Catégorisation automatique non-supervisée

Identification automatique des frontières de langues

Le cas des frontières floues

Pascal Vaillant (Univ. Paris-13)
Richard Nock (Univ. Antilles-Guyane)

Questionnements d’origine

- Les travaux sur la modélisation syntaxique du créole, et sur le développement d'outils d'informatique linguistique pour le créole, butent sur deux phénomènes très intéressants :
  - Labilité des catégories morpho-syntaxiques
  - Mélange quasi-permanent de créole et de français (sauf certains genres particuliers)

Mélanges de langues

- Corpus recueillis par deux étudiantes de maîtrise à l'UAG (Martinique) en 2006 (Lengrai, Moustin)
- Émissions de radio en créole
- Illustrent des phénomènes relevant du mélange de langues, et (hypothèse) de l'évolution du système lexico-syntaxique du créole

Corpus de créole à la radio

- Phénomènes intéressants :

C'est le fait que vèn a pé subi *des stress les plus* ***diverses***, *des stress répétés comme par exemple le fait que* ou ka manipulé *des charges plus ou moins* ***lourd***, *ou bien des charges* pli léjè *mais de manière fréquente*, ou pé ni an *environnement* ki pé, ki ka favorizé *mauvaise circulation* an, *par exemple, le fait que* ka fè cho, *le fait que* ni humidité, *le fait que* ou ka travay, *par exemple,* adan an navion éti ou ka ni *des différences de pression* ki ka *exercé* ko yo *au niveau* ***(Ø)*** *circulation et au niveau* ko'w, sé pou sa lé moun ki ka voyajé ében yo sav bien ke lè ou adan avion-a sa ka rivé'w *de manière épisodique, de manière rare*, mais ou ka santi *quand même que* lè ou asiz adan avion-a, lé ou asiz adan avion-a pendant witè d'tan *eh bien* jamb ou ka vini lou *et que souvent* si ou tiré soulié'w adan avion-a *eh bien* ou pa a rivé mété soulié-a an didan pié'w paske pié i gonflé *et que* soulié-a limenm i pa gonflé i pa dilaté ko'y.

Corpus observé : lexique

- Plan du lexique : utilisation massive du lexique français, notamment dans des domaines spécifiques (ex. émission médicale), mais aussi dans le secteur des connecteurs logiques et argumentatifs

  ⇒ banal … *mais* spécificité = lexique partagé

- *De ce point de vue*, le contact créole / langue lexificatrice a des propriétés communes avec le contact de dialectes apparentés

Corpus observé : syntaxe

- Plan de la morphosyntaxe : beaucoup de variation des formes observées sur certains axes (détermination du syntagme nominal, introduction des subordonnées)
- Sur certains aspects du système syntaxique, propriétés radicalement différentes dans les deux langues : *de ce point de vue*, le contact créole / langue lexificatrice est radicalement différent du contact entre dialectes
- Variation *dans les deux langues*

  ⇒ hypothèse de la « décréolisation » (entendu comme catégorie particulière d'attrition) trop simpliste

Problèmes

- Comment transcrire ?
  - La question de savoir s'il faut utiliser la norme orthographique française ou créole pour une unité est souvent indécidable
- Comment décrire ?
  - Comment un descriptiviste/modélisateur qui étudie le créole peut-il exploiter ce corpus ?
  - Comment un linguiste qui étudie ce type de corpus peut-il annoter les différents segments ?

Frontières floues

- Unités linguistiques apparentées dans deux langues : quels critères d'attribution ?
- Phonétique ? (*fimɛ / fymœχ*) → pas à chaque fois
- Morphologie figée (traces d'amalgame) ?
  - Mais on trouve : *latè-a, latè, tè-a, tè ;*
    avion, lavion, navion
- Contexte syntaxique ?
  - Mais on trouve : *les globules rouges, sé éléments ta'a,*
    écoulement urine a, an étanchéité de vessie a …

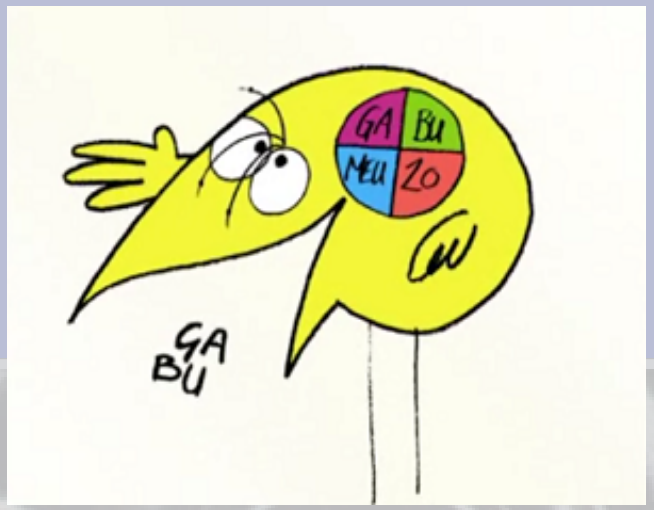
Discrimination automatique

- Recherche en cours : systèmes d'apprentissage pour la discrimination automatique de segments homogènes dans les corpus
- Principe : sur la base d'un corpus, retrouver des sous-ensembles d'unités lexicales qui sont stables du point de vue de l'enchaînement syntagmatique
- Résultat attendu : retrouver les différentes langues

Apprentissage non-supervisé

- Le problème a priori n'est pas l'identification d'une langue déjà connue et pour laquelle nous possédons un modèle de description
- Il s'agit de retrouver des enchaînements réguliers d'unités dans les corpus étudiés, sans savoir combien il y a de langues (1, 2, 3 ... + ?), quelles sont ces langues, et où en sont les frontières

Principe intuitif

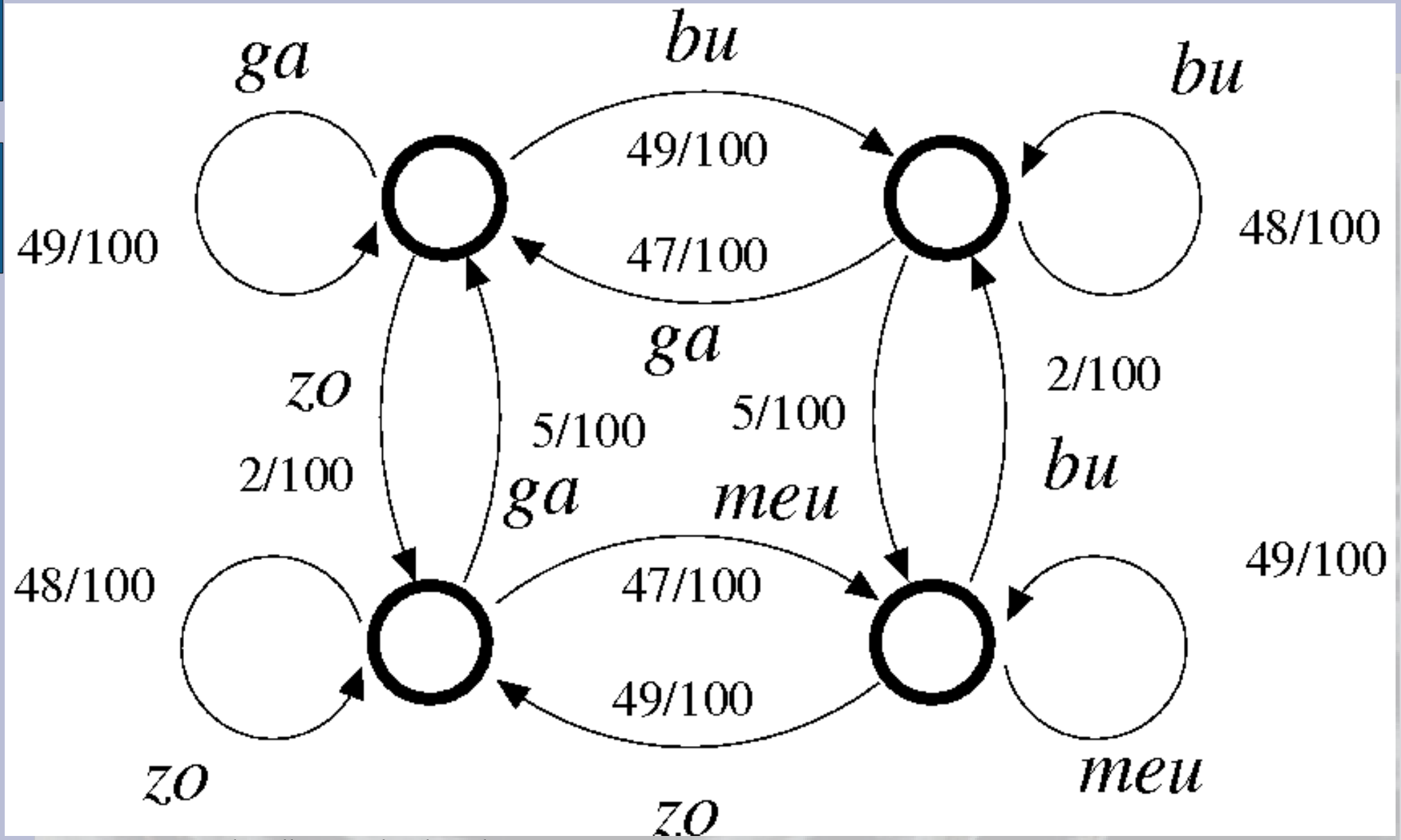


- On observe une série d'unités :

  « *ga bu ga ga bu ga bu ga ga bu ga bu ga bu meu zo meu zo zo meu zo zo meu zo meu zo zo meu zo ga bu ga bu ga ga bu ga bu zo …* »

- Les régularités observées témoignent de l'existence d'un système A (unités : *ga* et *bu*) et d'un système B (unités : *zo* et *meu*)
- On peut aussi repérer les mots près desquels se produisent les transitions (*bu* et *zo*)

Processus génératif



Observables

- Pour un niveau donné (ici, les mots) :
- Corpus de n mots-types (vocables) distincts (w_i, $i \in [1,n]$) et de N mots-occurrences
- On note $w_{i,j}$ le nombre de fois où le mot w_j vient juste après le mot w_i
- La matrice W ($n \times n$) qui contient les coefs. $w_{i,j}$ est la matrice des transitions (table de contingence)

Éléments du calcul

- On note d_i le nombre total d'occurrences de chaque mot w_i
- $P = D^{-1} W$ ($p_{i,j} = w_{i,j} / d_i$) est la matrice des probabilités de transition
- Dans le cas général, P n'est pas symétrique, mais on peut la rendre symétrique en postulant que le sens de lecture est aléatoire (i.e. $p_{i,j} = (w_{i,j} + w_{j,i}) / 2d_i$)

Matrice de transition

	ga	*bu*	*zo*	*meu*
ga	0,49	0,49	0,02	0,00
bu	0,47	0,48	0,00	0,05
zo	0,05	0,00	0,48	0,47
meu	0,00	0,02	0,49	0,49

Analyse spectrale

- P est la matrice de probabilité de transition de la chaîne de Markov à n états qui représente le langage engendrant le corpus (modèle de bigrammes)
- On peut faire une Analyse Spectrale sur P (comme on le fait, dans le cas d'une ACP, avec la matrice normalisée $Q = D^{-½} W D^{-½}$)
- On obtient des axes propres sur lesquels on observe la répartition des unités w_i

Classification spectrale discrète (1)

- On postule l'existence d'une fonction d'appartenance d'un « mot » i à un « cluster » k : $z_{i,k} = 0$ ou 1
- La matrice $Z = (z_{i,k})$ ($i \in [1,n]$, $k \in [1,q]$) définit la répartition des mots en clusters
- Sur chaque ligne, un seul 1 et plein de 0

Classification spectrale discrète (2)

- On peut définir une fonction de « coût » à minimiser (le *coût normalisé*) :

$$\mu(Z) = \Sigma_{k=1}^{q} \kappa_k(Z) / \alpha_k(Z)$$

où

$$\kappa_k(Z) = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} w_{i,j} (z_{i,k} - z_{j,k})^2 \qquad \text{« surface »}$$

$$\alpha_k(Z) = \Sigma_{i=1}^{n} d_i z_{i,k}^2 \qquad \text{« volume »}$$

(minimiser le poids des arcs traversant les frontières de clusters, maximiser le poids des arcs internes)

Classification spectrale discrète (3)

- **Problème n°1** : on n'est pas sûr d'obtenir une solution dans un temps de calcul raisonnable (problème « NP-complet »)
- **Problème n°2** : ce n'est pas vraiment le résultat que l'on veut obtenir !

Classification spectrale continue

- On ne recherche plus une fonction Z discrète mais une fonction *Y* continue
- On ne sait pas à l'avance combien de clusters on va avoir, mais on sait que *pour un nombre donné de clusters*, il suffit de minimiser $\Sigma_k \kappa_k(Y)$ (la « conductance » entre clusters)

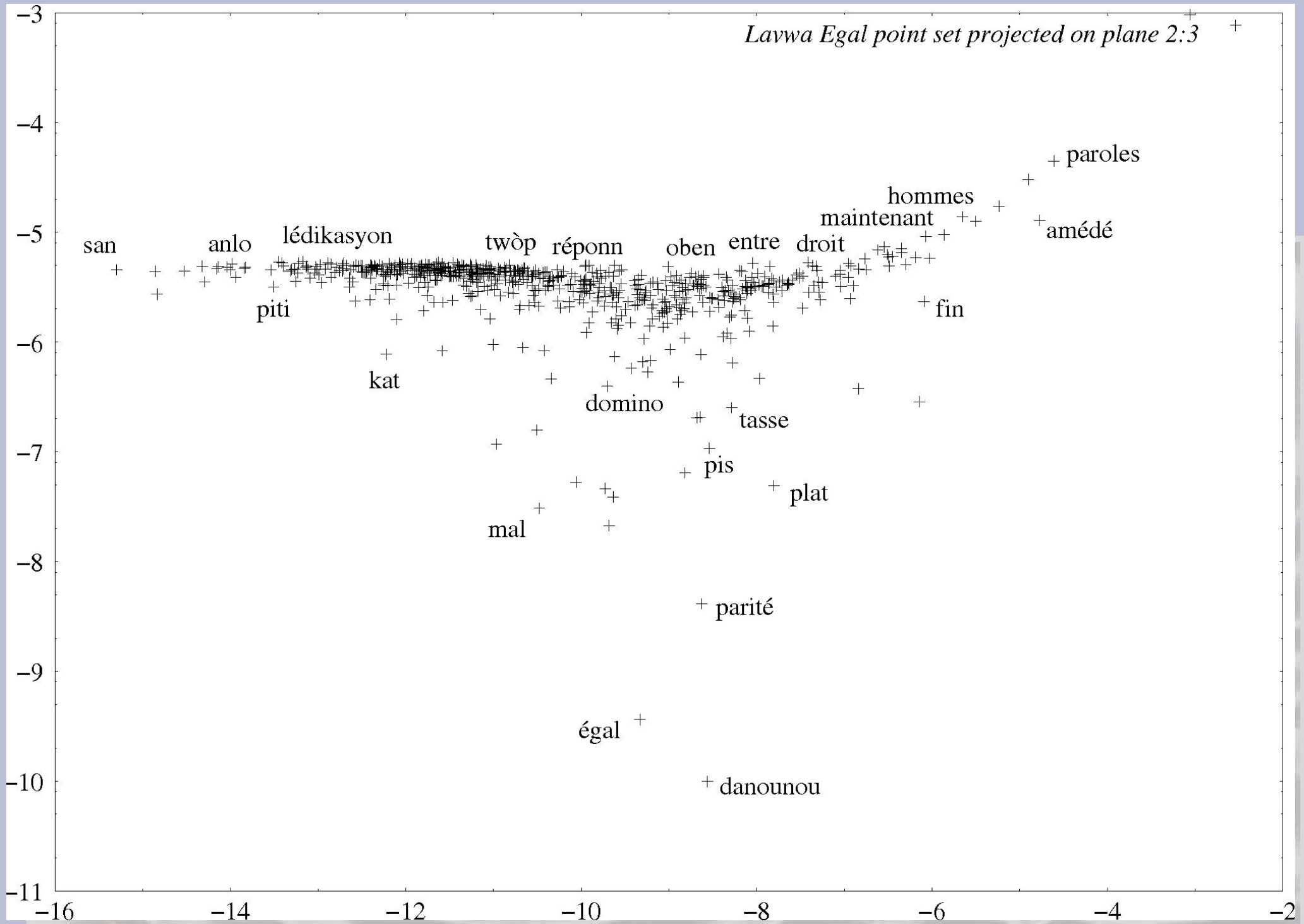
Problème de valeurs propres

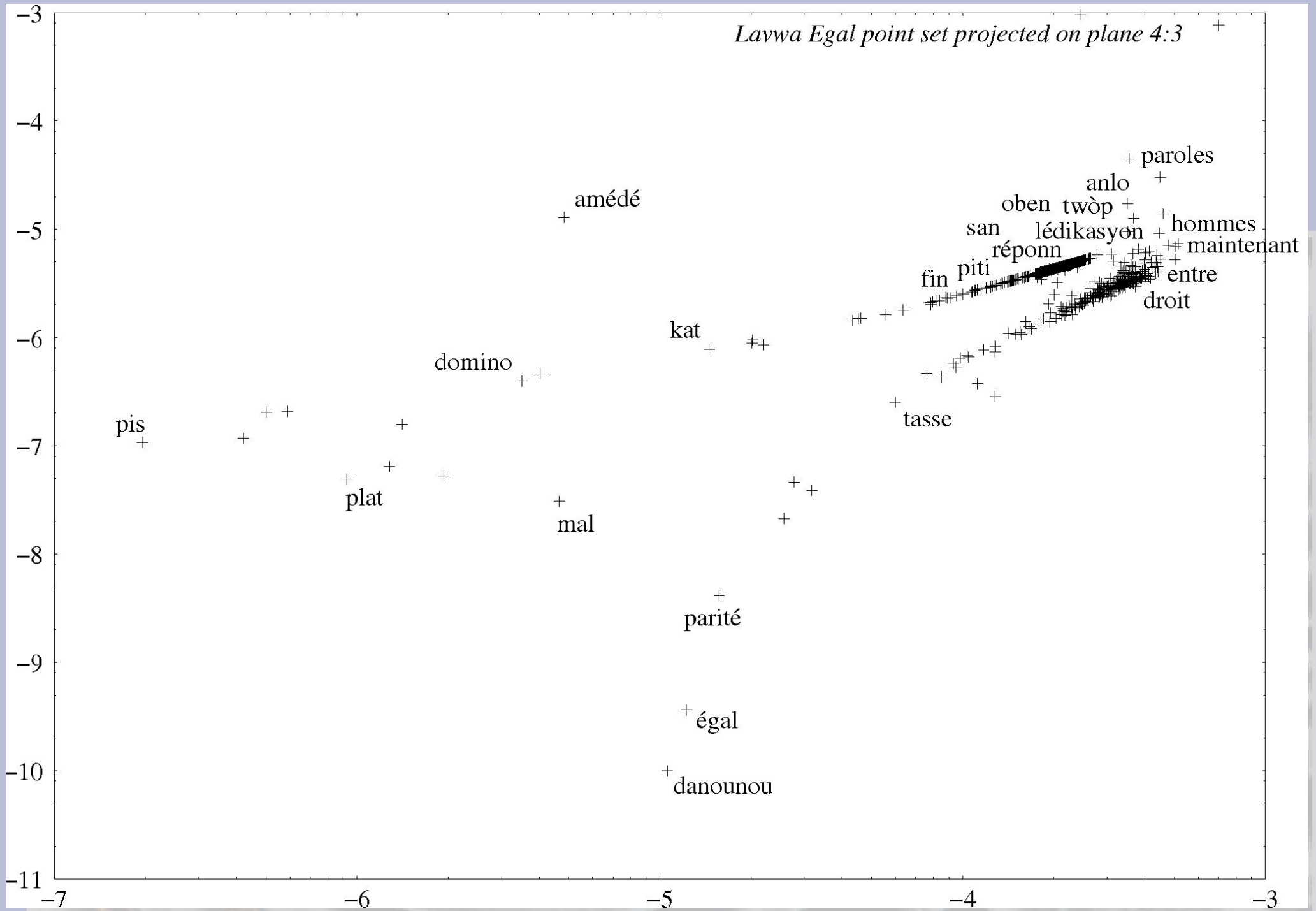
- Minimiser $\Sigma_k \kappa_k(\mathrm{Y}) = \Sigma_k \Sigma_{i,j} \mathrm{w}_{i,j}(\mathrm{y}_{i,k} - y_{i,k})^2$
- Trouver une solution à ce problème revient à minimiser λ_k dans $(\mathrm{D} - \mathrm{W})\, y_k = \lambda_k \mathrm{D}\, y_k$

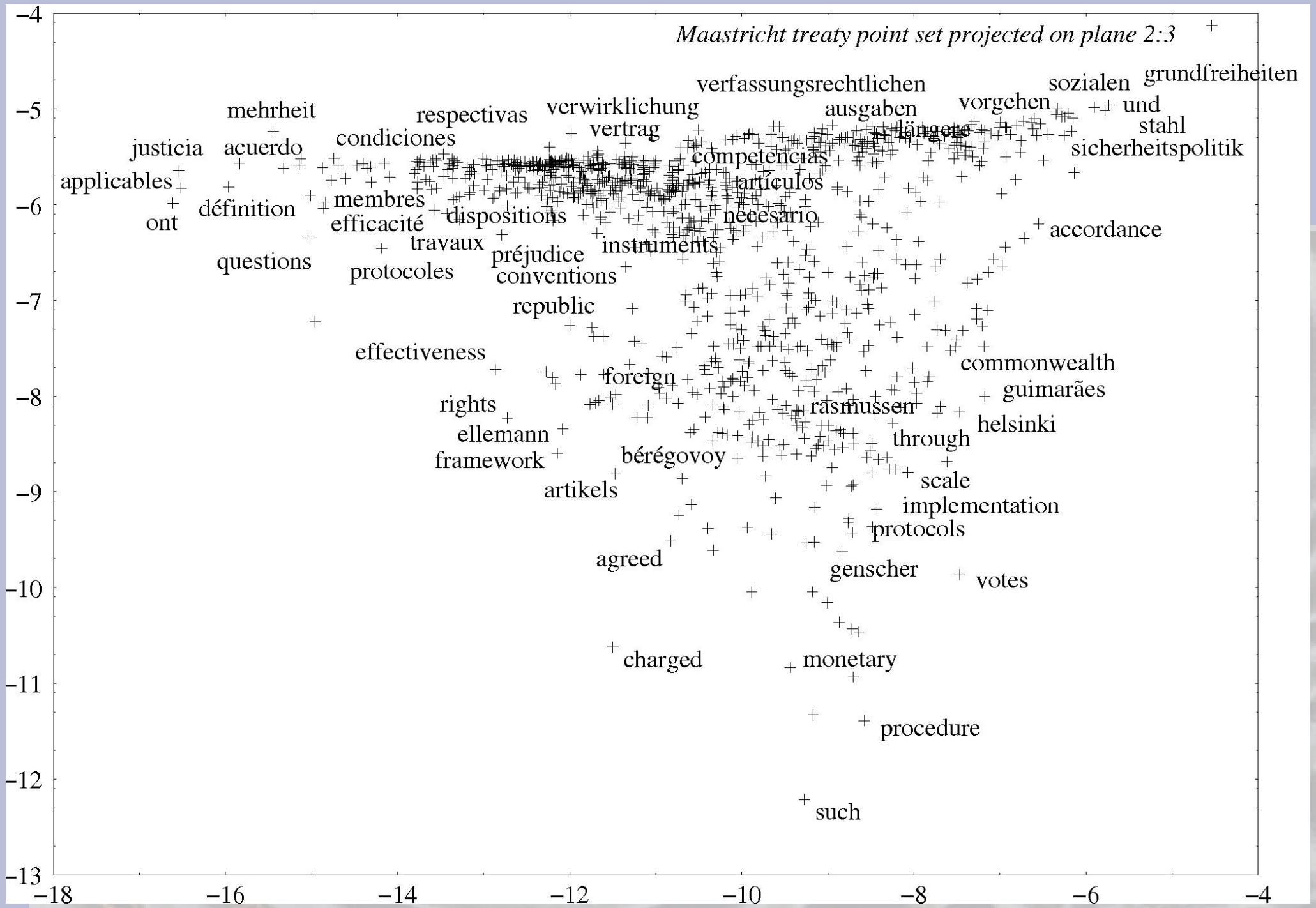
  (Pothen, Simon & Liou, 1990)
- donc à minimiser λ_k dans $\mathrm{P}\, y_k = (1 - \lambda_k)\ y_k$
- ⇒ donc à trouver les plus grandes valeurs propres de P

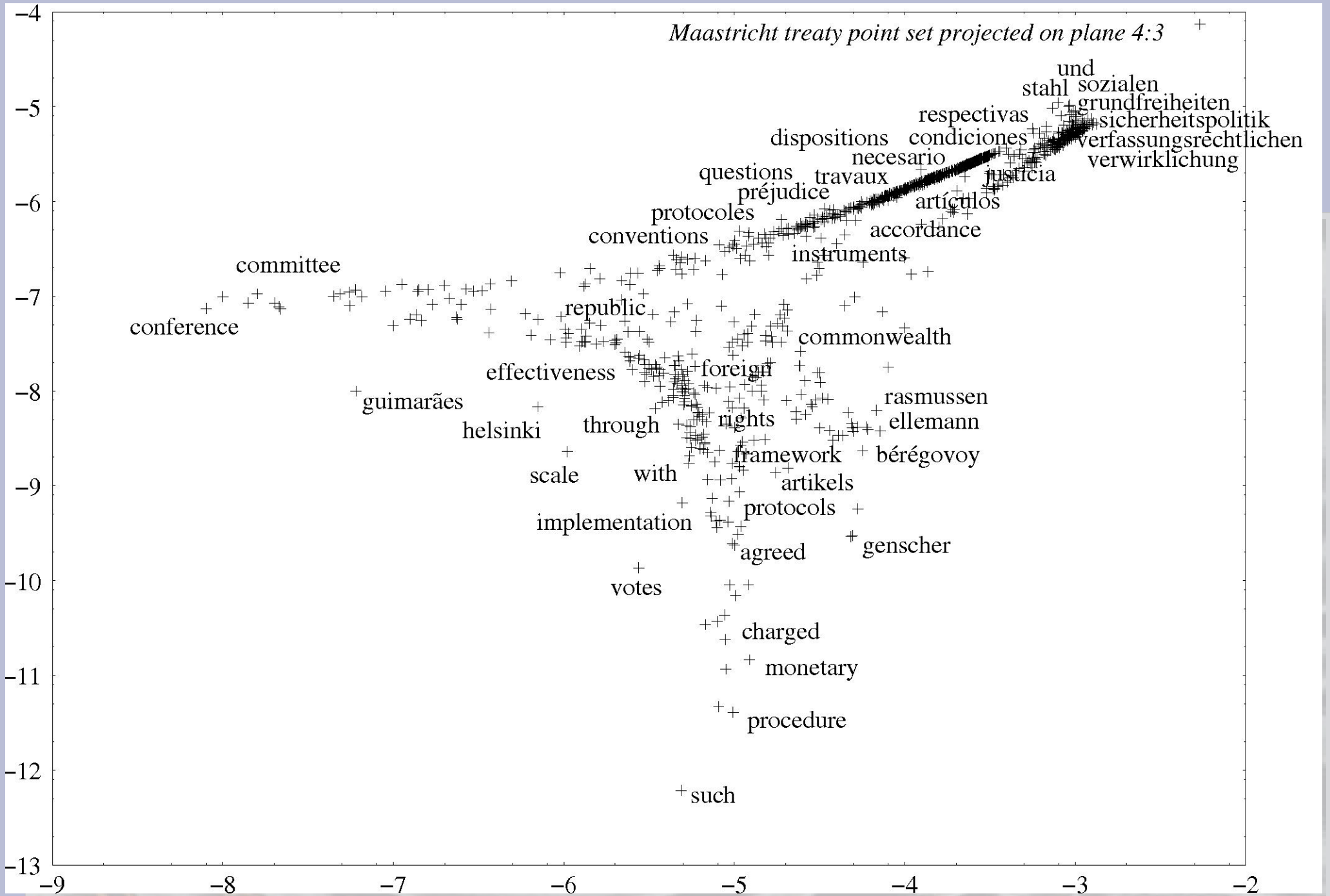
Interprétation probabiliste

- Si au lieu de considérer les n vecteurs propres de la matrice P (les vecteurs Y_k, composés chacun de n coordonnées $y_{k,i}$), on considère les vecteurs $\tilde{Y}_k$ définis par $\tilde{y}_{k,i} = d_i \, {y_{k,i}}^2$
- ... on a une distribution de probabilité sur l'ensemble des mots (ce qui ouvre la voie aux calculs bayésiens)
- et au niveau discrimination, ça marche encore


Lavwa Egal point set projected on plane 2:3
paroles
hommes
maintenant
amédé
san
anlo
lédikasyon
twòp
réponn
oben
entre
droit
fin
piti
kat
domino
tasse
pis
plat
mal
parité
égal
danounou
−3
−4
−5
−6
−7
−8
−9
−10
−11
−16
−14
−12
−10
−8
−6
−4
−2


Lavwa Egal point set projected on plane 4:3
−3
−4
−5
−6
−7
−8
−9
−10
−11
−7
−6
−5
−4
−3
paroles
anlo
oben
twòp
hommes
san
lédikasyon
maintenant
réponn
piti
fin
entre
droit
amédé
kat
domino
tasse
pis
plat
mal
parité
égal
danounou


Maastricht treaty point set projected on plane 2:3
verfassungsrechtlichen
sozialen
grundfreiheiten
mehrheit
respectivas
verwirklichung
ausgaben
vorgehen
und
vertrag
längere
stahl
justicia
acuerdo
condiciones
competencias
sicherheitspolitik
applicables
artículos
définition
membres
dispositions
necesario
ont
efficacité
accordance
travaux
instruments
questions
préjudice
protocoles
conventions
republic
effectiveness
commonwealth
foreign
guimarães
rights
rasmussen
helsinki
ellemann
through
framework
bérégovoy
scale
artikels
implementation
protocols
agreed
genscher
votes
charged
monetary
procedure
such
−4
−5
−6
−7
−8
−9
−10
−11
−12
−13
−18
−16
−14
−12
−10
−8
−6
−4

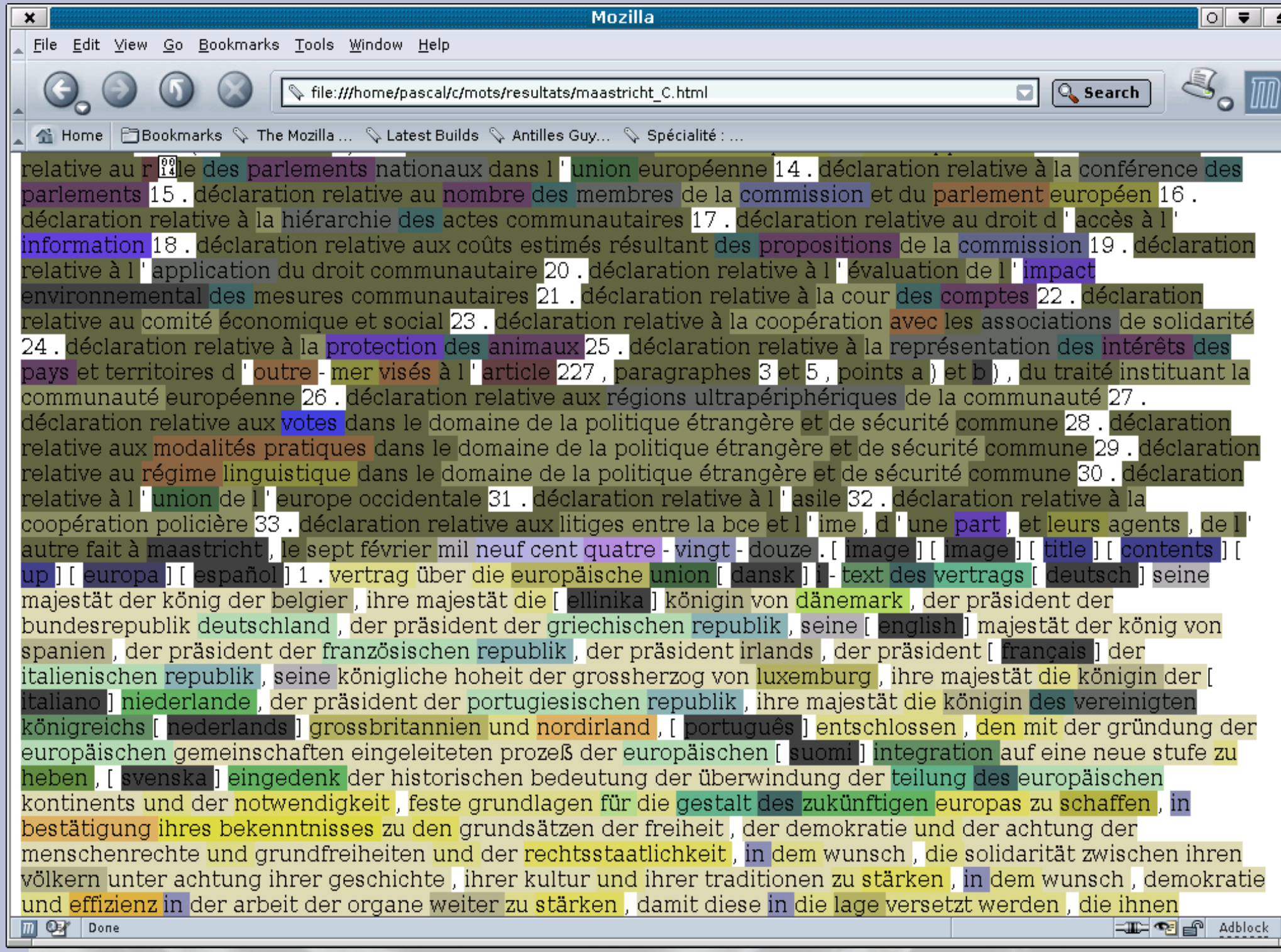

Maastricht treaty point set projected on plane 4:3
und
stahl
sozialen
grundfreiheiten
respectivas
sicherheitspolitik
dispositions
condiciones
verfassungsrechtlichen
necesario
verwirklichung
questions
travaux
justicia
préjudice
artículos
protocoles
accordance
conventions
instruments
committee
republic
conference
commonwealth
effectiveness
foreign
guimarães
rasmussen
helsinki
through
rights
ellemann
framework
bérégovoy
scale
with
artikels
protocols
implementation
genscher
agreed
votes
charged
monetary
procedure
such
−4
−5
−6
−7
−8
−9
−10
−11
−12
−13
−9
−8
−7
−6
−5
−4
−3
−2

Exemples résultats

- On peut également visualiser les résultats dans un espace de couleurs tridimensionnel (ici, dans l'ordre du texte) :
  - Lavwa Egal
  - Traité de Maastricht

Lavwa Egal

solisyon pou pwoblèm asou pwoblèm yo ka ban nou . nou , nou ni solisyon . yo , sé pwoblèm yo ni . sa zòt lé ! moun té ka di lan dé mil té ké mennen lafendimond . mwen menm lan pèsonèlman , an ka di zòt konsa , si sé fanm - lan , yo tou , yo ka mété kòyo mandé moun kous - kouri , olyé yo kouri fè kous pou mété kay yo , kon sa pou fèt , an ka di viraj - la danjéré , nèg , i brital menm menm . nou ka pati pa tèt . « quel malheur ! quel grand malheur pour nous ! » i bout èpi chanté taa . - sizànn bèl fanm * ! roven réponn . i montré épi i kwenyen an domino sis - twa , byen obéyisan , asou tab - la . - twa - zilé bagatèl * ! tokant réplitjé ' y lamenm lamenm ka fésé lanmen ' y tout fòs li toujou anlè tab - la , ki viré pwan ankò an bèl volé . i pozé domino doub - twa ' y la , dan lalinman sé lézòt - la . - lésé woulé , lésé woulé ! dilasouy fè an bout dan , jé bankal li , an lanmen ' y . même langage . je peux en donner mille autres exemples . elles naissent avec ces défauts , je te dis . elles ont là une marque de fabrique indélébile . le pire c ' est qu ' elles l ' ignorent . . . une femme , premièrement déjà c ' est une catastrophe , une calamité dans notre existence . nous nous devons de porter toutes les solutions aux problèmes qu ' elles nous procurent . nous devons surtout leur porter les solutions les gens disaient qu ' en l ' an deux mille , ce serait la fin du monde , moi , personnellement , je vous assure que si les femmes , elles aussi , se mettent dans la course , au lieu d ' aller faire leurs courses , comme il se doit , nous amorçons le virage de notre


Mozilla
File Edit View Go Bookmarks Tools Window Help
file:///home/pascal/c/mots/resultats/maastricht_C.html
Search
Home Bookmarks The Mozilla ... Latest Builds Antilles Guy... Spécialité : ...
relative au r le des parlements nationaux dans l ' union européenne 14 . déclaration relative à la conférence des parlements 15 . déclaration relative au nombre des membres de la commission et du parlement européen 16 . déclaration relative à la hiérarchie des actes communautaires 17 . déclaration relative au droit d ' accès à l ' information 18 . déclaration relative aux coûts estimés résultant des propositions de la commission 19 . déclaration relative à l ' application du droit communautaire 20 . déclaration relative à l ' évaluation de l ' impact environnemental des mesures communautaires 21 . déclaration relative à la cour des comptes 22 . déclaration relative au comité économique et social 23 . déclaration relative à la coopération avec les associations de solidarité 24 . déclaration relative à la protection des animaux 25 . déclaration relative à la représentation des intérêts des pays et territoires d ' outre - mer visés à l ' article 227 , paragraphes 3 et 5 , points a) et b) , du traité instituant la communauté européenne 26 . déclaration relative aux régions ultrapériphériques de la communauté 27 . déclaration relative aux votes dans le domaine de la politique étrangère et de sécurité commune 28 . déclaration relative aux modalités pratiques dans le domaine de la politique étrangère et de sécurité commune 29 . déclaration relative au régime linguistique dans le domaine de la politique étrangère et de sécurité commune 30 . déclaration relative à l ' union de l ' europe occidentale 31 . déclaration relative à l ' asile 32 . déclaration relative à la coopération policière 33 . déclaration relative aux litiges entre la bce et l ' ime , d ' une part , et leurs agents , de l ' autre fait à maastricht , le sept février mil neuf cent quatre - vingt - douze . [image] [image] [title] [contents] [up] [europa] [español] 1 . vertrag über die europäische union [dansk] 1 - text des vertrags [deutsch] seine majestät der könig der belgier , ihre majestät die [ellinika] königin von dänemark , der präsident der bundesrepublik deutschland , der präsident der griechischen republik , seine [english] majestät der könig von spanien , der präsident der französischen republik , der präsident irlands , der präsident [français] der italienischen republik , seine königliche hoheit der grossherzog von luxemburg , ihre majestät die königin der [italiano] niederlande , der präsident der portugiesischen republik , ihre majestät die königin des vereinigten königreichs [nederlands] grossbritannien und nordirland , [português] entschlossen , den mit der gründung der europäischen gemeinschaften eingeleiteten prozeß der europäischen [suomi] integration auf eine neue stufe zu heben , [svenska] eingedenk der historischen bedeutung der überwindung der teilung des europäischen kontinents und der notwendigkeit , feste grundlagen für die gestalt des zukünftigen europas zu schaffen , in bestätigung ihres bekenntnisses zu den grundsätzen der freiheit , der demokratie und der achtung der menschenrechte und grundfreiheiten und der rechtsstaatlichkeit , in dem wunsch , die solidarität zwischen ihren völkern unter achtung ihrer geschichte , ihrer kultur und ihrer traditionen zu stärken , in dem wunsch , demokratie und effizienz in der arbeit der organe weiter zu stärken , damit diese in die lage versetzt werden , die ihnen
Done
Adblock

Classification spectrale continue

- L'examen de la répartition des points sur les axes propres de l'analyse spectrale montre que les unités tendent à se grouper par langues
- On pourrait donc « seuiller » pour obtenir une classification binaire ...
- Mais on peut faire quelque chose de plus intelligent dans certains cas : *ne pas* seuiller
- On obtient un « taux d'appartenance à » la langue X ou Y

Utilisation : corpus non annotés

- Acquisition de segments de corpus homogènes à partir de corpus hétérogènes
- À partir de ces segments homogènes, on peut établir des modèles de langue homogènes et s'en servir pour faire de la reconnaissance (et non plus seulement de la discrimination)
- On peut réinsérer ces étiquetages de 1er ordre dans la dimension syntagmatique du texte pour avoir un étiquetage des occurrences en contexte

Limitations

- Fonctionne grâce à la stabilité des enchaînements au sein de sous-ensembles du vocabulaire du corpus
- … mais ne peut « voir » que la variabilité interne au corpus
- Avec un grain trop fin, mauvais résultats
- Il faut un corpus d'apprentissage contenant quelques grosses zones homogènes pour pouvoir traiter aussi les corpus très hachés

Pour aller plus loin

- Avec des corpus plurilingues annotés : apprentissage semi-supervisé
- Annotations utilisables pour l'essaimage de premières valeurs (mots dont on connaît la langue ou le caractère flottant entre *n* langues)
- Au-delà des catégories de langues, devenues solides, exploration des catégories de morphèmes sensibles à l'interférence