# Extraction de collocations à partir de textes

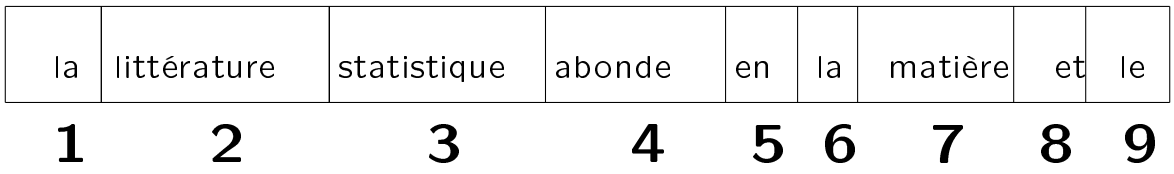
1. **Statistiques**
   $\Rightarrow$ critère d'« habituel »

2. **Statistiques + analyse linguistique**
   (a) analyse morphosyntaxique
   (b) analyse syntaxique
   $\Rightarrow$ critères d'« habituel » + « syntactiquement bien formé »

3. **Statistiques + analyse linguistique + ressources lexicales**
   $\Rightarrow$ critères d'« habituel » + « syntactiquement bien formé » + « lexicalement transparent »

# Mesures statistiques (1)

## Méthode de la fenêtre




⋆ **Fréquence**

Exemple extrait de (Manning et Schütze, 1999)

| **C (mot1, mot2)** | **mot1** | **mot2** |
|---|---|---|
| 80871 | of | the |
| 58841 | in | the |
| 26430 | to | the |
| 21842 | on | the |
| ... | | |

⋆ **Information mutuelle**
(Fano, 1961)

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

# Mesures statistiques (2)

(Church, 1990)

$$IM(x,y) \;=\; \log_2 N \frac{C(x,y)}{C(x)C(y)}$$

Exemple extrait de (Church, 1990) : corpus de 48 millions de mots, fenêtre de 5 mots, C(x,y) $\geq$ 5.

| **IM(x,y)** | **C(x,y)** | **x** | **y** |
|---|---|---|---|
| 11,3 | 12 | honorary | doctor |
| 11,3 | 8 | doctors | dentists |
| 10,7 | 30 | doctors | nurses |
| 9,4 | 8 | doctors | treating |
| 9,0 | 6 | examined | doctor |
| 8.9 | 11 | doctors | treat |
| ... | | | |
| 0,96 | 6 | doctor | with |
| 0,95 | 41 | a | doctors |
| 0,93 | 12 | is | doctors |

# Mesures statistiques (3)

⋆ **z-score**

Fréquence relative d'un mot ou groupes de mots par rapport à la fréquence moyenne de tous les autres mots du texte

$$z = \frac{C(x,y) - E}{\sqrt{Eq}} \text{ où } q = 1 - p$$

avec :

$N$ le nombre total d'éléments lexicaux du texte

$x$ un mot donné apparaissant $C(x)$ dans le texte

$y$ un collocatif de x apparaissant $C(y)$ dans le texte

$C(x,y)$ le nombre de co-occurrences entre $y$ et $x$

$S$ la taille de la fenêtre, c'est-à-dire la distance maximale autorisée entre $x$ et $y$

$p$ la probabilité d'occurreence de y n'importe où dans le texte sauf avec $x$ : $p = \frac{C(y)}{N - C(x)}$

$E$ le nombre probable de coocurrences entre $x$ et $y$ : $E = pC(x)S$

# Mesures statistiques (4)

Exemple extrait de (Berry-Rogghe, 1974) : livre de Dickens *A Christmas Carol*, “house”, $S$ non précisé.

| $z(house, y)$ | $y$ |
|---|---|
| 24.0500 | sold |
| 21.2416 | commons |
| 19.9000 | decorate |
| 13.3937 | this |
| 11.9090 | empty |
| 10.5970 | buying |
| 10.5970 | painting |
| ... | |
| 2.1721 | is |
| 2.0736 | every |
| ... | |
| -0.0385 | his |
| ... | |

**Mesure apparentée** : t-test ou t-score

# Mesures statistiques + analyse linguistique (1)

⋆ Classement en fonction des parties du discours :

1. Nom + Ajectif : *amour platonique*, *colère noire*
2. Nom + Nom : *bourreau des cœurs*
3. Verbe + Adverbe : *exploiter efficacement*
4. Adjectif + Adverbe : *sexuellement transmissible*
5. Nom + Verbe : *commettre une agression*, *retirer de l'argent*

## Mesures statistiques et analyse morphosyntaxique

Méthode :

– filtrage sur les parties du discours
– grammaires locales décrivant des syntagmes
– application de mesures pour ordonner les candidats

⋆ mwetoolkit : a Framework for Multiword Expression Identification (Ramish et al. 2010)

⋆ Application à l'extraction de combinaisons lexicales spécialisées ou termes complexes

Exemple : Livre Bleu du CCITT - 800 000 mots -

| Fag | **N Adj** |
|---|---|
| 0,885 | télégraphie harmonique |
| 0,852 | compte tenu |
| 0,849 | polynôme générateur |
| 0,828 | faute matérielle |
| 0,800 | période probatoire |
| 0,796 | assemblée plénière |
| 0,787 | accord bilatéral |
| 0,755 | signe diacritique |
| 0,750 | océan indien |
| 0,683 | étude ultérieure |
| ... | |

# XTRACT

**Définition :**
Une collocation est une séquence de mots, syntactiquement bien formé et spécifique au domaine. Trois étapes :

1. Extraction de bigrammes significatifs
2. Bigrammes $\rightarrow$ n-grammes
3. Ajout d'analyse syntaxique partielle

## 1. Extraction de bigrammes significatifs

**a** Production de concordances
Étant donné un mot w, rechercher toutes les phrases contenant w.

**b** Compilation et tri
Produit une liste de mots $w_i$ avec leur fréquence et l'information de comment w et $w_i$ cooccurrent.
$S(w, w_i) = (C(w_i), POS(w_i), -5 \leq j \leq 5\ C_j(w_i))$

Exemple : *takeover* (w) ; *possible* ($w_i$)
C(*possible*) = 178, POS(*possible*) = adj,

| $C_{-5}$ | $C_{-4}$ | $C_{-3}$ | $C_{-2}$ | $C_{-1}$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 13 | 4 | 23 | 138 | 0 | 0 | 0 | 0 | 0 |

# XTRACT

**c** Analyse

Fournit une liste de couples de mots associés à 3 indices :

- critère d'association : z-score($w_i$) $\geq z_0$
- variance des distances : $\cup_i \geq \cup_0$ (permet d'identifier des distances privilégiées)
- distance : $C_j(w_i) \geq \overline{C(w_i)} + (k_1 + \sqrt{\cup_i})$

($z_0$, $k_1$, $\cup_0$) = (1, 1, 10)

## 2. Bigrammes $\rightarrow$ n-grammes

**Objectifs :**

- Fournir des collocations de plus de 2 mots
- Filtrer de mauvais bigrammes

Les trois étapes de **1.** à partir d'un couple (w, $w_i$)

| Bigramme | N-gramme |
|---|---|
| average-industriel | the Dow Jones industrial average |
| composite-index | the NYSE's composite index of all its listed common stocks fell *NUMBER* to *NUMBER* |

# XTRACT

## 3. Ajout d'analyse syntaxique partielle

**Étapes :**
**a** Production de concordances sur le corpus étiqueté
**b** Analyse de chaque phrase CASS (Abney, 1990) pour obtenir des associations syntaxiques binaires :

VO verbe-objet
SV sujet-verbe
NJ nom-adjectif
NN nom-nom

Pour chaque bigramme (w, $w_i$), assignation d'étiquettes syntaxiques.

| Phrase | Étiquette |
|---|---|
| . . . when they **rose** pork prices 1.1 percent . . . | VO |
| Bond **prices rose** because many traders tool the report as a signal . . . | SV |
| Stock **prices rose** in moderate trading today with little news . . . | SV |

# XTRACT

**c** Filtrage et étiquetage des collocations
Ne sont conservés que les bigrammes (w, $w_i$) pour lequels la fréquence d'une étiquette syntaxique est supérieure à un certain seuil.

## Évaluation

1. 4000 collocations expertisées par un lexicographe produites après les 2 premières étapes : précision de 40 %
2. Sur les 4000 collocations, 40 % sont retenues à l'étape **3**, 80 % sont acceptées par le lexicographe.

## Conclusion

– Précision élevée car beaucoup de filtrage
– collocations ?

# Mesures statistiques et analyse syntaxique

⋆ Identification des collocations de type 5 : Nom + Verbe
◦ Arguments privilégiés d'un verbe
◦ Construction à verbe support

– Analyse syntaxique où les différents arguments sont correctement identifiés
– Définir des chemins dans l'arbre syntaxique

Extrait de l'analyse syntaxique de la phrase "Der Student führe komplexe Berechnungen durch" (*L'étudiant ferait des calculs complexes*) fourni par GEPARD (P. Ludewig, 2001)

**Arbre :**
(('S', 1), [('NPn', 2), ('VP', 7)])
(('NPn', 2), [('NP', 3)])
(('NP', 3), [('DET', 4), ('N1', 5)])
(('DET', 4), [('WORT', 'der', 0)])
(('N1', 5), [('N', 6)])
(('N', 6), [('WORT', 'Student', 1)])
*(('VP', 7), [('Vstamma', 8), ('VCOMPa', 9), ('Vpfx', 16)])*

**Structures d´attribut :**
*(('S',1)*, {'SUBJNUM' :['SING'], 'STYPE' :['MAIN'],
'VSUBCAT' :['SUBCATA'],
*'DIATHESE' :['AKTIV']*,
'WORTSTELLUNG' :['SVO'],
'TEMP' :['PRES'], 'MODUS' : ['KONJUNKTIV'],
'SUBJCASE' :['NOM'],
'SUBJCONSTR' :['NPLEVELNORMAL'],
'SUBJPER' : ['THIRD']})
*(('NPn',2)*, {'PER' :['THIRD'],
'ZUINFNOUN' :['NOTZUINF'],
*'DETERMINED' :['ISDETERMINED'],*
*'NUM' :['SING']*,'RELNUM' :['NONUM'],
'GENDER' :['MASC'],
'NPLEVEL' :['NPLEVELNORMAL'],
'RELGENDER' :['NOGENDER'], *'CASE' :['NOM']*})

**chemin verbal :** [('WORT','führe','V'),
('Vstamma',8),
*('VP',7)*,
('S',1)]

**chemin nominal :** [('WORT','Berechnungen','N'),
('N',15),
('N1',11),
('NPa',10),
('VCOMPa',9),
*('VP',7)*,
('S',1)]

RANLP 2013 Cours
Violetta Seretan : Collocation extraction based on Syntactitc criteria

# Acquisition de verbe support

Travaux de Grefenstette et Teufel (1995)

1. liste des formes nominalisées dont il faut découvrir le Vsup
2. analyse syntaxique à l'aide de SEXTANT
3. extraction de 2 types de schémas :
   - Vsup + nominalisation + GP
   - Verbe + GP
4. conservation des formes nominalisées qui sélectionnent le même GP que le verbe
5. extraction des verbes qui accceptent une forme nominalisée en OD
6. conservation des verbes les plus fréquents

**Résultats :**

| | | |
|---|---|---|
| *demand-demand* | *for, in, of* | meet(58), press(34) increase(22) |
| *propose-proposal* | *in, for, to* | make(28), reject(26) submit(19) |

# Mesures statistiques, analyse syntaxique sur corpus alignés

« syntactiquement bien formé » + « arbitraire »

**Proposition :**
Les combinaisons collocatives sont des expressions qui sont difficiles à reproduire pour un locuteur étranger :

1. Cas des constructions à verbe support : pas de traduction évidente du verbe
   - *N0 faire un résumé de N1* $\rightarrow$ *N0 make a summary of N1*
   - *N0 faire une promenade* $\rightarrow$ *N0 take a walk*
   - *N0 faire un baiser à N1* $\rightarrow$ *N0 give a kiss to N1*
2. Cas de transfert complexe entre deux langues : non correspondance lexicale entre les têtes prédicatives
   - *porter un jugement* $\rightarrow$ *beurteilen*
   - *apprendre par cœur* $\rightarrow$ *memorize*