Introduction à l'extraction de séquences pertinentes

felipe@iro.umontreal.ca

RALI
Dept. Informatique et Recherche Opérationnelle
Université de **Montréal**

mars 2016, V0.1

Plan

- Motivation
- Suffix Arrays
- Métriques
- Applications monolingues
  - AL
  - Dagan
  - Thesaurus
- Applications bilingues
  - Trad
  - Noisy
  - Comp
- Sélection d'études récentes
  - Sangati & van Cranenburgh, MWE 2015
  - Rondon & al, MWE 2015
  - Huet et Langlais, 2010

rali

Plan

rali

Phrasème / Multi Word Expressions

- *[Jackendoff, 1997]* : autant de MWE que de mots dans notre lexique
- Dans WordNet, plus de 40% des entrées sont des MWE
- Profusion dans les domaines spécialisés (termes)
  - deep neural network
  - monoxyde de phosphore
- Lire *[Mel'cuk, 2011]* pour une typologie des phrasèmes

rali

See this image

Grand Druide des cooccurrences (French) Hardcover – Aug 21 2012

by Collectif (Author)

★★★★★ 5 customer reviews

See all formats and editions

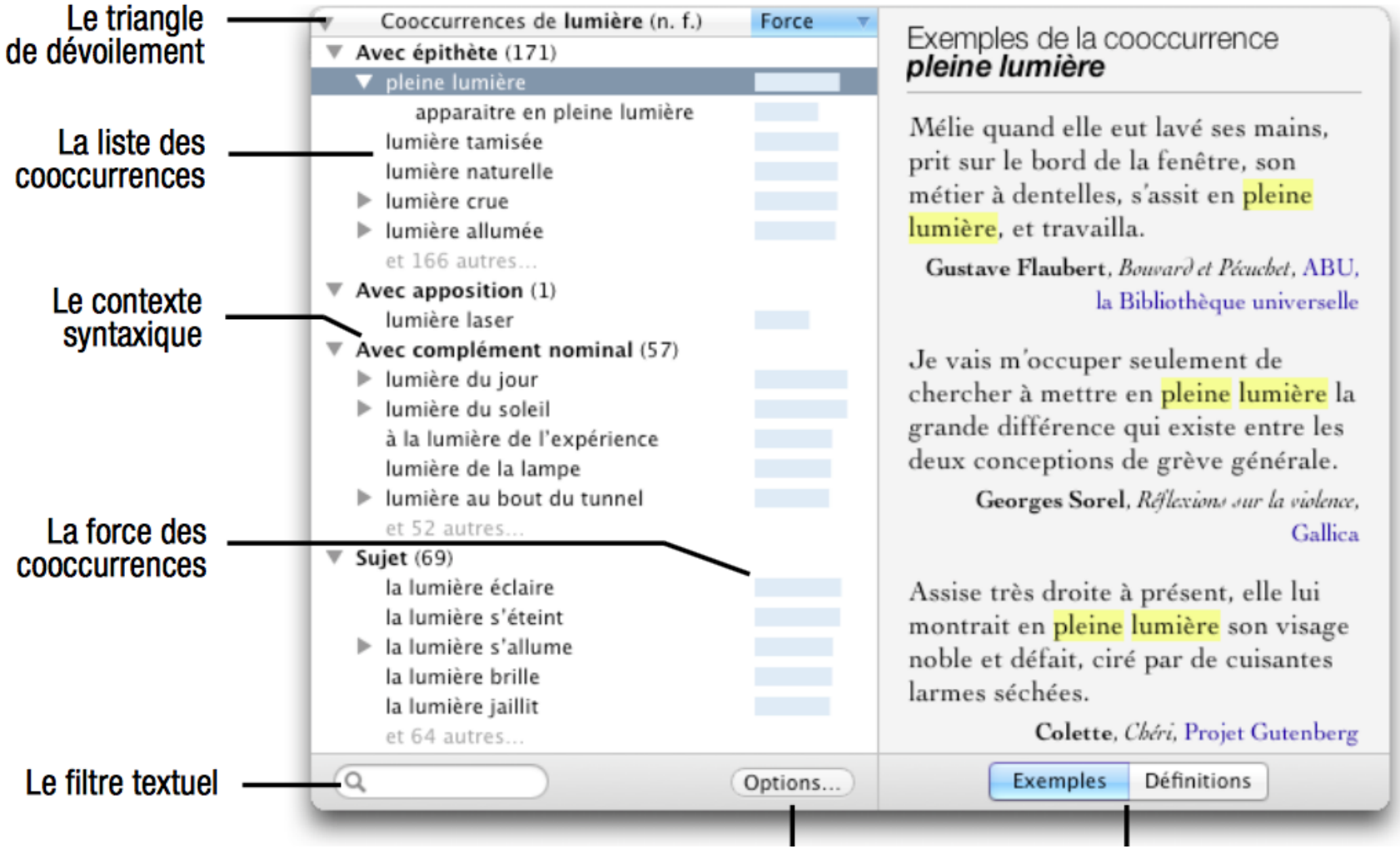
Hardcover
CDN$ 44.96

2 Used from CDN$ 198.27
5 New from CDN$ 36.00

Avec 10 000 entrées et près d'un million de coocurrences, le Grand Druide des coocurrences offre l'image la plus complète et la plus précise du français sous son angle associatif. Quel adverbe décrit le mieux l'intensité de l'amour? Quel verbe exprime le mieux la satisfaction d'une passion? Quel adjectif qualifie le mieux l'abondance d'un trésor? Grâce à des algorithmes d'analyse avancés appliqués à un corpus de 92 millions de phrases, Druide a colligé un étonnant répertoire de combinaisons de mots qui répond à ces questions familières aux habitués de l'écriture. Au rédacteur éperdument amoureux des mots qui veut assouvir sa passion de l'expression juste, le Grand Druide des coocurrences offre un inépuisable trésor langagier. - Plus de 450 000 coocurrences, la plus grand dictionnaire du genre publié à ce jour. - Un classement par force

Read more

rali


Le triangle de dévoilement
La liste des cooccurrences
Le contexte syntaxique
La force des cooccurrences
Le filtre textuel
Cooccurrences de lumière (n. f.)
Force
Avec épithète (171)
pleine lumière
apparaitre en pleine lumière
lumière tamisée
lumière naturelle
lumière crue
lumière allumée
et 166 autres...
Avec apposition (1)
lumière laser
Avec complément nominal (57)
lumière du jour
lumière du soleil
à la lumière de l'expérience
lumière de la lampe
lumière au bout du tunnel
et 52 autres...
Sujet (69)
la lumière éclaire
la lumière s'éteint
la lumière s'allume
la lumière brille
la lumière jaillit
et 64 autres...
Options...
Exemples de la cooccurrence
pleine lumière
Mélie quand elle eut lavé ses mains, prit sur le bord de la fenêtre, son métier à dentelles, s'assit en pleine lumière, et travailla.
Gustave Flaubert, Bouvard et Pécuchet, ABU, la Bibliothèque universelle
Je vais m'occuper seulement de chercher à mettre en pleine lumière la grande différence qui existe entre les deux conceptions de grève générale.
Georges Sorel, Réflexions sur la violence, Gallica
Assise très droite à présent, elle lui montrait en pleine lumière son visage noble et défait, ciré par de cuisantes larmes séchées.
Colette, Chéri, Projet Gutenberg
Exemples
Définitions

MWE : typologie selon *[Sag et al., 2002]*

- lexicalized phrases
  - fixed expressions : by and large (en général) , every which way (en désordre total)
  - semi-fixed expressions
    - non-decomposable idioms : kick the bucket (mourir), shoot the breeze (parler pour ne rien dire ?)
    - compound nominals : car **park**, **attorney** general
    - proper-names
  - syntactically flexible expressions
    - Verb-particle constructions : write up, look up, brush up on
      fall of a truck / ⋆ fall a truck of
      call Kim up / call up Kim
      call/ring/phone/telephone, mais ⋆ telephone up
    - Decomposable idioms : let the cat out of the bag (révéler)
    - Light verbs : **make** a **mistake** / ⋆ **make** a **demo** (**give** a **demo**)
- institutional phrases : traffic light, telephone booth/box
  - statistiquement saillantes (collocations)
  - syntaxiquement et sémantiquement compositionnelles

rali

Séquence pertinente ?

1 = vraiment pertinente
2 = plutôt pertinente
3 = plutôt pas pertinente
4 = vraiment pas pertinente

- intérêts du canada
- intérêts des sociétés multinationales
- fin du printemps
- enfreint le règlement
- les lignes directrices sur les conflits d' intérêts
- n' a pas
- le projet de loi
- monsieur le président , je
- le gouvernement ne
- et des
- apprentissage profond

rali

Collocations

- Lire *[Smadja, 93]* pour une bonne caractérisation
- **Propriété 1 :** Les collocations sont arbitraires.
  - ex : to break down/force the door

langue	expression	équivalent anglais
français	enfoncer la porte	to push the door through
allemand	die Tür aufbrechen	to break the door
italien	sfondare la porta	to hit/demolish the door
espagnol	tumbar la puerta	to fall the door
turc	kapiyi kirmak	to break the door

  - mais to see the door se “traduit” dans ces 5 langues de manière compositionnelle.

Collocations

- **Propriété 2 :** Les collocations sont dépendantes du domaine
  - connu du plongeur québécois, mais pas nécessairement du plongeur marseillais : dry suit
- **Propriété 3 :** Les collocations sont récurrentes
  - on retrouve normalement plusieurs fois une collocation donnée dans un passage
- **Propriété 4 :** Les collocations sont des unités cohérentes
  - La donnée d'un mot (ou de plusieurs) d'une collocation permet de reconstituer la collocation
    - ex : traffic ? dans le sens *embouteillage*
  - c'est sur des études perceptives de ce genre que se basent les lexicographes pour décider de la nature collocative d'une séquence de mots

rali

Testez votre niveau d'anglais [1]

- If a fire breaks out, the alarm will ??
- The boy doesn't know how to ?? his bicycle
- The American congress can ?? a presidential veto
- Before eating your bag of microwavable popcorn, you have to ?? it

1. Merci à Graham Russell pour les réponses

Testez votre niveau d'anglais [1]

- If a fire breaks out, the alarm will ??
  ↪ **ring** / go off / **sound** / start
- The boy doesn't know how to ?? his bicycle
- The American congress can ?? a presidential veto
- Before eating your bag of microwavable popcorn, you have to ?? it

1. Merci à Graham Russell pour les réponses

Testez votre niveau d'anglais [1]

- If a fire breaks out, the alarm will ??
  ↪ **ring** / go off / **sound** / start
- The boy doesn't know how to ?? his bicycle
  ↪ drive / **ride** / conduct
- The American congress can ?? a presidential veto
- Before eating your bag of microwavable popcorn, you have to ?? it

1. Merci à Graham Russell pour les réponses

Testez votre niveau d'anglais [1]

- If a fire breaks out, the alarm will ??
  ↪ **ring** / go off / **sound** / start
- The boy doesn't know how to ?? his bicycle
  ↪ drive / **ride** / conduct
- The American congress can ?? a presidential veto
  ↪ ban / cancel / delete / reject / turn down / abrogate / **overrule**
- Before eating your bag of microwavable popcorn, you have to ?? it

1. Merci à Graham Russell pour les réponses

Testez votre niveau d'anglais [1]

- If a fire breaks out, the alarm will ??
  ↪ **ring** / go off / **sound** / start
- The boy doesn't know how to ?? his bicycle
  ↪ drive / **ride** / conduct
- The American congress can ?? a presidential veto
  ↪ ban / cancel / delete / reject / turn down / abrogate / **overrule**
- Before eating your bag of microwavable popcorn, you have to ?? it
  ↪ **cook** / **nuke** / broil / fry / bake

1. Merci à Graham Russell pour les réponses

Différentes formes de collocations

- Collocations prédicatives

N-Adj	▸ heavy/light ⋈ trading/smoker/traffic
	▸ strong ⋈ tea
S-V	▸ stock ⋈ rose/fell/jumped
V-Part	▸ take ⋈ from
	▸ raise ⋈ by
	▸ mix ⋈ with

- Collocations rigides
  - souvent des groupes nominaux
    traffic jam, foreign exchange, New York Stock Exchange, Stock Market, White House Spokesman Marlin Fitwater
  - inséparables sans perte de sens

Plan

rali

Tableaux de suffixes (*Suffix arrays*)

- indexer **toutes** les séquences d'un texte
- afin de calculer efficacement des statistiques sur ces séquences

- exploité dans un cadre monolingue ou bilingue
  [Nagao and Mori, 1994, Ikehara et al., 1996, Haruno et al., 1996, Shimohata et al., 1997, Russell, 1998].
- **Note :** les algorithmes qui suivent sont extraits de ***[Russell, 1998]***. Pour une implémentation efficace : ***[Gonnet et al., 1992, Manber and Myers, 93]***

rali

Some_0 of_1 the_2 words_3 of_4 the_5 sentence_6 are_7 the_8 same_9

- Ce texte contient 10 mots et 7 types
- Soit un ordre sur les types (Some=0, words=6) :

$$\text{Some} < \text{are} < \text{of} < \text{same} < \text{sentence} < \text{the} < \text{words}$$

- un suffix array `SFX` est un tableau représentant l'ensemble **ordonné** des suffixes d'un texte

0	7	4	1	9	6	8	5	2	3

- `SFX[i]` désigne une position dans le texte
  - ex : `SFX[6]` encode le suffixe the same

Some$_0$ of$_1$ the$_2$ words$_3$ of$_4$ the$_5$ sentence$_6$ are$_7$ the$_8$ same$_9$

SFX:	0	7	4	1	9	6	8	5	2	3

- `LPC` : la table des plus longs préfixes communs de 2 suffixes
- `LPC[i]` = nb. de mots communs des suffixes qui commencent respectivement aux positions `SFX[i]` et `SFX[i+1]`

0	0	2	0	0	0	1	1	0

  - ex : of the words et of the sentence ont un préfixe commun de deux mots
  - requiert un passage sur `SFX` en comparant les suffixes deux à deux

Retrouver les occurrences d'une séquence

- **in :** une occurrence d'une séquence `key` a été trouvée en position `found` dans `SFX`
- **out :** le nombre d'occurrences de cette séquence, `left` indique la première position d'occurrence dans `SFX`

```
left ← right ← found
while left > 0 ∧ lcps[left-1] ≥ |key| do
  left ← left − 1
while right < |lcps| ∧ lcps[right] ≥ |key| do
  right ← right + 1

return right - left +1
```

- **Note :** pas besoin de la table `SFX`

Extraire les séquences

Extraire toutes les séquences d'au moins `minlength` mots et de fréquence minimale `minfreq`

- une séquence s est caractérisée par un triplet $\langle l, f, p \rangle$, où :
  - l est la longueur d'une séquence,
  - f sa fréquence,
  - p la première position dans `SFX` qui pointe (dans le texte) sur une séquence dont le préfixe est s
  - of the est caractérisée par le triplet $\langle 2, 2, 2 \rangle$
  - the same est caractérisée par le triplet $\langle 2, 1, 6 \rangle$
- en maintenant deux ensembles de triplets `active` et `result`
  - `active` = ensemble qui contient les séquences potentiellement intéressantes en cours d'analyse
  - `result` = la réponse

Extraction des séquences

```
results ← active ← ∅
new_length ← 0
prev_length ← 0
min_length ← 0
this_pos ← 0

while this_pos < |lcps| do
    this_length ← lcps[this_pos]
    if this_length < min_length then
        keepGoodFromActive()
        new_length ← min_length
    else if this_length ≥ prev_length
    then
        updateAndAdd()
        new_length ← this_length + 1
    else
        updateOrFlush()
        new_length ← this_length + 1
    prev_length ← this_length
    this_pos ← this_pos + 1

keepGoodFromActive()

return results
```

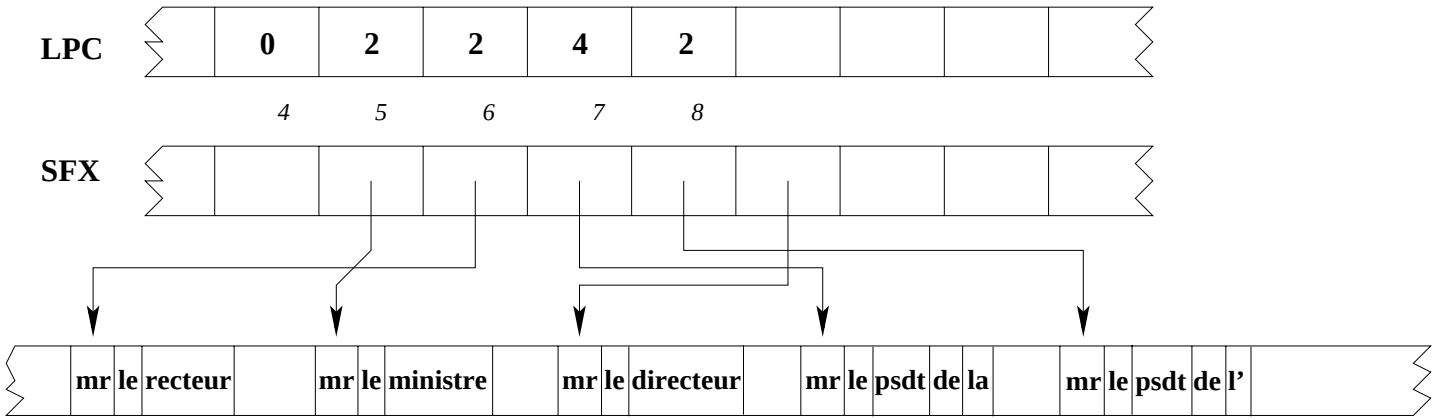
```
flush(t ≡ ⟨len, freq, pos⟩)
if freq ≥ min_freq then
    results ← results ∪ {t}
active ← active − {t}

keepGoodFromActive:
for all t ∈ active do
    flush(t)

updateAndAdd:
for all triples ⟨len, freq, pos⟩ ∈ active do
    freq ← freq + 1
while new_length ≤ this_length do
    active ← active ∪ {⟨new_length, 2, this_pos⟩}
    new_length ← new_length + 1

updateOrFlush:
for all t ≡ ⟨len, freq, pos⟩ ∈ active do
    if len ≤ this_length then
        freq ← freq + 1
    else
        flush(t)
```

Extraction des séquences

- *this_length* $<$ *min_length*
  - vide `active` en gardant (dans `result`) les séquences qui vérifient le critère de fréquence.
- *this_length* $\geq$ *prev_length*
  - incrémente la fréquence de toutes les séquences dans `active`
  - boucle sur `new_length` : “lorsqu’on voit une séquence de taille n, on voit également une séquence de taille $n-1$”
- *this_length* $<$ *prev_length* ***et*** *this_length* $\geq$ *min_length*
  - incrémente la fréquence de toutes les séquences d’au plus `this_length` mots dans `active`
  - retire de `active` les autres séquences en gardant dans `result` celles qui vérifient le critère de fréquence

rali

Illustration



- Évolution de l'ensemble `active` lorsque `min_length` vaut 2 :

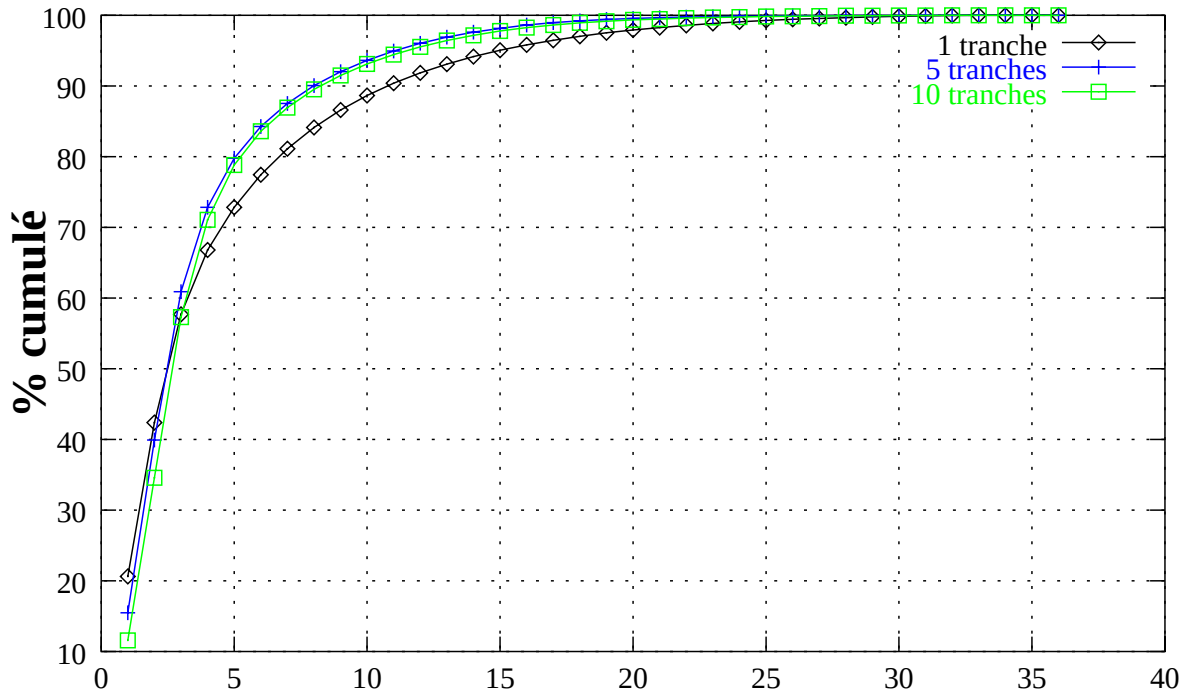
this_pos	this_length	new_length	active	cas
4	0	2	$\{\}$	1
5	2	3	$\{< 2, 2, 5 >\}$	2
6	2	3	$\{< 2, 3, 5 >\}$	2
7	4	5	$\{< 2, 4, 5 >, < 3, 2, 7 >, < 4, 2, 7 >\}$	2
8	2	3	$\{< 2, 5, 5 >\}$	3

Triplet $\langle l, f, p \rangle$

$$\{\langle T[SFX[j]], \ldots T[SFX[j] + l - 1]\rangle : p \leq j < p + f\}$$

- ensemble de séquences identiques dans le texte `T`
- en pratique, on ne recherche que les séquences contenues à l'intérieur des phrases de `T`

$\texttt{min\_length} = 1, \texttt{min\_freq} = 2$



tranches hansard	types	séquences
1	4 629	22 457
5	10 172	65 715
10	14 832	128 253

Plan

Sélection des “bonnes” séquences

par fréquence

lg.	fréq.	séquence	lg.	fréq.	séquence
2	1760	de la	2	893	) :
2	1594	de l’	2	867	qu’ il
2	1391	le président	2	778	, le
2	1083	monsieur le	2	724	: monsieur
3	1076	monsieur le président	3	724	: monsieur le
2	1069	le gouvernement	4	723	: monsieur le président
2	1040	à la	5	720	: monsieur le président ,
2	988	c’ est	2	701	à l’
2	985	président ,	2	643	le député
3	983	le président ,	2	640	, je
4	971	monsieur le président ,	3	608	) : monsieur
2	913	que le	5	608	) : monsieur le président

Sélection des “bonnes” séquences

par information mutuelle (ponctuelle)
[Church and Hanks, 1989]

- voir ***[Manning and Schütze, 1999]*** chapitre 5 pour un vaste panorama de statistiques

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{P(x|y)}{P(x)} = \log_2 \frac{P(y|x)}{P(y)}$$

Ex.[2] :

	chambre	¬ chambre
house	31 950	12 004
¬ house	4 793	848 330

	communes	¬ communes
house	4 974	38 980
¬ house	441	852 682

$$\log \frac{P(house|chambre)}{p(house)} = \log \frac{\frac{31\,950}{31\,950+4\,793}}{P(house)} \approx \log \frac{0.87}{P(house)}$$

$$\log \frac{P(house|communes)}{p(house)} = \log \frac{\frac{4\,974}{4\,974+441}}{P(house)} \approx \log \frac{0.92}{P(house)}$$

2. [Manning and Schütze, 1999], page 179

Sélection des “bonnes” séquences

rapport de vraisemblance [Dunning, 1993]

$$
\begin{aligned}
-\log\lambda \quad = \quad & a\log a + b\log b + c\log c + d\log d + N\log N \\
& -(a+c)\log(a+c) - (a+b)\log(a+b) \\
& -(c+d)\log(c+d) - (d+b)\log(d+b)
\end{aligned}
$$

- a : est le nombre de fois où A et B se suivent
- b : est le nombre de fois où A apparait non suivi de B
- c : est le nombre de fois où B est non précédé de A
- d : est le nombre de fois où ni A ni B n’apparaissent, dans cet ordre, dans le corpus
- N : a + b + c + d = taille du corpus

- plus $-\log\lambda$ est grand, plus les mots testés sont dépendants

Rapport de vraisemblance *[Dunning, 1993]*

- Le rapport de vraisemblance d'une hypothèse particulière (H_0) :

$$\lambda = \frac{max_{\omega \in \Omega_0} H(\omega; k)}{max_{\omega \in \Omega} H(\omega; k)}$$

où

  - Ω est l'espace des paramètres
  - Ω_0 est l'espace des paramètres correspondant à l'hypothèse

- ***hypothèse*** les événements (mots) sont distribués selon une loi binomiale

$$H(p; k, n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Rapport de vraisemblance

- La comparaison de deux binomiales de paramètres respectifs p_1 et p_2 peut se faire en prenant comme hypothèse H_0 : $p = p_1 = p_2$
- Dans ce cas le rapport de vraisemblance s'écrit :

$$\lambda = \frac{max_p H(p, p; k_1, n_1, k_2, n_2)}{max_{p_1,p_2} H(p_1, p_2; k_1, n_1, k_2, n_2)}$$

avec $H(p_1, p_2; k_1, n_1, k_2, n_2) =$
$\binom{n_1}{k_1} p_1^{k_1} (1 - p_1)^{n_1 - k_1} \times \binom{n_2}{k_2} p_2^{k_2} (1 - p_2)^{n_2 - k_2}$

- Le maximum du dénominateur est obtenu pour $p_1 = \frac{k_1}{n_1}$ et $p_2 = \frac{k_2}{n_2}$
- Le maximum du numérateur est obtenu pour $p = \frac{k_1 + k_2}{n_1 + n2}$

rali

Rapport de vraisemblance

- maximum si :

$$\frac{\delta H(p_1, p_2; k_1, n_1, k_2, n_2)}{\delta p_1} = 0$$

$$\Rightarrow \overbrace{\binom{n_1}{k_1}\binom{n_2}{k_2}p_2^{k_2}(1-p_2)^{n_2-k_2}}^{cst} \frac{\delta p_1^{k_1}(1-p_1)^{n_1-k_1}}{\delta p_1} = 0$$

$$\Rightarrow p_1^{k_1-1}\left[k_1(1-p_1)^{n_1-k_1} - p_1(n_1-k_1)(1-p_1)^{n_1-k_1-1}\right] = 0$$

$$(1-p_1)^{n_1-k_1-1}\left[k_1(1-p_1) - p(n_1-k_1)\right] = 0$$

$$k_1(1-p_1) - p_1(n_1-k_1) = 0$$

$$k_1 - k_1p_1 - p_1n_1 + k_1p_1 = 0$$

$$p_1 = \frac{k_1}{n_1}$$

- Idem pour p_2 et p

Rapport de vraisemblance

- Le rapport de vraisemblance s'écrit alors :

$$\lambda = \frac{L(p, k_1, n_1) \times L(p, k_2, n_2)}{L(p_1, k_1, n_1) \times L(p2, k_2, n_2)}$$

où $L(p, k, n) = p^k(1-p)^{n-k}$

- Soit, en prenant le logarithme :

$$\begin{aligned} -\log \lambda = \quad & \log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) \\ & - \log L(p, k_1, n_1) - \log L(p, k_2, n_2) \end{aligned}$$

avec $p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$, et $p = \frac{k_1+k_2}{n_1+n2}$

rali

Sélection des “bonnes” séquences

rapport de vraisemblance [Dunning, 1993]

- A et B deux mots dont nous voulons mesurer le degré de dépendance
- $H_0 \equiv$ indépendance de A et B
  - $P(B|A) = P(B|\neg A) = P(B)$
- Soit la table de contingence (où par exemple) a (resp. b) indique le nombre de fois où B suit (resp. ne suit pas) A dans un corpus

	B	$\neg B$
A	a	b
$\neg A$	c	d

- alors $p(B|A) = p_1 = \frac{a}{a+b}$ et $p(B|\neg A) = p_2 = \frac{c}{c+d}$
- En développant, on retombe sur la formulation initiale du rapport de vraisemblance

Sélection des “bonnes” séquences

- Les 16 premières (et les 4 dernières) séquences de deux mots selon le score $-\log\lambda$ sur les 10 tranches du Hansard :

fréq.	$-\log\lambda$	séquence	fréq.	$-\log\lambda$	séquence
1391	4690.07	le président	985	2502.40	président ,
988	4400.53	c’ est	307	2387.39	p. 100
1083	3935.63	monsieur le	592	2117.01	la chambre
893	3504.88	) :	1594	1943.32	de l’
724	2930.47	: monsieur	1760	1896.21	de la
540	2899.87	j’ ai	643	1679.89	le député
1069	2734.26	le gouvernement	385	1670.77	nous avons
867	2650.22	qu’ il	293	1509.37	premier ministre
2	3.7e-6	maintenant un	10	2.2e-6	loi que
2	2.3e-6	là a	4	7.9e-7	jamais .

Étendre une mesure d'association binaire

- **Idée :** Une séquence est non pertinente si on la retrouve comme préfixe ou suffixe de séquences pertinentes ***[Ries et al., 1995]***
- Le score ρ d'une séquence w_1^n peut être donné par :

$$\rho_A(w_1^n) = \min_{i \in [1, n-1]} A(w_1^i, w_{i+1}^n)$$

  - pour une mesure d'association binaire A

- Ce score mesure la "résistance d'une séquence à la division", et reflète d'une certaine manière son degré de cohésion

rali

Sélection des “bonnes” séquences

- Les 12 séquences les mieux notées selon ρ sur 10 tranches du Hansard, et les 4 séquences les moins bien notées

fréq.	ρ	séquence	fréq.	ρ	séquence
1076	6297	monsieur le président	608	3120	) : monsieur
1391	4690	le président	971	3076	monsieur le président ,
988	4400	c’ est	723	2937	: monsieur le président
1083	3936	monsieur le	724	2930	: monsieur
893	3505	) :	540	2900	j’ ai
608	3122	) : monsieur le président	459	2875	projet de loi
3	2e-6	est sur le	3	6e-7	des mesures .
4	8e-7	jamais .	2	4e-7	la souveraineté .

Sélection des “bonnes” séquences

score d’entropie [Shimohata et al., 1997]

$$
\begin{aligned}
e(w_1^n) &= (e_{left}(w_1^n) + e_{right}(w_1^n))/2 \\
e_{left}(s) &= \sum_{w/ws \in T} h(\frac{|ws|}{|s|}) \\
e_{right}(s) &= \sum_{w/sw \in T} h(\frac{|sw|}{|s|}) \\
h(x) &= x \log(x)
\end{aligned}
$$

- $e_{left}(s)$ (resp. $e_{right}(s)$) est nul quand seulement une forme suit (resp. précède) s dans toutes les occurrences de s
- $e(s)$ est maximal lorsqu’il y a exactement $freq(s)$ types qui suivent (resp. précèdent) s
- **Intuitivement :** une séquence cohérente devrait apparaître dans un nombre varié de contextes (entropie élevée)

rali

Illustration du score d'entropie

$$(4) \left\{ \begin{array}{c} - \\ , \\ : \\ \text{ce} \end{array} \right\} \text{monsieur} \left\{ \begin{array}{c} \text{a} \\ \text{le} \end{array} \right\} (2)$$

faible entropie

$$(364) \left\{ \begin{array}{c} \text{abandon} \\ \text{accepté} \\ \vdots \\ \text{vraiment} \end{array} \right\} \text{le} \left\{ \begin{array}{c} 10 \\ \text{baril} \\ \vdots \\ \text{yukon} \end{array} \right\} (440)$$

forte entropie

Sélection des “bonnes” séquences

fréq.	$e(s)$	séquence	fréq.	$e(s)$	séquence
1760	2.5362	de la	393	2.39531	d’ un
333	2.53098	a été	517	2.34532	que les
370	2.46355	, les	371	2.34049	d’ une
256	2.44128	et les	156	2.2844	ont été
385	2.4374	nous avons	379	2.28096	il est

Note : 40% des séquences de deux mots ou plus ont un score nul avec cette métrique.

rali

Séq. ($f \geq 2$) contenant aviation safety (hansard)

	$\rho(s)$	$e(s)$	freq.	l	s
1	113.32	0.23	16	2	aviation safety
2	109.19	1.14	15	3	aviation safety board
3	92.02	0.24	15	3	canadian aviation safety
4	85.69	1.13	14	4	canadian aviation safety board
5	35.33	1.23	14	4	the canadian aviation safety
6	32.62	1.99	13	5	the canadian aviation safety board
7	12.91	0.69	2	4	aviation safety board recommendations
8	6.29	1.39	4	6	the canadian aviation safety board ,
9	5.97	0.69	4	5	canadian aviation safety board ,
10	5.67	0.69	4	4	aviation safety board ,
11	5.65	0.69	2	6	by the canadian aviation safety board
12	5.49	0.35	2	5	by the canadian aviation safety
13	5.28	0.35	2	7	of the canadian aviation safety board .
14	3.81	1.10	3	6	, the canadian aviation safety board
15	3.59	0.55	3	5	, the canadian aviation safety
16	2.71	0.32	3	6	the canadian aviation safety board .
17	2.51	0	3	5	canadian aviation safety board .
18	2.32	0	3	4	aviation safety board .
19	1.89	0.35	2	6	of the canadian aviation safety board
20	1.76	0.35	2	5	of the canadian aviation safety

rali

Exemple de filtre entropique

- éliminer une séquence si c'est le préfixe d'une séquence de plus grande entropie

```
Filtering of tv=113.3 e=0.23 f=16 lg=2 seq=aviation safety
coz: tv=109.2 e=1.13 f=15 lg=3 seq=aviation safety board

Filtering of tv=92.0 e=0.24 f=15 lg=3 seq=canadian aviation safety
coz: tv=85.7 e=1.13 f=14 lg=4 seq=canadian aviation safety board

Filtering of tv=86.9 e=0.23 f=16 lg=2 seq=canadian aviation
coz:  tv=85.7 e=1.13 f=14 lg=4 seq=canadian aviation safety board

Filtering of tv=38.0 e=1.27 f=15 lg=3 seq=the canadian aviation
coz: tv=32.6 e=1.99 f=13 lg=5 seq=the canadian aviation safety board

Filtering of tv=35.3 e=1.23 f=14 lg=4 seq=the canadian aviation safety
coz: tv=32.6 e=1.99 f=13 lg=5 seq=the canadian aviation safety board
```

Plan

rali

Séquences non contigües

- *[Maarek et al., 1991]* décrivent une approche permettant de sélectionner des affinités lexicales (collocations non contigües)
- *[Martin and van Sterkenburg, 1983]* : 98% des relations lexicales en anglais mettent en jeu des mots qui sont distants d'au plus 5 mots dans une phrase (sans compter les mots outils).
- algorithme :
  - faire glisser une fenêtre de taille fixe et accumuler les fréquences de tous les couples de mots apparus dans ces fenêtres.
  - sélectionner les bigrammes représentatifs

rali

Séquences non contigües *[Maarek et al., 1991]*

- score d'une affinité lexicale (m_1, m_2) :

$$\rho(m_1, m_2) = -freq_{\mathcal{D}}(m_1, m_2) \times log_2(p(m_1)p(m_2))$$

où :

  - $freq_{\mathcal{D}}(m_1, m_2)$ est la fréquence de l'affinité lexicale dans le document,
  - $p(x)$ calculée à partir d'un grand corpus représentatif de la langue étudiée

- gardent les affinités lexicales dont le score est supérieur à la moyenne de la distribution plus son écart type (z-score ≥ 1)

Application des affinités lexicales

- calculer du profil d'un document (AL les mieux notées)
- exemple de profils calculés sur des *group-news* :

group	affinités lexicales
hardware	[drive, hard] [clock speed] [clock oscillator]
baseball	[game, save] [baseball, player] [game team] [bolick, frank
hockey	[gargle, howl] [game, play] [player team] [good team]
politics	[homosexual, male] [care health] [jews, mormons]
medical	[antibiotic, chapter] [bloom candida] [chronic, hepatitis]

- Lire [Alvarez et al., 2004] pour une application des affinités lexicales en recherche d'information

rali

Probabilité d'une co-occurrence non vue

[Dagan et al., 1993]

- **Idée :**
  - eat bread n'a jamais été vue à l'entraînement, contrairement à eat toast
  - eat toast est un bon bigramme pour estimer p(eat bread)
- (rappel) les techniques de lissage se rabattant sur l'unigramme perdent le lien qu'entretiennent ces mots :
  - eat bread et eat car peuvent avoir la même estimée si bread et car sont aussi fréquents dans le corpus d'entraînement

Probabilité d'une co-occurrence non vue

- **Hypothèse :** des co-occurrences similaires ont des valeurs d'information mutuelle similaires.
- **Définition** d'une co-occurrence dans [Dagan et al., 1993] :
  - toute paire de mots qui co-occurent dans une fenêtre de d mots (3 dans leur cas), abstraction faite des mots outils. Les paires sont directionnelles : $(x, y) \neq (y, x)$.
- on peut mesurer l'association des mots d'une paire (x, y) par l'information mutuelle (N est la taille du corpus) :

$$I(x, y) = \log_2 \left(\frac{N}{d} \frac{f(x, y)}{f(x) f(y)} \right)$$

rali

Similarité de deux co-occurrences

- deux co-occurrences (w_1, w_2) et (w'_1, w'_2) sont similaires **ssi** w_1 est similaire à w'_1 et w_2 est similaire à w'_2
- Ex : (chapter,describes) n'est pas observée dans un corpus de 9 million de mots postés à USENET. Les co-occurrences suivantes ont cependant été vues dans ce même corpus et sont similaires à cette paire (au sens d'une métrique à préciser) : *(introduction,describes)*, *(book,describes)*, *(section, describes)*.
- **Idée** : l'information mutuelle d'une co-occurrence inconnue (w_1, w_2) est estimée par la moyenne de l'information mutuelle des paires similaires : $\hat{I}(w_1, w_2)$. Et alors :

$$\hat{f}(w_1, w_2) = \frac{d}{N} f(w_1) f(w_2) 2^{\hat{I}(w_1, w_2)}$$

rali

Similarité de deux co-occurrences

$$
\begin{aligned}
I(introduction, describes) &= 6.85 \\
I(book, describes) &= 6.27 \\
I(section, describes) &= 6.12 \\
\\
\Rightarrow \quad \hat{I}(chapter, book) &= 6.41 \\
\Rightarrow \quad \hat{f}(chapter, book) &= 0.124
\end{aligned}
$$

- c'est plus élevé que ce qu'on obtiendrait à partir des fréquences des mots considérés comme indépendants.
- **Détail d'implémentation :** pour calculer $\hat{I}(w_1, w_2)$, les auteurs prennent les 6 mots les plus similaires à w_1 et les 6 mots les plus similaires à w_2, et la moyenne est calculée sur toutes les co-occurrences d'une combinaison de ces (au plus) 6x6 mots.

Similarité entre deux mots

- **idée :** deux mots w_1 et w_2 sont similaires s'ils co-occurrent de manière semblable avec toutes sortes de mots
(vive la récursivité)
  - s'ils ont des valeurs semblables d'information mutuelle avec d'autres mots du vocabulaire.
- les paires étant directionnelles, il y a un ratio à droite et un ratio à gauche :

$$sim_L(w_1, w_2, w) = \frac{min(I(w,w_1),I(w,w_2))}{max(I(w,w_1),I(w,w_2))}$$
$$sim_R(w_1, w_2, w) = \frac{min(I(w_1,w),I(w_2,w))}{max(I(w_1,w),I(w_2,w))}$$

  - à valeur dans $[0, 1]$, 1 = très similaire, 0 = pas similaire
- ces ratios sont à calculer pour chaque mot w du vocabulaire (gasp !)

rali

Similarité entre deux mots

- mis bout à bout, voici la mesure de similarité entre deux mots :

$$sim(w_1, w_2) = \frac{\sum_{w \in V} min(I(w, w_1), I(w, w_2)) + min(I(w_1, w), I(w_2, w))}{\sum_{w \in V} max(I(w, w_1), I(w, w_2)) + max(I(w_1, w), I(w_2, w))}$$

- ex. des mots les plus similaires au mot aspects :

recherche exhaustive		approximation	
mots similaires	sim	mots similaires	sim
aspects	1.0	aspects	1.0
topics	0.1	topics	0.1
areas	0.09	areas	0.09
expert	0.079	expert	0.079
issues	0.076	issues	0.076
approaches	0.072	concerning	0.069

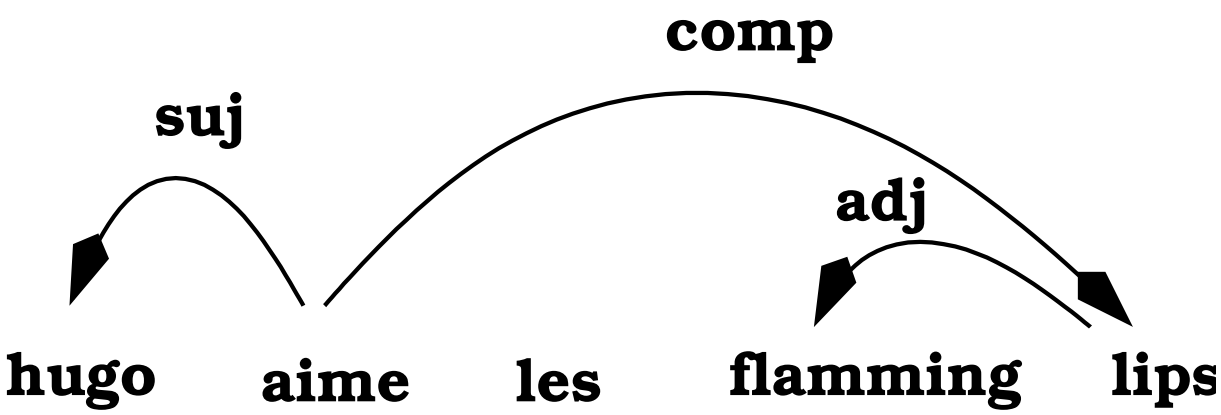
Similarité entre deux mots (détails)

- La recherche approximative des mots similaires se fait par indirection et seuillage : les plus similaires des plus similaires du mot w dont on cherche les mots similaires
  - réduction du temps de 17 minutes à 7 secondes pour trouver les mots similaires d'un mot
- Est-ce que ça marche ?
  - lire ***[Dagan et al., 1993, Lee and Pereira, 1999, Lee, 1999]***

rali

Identification d'expressions non compositionnelles *[Lin, 1999]*

- **input** : analyseur grammatical en dépendances (Minipar[3]) :



  - produisant des triplets *(tête, relation, modifieur)*
    - (aime,sujet,hugo), (aime,comp,lips), (lips,adj,flammings)
- Lire aussi [Melamed, 1997] pour une identification bilingue

3. `http://www.cs.ualberta.ca/~lindek/minipar.htm`

Hypothèse de *[Lin, 1999]*

- une collocation (H, R, M) est non compositionnelle si son information mutuelle diffère de manière significative de l'information mutuelle de collocations proches
  - (H', R, M') est proche de (H, R, M) si $H' \in sim(H), M' \in sim(M)$ et $\neg\,(H' = H \wedge M' = M)$
  - **Def.** $I(H, R, M) = \log \frac{P(H,R,M)}{P(H|R)P(M|R)P(R)}$
  - où chaque distribution est apprise par fréquence relative sur un gros corpus (125 millions de mots, 80 millions de dépendances, filtrées par seuillage)

Similarité entre deux mots *[Lin, 1998]*

- deux mots w_1 et w_2 sont similaires s'ils partagent des relations proches avec les autres mots
  - même idée que *[Dagan et al., 1993]*, mais avec les relations en plus

$$sim(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \bigcap T(w_2)} I(w_1, r, w) + I(w_2, r, w)}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

où $T(w) = \{(r, w')/I(w, r, w') > 0\}$

  - s'approche d'autant plus de 0 que w_1 et w_2 partagent peu de relations

Similarité entre deux mots [Lin, 1998]

- En gardant les N meilleurs associations ($N = 200$) de chaque mot, l'auteur obtient un thésaurus dont voici quelques exemples [4] :

brief(noun) : affidavit, 0.13, petition 0.05, memorandum 0.05, motion 0.05, document 0.04, paper 0.04,. . .
brief(verb) : tell 0.09, urge 0.07, ask 0.07, meet 0.06, appoint 0.06, elect 0.05,. . .
brief(adj) : lengthy 0.13, short 0.12, recent 0.09, prolonged 0.09, long 0.09, stormy 0.07,. . .

4. Extrait de l'article.

À propos des thésaurus

- **Def :** répertoire de termes normalisés classés alphabétiquement et répartis en structures correspondant aux divers champs de la connaissance [5].
  - `http://thesaurus.reference.com/search?q=flower`

```
Entry:flower
Function:noun
Definition:bloom
Synonyms:annual, blossom, bud, cluster, efflorescence, floret,
      floweret, head, herb, inflorescence, perennial, pompon,
      posy, shoot, spike, spray, vine

Source:Roget's New Millennium Thesaurus, First Edition (v 1.0.5)
Copyright 2004 by Lexico Publishing Group, LLC. All rights reserve
```

5. Pris du dictionnaire encyclopédique de la langue française, 1996

Non compositionnalité et IM *[Lin, 1999]*

- **spill** one's **guts** *(se plaindre, geindre)*

  spill(verb) : leak 0.153, pour 0.127, spew 0.125, dump 0.098, seep 0.096, . . .
  gut(noun) : intestine 0.091, instinct 0.089, foresight 0.085, creativity 0.082, heart 0.079, . . .

- $(spill, comp, gut)$ est apparu 13 fois dans le corpus (I=6.24)
- mais aucune co-occurrence n'est apparue en remplaçant l'un des mots par un mot similaire (ex : leak gut).
- collocation non compositionnelle

Non compositionnalité et IM *[Lin, 1999]*

- **red tape** *(tracasseries administratives ?)*

  red : yellow 0.164, purple 0.149, pink 0.146, green 0.136, . . .
  tape : videotape 0.196, cassette 0.177, videocassette 0.168, . . .

- d'autres collocations existent mais avec des informations mutuelles différentes

verb	object	freq	I
red	tape	259	5.87
yellow	tape	12	3.75
orange	tape	2	2.64
black	tape	9	1.07

- collocation non compositionnelle

Non compositionnalité et IM *[Lin, 1999]*

- **economic impact** *(impact économique)*

  economic : financial 0.305, political 0.243, social 0.219, fiscal 0.209, cultural 0.202, . . .
  impact : effect 0.227, implication 0.163, consequence 0.156, significance 0.146, . . .

- beaucoup de collocations ont le même genre d'information mutuelle

verb	object	freq	I
economic	impact	171	1.85
economic	consequence	59	1.88
economic	repercussion	7	1.84

- collocation compositionnelle

Parenthèse : Intervalle de confiance

Soit : $X_i \sim D(\mu, \sigma^2), \quad i \in [1, N]$

- et un estimateur de la moyenne μ à partir des N observations $X_{i=1}^n$: $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$
- On veut trouver l'intervalle de confiance à $(1 - \alpha) \times 100\%$, cad l'intervalle à l'intérieur duquel on est certain à $(1 - \alpha) \times 100\%$ que la vrai valeur est.

On sait que $\overline{X} \sim N(\mu, \sigma^2/n)$, donc $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

Détour rapide : Intervalle de confiance

Donc :

$$P(-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1-\alpha$$
$$P(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Si $\overline{X}$ est notre estimée, alors on est certain à $(1-\alpha)\times 100\%$ que la moyenne μ de la distribution est dans $(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$

- Intervalle de confiance à 95%

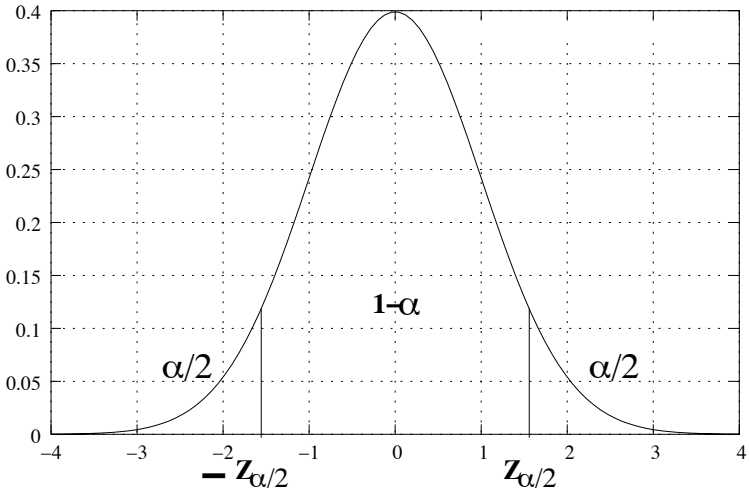
$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

- Intervalle de confiance à 99%

$$(\overline{X} - 2.58\frac{\sigma}{\sqrt{n}}, \overline{X} + 2.58\frac{\sigma}{\sqrt{n}})$$

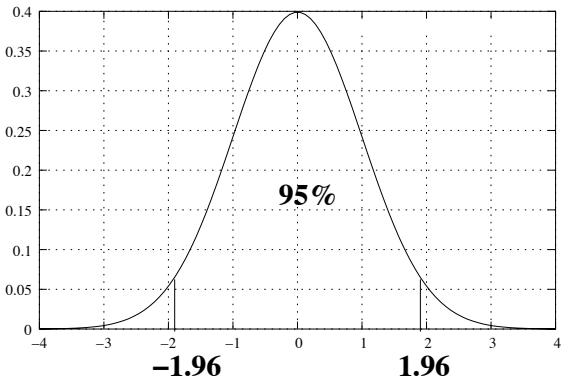
rali

Intervalle de confiance

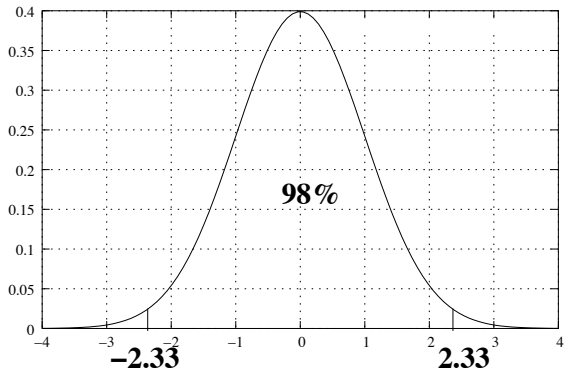


$$P(Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$
$$P(Z > Z_{\alpha/2}) = \alpha/2 = 1 - \phi(Z_{\alpha/2})$$

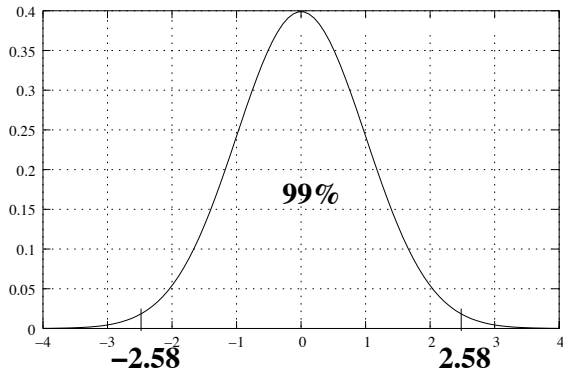
α	$\alpha/2$	%	$Z_{\alpha/2}$
0.01	0.005	99	2.58
0.02	0.01	98	2.33
0.05	0.025	95	1.96



$95\% \in [-1.96, 1.96]$



$98\% \in [-2.33, 2.33]$



$99\% \in [-2.58, 2.58]$

rali

Intervalle de confiance

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy$$

x	.00	.01	.02	.03	.04	.05	...
0.0	.5000	.5040	.5080	.5120	.5160	.5199	
0.1	.5398	.5438	.5478	.5517	.5557	.5596	
0.2	.5793	.5832	.5871	.5910	.5948	.5987	
...							
1.0	.8413	.8438	.8461	.8485	.8508	.8531	
...							
1.5	.9332	.9345	.9357	.9370	.9382	.9394	
1.6	.9452	.9463	.9474	.9484	.9495	.9505	
...							
1.9	.9713	.9719	.9726	.9732	.9738	.9744	
2.0	.9772	.9778	.9783	.9788	.9793	.9798	
...							
2.3	.9893	.9896	.9898	.9901	.9904	.9906	
...							
2.5	.9938	.9940	.9941	.9943	.9945	.9946	
...							
3.4	.9997	.9997	.9997	.9997	.9997	.9997	

rali

Que veut dire même genre d'information mutuelle ?

Réponse de [Lin, 1999] :

- La fréquence k d'un triplet est une variable aléatoire $B(n, \underline{p})$, n le nombre total de triplets observés en corpus. Pour des comptes élevés (le cas ici), on peut approcher la binomiale par une gaussienne, pour laquelle on peut calculer un intervalle de confiance :

$$\frac{k}{n} \pm z_N \frac{\sqrt{k(1-\frac{k}{n})}}{n}$$

- On a donc deux bornes pour le calcul de l'information mutuelle (on suppose que l'estimation de $p(M|R)$ et de $p(H|R)$ est juste et que seule l'estimation de la distribution jointe $p(H, R, M)$ est bruitée).

Que veut dire même genre d'information mutuelle ?

95% verb-object	freq.	IM	lower bound	upper bound
make difference	1489	2.928	2.876	2.978
make change	1779	2.194	2.146	2.239

- Pas de recouvrement $\Rightarrow$ pas la même information mutuelle
- **Def :** Une collocation est non compositionnelle s'il n'existe pas d'autre collocation proche dont les intervalles de confiance (à 95%) se chevauchent.

  Ex : take a bit, take advantage, take a look, take part

Plan

rali

Ancêtre du modèle de segments

- Déterminer séparément pour chaque langue les séquences pertinentes :
  - soit $\mathcal{L}_s$ la liste source et $\mathcal{L}_c$ la liste cible
- Une approche simple pour obtenir des associations bilingues de ces séquences consiste à “retokeniser” (découper à nouveau en “mots”) le corpus d’entraînement au regard des listes $\mathcal{L}_s$ et $\mathcal{L}_c$.
- On peut alors entraîner un modèle de traduction dont les “mots” sont soit de véritables mots, soit des séquences de mots.

rali

Exemples ainsi obtenues

source unit (s)	$f(s)$	target units
we have	1748	[nous,0.49] [avons,0.41] [, nous avons,0.07]
we must	720	[nous devons,0.61] [il faut,0.19] [nous,0.14]
this bill	640	[ce projet de loi,0.35] [projet de loi .,0.21] [projet de loi,0.18]
people of canada	282	[les canadiens,0.26] [des canadiens,0.21] [la population,0.07]
mr. speaker :	269	[m. le président :,0.80] [a,0.07] [à la,0.06]
what is happening	190	[ce qui se passe,0.21] [ce qui se,0.16] [et,0.15]
of course ,	178	[évidemment ,,0.26] [naturellement,0.08] [bien sûr,0.08]
is it the pleasure of the house to adopt the	14	[plaît-il à la chambre d' adopter,0.49] [la motion ?,0.42] [motion ?,0.04]

Mais . . .

- il est parti en train de banlieue / he left with a commuter train
- $\mathcal{L}_s = \{\text{en train de}, \text{train de banlieu}\}$
  $\mathcal{L}_t = \{\text{commuter train}\}$

```
src: [he] [left] [with] [a] [commuter train]
-------------------------------------------
trg: [il] [est] [parti] [en] [train de banlieu]
   : [il] [est] [parti] [en train de] [banlieu]
```

- aucune garantie que les unités sources ont un correspondant dans la liste des unités cibles

Éliminer du bruit

- De nombreux auteurs ont proposé de restreindre leur étude à des groupes nominaux (*noun phrase*)
  [Gaussier, 1995, Kupiec, 1993, hua Chen and Chen, 1994, Fung, 1995, Evans and Zhai, 1996].
- Un filtre à expressions régulières peut faire l'affaire :

```
(NomC|Ordi|NomP|AdjQ|Quan)
( (Quan|Dete|NomC|Ordi|NomP|AdjQ|Prep|VPPR))*
(Quan|NomC|Ordi|NomP|AdjQ|VPPS)
```

Modèle appris sur des mots et des NPs [6]

boom → prospérité,0.32 essor,0.27 explosion démographique,0.2 explosion,0.11 vague de prospérité,0.11
fbdb → banque fédérale de développement,1
rights of women → droits des femmes,1
canadian aviation safety board → bureau canadien de la sécurité aérienne,1
office of the superintendent of financial institutions → bureau du surintendant des institutions financières,1
newfoundland unemployment → taux de chômage à terre-neuve,1
small craft harbours → ports pour petits bateaux,0.53 ports pour petites embarcations,0.47
airline industry → industrie du transport aérien,0.73 secteur du transport aérien,0.13 industrie aérienne,0.13
food processing industry → secteur de la transformation des aliments,1
ordinary canadians → canadiens ordinaires,0.72 canadiens moyens,0.19 simples canadiens,0.082

- fbdb is an acronym for Federal Business Development Bank

6. Obtenu à partir d'un extrait du Hansard après filtrage

Trouver des traductions dans des corpus bilingues parallèles **bruités**

- les K-vecs ***[Fung and Church, 1994]***
- **Idée :**
  - diviser chaque texte du corpus bilingue en K régions (uniformes).
  - associer à chaque mot w (source et cible), un vecteur de positions p où $p[i]$ vaut 1 si w est dans la région $i \in [1, K]$ du texte.
  - la similarité de deux mots est donnée par la similarité de leur K-vec.
- les auteurs proposent de représenter des mots par des vecteurs de positions relatives v :

$$v[i] = p[i+1] - p[i] \quad \forall i \in [1, K-1]$$

où p est ici un vecteur positionnel dont la dimension correspond à la taille du texte (comptée en mots).

- un algorithme de programmation dynamique est proposé pour aligner deux vecteurs (car ils n'ont pas forcément la même dimension)

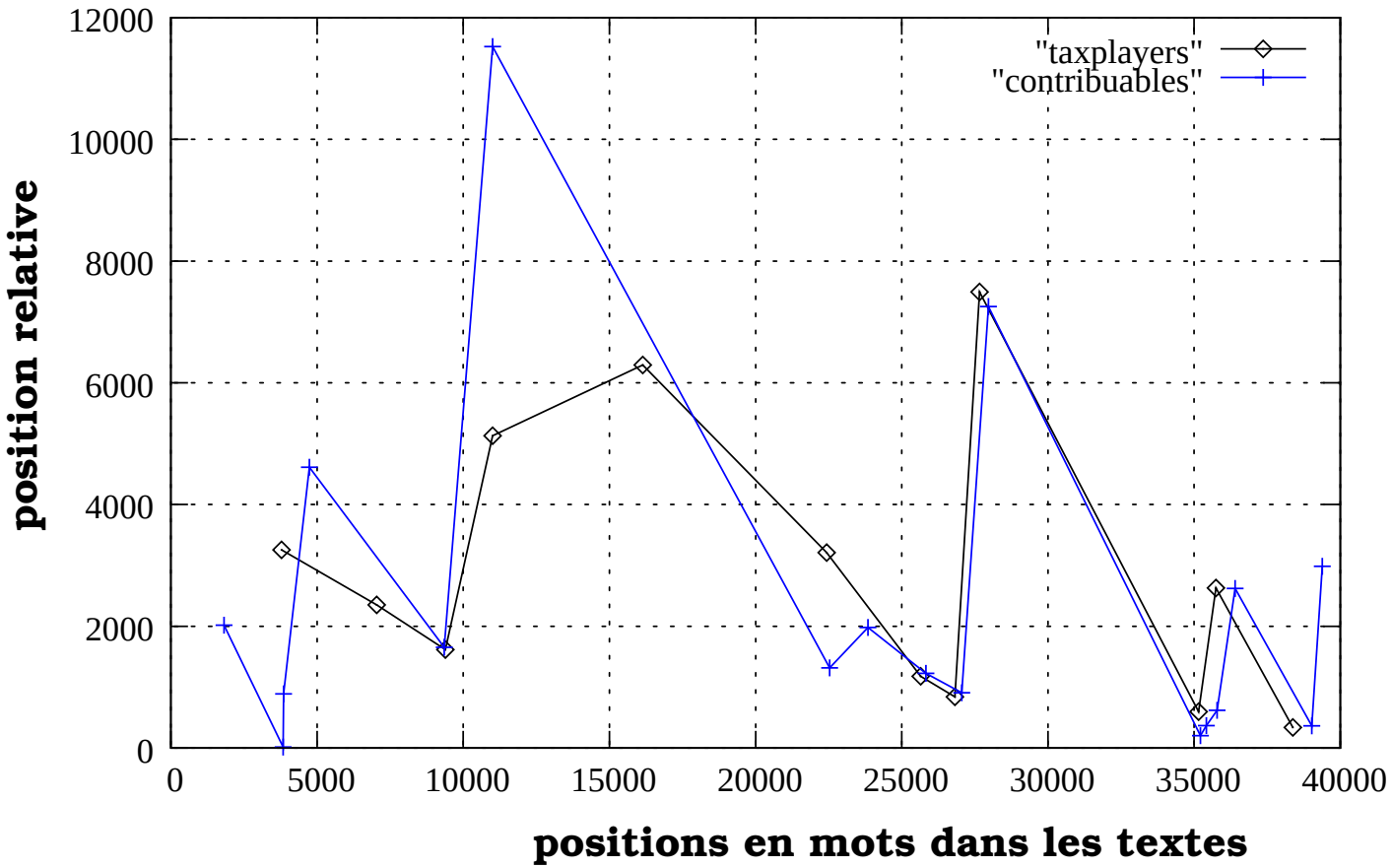
rali

Exemple de représentation des mots

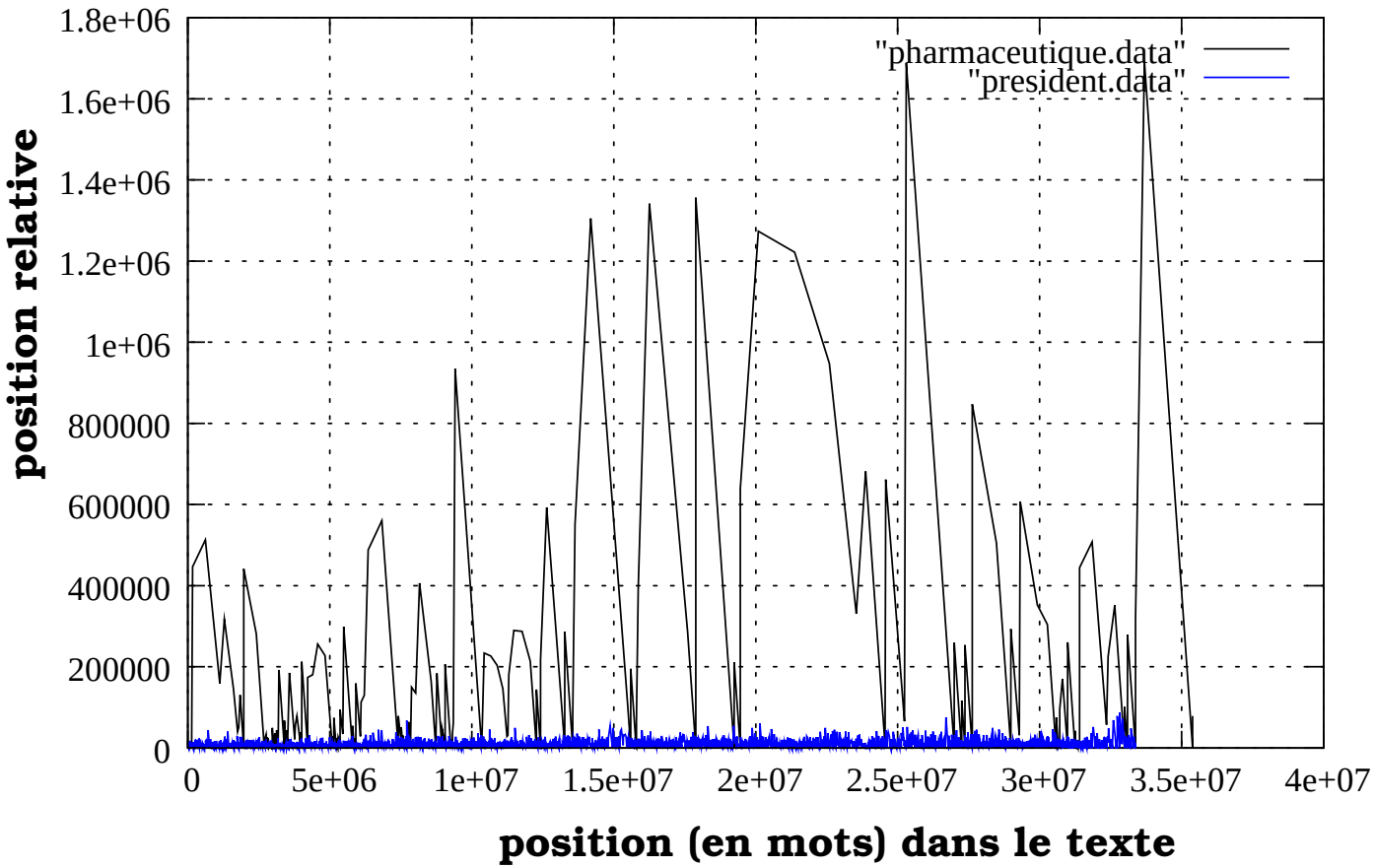
- dans une tranche du Hansard de 42 799 mots anglais et sa traduction française de 43 607 mots (1 867 phrases) :
  - contribuables : 1821 3839 3853 4740 9353 11006 22529 23846 25826 27054 27961 35213 35416 35782 36400 39023 39385 42369
  - taxpayers : 3785 7041 9393 11008 16139 22433 25643 26819 27660 35153 35749 38376 38712
- leur encodage relatif :
  - contribuables : 2018 14 887 4613 1653 11523 1317 1980 1228 907 7252 203 366 618 2623 362 2984
  - taxpayers : 3256 2352 1615 5131 6294 3210 1176 841 7493 596 2627 336

rali

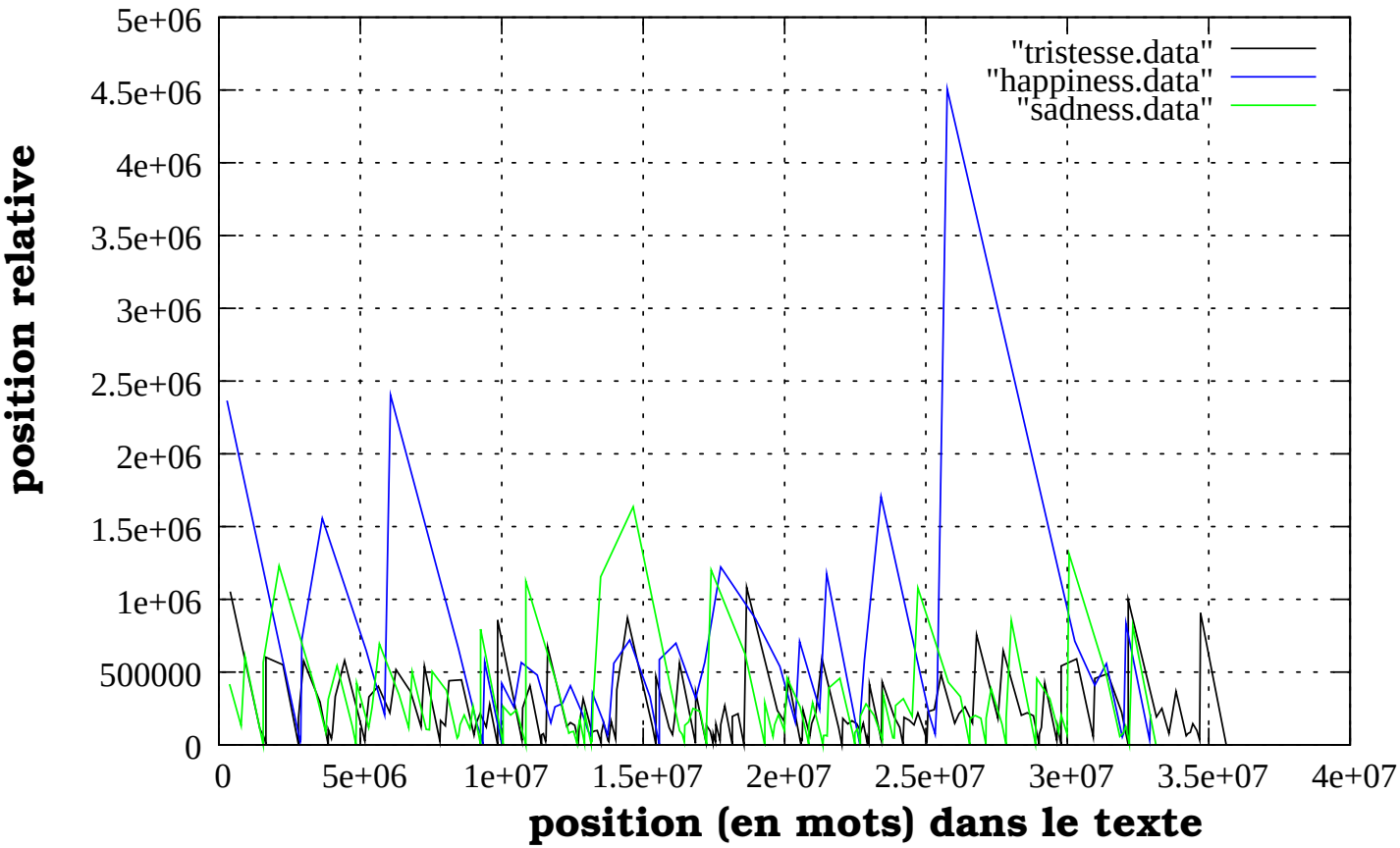
Exemple de représentation des mots



Exemple de représentation des mots



Exemple de représentation des mots



Identifier des traductions dans des corpus bilingues non parallèles

- **Challenge :** étant donnés deux textes de deux langues différentes, non corrélés (*à priori*). Peut-on découvrir automatiquement des paires de mots source/cible qui sont traduction l'un de l'autre ?
- **Quiz :** avez-vous une idée du type d'indice que l'on pourrait tenter d'exploiter ?

SRC Le ciel est bleu, pas rouge.

TGT All of a sudden, with a reassessment out of the clear blue sky this older couple is now faced with a $72,000 tax bill.

rali

Trouver des traductions dans des corpus bilingues non parallèles

- *[Rapp, 1995]* : les co-occurrences fortes dans une langue sont également des co-occurrences fortes dans une autre langue
  - si nous savons que bleu se traduit par blue

    SRC Le ciel est bleu, pas rouge.

    TGT All of a sudden, with a reassessment out of the clear blue sky thi couple is now faced with a $72,000 tax bill.
  - alors on peut (peut-être) en déduire que ciel et sky se correspondent
- voir *[Fung, 1995, Tanaka and Iwasaki, 1996, Tanaka and Matsuo, 1999, Ohmori and Higashida, 1999]* pour des variantes

rali

Illustration de l'idée développée dans [Rapp, 1995]

	1	2	3	4	5	6
$blue_1$		x			x	
$green_2$	x		x			
$plant_3$		x				
$school_4$						x
sky_5	x					
$teacher_6$				x		

	1	2	5	6	3	4
$blue_1$		x	x			
$green_2$	x				x	
sky_5	x					
$teacher_6$						x
$plant_3$		x				
$school_4$				x		

	1	2	3	4	5	6
$blau_1$		x	x			
$grün_2$	x				x	
$Himmel_3$	x					
$Lehrer_4$						x
$Pflanze_5$		x				
$Schule_6$				x		

Ressources utilisées dans [Rapp, 1999]

1. un corpus allemand (135 millions de mots du *Frankfurter Allgemeine Zeitung* couvrant une période de 1993 à 1996)
2. un corpus anglais (163 millions de mots du *Guardian* couvrant une période de 1990 à 1994)
3. un lexique bilingue de base allemand $\rightarrow$ anglais (16380 entrées extraites du *Collins Gem German Dictionnary*)
4. une liste de mots allemands (test) dont on cherche la traduction (100 mots) ; une traduction privilégiée a été associée à la main pour des fins d'évaluation

Note : Les corpus allemand et anglais ont été lemmatisés pour limiter le vocabulaire. Les mots outils sont également retirés.

rali

Étude décrite dans [Rapp, 1999]

- **Étape 1 :** calcul d'une matrice de co-occurrence pour la langue anglaise.

  lignes : les différents types du corpus anglais de fréquence 100 ou plus
  colonnes : les mots anglais qui apparaissent dans le lexique de base
  score : rapport de vraisemblance, les lignes sont ensuite normalisées

Étude décrite dans [Rapp, 1999]

- **Étape 2 :** calcul pour chaque mot test allemand d'un vecteur de co-occurrence. Même méthode que précédemment.
  - X un vecteur pour chaque mot allemand
- **Étape 3 :** Comparer X avec les vecteurs Y_j de la matrice de co-occurrence et sélectionner le mot anglais ayant le meilleur taux de similarité.
  - La similarité entre deux vecteurs A et B est ici :

$$s = \sum_i |A_i - B_i|$$

rali

Étude décrite dans [Rapp, 1999]

allemand	attendu	rang	top-5	
Baby	baby	1	baby	child mother daughter father
Brot	bread	1	bread	cheese meat food butter
Frau	woman	2	man	woman boy friend wife
gelb	yellow	1	yellow	blue red pink green
Haüschen	cottage	2	bungalow	cottage house hut village
Kind	child	1	child	daughter son father mother
Kohl	cabbage	17074	Major	Kohl Thatcher Gorbachev Bush
Musik	music	1	music	theatre musical dance song
Tabak	tobacco	1	tobacco	cigarette consumption nicotine drin
weiss	white	46	know	say thought see think
Whisky	whiskey	11	whisky	beer Scotch bottle wine

- 65% de traductions correcte en 1ère position
- 72% des traductions en tête jugées correctes
- 89% des mots traduits si on considère les 10 premières traductions

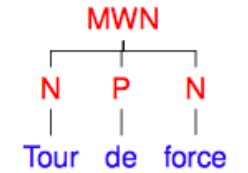
rali

Plan

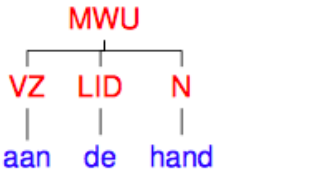
rali

[Sangati and van Cranenburgh, 2015]

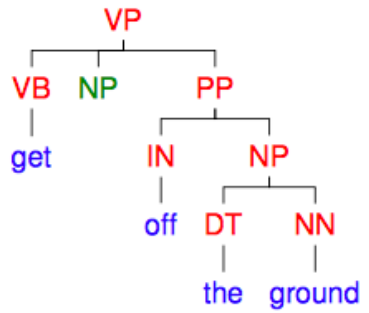
- grammaire DOP entrainée sur un corpus arboré (*Treebank*)
  - ***[Abeillé et al., 2000]*** Français, arbres : 13K, annotation (à plat) des MWEs par catégorie
  - ***[van Noord, 2009]*** Hollandais, arbres : 52K, annotation (à plat) des MWEs
  - GigaWord annoté : Anglais, arbres : 180K (automatique), pas d'annotation des MWEs (`http://catalog.ldc.upenn.edu/LDC2012T21`)



(a) French treebank



lit.: *on the hand*, "going on."
(b) Dutch Lassy treebank



(c) Annotated English Gigaword

- détection avec parsing (lorsque le *Treebank* marque les MWE), avec mesures d'association sur les fragments d'arbres (sinon)

[Sangati and van Cranenburgh, 2015]

Exp 1 – parsing

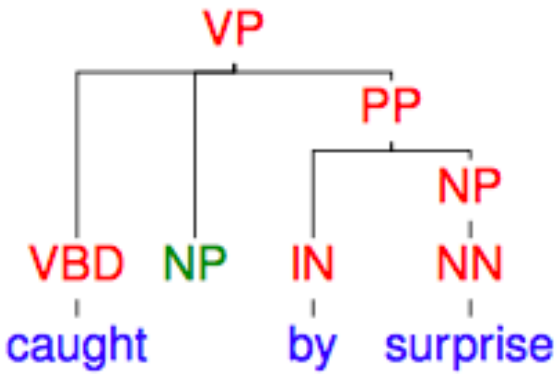
- fragments d'arbres extraits des treebanks à l'aide de `FragmentSeeker` (recherche des sous-arbres récurrents)
- parsing avec modèle 2DOP implémenté dans `disco-dop` (`https://github.com/andreasvc/disco-dop`)
- expériences sur le français et le hollandais
  - F1f-measure, EX tree accuracy, MWE-F1 f-measure sur l'identification des MWEs

Parser	F1	EX	MWE-F1
FRENCH			
Green et al. (2013): DP-TSG	76.9	16.0	71.3
Green et al. (2013): Stanford	79.0	17.6	70.5
disco-dop, 2DOP	**79.3**	**19.9**	**71.9**
DUTCH			
disco-dop, PCFG baseline	63.9	21.8	50.4
disco-dop, 2DOP	**77.0**	**35.2**	**75.3**

Table 3: Performance of the parsing models on the French and Dutch treeba

[Sangati and van Cranenburgh, 2015]
Exp 2 : mesures d'associations

- chaque fragment est noté par des mesures d'association



- où un fragment est vu comme la séquence $s = s_1, \ldots, s_n$ des mots/tags à la frontière (caught/VBD ◦/NP by/IN surprise/NN)
- et $\Sigma(s)$ les partitions contiguës de s
  - ex : $\Sigma(s_1, s_2, s_3) = \{((s_1, s_2), s_3), (s_1, s_2, s_3), (s_1, (s_2, s_3))\}$

rali

[Sangati and van Cranenburgh, 2015]

Exp 2 : mesures d'associations

PMI (multivariée)

$$PMI(s_1, \ldots, s_n) = \log \frac{p(s_1, \ldots, s_n)}{\prod_i p(s_i)}$$

LLR où σ indique une partitions de $s \equiv s_1, \ldots, s_n$

$$LLR(s_1, \ldots, s_n) = \log \frac{p(s_1, \ldots, s_n)}{\sum_{\sigma \in \Sigma(s)} \prod_{s \in \sigma} p(s)}$$

LIR où $inside$ est la somme des prob. de toutes les dérivations menant à ce fragment

$$\log \frac{p(frag)}{inside(frag)}$$

rali

[Sangati and van Cranenburgh, 2015]

Exp 2 : mesures d'associations

PMI	Freq.	Sequence Pattern
15.3	13	VB_take IN_into NN_account
9.8	5	VB_take NN_responsibility IN_for
9.7	8	VB_take NN_credit IN_for
9.3	12	VB_take DT_a NN_look
8.4	88	VB_take NN_advantage IN_of
8.4	7	VB_take NN_place IN_on
8.3	6	VB_take NN_effect IN_in
8.1	14	VB_take NNS_steps TO_to
...	...	...
4.6	6	VB_take DT_the NN_money

Table 6: A sample of English fragments conforming to the sequence pattern VB_take L L, sorted by PMI.

rali

[Sangati and van Cranenburgh, 2015]

Exp 2 : mesures d'associations

PMI	Freq.	Sequence Pattern
18.0	6	VB_take NP IN_into NN_account
14.6	6	VB_take NP IN_for VBN_granted
13.6	7	VB_take DT NN_look IN_at
12.9	6	VB_take NP TO_to NN_court
12.5	6	VB_take NN RB_away IN_from
12.4	17	VB_take NP RB_away IN_from
12.0	6	VB_take JJ NN_action TO_to
11.2	5	VB_take NP RB_away IN_from
10.5	6	VB_take QP NNS_years TO_to
8.3	10	VB_take DT NN_time TO_to

Table 5: List of English fragments conforming to the sequence pattern VB_take X L L, sorted by PMI.

rali

[Rondon et al., 2015]

- récupère 40 nouvelles journalistiques par jour
- extraction des MWEs nominales (ex : business hours) à l'aide de `mwetoolkit`
  - exp. régulières sur les POS (TreeTagger)
- classificateur pour promouvoir certaines MWEs candidates
- testé sur plusieurs itérations
  - init : 1100 MWEs candidates annotées par 2 annotateurs (kappa de 0.85)

rali

TSRali.com *[Huet and Langlais, 2012]*

- Aptitude de TSRALI.com à identifier des traductions d'idiomes
  - 1 467 idioms (fréquents) et leur traduction extraits de *[Piat, 2008]*

French		English
Il est agile comme un singe	R	*He's as nimble as a goat*
	G	*He is agile as a monkey*
Elle était sur son trente et un	R	*She was dressed to kill*
	R	*She was all dressed up*
	G	*She was on her thirty-one*
(Je vais d'abord) me rincer la dalle — familiar —	R	*(I'm going to) wet my whistle (first)*
	G	*First I'll rinse my slab*
(Il aime) rouler des mécaniques — familiar —	R	*(He likes) flexing his muscles*
	R	*(He likes) playing the tough guy*
	G	He loves rolling mechanical
J'ai vu trente-six chandelles	R	*I saw stars*
	G	*I saw thirty-six candles*

TSRali.com *[Huet and Langlais, 2012]*

- % de match dans la mémoire de TSRali selon le thème de l'expression idiomatique

	English queries		French queries	
1	Behaving	69 %	Behaving	68 %
2	Discussion	68 %	Feelings and emotions	62 %
3	Time, age and experience	65 %	Discussion	61 %
	...		...	
18	Love, sex and seduction	14 %	Human body and physical activity	22 %
19	Weather	14 %	Clothing and fashion	13 %
20	Drinking and eating	8 %	Weather	10 %

TSRali.com *[Huet and Langlais, 2012]*

- rappel (% à la référence) mesuré sur 238 requêtes formes idiomatiques où la requête et la traduction de référence existent dans la mémoire (pas forcément dans une même paire de phrases)

k	1	2	3	5	10	∞
English queries	41.6	56.3	59.2	65.1	69.3	74.8
French queries	41.8	49.8	54.9	62.9	69.6	76.8

$\Rightarrow$ algorithme de transpotting ok.

TSRali.com *[Huet and Langlais, 2012]*

- rappel (% à la référence) mesuré sur toutes les requêtes :

		k	1	2	3	5	10
English queries	GOOGLE		12.3				
	TSRALI		8.0	10.6	11.3	12.4	13.3
French queries	GOOGLE		12.6				
	TSRALI		7.6	9.1	10.3	11.9	13.2

- 50% des requêtes sans match dans la mémoire

TSRali.com *[Huet and Langlais, 2012]*

- rappel (% référence) mesuré sur les requêtes (700) avec un moins un match dans la mémoire

	k	1	2	3	5	10
English queries	GOOGLE	15.7				
	TSRALI	16.7	22.3	23.6	26.0	27.9
French queries	GOOGLE	16.7				
	TSRALI	15.9	18.9	21.4	24.7	27.5

- 50% des requêtes sans match dans la mémoire

TSRali.com *[Huet and Langlais, 2012]*

appeler un chat un chat	J1	J2	J5
▷ we should call it what it is	correct	correct	correct
▷ we can say the d word and the m word	correct	wrong	partial
▷ calling manure a rose doesn't change the smell	correct	wrong	partial

manger à tous les râteliers	J1	J2	J5
▷ slurps at everyone 's trough	correct	correct	correct
▷ double - dipper	partial	correct	partial
▷ them pot lickers and accusing them of being at the trough and pork barrelling	wrong	partial	wrong

- TSRali identifie une traduction identifiée correcte par au moins un juge dans 97 de 100 requêtes (évaluation manuelle)

correct	partial	wrong	avr	$rank$
42%	22%	36%	13.4	1.4

Abeillé, A., Clément, L., and Kinyon, A. (2000).
Building a treebank for french.
In *In Proceedings of the LREC 2000*.

ACL-31 (1993).
Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL), Columbus, Ohio.

ACL-33 (1995).
Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), Cambridge, Massachusetts.

ACL-37 (1999).
Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), College Park, Maryland.

Alvarez, C., Langlais, P., and Nie, J.-Y. (2004).
Word pairs in language modeling for information retrieval.
In *7th Conference on Computer Assisted Information Retrieval (RIAO)*, pages 686–705, Avignon, France.

rali

Church, K. and Hanks, P. (1989).
Word association norms, mutual information, and lexicography.
In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 76–83, Vancouver, British Columbia.

COLING-94 (1994).
Proceedings of the International Conference on Computational Linguistics (COLING) 1994, Kyoto, Japan.

COLING-96 (1996).
Proceedings of the International Conference on Computational Linguistics (COLING) 1996, Copenhagen, Denmark.

Dagan, I., Marcus, S., and Markovitch, S. (1993).
Contextual word similarity ans estimation from sparse data.
In [ACL-31, 1993], pages 164–171.

Dunning, T. (1993).
Accurate methods for the statistics of surprise and coincidence.
Computational Linguistics, 19(1).

Evans, D. A. and Zhai, C. (1996).
Noun-phrase analysis in unrestricted text for information retrieval.
In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 17–24, Santa Cruz, California.

Fung, P. (1995).
A pattern matching method for finding noun and proper noun translations from noisy parallel corpora.
In [ACL-33, 1995], pages 236–243.

Fung, P. and Church, K. (1994).
K-vec : A new approach for aligning sentences in bilingual corpora.
In [COLING-94, 1994], pages 1096–1102.

Gaussier, E. (1995).
Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues.
PhD thesis, Université de Paris 7.

rali

Gonnet, G., Baeza-Yates, R., and Snider, T. (1992).
New Indices for Text : PAT trees and PAT arrays, chapter Information Retrieval : Data Structures and Algorithms.
Information Retrieval : Data Structures and Algorithms. B. Frakes and R. Baeza-Yates (eds.), Englewood Cliffs : Prentice Hall.

Haruno, M., Ikehara, S., and Yamazaki, T. (1996).
Learning bilingual collocations by word-level sorting.
In [COLING-96, 1996], pages 525–530.

hua Chen, K. and Chen, H.-H. (1994).
Extracting noun phrases from large-scale texts : A hybrid approach and its automatic evaluation.
In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 234–241, Las Cruces, New Mexico.

Huet, S. and Langlais, P. (2012).
Translation of Idiomatic Expressions across Different Languages : A Study of the Effectiveness of TransSearch,

rali

chapter Where Humans Meet Machines. Innovative Solutions for knotty Natural-Language Problems.
Springer.

Ikehara, S., Shirai, S., and Uchino, H. (1996).
A statistical method for extracting uninterrupted and interrupted collocations from very large corpora.
In [COLING-96, 1996], pages 574–579.

Jackendoff, R. (1997).
The Architecture of the Language Faculty.
MIT Press.

Kupiec, J. (1993).
An algorithm for finding noun phrase correspondences in bilingual corpora.
In [ACL-31, 1993], pages 17–22.

Lee, L. (1999).
Measures of distributional similarity.
In [ACL-37, 1999], pages 25–32.

rali

Lee, L. and Pereira, F. (1999).
Distributional similarity models : Clustering vs. nearest neighbors.
In [ACL-37, 1999], pages 33–40.

Lin, D. (1998).
Automatic retrieval and clustering of similar words.
In *COLING/ACL*, Montreal.

Lin, D. (1999).
Automatic identification of non-compositional phrases.
In *ACL*, pages 317–324.

Maarek, Y. S., Berry, D. M., and Kaiser, G. E. (1991).
An information retrieval approach for automatically constructing software librairies.
In *IEEE Transactions on Software Engineering*, volume 17(8), pages 800–813.

Manber, U. and Myers, G. (93).
Suffix arrays : a new method for on-line string searches.

rali

SIAM Journal on Computing, 22(5) :935–948.

Manning, C. D. and Schütze, H. (1999).
Foundations of Statistical Natural Language Processing.
MIT Press.

Martin and van Sterkenburg (1983).
On the processing of a text corpus : from textual data to lexicographic information.
In Ed., H., editor, *Lexicography : Principles and Practices*.
London Academic.

Melamed, I. D. (1997).
Automatic discovery of non-compositional compounds in parallel data.
In *EMNLP*, Providence, RI.

Mel'cuk, I. (2011).
Le figement lingusitique : la parole entrâvée, chapter Phrasèmes dans le dictionnaire, pages 41–61.
Paris : Honoré Champion.

rali

Nagao, M. and Mori, S. (1994).
A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese.
In [COLING-94, 1994], pages 611–615.

Ohmori, K. and Higashida, M. (1999).
Extracting bilingual collocations from non-aligned parallel corpora.
In [TMI-8, 1999], pages 88–97.

Piat, J.-B. (2008).
It's raining cats and dogs et autres expressions idiomatiques anglaises.
Librio. J'ai lu.

Rapp, R. (1995).
Identifying word translation in non-parallel texts.
In [ACL-33, 1995], pages 320–322.

Rapp, R. (1999).

Automatic identification of word translations from unrelated english and german corpora.
In [ACL-37, 1999], pages 519–526.

Ries, K., Buo, F. D., and Wang, Y.-Y. (1995).
Improved language modeling by unsupervised acquisition of structure.
In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1995*, pages 193–196, Detroit, Michigan. IEEE.

Rondon, A., Caseli, H., and Ramisch, C. (2015).
Never-ending multiword expressions learning.
In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 45–53, Denver, Colorado.

Russell, G. (1998).
Identification of salient token sequences.
Internal Report, RALI.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002).

Multiword expressions : A pain in the neck for nlp.
In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15.

Sangati, F. and van Cranenburgh, A. (2015).
Multiword expression identification with recurring tree fragments and association measures.
In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 10–18, Denver, Colorado.

Shimohata, S., Sugio, T., and Nagata, J. (1997).
Retrieving collocations by co-occurrences and word order constraints.
In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 476–481, Madrid, Spain.

Smadja, F. (93).
Retrieving collocations from text : Xtract.
Computational Linguistics, 19(1) :144–177.

rali

Tanaka, K. and Iwasaki, H. (1996).
Extraction of lexical translations from non-aligned corpora.
In [COLING-96, 1996].

Tanaka, T. and Matsuo, Y. (1999).
Extraction of translation equivalents from non-parallel corpora.
In [TMI-8, 1999], pages 109–119.

TMI-8 (1999).
Proceedings of the 8th Conference on Theoretical and Methodological Issues in Machine Translation (TMI), Chester, England.

van Noord, G. (2009).
Huge Parsed Corpora in LASSY, volume 12, pages 115 – 126.
Lot occasional series edition.