Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales

Romain Brixtel[1], Charlotte Lecluze[2], Gaël Lejeune[3]

(1) HEC - Dpt. de comportement organisationnel, Université de Lausanne

(2) GREYC, Université de Caen

(3) LINA, Université de Nantes

prenom.nom@unil.ch, unicaen.fr, univ-nantes.fr

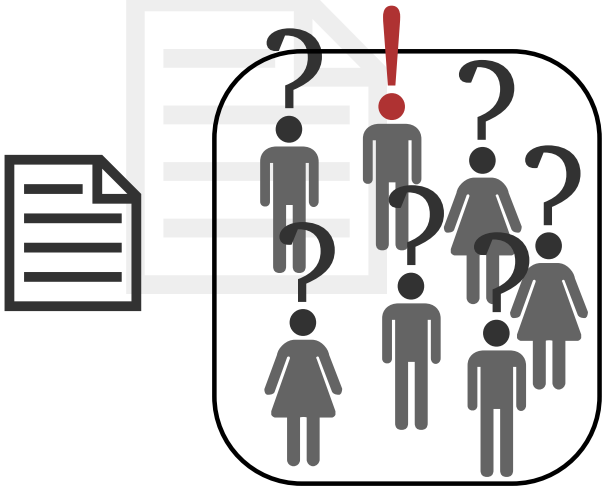
23 *Juin* 2015

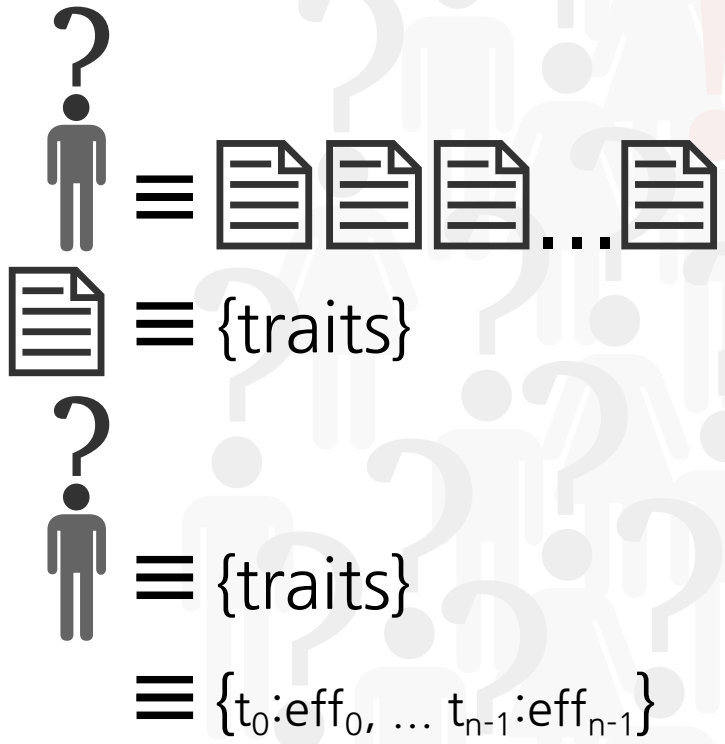
Attribution d'Auteur

Définition

Caractériser un auteur en tant qu'individu à partir de ses écrits

Monde clos





Étude contrastive
Détection de Doppelgänger

Attribution d'Auteur

Définition : état de l'art

Utilisation des traits en contexte multilingue

- Apprentissage automatique : SVM
- Multilinguisme : n-grammes de caractères

Mhmm... n-grammes de caractères...

- ... n ?
- traits redondants, non ?

Proposition :

Répétitions maximales

Répétitions maximales (motifs)

Définition

Les répétitions maximales (*motifs* dans (Ukkonen,2009)) sont des chaînes de caractères :

- **repétées** : apparaissent deux fois ou plus.
- **maximales** : pas inclues dans une autre chaîne de même effectif.

Répétitions maximales (motifs)

Exemples

motifs(A^n)

{A, AA, AAA, AAA, ..., A^{n-1}}

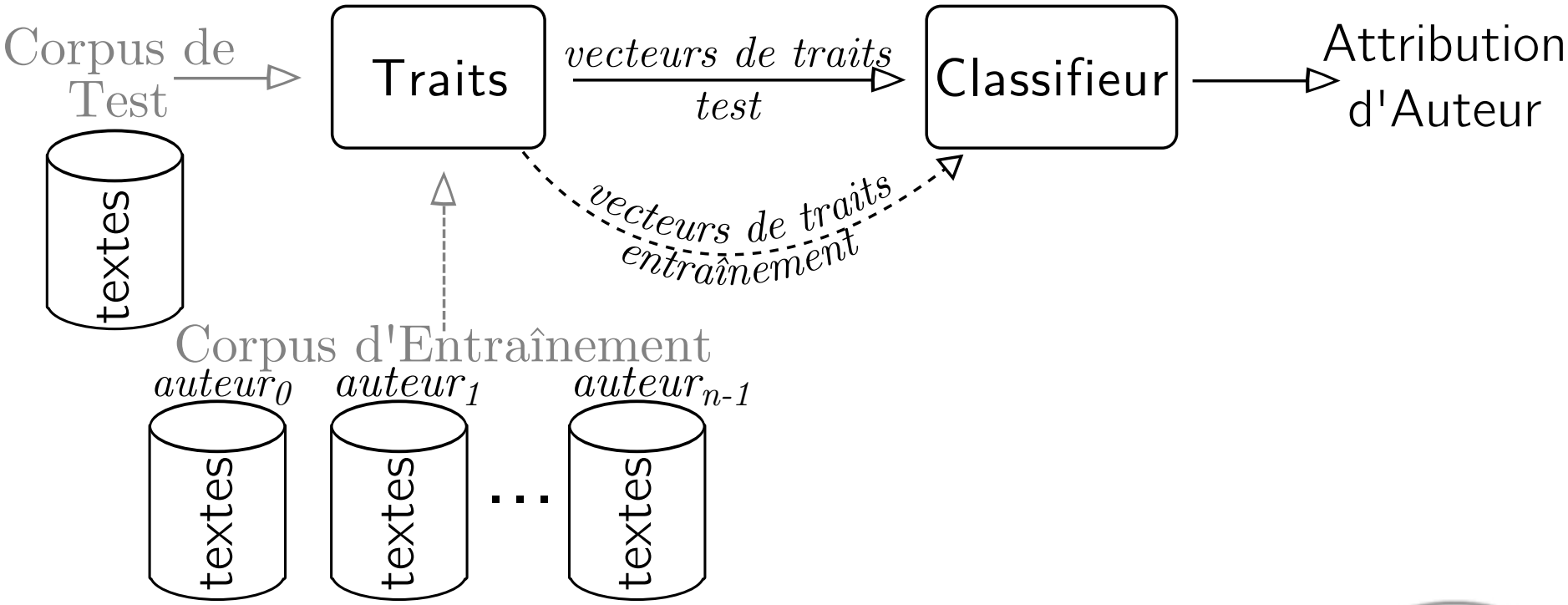
motifs(HATTIVATTIAA)

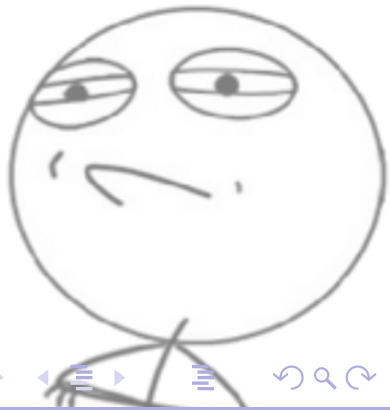
{T, A, ATTI}
répétées : ni H ni V (hapaxes)
maximales : pas TT (inclu dans ATTI)

motifs(HATTIV, ATTIAA)

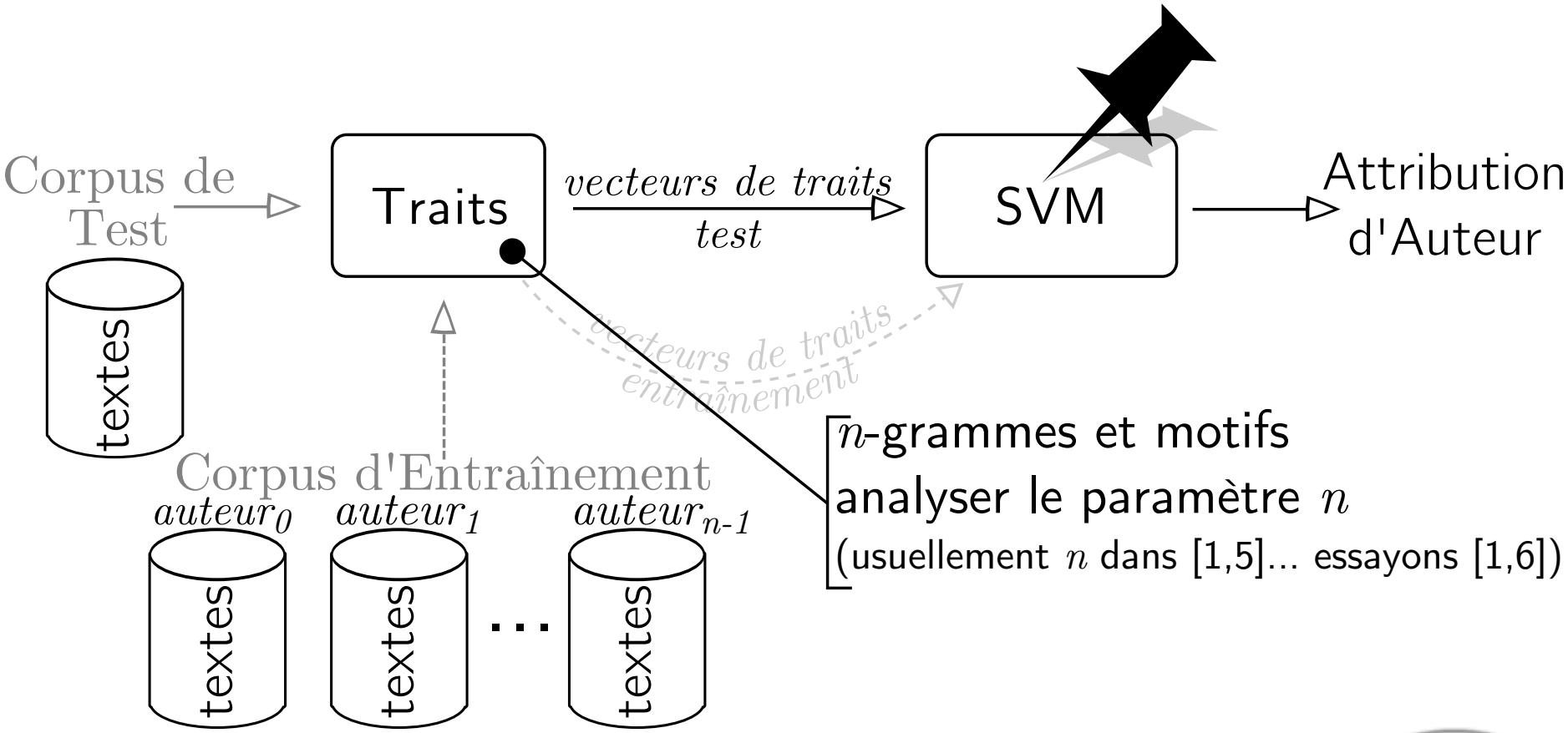
$\equiv$ *motifs*(HATTIV$\$_0$ATTIAA$\$_1$)
$\not\equiv$ *motifs*(HATTIV) $\cup$ *motifs*(ATTIAA)

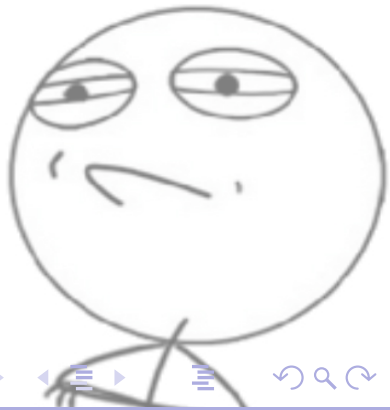
Chaîne de traitement expérimentale





Chaîne de traitement expérimentale





Corpus

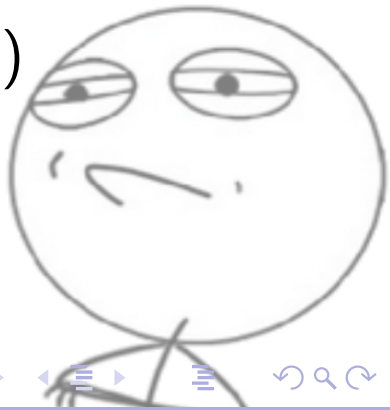
EBG, LIB et MIXT

	#caractères	#textes	#auteurs
corpus	$1.9\ 10^6$	631	40
auteurs (*moy.* $\pm$ *ecart*)	$4.6\ 10^4 \pm 8075$	15.8 ± 2.6	
textes (*moy.* $\pm$ *ecart*)	2945.1 ± 178.5		

Table : Caractéristiques du corpus EBG (anglais)

	#caractères	#textes	#auteurs
corpus	$5.1\ 10^6$	1247	40
auteurs (*moy.* $\pm$ *ecart*)	$1.3\ 10^5 \pm 2.6\ 10^4$	31.2 ± 4.2	
textes (*moy.* $\pm$ *ecart*)	4070.6 ± 1524.2		

Table : Caractéristiques du corpus LIB (français)



MIXT = EBG + LIB

Répétitions maximales et n-grammes

Évaluation

Classifieur : SVM, Noyau linéaire, régularisation C=1.0

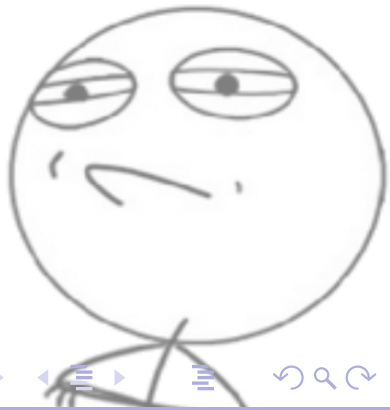
Traits : n-grammes et motifs, longueurs $\in [1, 6]$

Corpus :

- EBG (40 auteurs, anglais)
- LIB (40 auteurs, français)
- MIXT (80 auteurs, anglais et français)

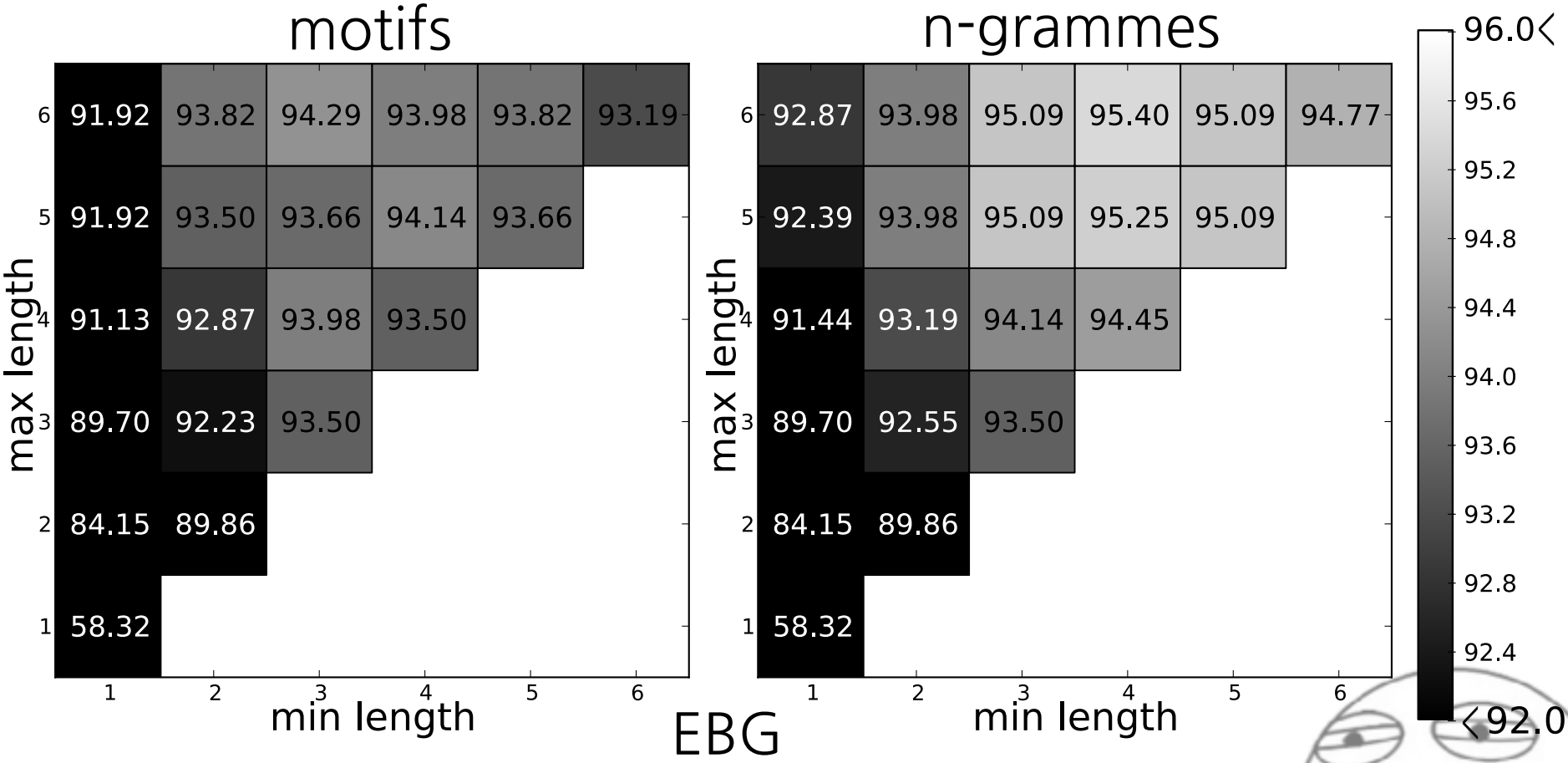
Validation croisée : stratified 10-fold

Score : $\frac{\#\text{couples (auteur,texte) correctement détectés}}{\#\text{couples (auteur,texte)}}$



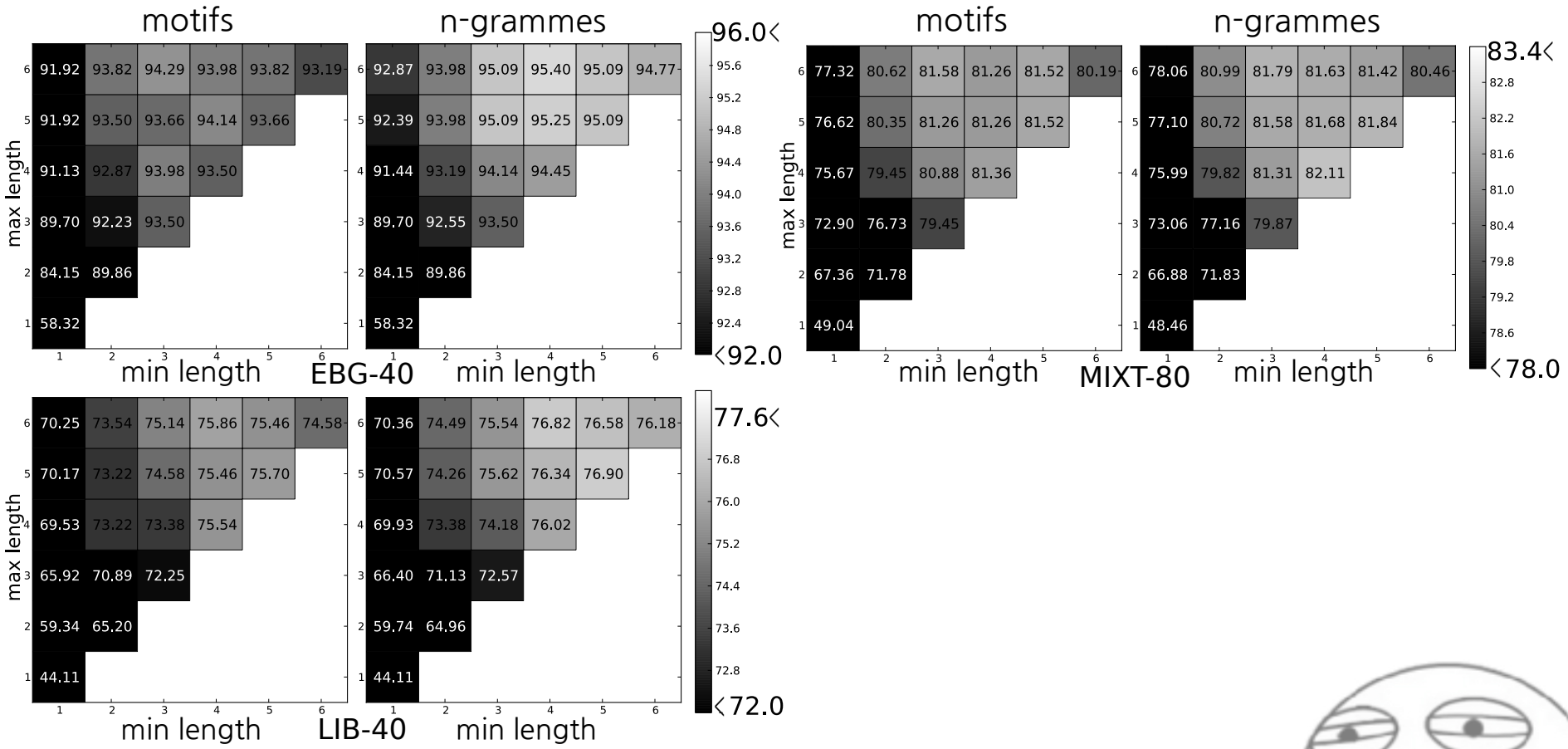
Motifs et n-grammes

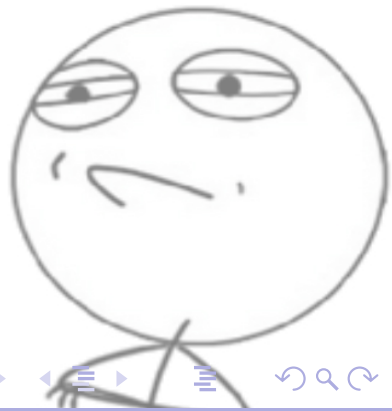
Premiers résultats...



Motifs et n-grammes

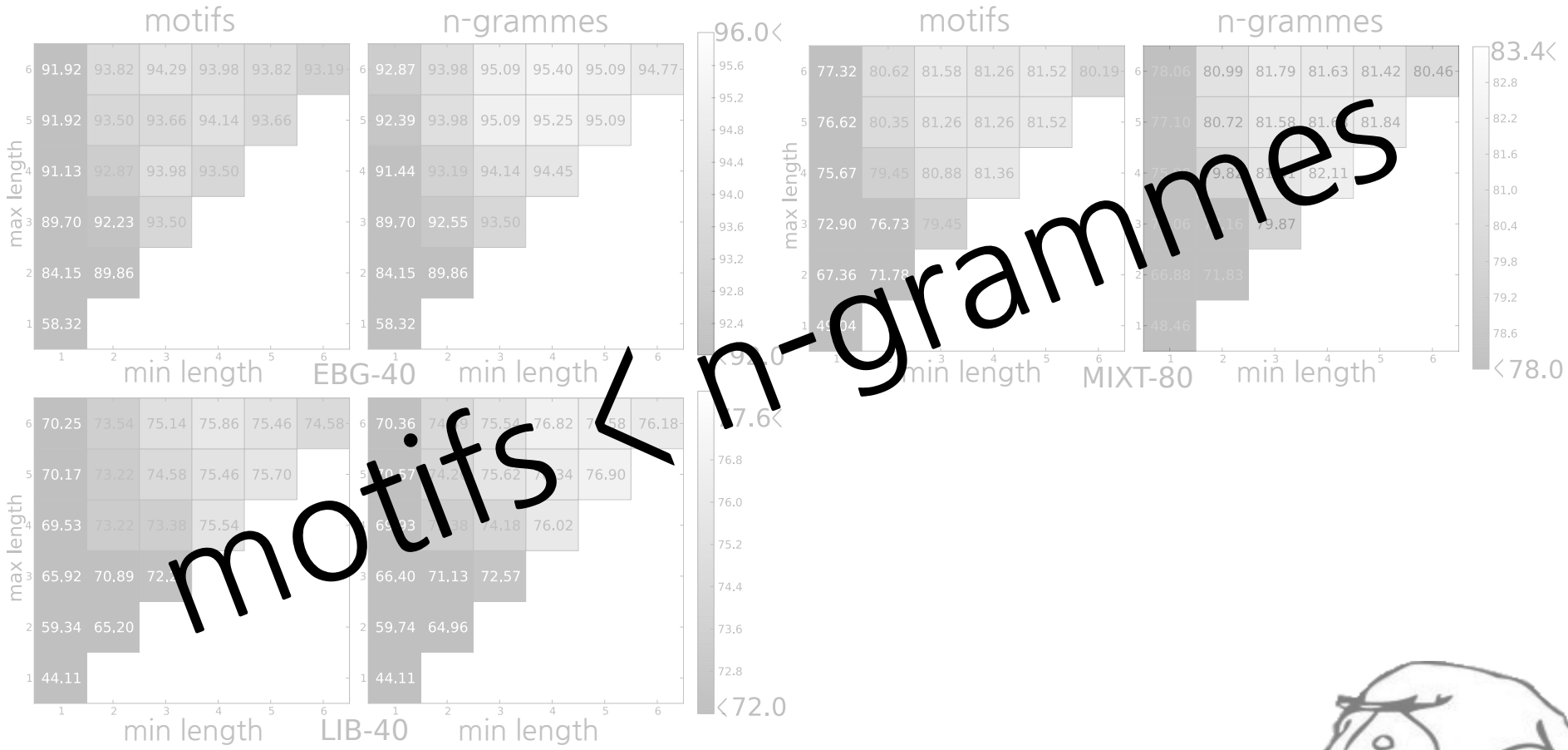
Premiers résultats...





Motifs et n-grammes

Premiers résultats... et premières déconvenues



motifs < n-grammes

Motifs et n-grammes

Différences

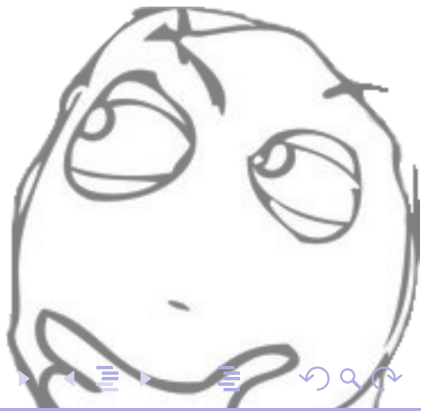
Pour des longueurs comprises entre $[min, max]$:

- les motifs constituent un sous-ensemble des n-grammes
- les n-grammes « représentent » des chaînes de longueur $> max$

(min, max)-grammes :

motifs
+ répétitions hors motifs
+ hapaxes

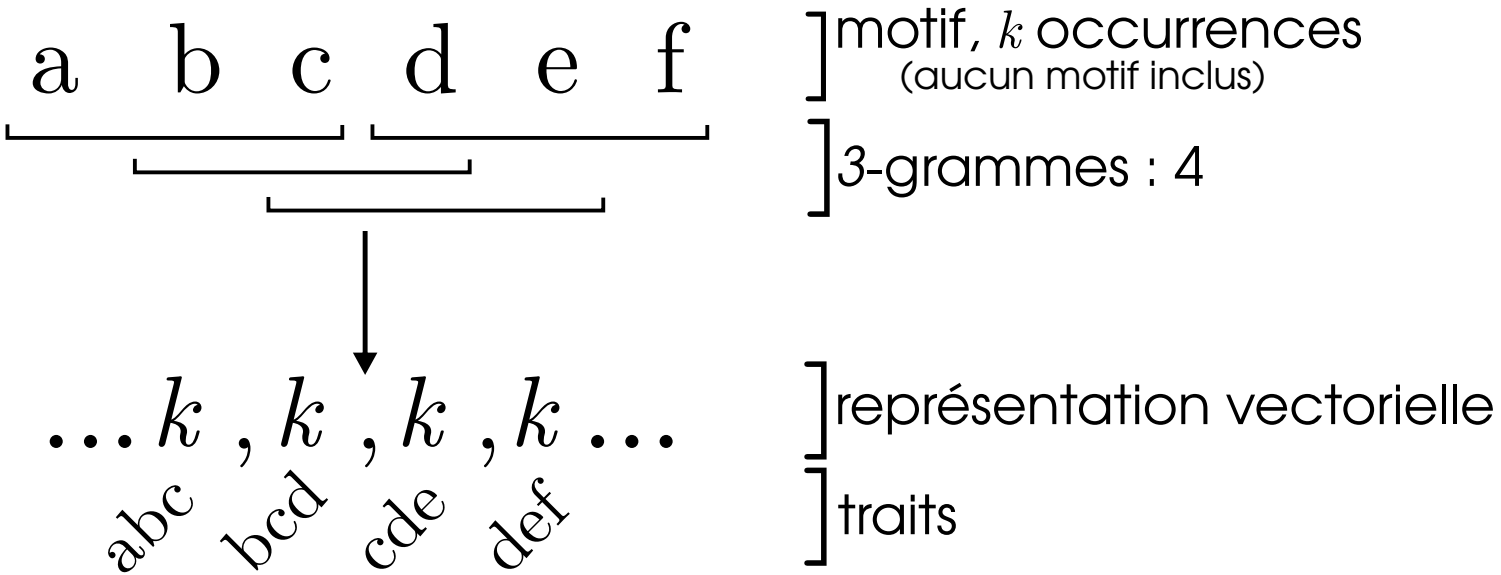
] longueur $n \in$ (min, max)

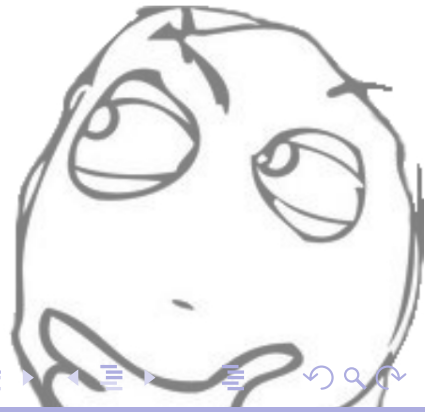


Motifs et n-grammes

Différences

- les motifs constituent un sous-ensemble des n-grammes
- les n-grammes « représentent » des chaînes de longueur $>$ max



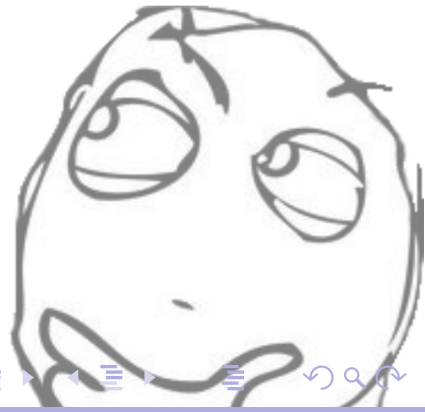


Motifs et n-grammes

Exploitation des différences

La taille compte : deux traits de même longueur ne sont pas pour autant d'égale importance

Les chaînes sont affectées par leurs sous-chaînes : les n-grammes « représentent » des chaînes de taille $> n$



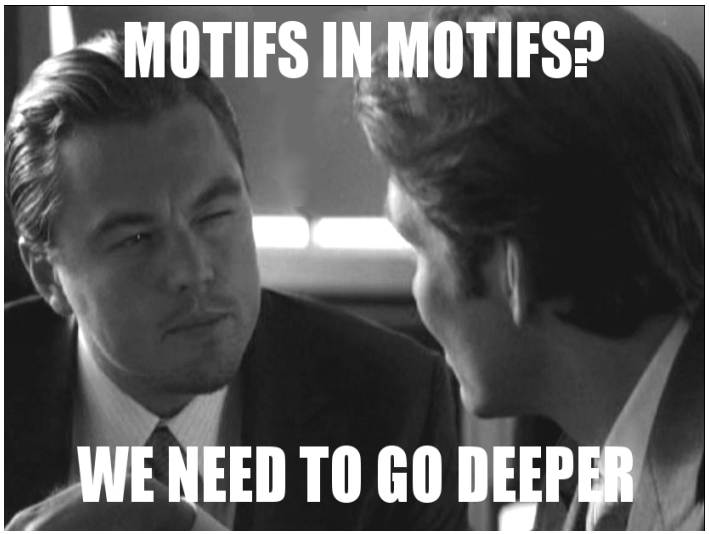
Motifs et n-grammes

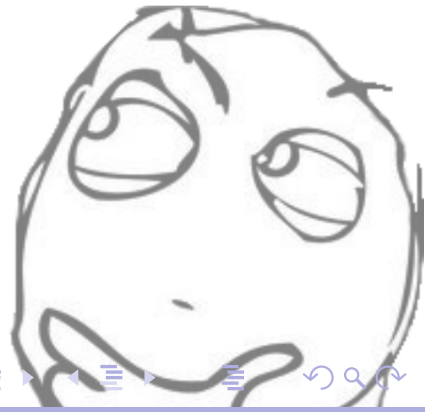
Exploitation des différences

La taille compte : deux traits de même longueur ne sont pas pour autant d'égale importance

Les chaînes sont affectées par leurs sous-chaînes : les n-grammes « représentent » des chaînes de taille $> n$

Motifs de $2^{\text{ème}}$ ordre



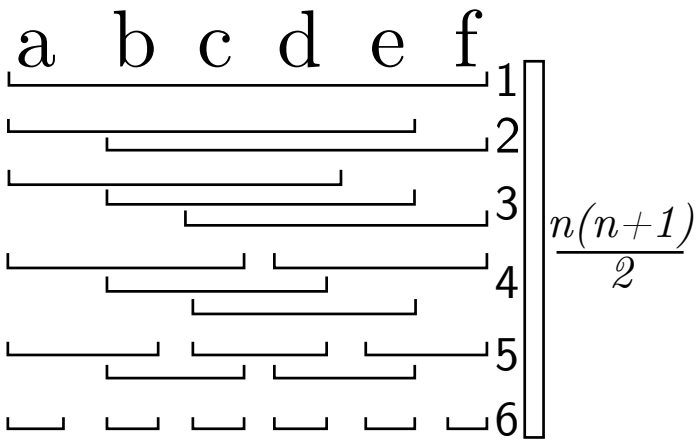


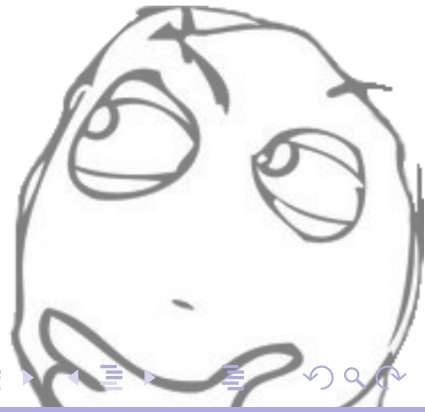
Motifs

pondérés par des motifs du 2ème ordre

$w_{2^{nd}\,order}$(trait) = #sous-chaînes potentielles - #motifs du 2ème ordre

sous-chaînes potentielles :



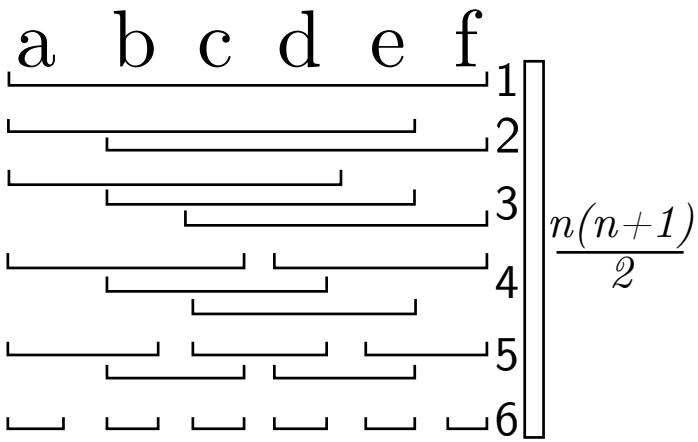


Motifs

pondérés par des motifs du 2ème ordre

$w_{2nd\,order}$(trait) = #sous-chaînes potentielles - #motifs du 2ème ordre

sous-chaînes potentielles :



sous-motifs (motifs du 2ème ordre) :

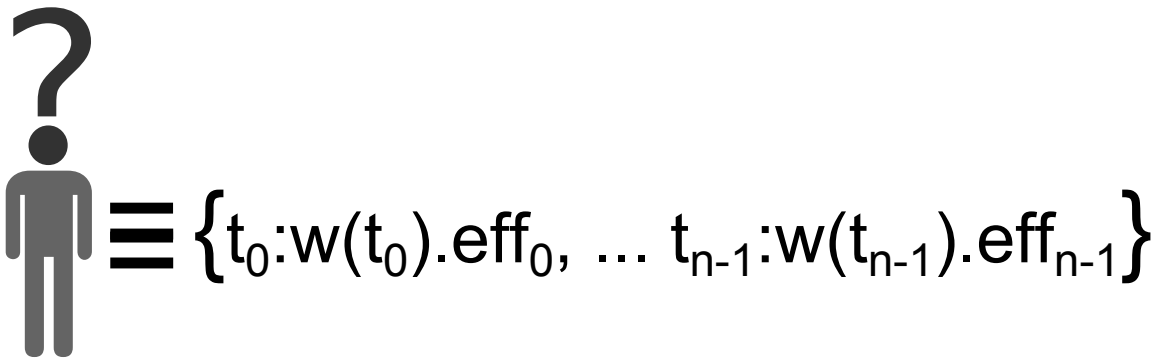
motifs (📄,📄,📄...📄) = $\{m_0, m_1, \ldots m_{n-1}\}$

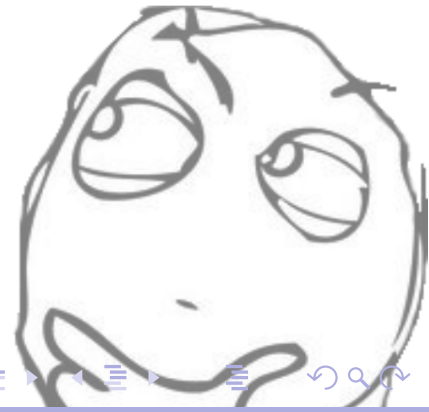
motifs $(m_0, m_1, m_2 \ldots m_{n-1}) = \{r_0, r_1, \ldots r_{p-1}\}$

("abcdef" est un des m_i)

$\{r_0, r_1, \ldots r_{p-1}\}$ un sous-ensemble de $\{m_0, m_1, \ldots m_{n-1}\}$

$p < n$





Motifs (2ème ordre) et n-grammes

Évaluation

Classifieur : SVM, Noyau linéaire, régularisation C=1.0

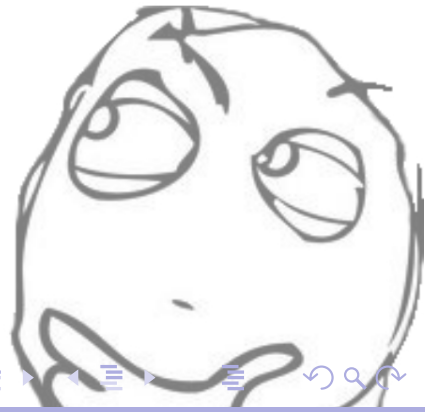
Traits : n-grammes, motifs (motifs du 2ème ordre, longueur), longueur $\in [1, 6]$

Corpus :

- EBG (40 auteurs, anglais)
- LIB (40 auteurs, français)
- MIXT (80 auteurs, anglais et français)

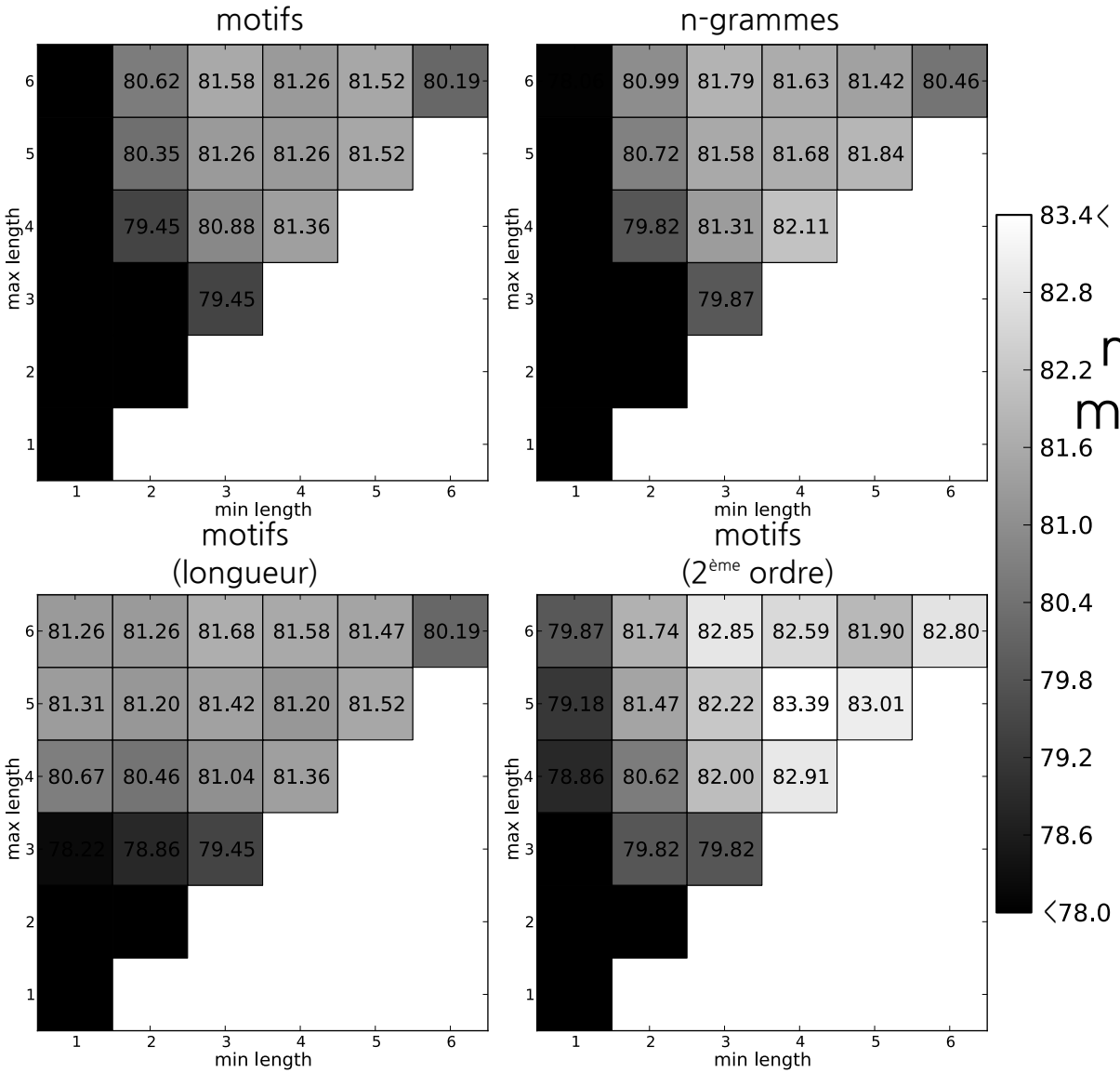
Validation croisée : stratified 10-fold

Score : $\frac{\#\text{couples (auteur,texte) correctement détectés}}{\#\text{couples (auteur,texte)}}$



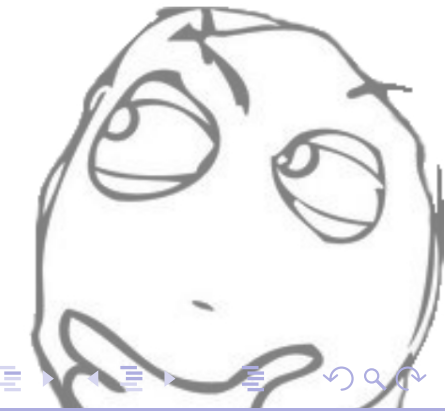
Motifs (2ème ordre) et n-grammes

Seconde manche (MIXT)



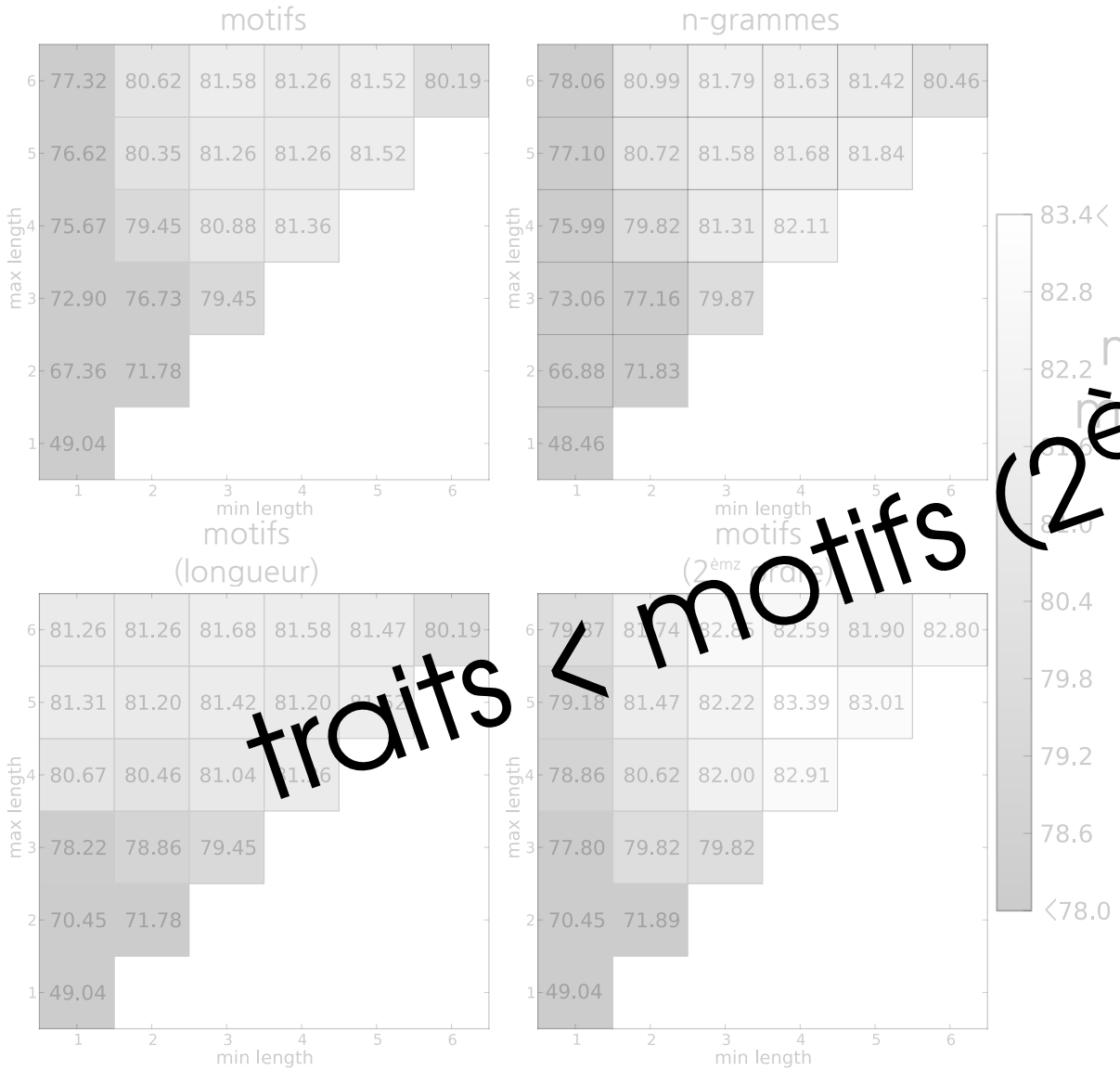
score moyen EBG / LIB / MIXT

n-grammes	(4,6)	84.61%
motifs	(4,6)	83.69%
motifs (longueur)	(4,6)	83.88%
motifs (2ème ordre)	**(4,5)**	**85.96%**



Motifs (2ème ordre) et n-grammes

Seconde manche (MIXT)



score moyen EBG / LIB / MIXT

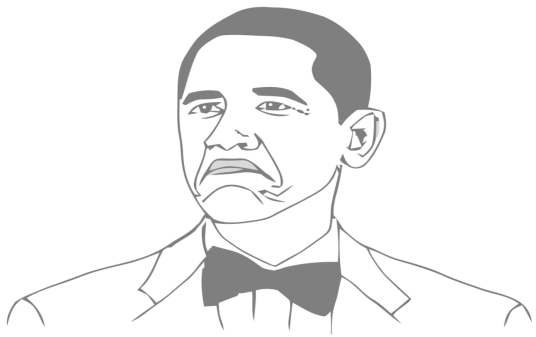
n-grammes (4,6) 84.61%

motifs (4,6) 83.69%

motifs (longueur) (4,6) 83.88%

motifs (2ème ordre) **(4,5) 85.96%**

traits < motifs (2ème ordre)



n-GRAMMES

PONDÉRÉS PAR DES MOTIFS DU 2ÈME ORDRE

Pas immédiat...

Les chevauchements entre n-grammes consécutifs créent naturellement des sous-motifs.

$3\text{-}grammes(\texttt{abcdef}) = \{\texttt{abc}, \texttt{bcd}, \texttt{cde}, \texttt{def}\}$

$motifs(\texttt{abc}, \texttt{bcd}, \texttt{cde}, \texttt{def}) = \{\texttt{c}, \texttt{d}, \texttt{de}, \texttt{bc}, \texttt{cd}\}$

Motifs (2ème ordre) et n-grammes

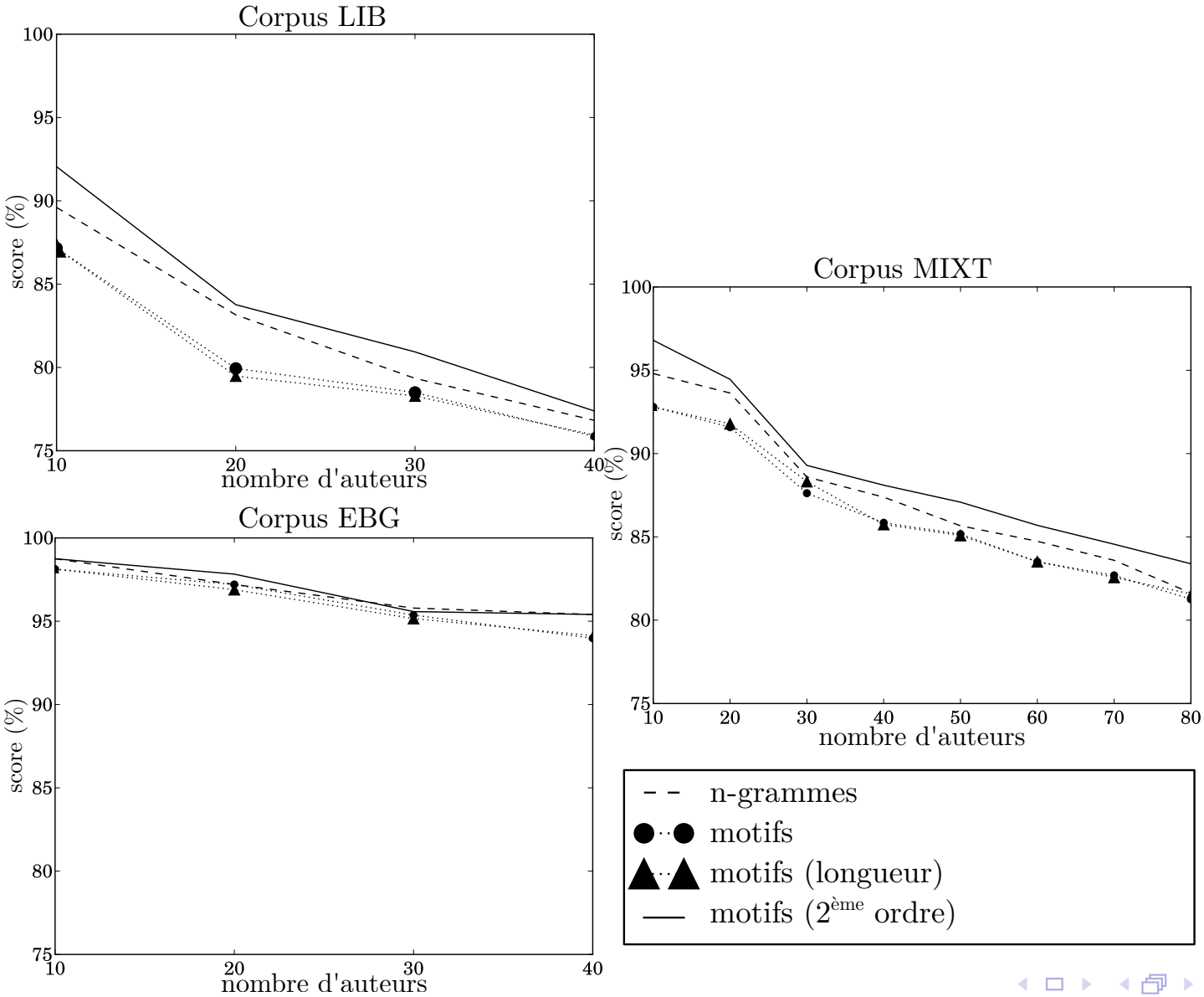
Résultats

Évaluation en fonction du nombre d'auteurs

- le meilleur score est le plus élevé -

Motifs (2ème ordre) et n-grammes

Résultats



Motifs (2ème ordre) et n-grammes

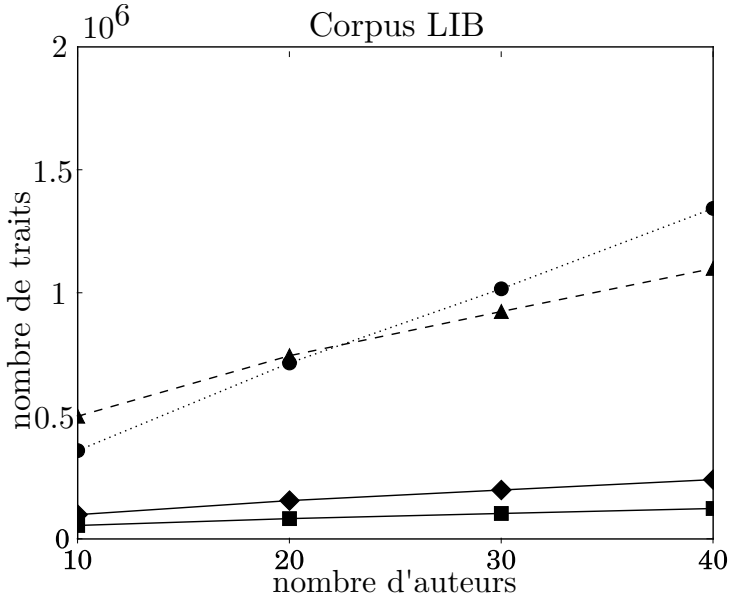
Résultats

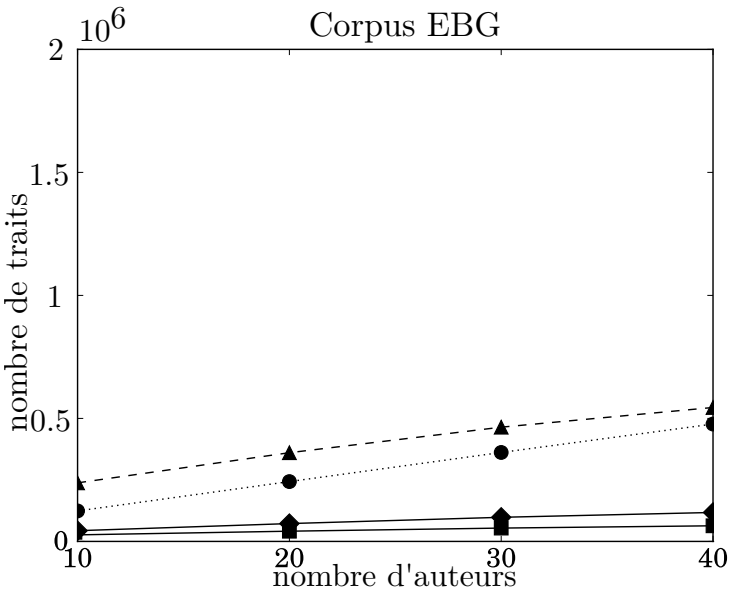
Taille de l'espace de recherche décrit par les traits

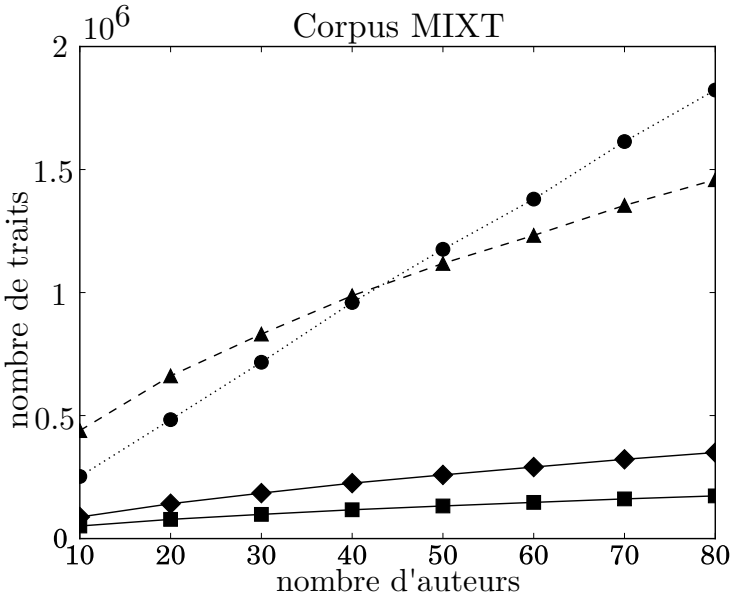
- le meilleur score est le plus bas -

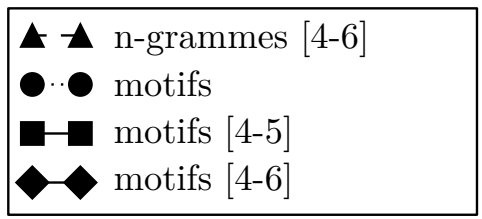
Motifs (2ème ordre) et n-grammes

Évolution de la taille de l'espace de recherche









Motifs (2ème ordre) et n-grammes

Perspectives multilingues

Quel trait pour un corpus multilingue ?

(LIB & EBG) et (LIB & EBG à partir de MIXT)

- le meilleur score est le plus élevé -

Motifs ($2^{\text{ème}}$ ordre) et n-grammes

Perspectives multilingues

n-grammes de longueur $[4, 6]$

nb. d'auteurs	EBG	EBG de MIXT	LIB	LIB de MIXT
10	**98.75%**	**98.75%**	89.60%	**91.13%**
20	**97.20%**	96.89%	**83.15%**	82.69%
30	**95.79%**	94.85%	**79.34%**	78.65%
40	**95.40%**	94.10%	**76.82%**	75.03%

Motifs de longueur $[4, 5]$
pondérés par des motifs du $2^{\text{ème}}$ ordre

nb. d'auteurs	EBG	EBG de MIXT	LIB	LIB de MIXT
10	**98.75%**	**98.75%**	92.01%	**92.35%**
20	**97.83%**	97.52%	**83.77%**	83.46%
30	95.59%	**96.84%**	**80.93%**	80.08%
40	**95.40%**	95.09%	77.38%	**77.47%**

Table : Scores sur LIB et EBG seuls et à partir de MIXT.

Motifs et Attribution d'Auteur

Conclusion

Motifs et *n*-grammes.

Mise en évidence de l'effet positif de la redondance des *n*-grammes

Pondération qui en découle adaptée aux motifs

Pour l'Attribution d'Auteur... et certainement ailleurs.

questions ?