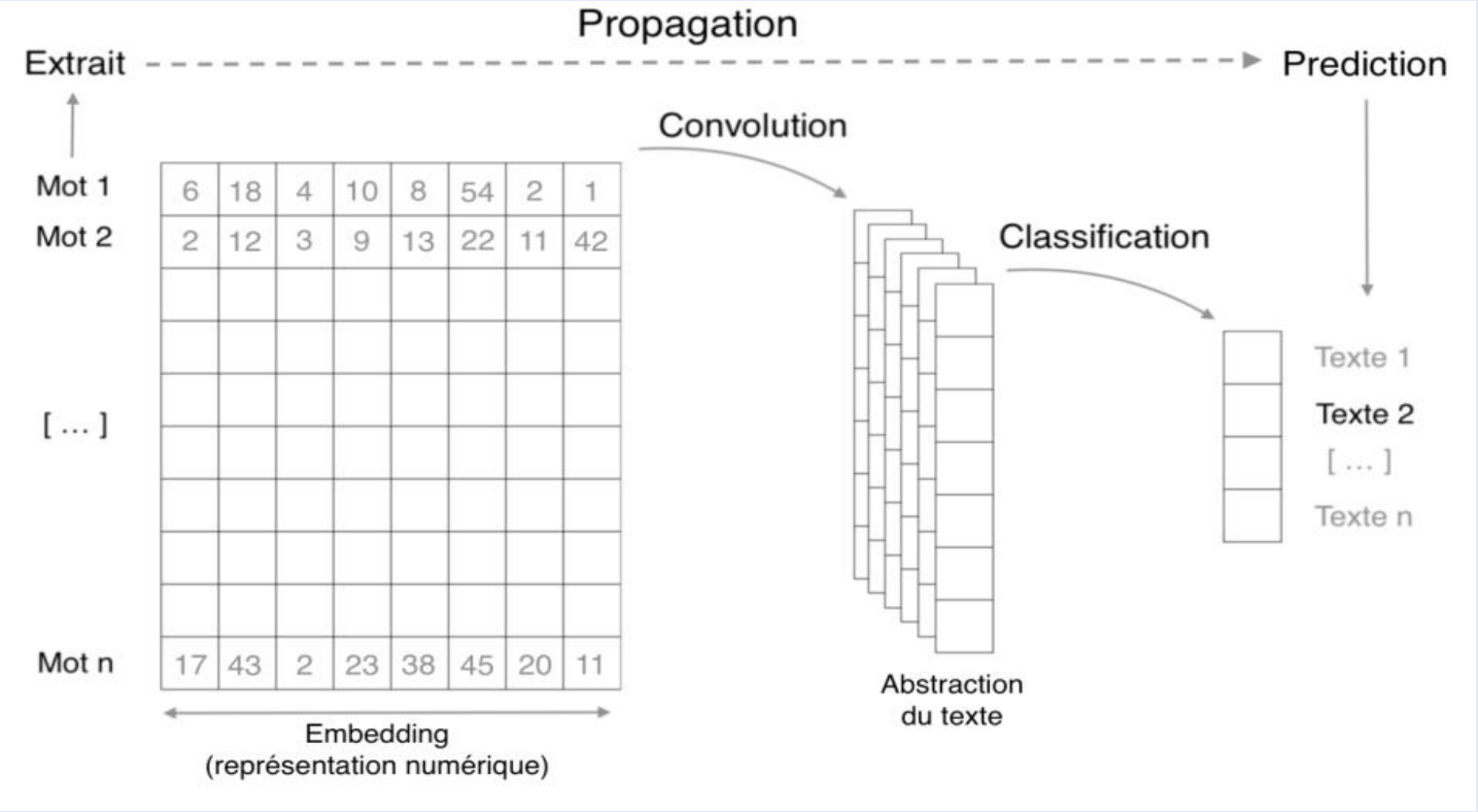
Résultats : l'intertexte macronien ou le repérable de « motifs » (Mellet) qui sont empruntés/empreintés

Le texte profond c'est l'intertexte (Kristeva, Barthes et al…). L'IA peut être un révélateur (chimique) du « texte palimpseste » (Genette, Heidmann, Adam… ; Mayaffre et al. 2020)

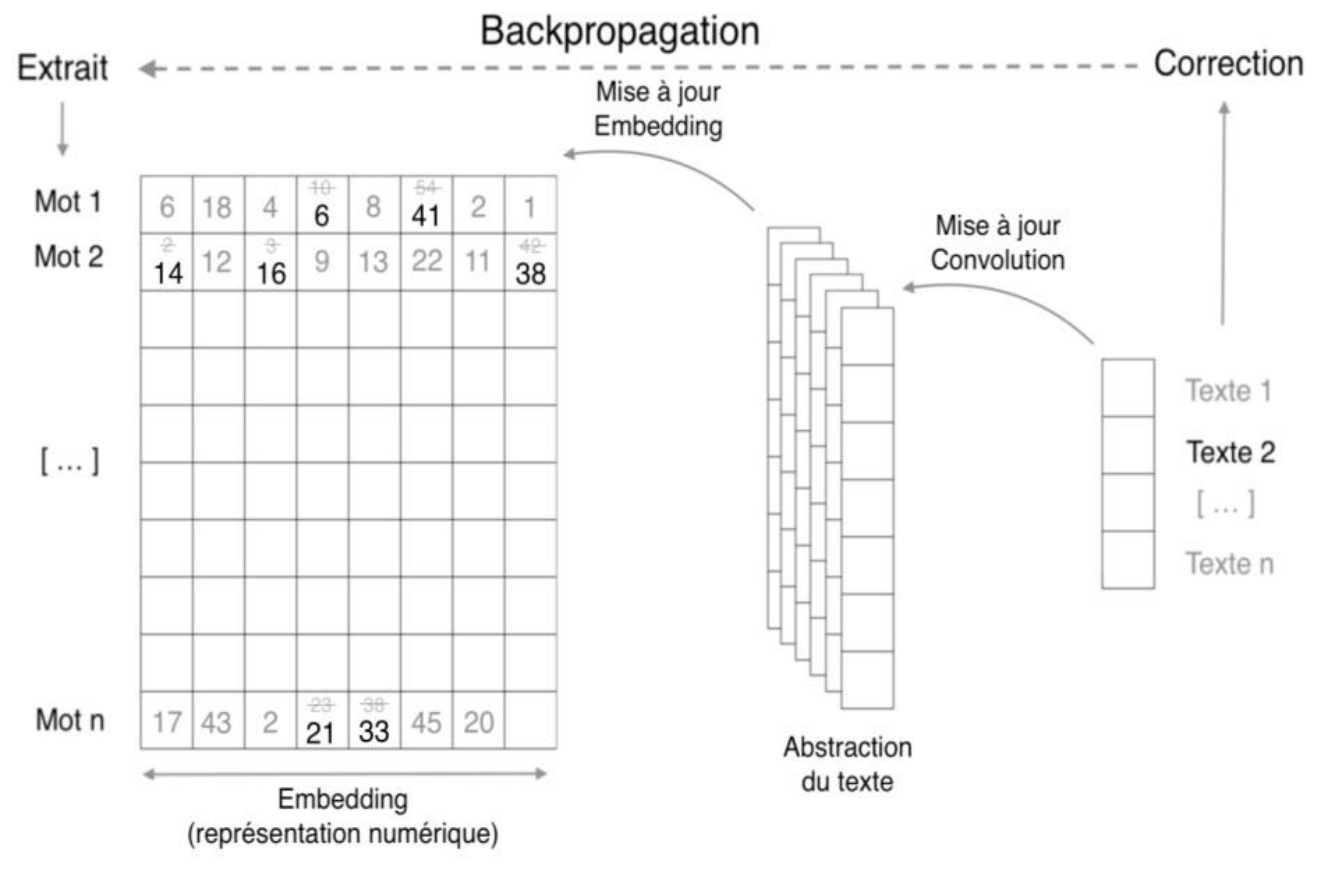
- <<… d'une NOM *européenne*. Restons **ce** *peuple uni* VIRGULE ADJ VIRGULE fier de son *histoire*, de ses NOM VIRGULE, de sa **culture**, confiant *dans* l'*avenir* et le **progrès**, sûr de son *talent* et de son énergie et *ambitieux* pour lui-même…>>

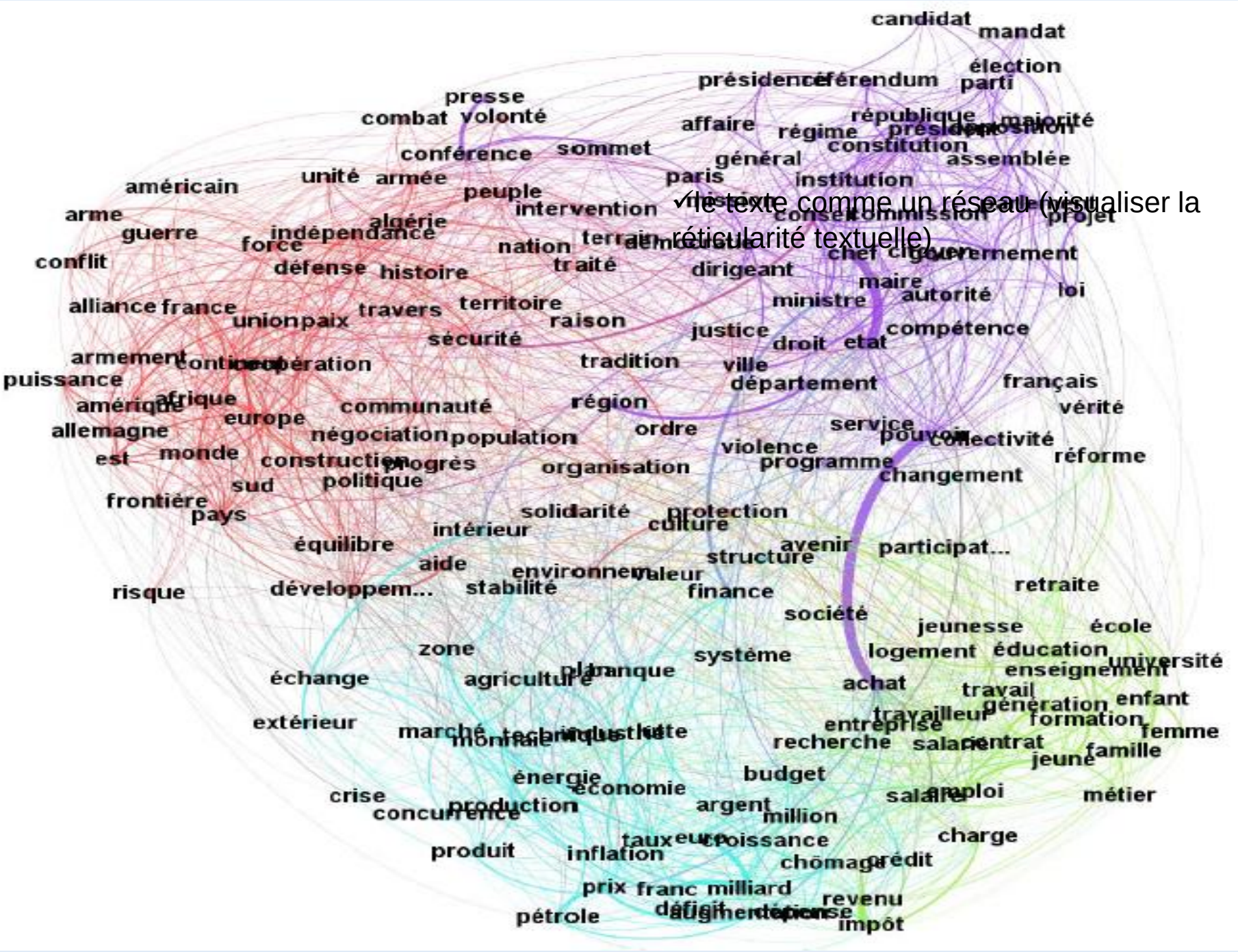
Extrait de Macron attribué à Pompidou dans le discours des vœux du 31 décembre 2020 (en vert les codes, en rouge les lemmes, en bleu les formes qui ont été activés par le réseau de neurones artificiels pour cette attribution à Pompidou)

Perspective : vers une implémentation du cercle herméneutique ? Rêvons un peu =☺



Perspective : vers une implémentation du cercle herméneutique ? Rêvons un peu =☺





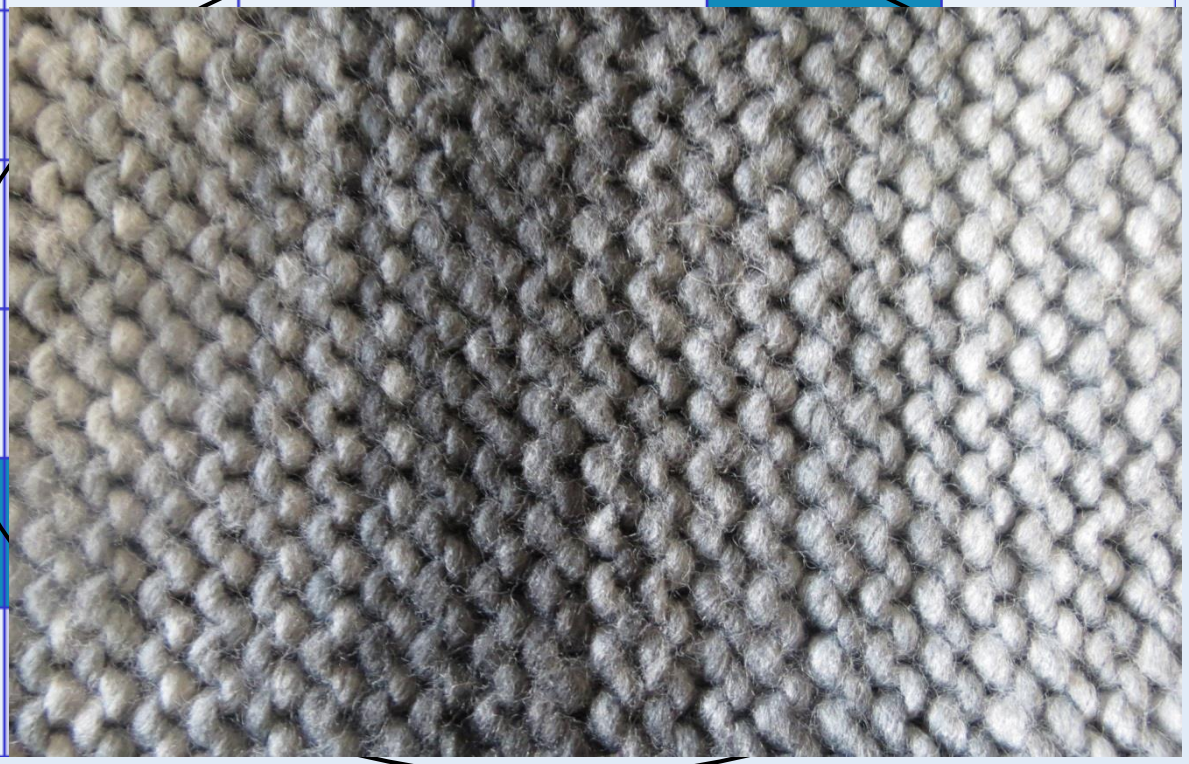
- le texte comme un réseau (visualiser la réticularité textuelle)

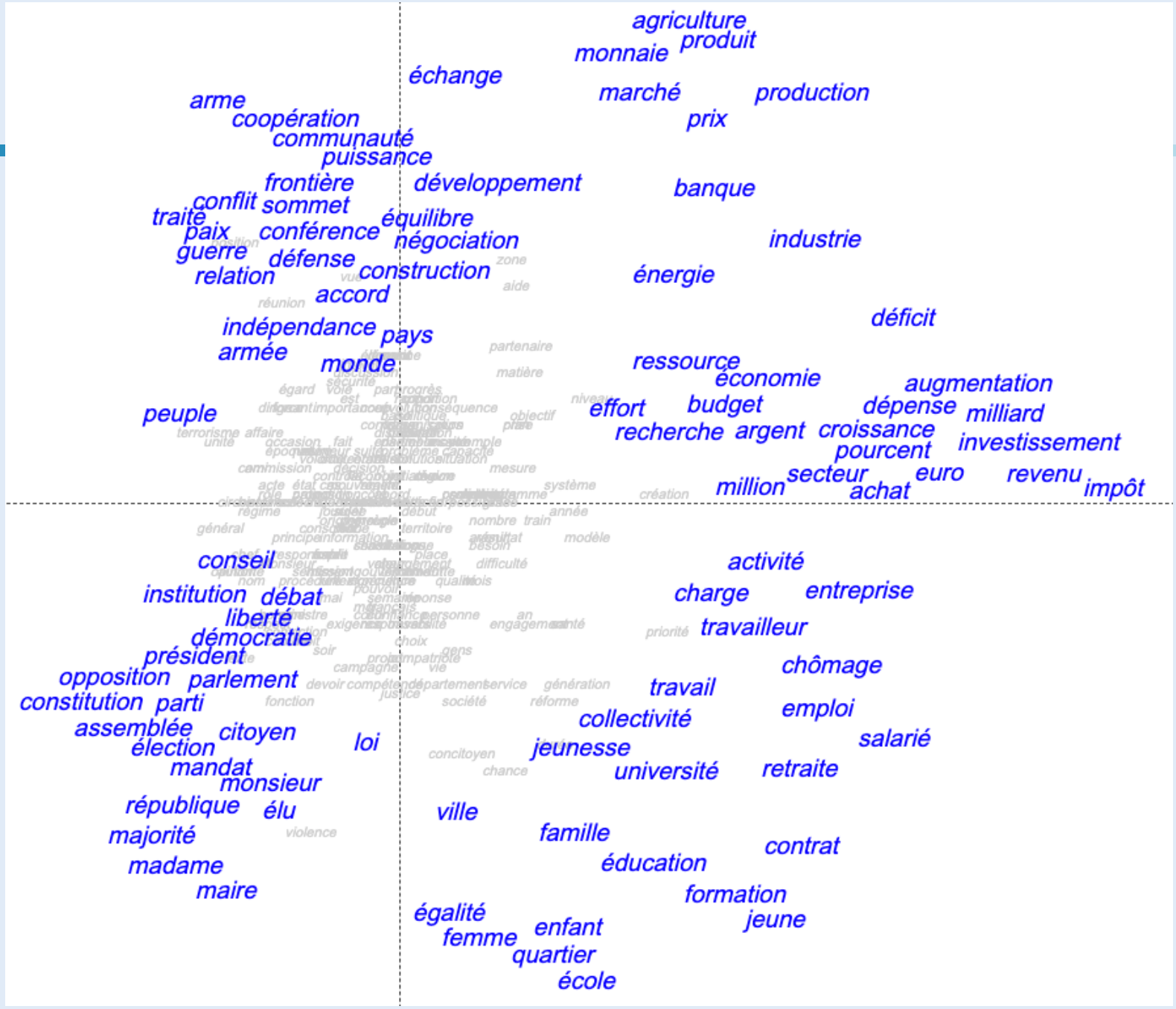
Les matrices de cooccurrences

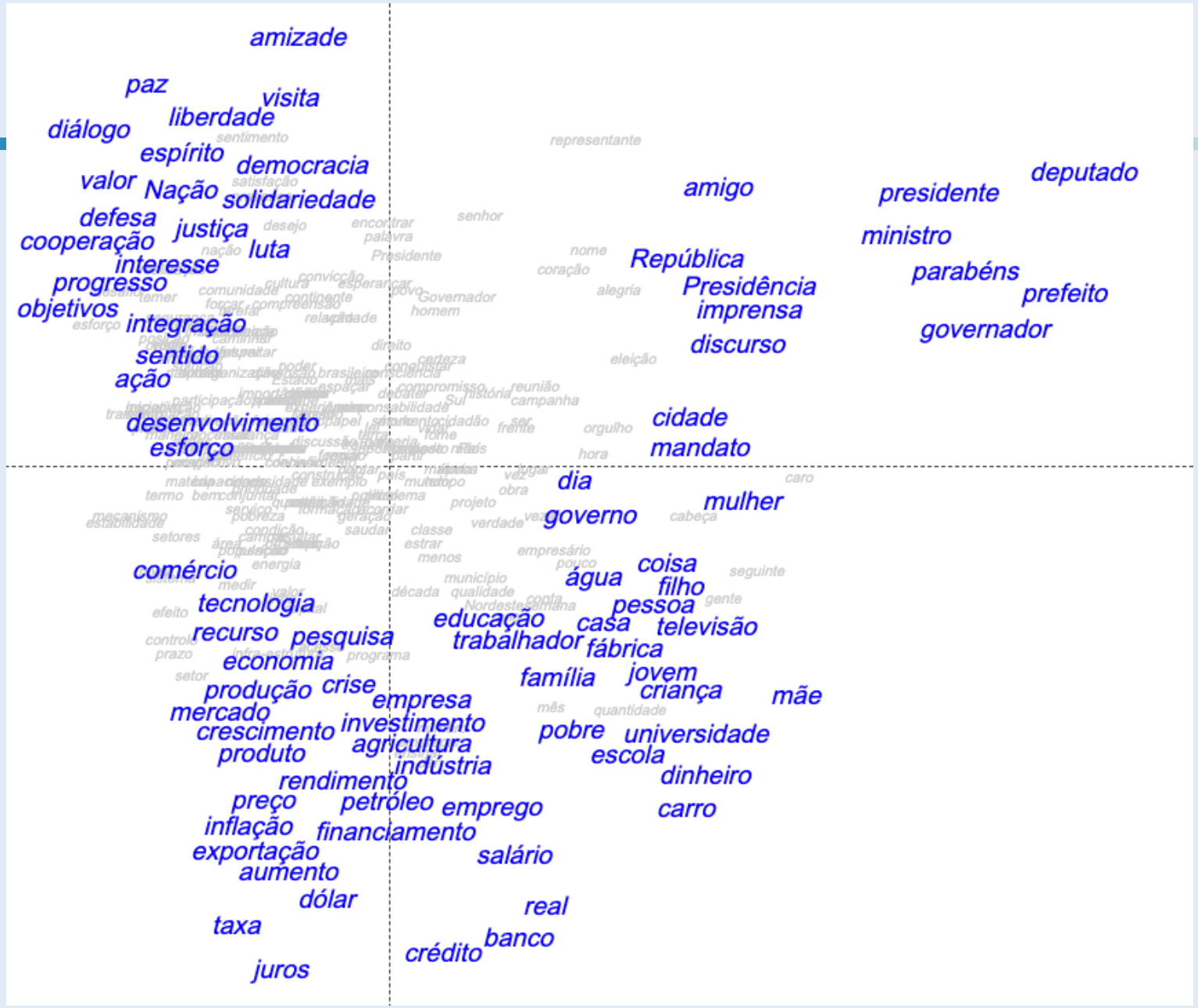
	Mot A	Mot B	Mot C	Mot D	Etc.
Mot A	***	*x* (cooc A_B)	*y* (cooc A_C)	*z* (cooc A_D)	…
Mot B	*x* (cooc A_B)	***	*v* (cooc B_C)	*w* (cooc B_D)	…
Mot C	*y* (cooc A_C)	*v* (cooc B_C)	***	*u* (cooc C_D)	…
Mot D	*z* (cooc A_D)	*w* (cooc B_D)	*u* (cooc C_D)	***	…
Etc.	…	…	…	…	***

Sens du mot =>

	‘peuple’	‘nation’	‘liberté’	‘ouvrier’	…
‘peuple’					
‘Nation’					
‘liberté’					
‘ouvrier’					
…					




agriculture
produit
monnaie
échange
marché
production
arme
coopération
prix
communauté
puissance
développement
banque
frontière
conflit
sommet
équilibre
traité
paix
conférence
négociation
industrie
guerre
défense
construction
énergie
relation
accord
déficit
indépendance
pays
armée
monde
ressource
économie
augmentation
peuple
effort
budget
dépense
milliard
recherche
argent
croissance
investissement
pourcent
secteur
euro
revenu
million
achat
impôt
conseil
activité
institution
débat
charge
entreprise
liberté
travailleur
démocratie
président
chômage
opposition
parlement
travail
constitution
parti
emploi
collectivité
assemblée
citoyen
loi
salarié
élection
jeunesse
mandat
université
retraite
monsieur
république
élu
ville
majorité
famille
contrat
madame
éducation
maire
formation
égalité
jeune
femme
enfant
quartier
école


amizade
paz
visita
liberdade
diálogo
espírito
democracia
valor
Nação
solidariedade
defesa
justiça
cooperação
luta
interesse
progresso
objetivos
integração
sentido
ação
desenvolvimento
esforço
comércio
tecnologia
recurso
pesquisa
economia
produção
crise
mercado
crescimento
produto
rendimento
preço
petróleo
inflação
financiamento
exportação
aumento
dólar
taxa
juros
amigo
presidente
deputado
ministro
República
parabéns
Presidência
prefeito
imprensa
governador
discurso
cidade
mandato
dia
mulher
governo
coisa
água
filho
pessoa
educação
casa
televisão
trabalhador
fábrica
família
jovem
criança
mãe
empresa
investimento
pobre
universidade
agricultura
escola
indústria
dinheiro
emprego
carro
salário
real
banco
crédito