

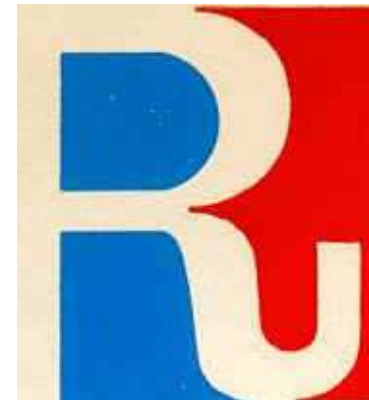
# ***Découverte automatique de mots similaires***

***Exemple de techniques de fouille de textes***

Pierre Senellart



Université  
catholique  
de Louvain



1. Introduction
- ✓ Problématique
2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion

# Introduction : Problématique

- ⑥ **But ultime** : recherche automatique de synonymes.
- ⑥ **Problème plus réaliste** : on recherche des *mots similaires* ou des *quasi-synonymes*
- ⑥ **Applications** :
  - △ Langage naturel
  - △ Recherche documentaire
  - △ Aide aux lexicographes

# 1. Introduction

## ✓ Plan

2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion

# Plan de l'exposé

## 1. Introduction

## 2. Grands corpus

- ✓ Principe
- ✓ Espace vectoriel / documents
- ✓ Thésaurus de mots rares
- ✓ SEXTANT

## 3. Web

- ✓ Particularités
- ✓ Test de synonymie

## 4. Dictionnaires monolingues

- ✓ Graphe du dictionnaire
- ✓ Algorithmes considérés
- ✓ Résultats expérimentaux

## 5. Conclusion

1. Introduction
2. Grands corpus  
✓ Principe
3. Web
4. Dictionnaires
5. Conclusion

# Principe

- ⑥ **Principe général** : *Deux mots similaires sont deux mots qui apparaissent dans les mêmes contextes.*
- ⑥ Définition de contexte ?
- ⑥ Quelle mesure de similarité ?

1. Introduction

2. Grands corpus

✓ EV orienté documents

3. Web

4. Dictionnaires

5. Conclusion

# Espace vectoriel orienté documents

⑥ **Dimensions** : documents  
**Vecteurs** : termes

⑥ Deux mesures de similarité envisagées :

$$\cos(\mathbf{i}, \mathbf{j}) = \frac{\mathbf{i} \cdot \mathbf{j}}{\sqrt{\mathbf{i} \cdot \mathbf{i} \times \mathbf{j} \cdot \mathbf{j}}}$$

$$\text{cluster}(\mathbf{i}, \mathbf{j}) = \frac{\mathbf{i} \cdot \mathbf{j}}{\|\mathbf{i}\|_1}$$

1. Introduction
2. Grands corpus  
✓ Mots rares
3. Web
4. Dictionnaires
5. Conclusion

# Thésaurus de mots rares

- ⑥ Dimensions : termes
- ⑥ Vecteurs : documents
- ⑥ Clustering des documents suivant une mesure de similarité.
- ⑥ Recherche des **discriminants indifférents** à l'intérieur de chaque classe
- ⑥ Approximation : les mots rares sont des discriminants indifférents

1. Introduction
2. Grands corpus  
✓ SEXTANT
3. Web
4. Dictionnaires
5. Conclusion

# SEXTANT

**Contextes** : relations syntaxiques entres mots (un nom **est le sujet** d'un verbe, un nom **est modifié par** un adjectif, etc.)

$$poids(att) = 1 + \sum_{\text{nom } i} \frac{p_{att,i} \log(p_{att,i})}{\log(\text{nb total de relations})}$$

$$p_{att,i} = \frac{\text{nb de fois } att \text{ app. avec } i}{\text{nb total d'attributs de } i}$$

$$jac(i, j) = \frac{\sum_{att \text{ attribut des deux } i \text{ et } j} poids(att)}{\sum_{att \text{ attribut de soit } i \text{ soit } j} poids(att)}$$

1. Introduction
2. Grands corpus
3. Web  
✓ Particularités
4. Dictionnaires
5. Conclusion

## Particularités du Web

- ⑥ Taille énorme, intraitable.
- ⑥ Longs temps d'accès mais bonne indexation.
- ⑥ Pourtant : corpus le plus **vivant** et le plus **riche** qui existe

1. Introduction
2. Grands corpus
3. Web  
✓ Test de synonymie
4. Dictionnaires
5. Conclusion

## Test de synonymie sur le Web

j est-il un bon synonyme de i ?

$$score_1(\mathbf{j}) = \frac{hits(\mathbf{i} \text{ AND } \mathbf{j})}{hits(\mathbf{j})}$$

$$score_2(\mathbf{j}) = \frac{hits(\mathbf{i} \text{ NEAR } \mathbf{j})}{hits(\mathbf{j})}$$

$$score_3(\mathbf{j}) = \frac{hits((\mathbf{i} \text{ NEAR } \mathbf{j}) \text{ AND NOT } ((\mathbf{i} \text{ OR } \mathbf{j}) \text{ NEAR "not"}))}{hits(\mathbf{j} \text{ AND NOT } (\mathbf{j} \text{ NEAR "not"}))}$$

$$score_4(\mathbf{j}) = \frac{hits((\mathbf{i} \text{ NEAR } \mathbf{j}) \text{ AND context AND NOT } ((\mathbf{i} \text{ OR } \mathbf{j}) \text{ NEAR "not"}))}{hits(\mathbf{j} \text{ AND context AND NOT } (\mathbf{j} \text{ NEAR "not"}))}$$

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
  - ✓ Graphe
5. Conclusion

## Graphe du dictionnaire

- ⑥ **Nœuds** : mots définis dans le dictionnaire
- ⑥ **Arêtes** : il y a une arête entre le mot **a** et le mot **b** si **b** apparaît dans la définition de **a**
- ⑥ **Graphe de voisinage** : le graphe de voisinage d'un mot **a** est le sous-graphe formé de **a**, des fils de **a** et des parents de **a**.

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires  
✓ Algorithmes
5. Conclusion

## Distance entre vecteurs

- ⑥  $A$  matrice d'adjacence du graphe.
- ⑥ Distance entre le mot d'indice  $i$  et le mot d'indice  $j$  :

$$\|(A_{i,\cdot} - A_{j,\cdot})\| + \|(A_{\cdot,i} - A_{\cdot,j})^T\|$$

( $\| \cdot \|$  est une norme vectorielle donnée).

- ⑥ Proche des méthodes classiques de fouille de textes avec mesure de similarité.

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires  
✓ Algorithmes
5. Conclusion

⑥ **PageRank** (Google) : distribution stationnaire des poids des nœuds du graphe correspondant au vecteur propre principal de la matrice d'adjacence.

⑥ **ArcRank** :

$$r_{s,t} = \frac{p_s / |a_s|}{p_t}$$

$|a_s|$  est le degré sortant de  $s$ .  $p_t$  est le pagerank de  $t$ .

⑥ *Les meilleurs synonymes de  $i$  sont les extrémités des meilleurs arcs arrivant sur ou quittant le noeud  $i$ .*

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires  
✓ Algorithmes
5. Conclusion

# Comparaison de graphes

- ⑥ On compare le graphe du dictionnaire avec le graphe  $1 \longrightarrow 2 \longrightarrow 3$ .
- ⑥ Les bons nœuds 2 sont les nœuds pointés par des bons nœuds 1 et pointant vers des bon nœuds 3 :  
**définition mutuellement récursive**
- ⑥ **Mots similaires** : mots semblables à 2 dans le graphe du voisinage

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
- ✓ Résultats expérimentaux
5. Conclusion

## Résultats expérimentaux - disappear

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	vanish	vanish	epidemic	vanish	vanish
2	wear	pass	disappearing	go away	cease to exist
3	die	die	port	end	fade away
4	sail	wear	dissipate	finish	die out
5	faint	faint	cease	terminate	go
6	light	fade	eat	cease	evaporate
7	port	sail	gradually		wane
8	absorb	light	instrumental		expire
9	appear	dissipate	darkness		withdraw
10	cease	cease	efface		pass away
Mark	3.6	6.3	1.2	7.5	8.6
Std dev.	1.8	1.7	1.2	1.4	1.3

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion

## Résultats expérimentaux - majesté

1. grandeur (0.382138)
2. titre (0.366114)
3. dignité (0.274143)
4. noblesse (0.267147)
5. noble (0.243879)
6. altesse (0.14966)
7. sire (0.147181)
8. gloire (0.134615)
9. pouvoir (0.125155)

1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion  
✓ Résumé

*En résumé...*



1. Introduction
2. Grands corpus
3. Web
4. Dictionnaires
5. Conclusion  
✓ Perspectives

# *Perspectives*

