VOILADIS : des relations lexicales aux structures de discours

Clémentine Adam, Cécile Fabre, Philippe Muller
CLLE/IRIT

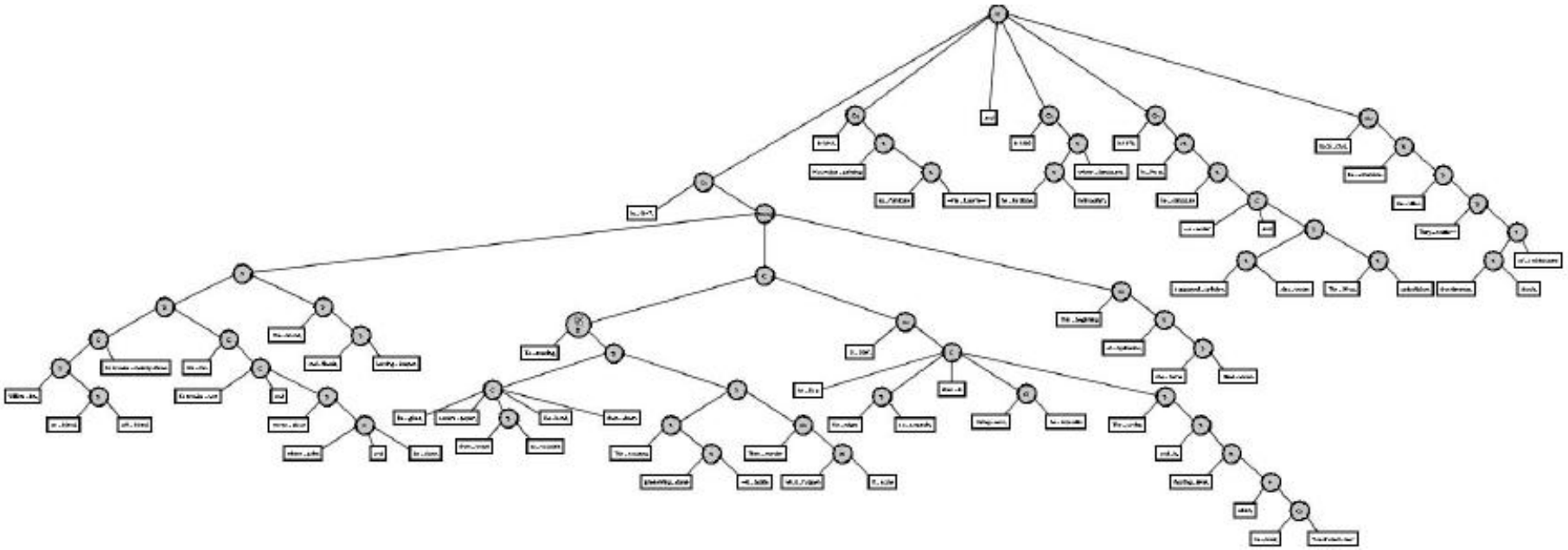
16 Décembre 2008

Analyse des structures de discours

Marcu/RST, Penn Discourse TreeBank, DISCOR/SDRT, Wolf & Gibson, Annodis

- un discours = un ensemble de segments et de relations entre segments
- segments = proposition, phrase, paragraphe, ...
- structure hiérarchique (segments élémentaires + segments englobants)
- relations : explication, contraste,

Structure hiérarchique du discours



Segmentation discursive / Topique

- découpage en segments à plus gros grain
- hypothèse : correspond aux étages supérieurs de l'analyse complète
- structure linéaire : partition du texte en segment convexe non emboîtés

Applications

- recherche d'informations/ extraction d'informations
- résumé automatique
- navigation documentaire (intra ou inter textes)
- filtrage sur d'autres tâches (ex: OCR)

Procédure

- ingrédient de base : trouver des segments à regrouper
  - par opposition aux segments voisins
  - parce qu'ils sont eux-même cohérents.
- à définir :
  - ce qu'est un segment de base
    - phrases
    - paragraphes typographiques
    - fenêtre de N mots (taille prédéterminée)
  - la mesure du lien (cohésion)
  - une méthode de regroupement/séparation
    - clusterisation hiérarchique
    - partitionnement

Méthodes computationnelles

- supervisé vs non-supervisé
- supervisé/catégorisation
  traits utilisables: lexique commun, présence de marqueurs, reprises anaphoriques,
- supervisé/modèles séquentiels : prend aussi en compte la présence des autres frontières
  (il est plus probable de ne pas avoir de borne juste après une borne)
- non supervisé : mesures de similarité entre phrases + critère fixé de coupure

Mesures de similarité/dissimilarité

- entre mots :
  - identiques
  - synonymes
  - collocations, distribution syntaxique,
  - similarité sémantique au sens large (LSA, ...)
- entre phrases
  - similarité lexicale (total de tous les mots, ou bien que les N, etc)
  - coréférences
  - n-grams
- entre paragraphes
  - somme des similarités de phrase
  - variations de distributions du lexique
- dissimilarité
  - marqueurs explicites de rupture
  - termes centraux (variations de distributions)

Les attentats du 11 septembre 2001 frappèrent New York et Washington à l'aide d'avions de ligne détournés, dans la matinée du jour éponyme.
Le terme regroupe une série d'évènements synchronisés qui se déroulèrent dans le nord-est des États-Unis d'Amérique : trois avions commerciaux (sur quatre détournés) furent précipités sur des immeubles représentatifs de la puissance américaine, économique pour les tours jumelles du World Trade Center à Manhattan, New York, et militaire pour le Pentagone, siège du ministère de la Défense des États-Unis, à Washington.
Les "Twin Towers" s'effondrèrent spectaculairement moins de deux heures après les impacts, ainsi qu'une troisième tour proche dite WTC7 , le Pentagone fut endommagé.

Similarités

- segment/phrase s_i = vecteurs de mots avec dimension = taille du lexique
- $w(s_i, s_j) = \frac{s_i \cdot s_j}{||s_i|| \times ||s_j||}$
  = nombre de mots en commun
- ajout contexte $s'_i = s_i + \sum_{s_j \in contexte} f(s_j, s_i)$
- pondération lexicale (du genre tf.idf).
  w= vecteur de pondération
  $s'_i = s_i \cdot w$
- généralisation : remplacer le produit scalaire par mesure de similarité lexicale
- groupes de segments : somme des similarités des segments

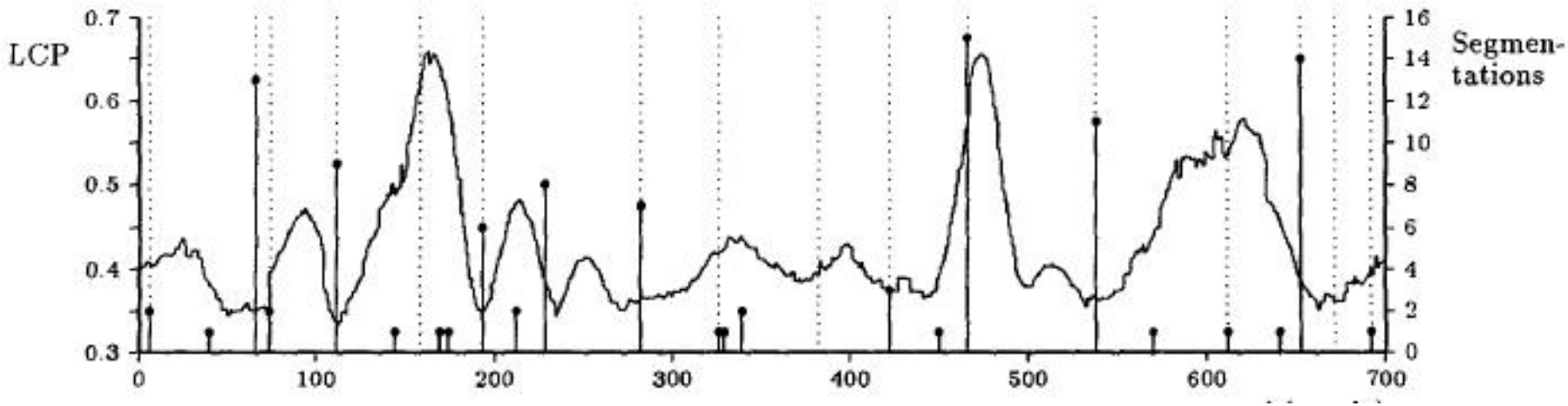
Prendre en compte les similarités

vue de façon :

- linéaire : similarité d'une phrase avec la précédente
- matricielle : similarité d'une phrase avec toutes les autres
- hiérarchique : similarité entre groupes de segments (récursivement)

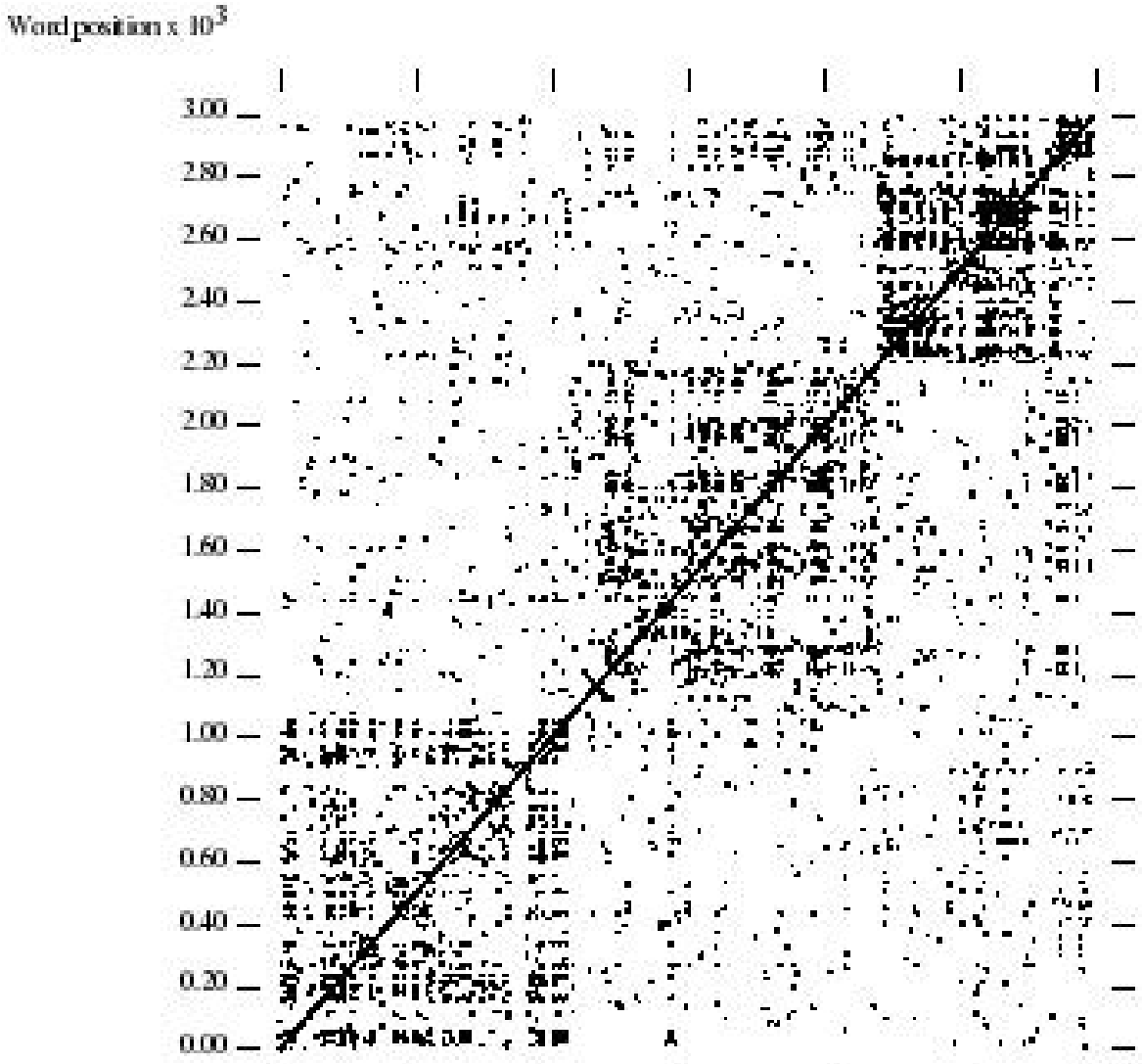
Segmentation Linéaire

Abscisse = numéro de la phrase
Ordonnée = similarité avec la phrase précédente
Frontière : changements "brusques"

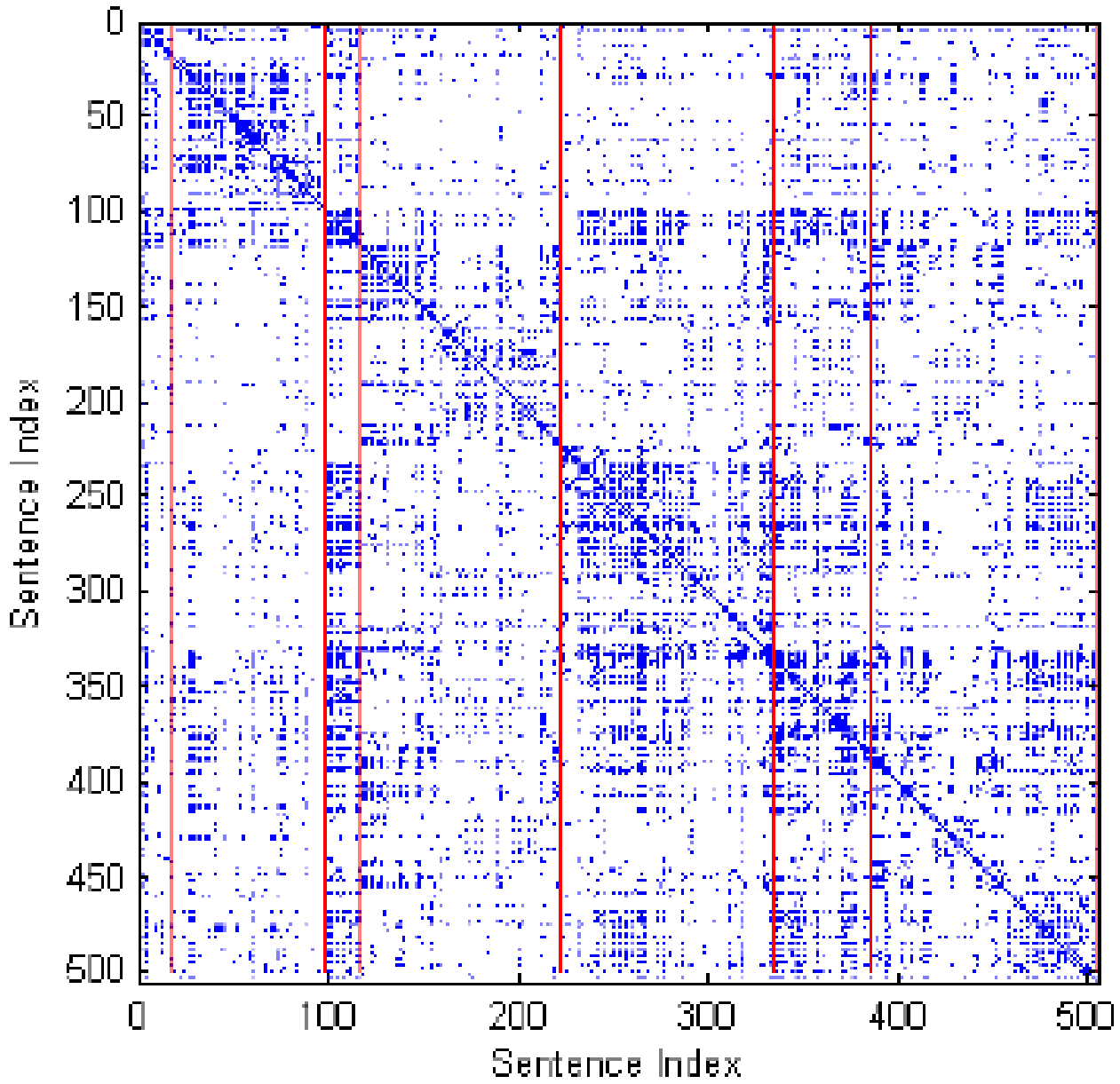


Représentation matricielle

Ligne / Colonne = numéros de phrase
valeur au point (i, j) = similarité de la phrase i / la phrase j

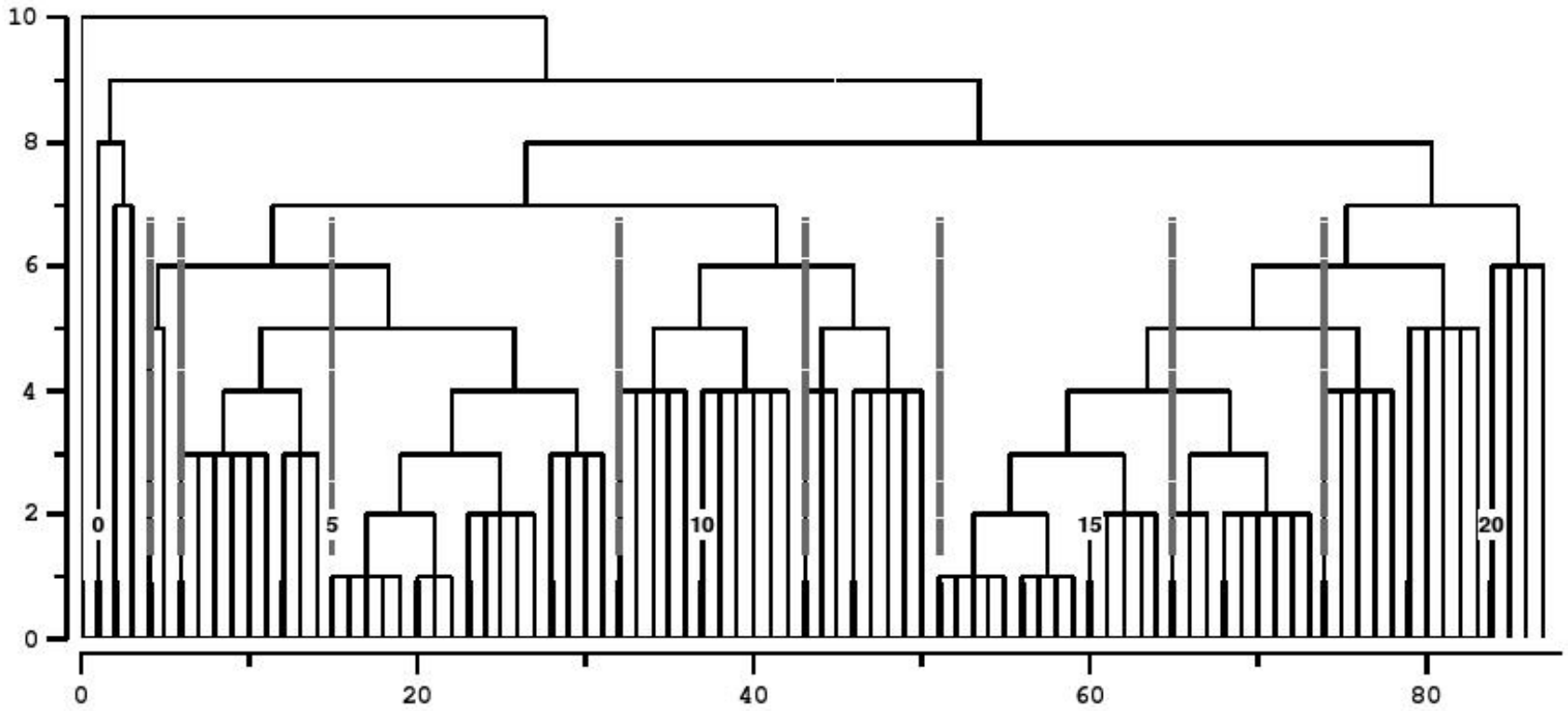


Représentation matricielle (2)



Segmentation hiérarchique ("clustering")

Abscisse = numéro de la phrase
Ordonnée = niveau de proximité dans la hiérarchie



Segmentation de groupes

- considérer chaque sous-ensemble (convexe) de phrases comme un segment potentiel
- optimiser directement la partition globale au lieu de prendre des décisions locales
- mesurer la cohérence interne d'un groupe et la distance avec les autres groupes
- = texte vu comme un graphe de distance qu'il faut couper en grappes
- complexité algorithmique beaucoup plus grande que de décider localement
- exemple de calcul : coupe minimale normalisée + taille de la partition fixée (Barzilay & Malioutov)

Lissage

- méthodes basées sur très peu d'occurrences
- beaucoup de variances
- nécessité d'un lissage des variations pour repérer des tendances significatives
- méthodes
  - vue linéaire : moyenne d'un point avec contexte avant et après (convolution)
  - vue matricielle : moyenne d'une zone entourant le point de décision
  - alternative : agrandir les segments considérés

Lissage en dimension 1

Pour chaque point d'abscisse x_i, on fait une moyenne pondérée sur le contexte (phrase avant, phrase courante, phrase après) = produit de convolution
$val(x_i) = \text{similarite}(i, i-1)$

$$val'(x_i) = \sum_{k \in [-1,+1]} w_k * val(x_{i+k})$$

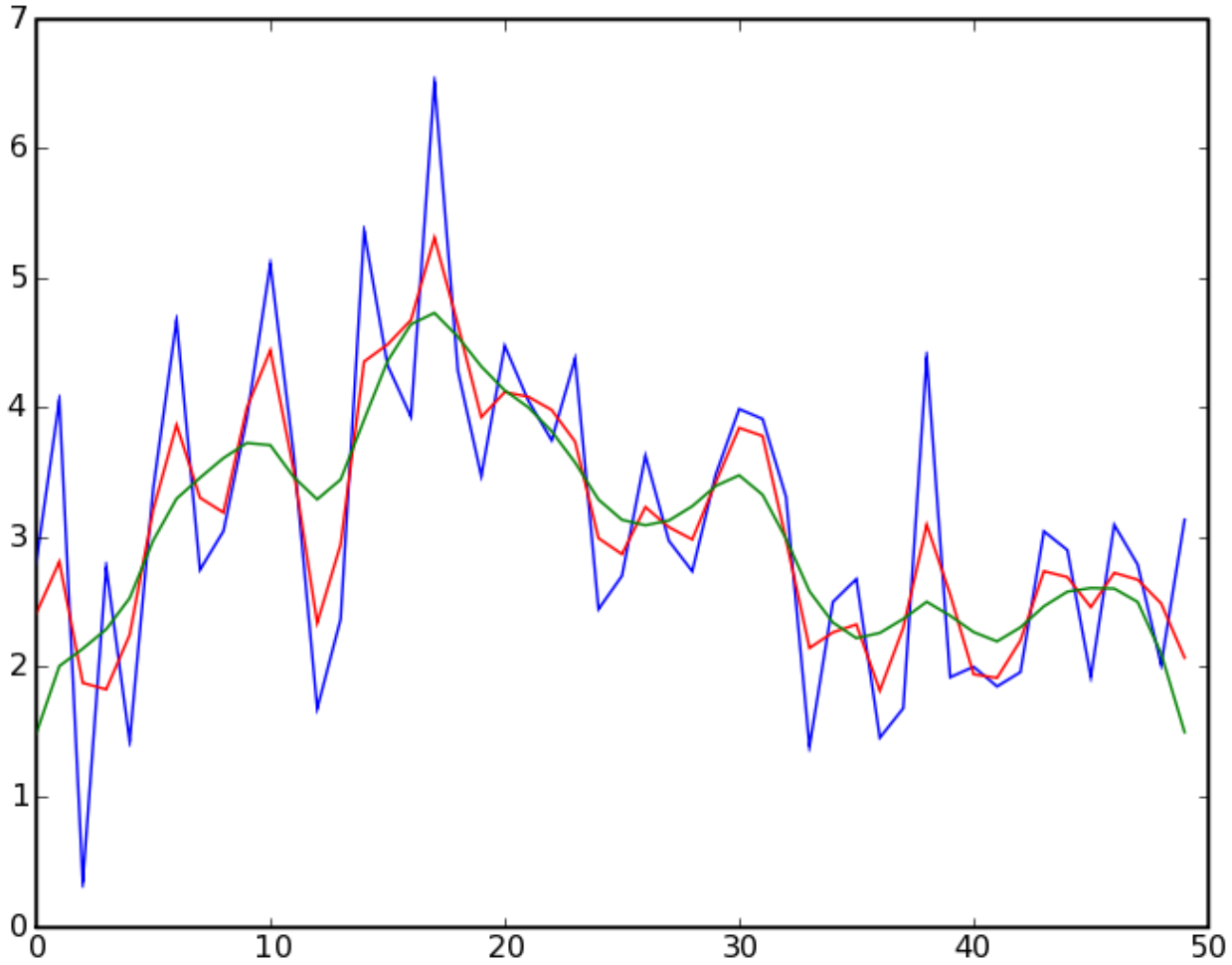
avec

$$\sum_{k \in [-1,+1]} w_k = 1$$

processus **itérable**
on peut aussi élargir à $k \in [-n, n]$ pour un n quelconque

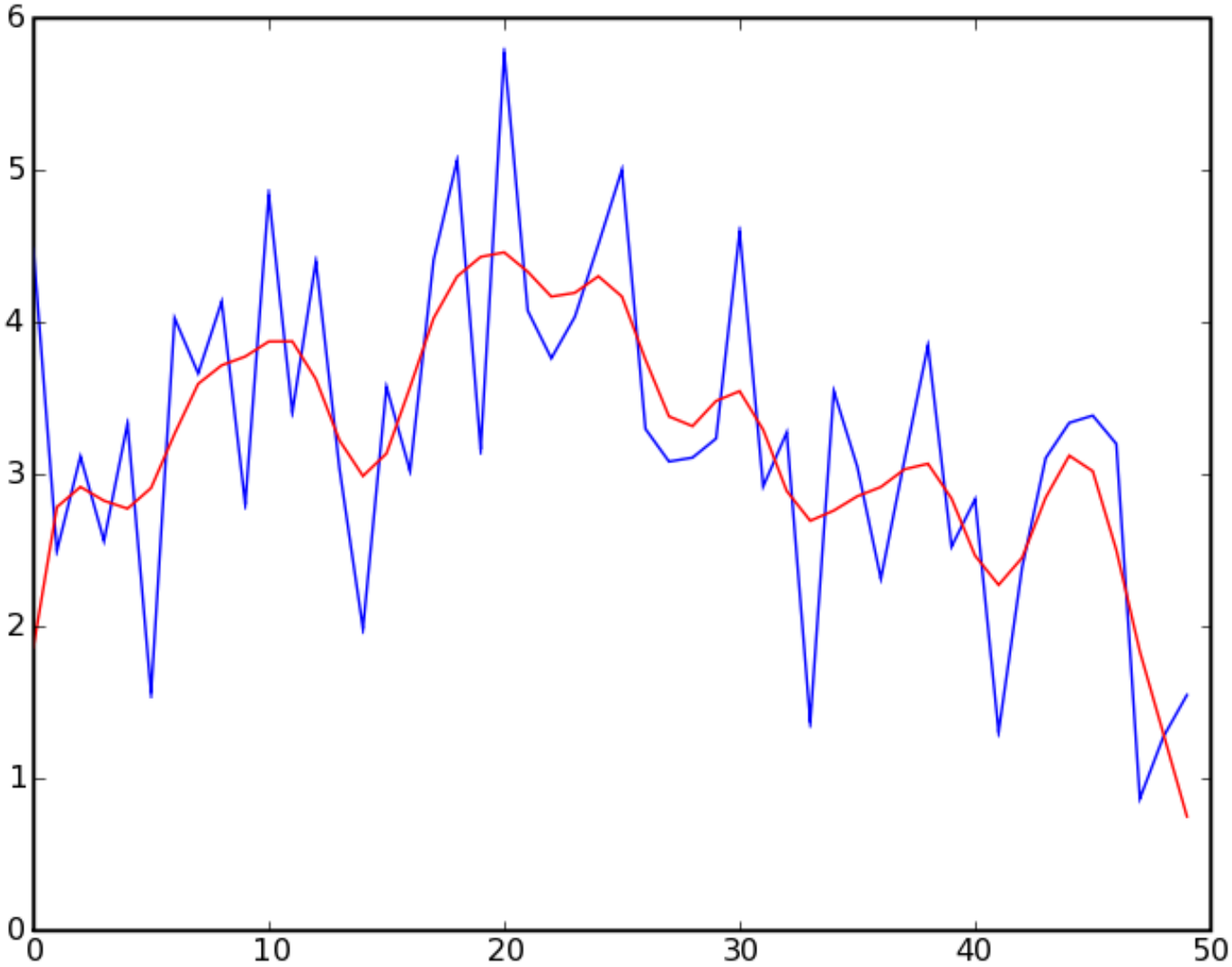
Exemple

bleu = origine ; (rouge=3 phrases ou vert=5 phrases) de contexte

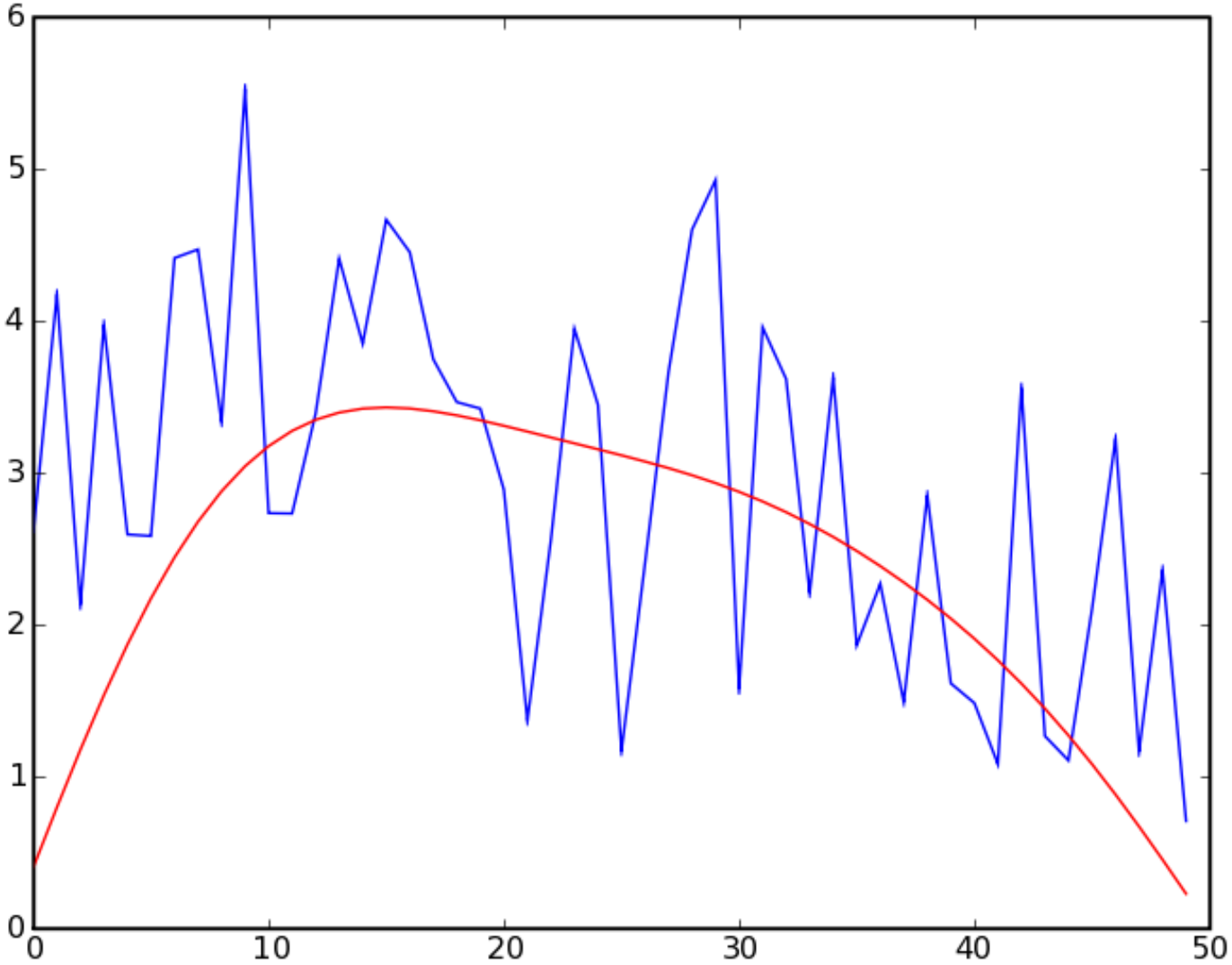


Exemple : lissage répété

contexte : 1 phrase autour + itération



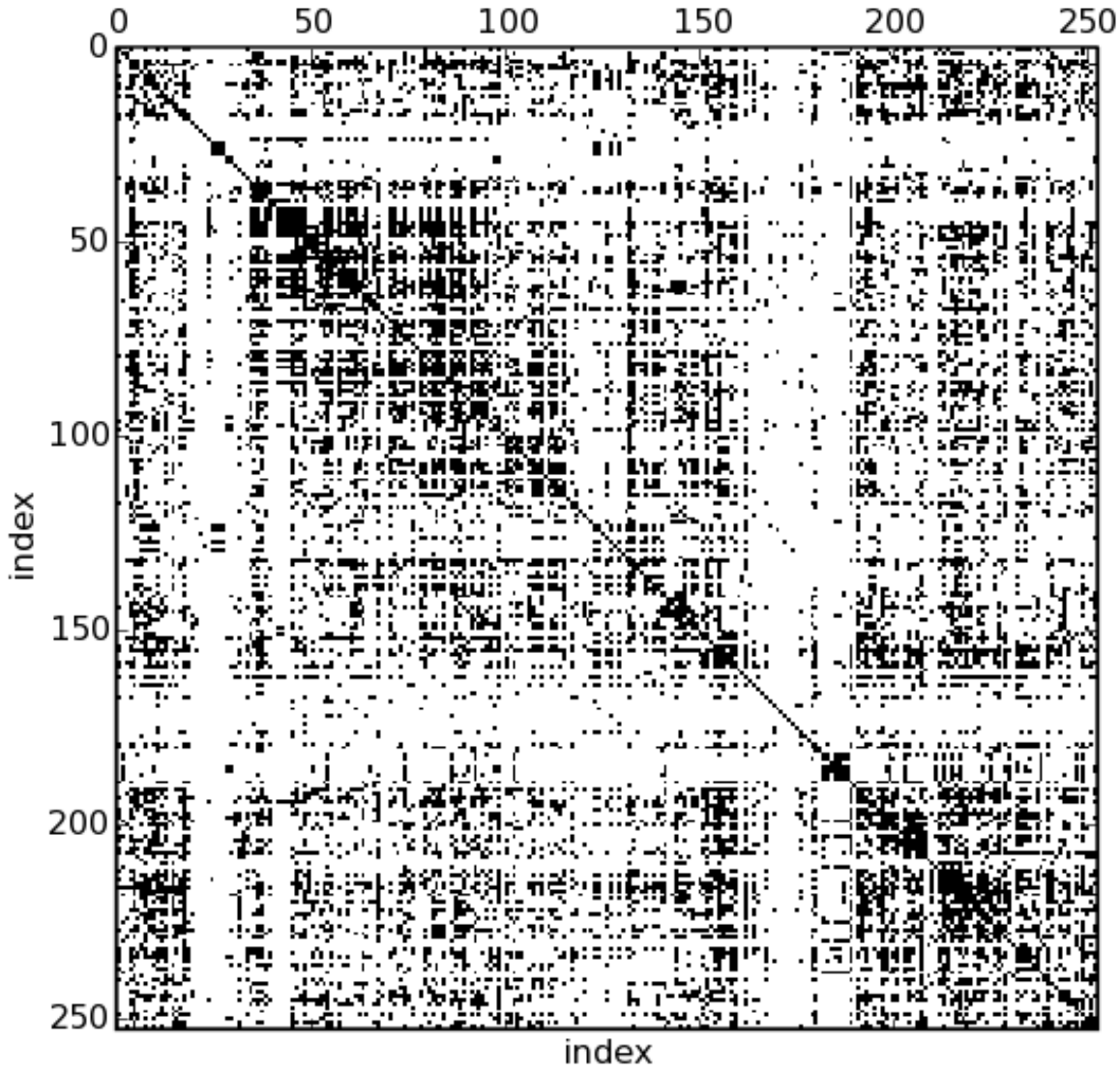
Exemple: trop de lissage tue le lissage



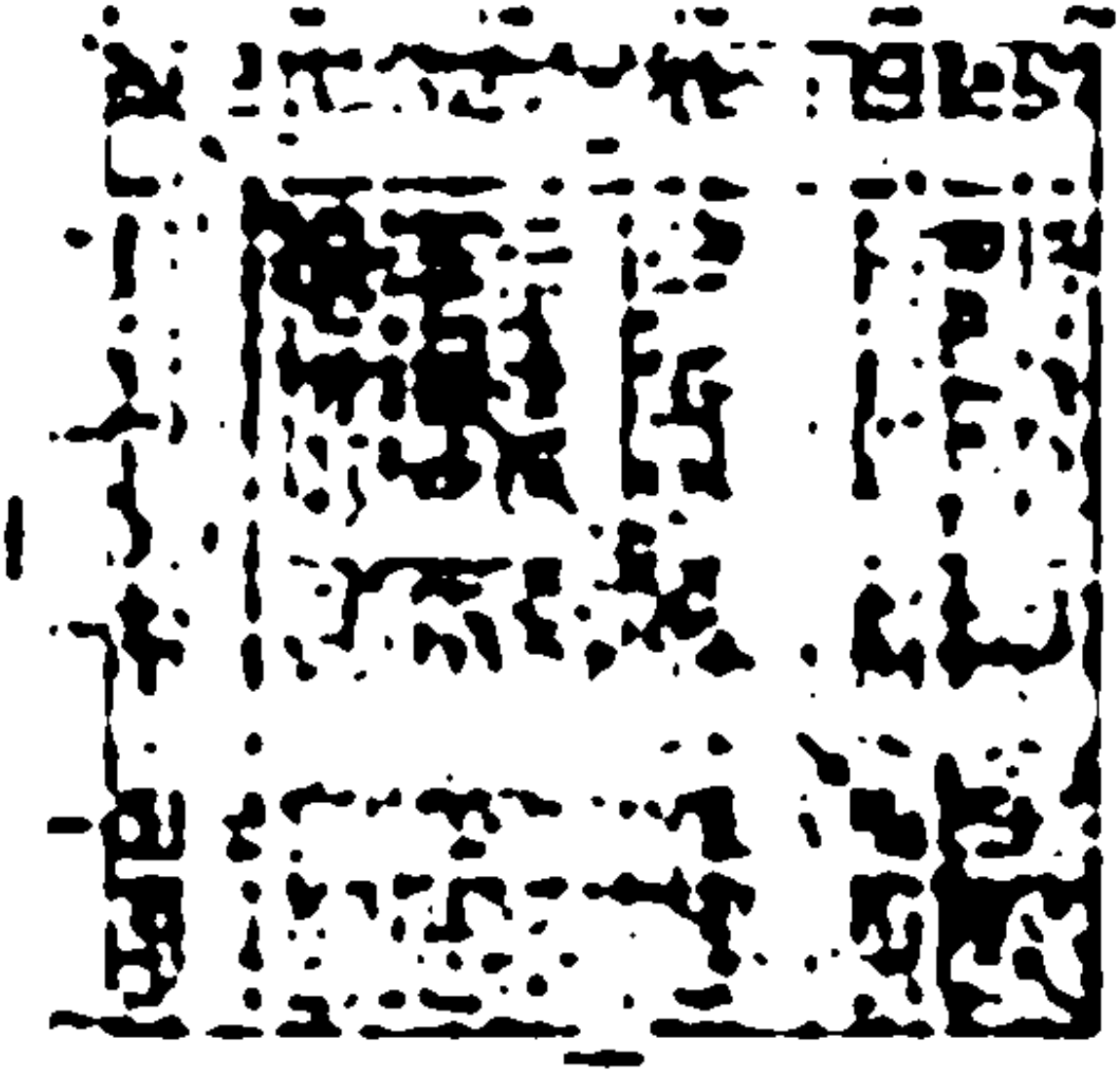
Lissage en deux dimensions

- Contexte = non plus seulement la similarité avec la phrase précédente, mais aussi les similarités à plus longue distance
- Par exemple : matrice 3x3 autour de la phrase courante i incluant donc
  similarité(i,i-1), similarité(i-1,i-2), similarité(i+1,i)
  similarité(i-1,i+1)
  auto-similarité i, i-1, i+1 ?
- variantes possibles (dérivées locales, taille variable du contexte, etc)

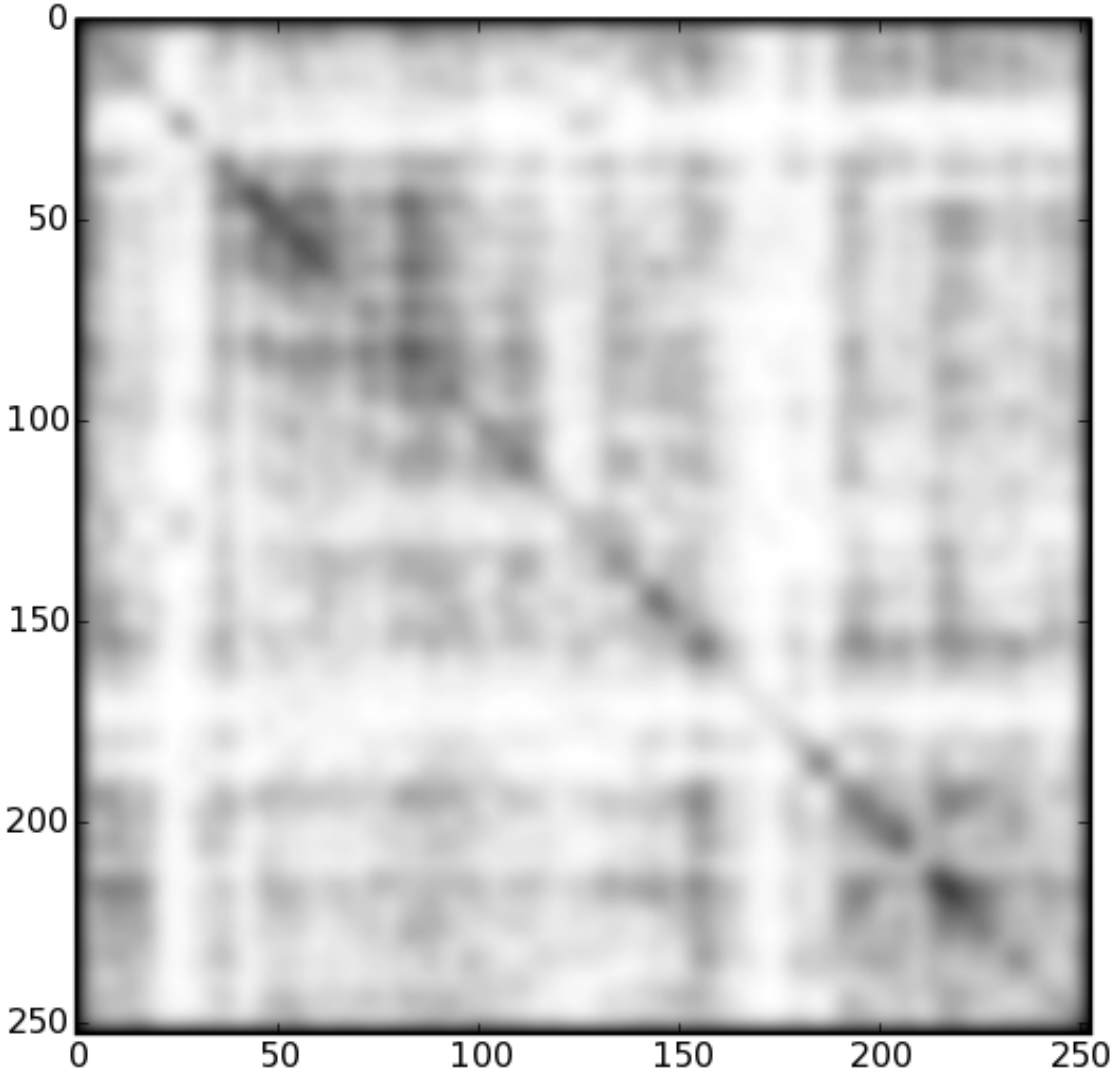
Lissage en 2D : exemple (1)



Lissage en 2D : exemple (2)



Lissage en 2D : exemple (2)



Evaluations

- zones de rupture vs marquage typographie (titre, paragraphe)
- zones de rupture vs marquage manuel

mesures de comparaison de segmentation

- précision/rappel des frontières (taux de frontières correctes, taux de frontières trouvées)
- P_k : nb de fois où deux mots pris au hasard à une distance k sont dans le même segment (ou non) à la fois dans la référence et l'hypothèse

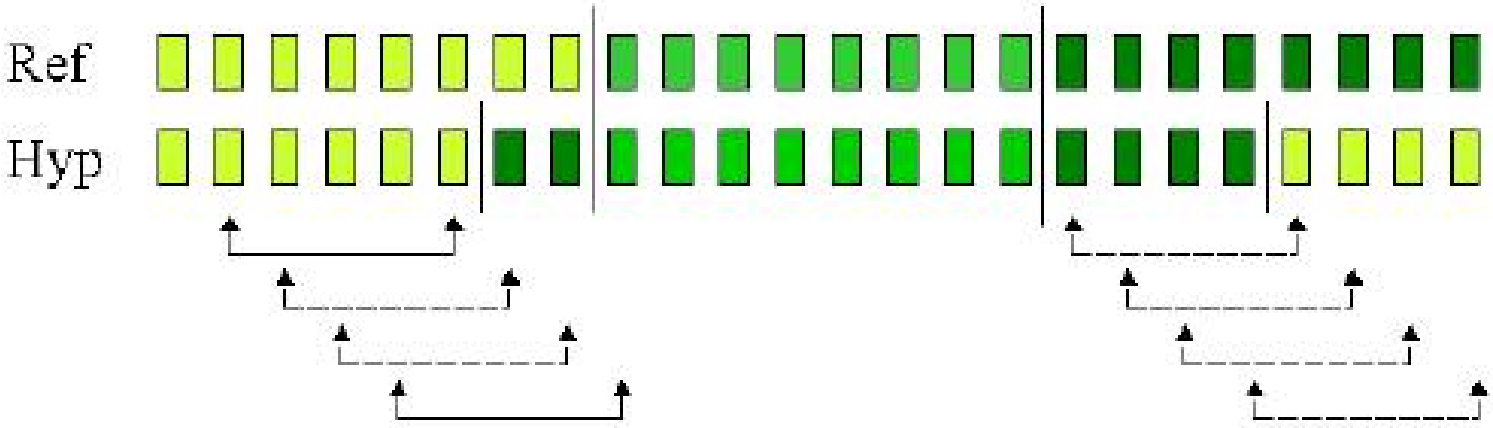
$$P_k = \frac{1}{n-k} \sum_{i \in [0, n-k]} (\delta_{ref}(i, i+k) = \delta_{hyp}(i, i+k))$$

- variante: Windowdiff = différences du nb de bornes dans une fenêtre glissante
- on pourrait en imaginer d'autres : distance d'édition, fonctions d'alignement

Illustration

P_k

fenêtre glissante, trait droit = correct



Problèmes de ces approches

- peu de données
- nombre de sujets souvent présupposé
- ou bien benchmarks artificiels (textes concaténés)
- la plupart des approches prennent des décisions très locales
- tâche de segmentation mal définie : accord inter-annotateurs assez mauvais

References