

# Linguistique de corpus

## Sémantique lexicale (suite)

Patrick Paroubek

LIMSI-CNRS  
Dépt. CHM - Groupe LIR  
Bât. 508 Université Paris XI, 91403 Orsay Cedex  
[pap@limsi.fr](mailto:pap@limsi.fr)

mercredi 26 avril 2017 / Semestre 2 - Cours 10

« *term frequency - inverse document frequency* » peut être utilisé pour trouver les termes utiles pour l'indexation d'un document dans un corpus.

mesure la spécificité du terme  $t$  par rapport au document  $d_i$  dans le corpus considéré.

$$\frac{\text{fréquence du terme } t \text{ dans le document } d_i}{\text{fréquence des documents contenant le terme } t}$$

Le terme « ballon » apparaît 2 fois dans le texte  $x_{fw}$  qui contient 86 mots, ce terme est présent dans 154 documents textes sur les 996 textes du corpus ballon, sa spécificité ( $tf-idf$ ) par rapport au texte  $x_{fw}$  dans le corpus ballon est donc de  $\frac{2}{\frac{154}{996}} = 0.14847$ .

Le même calcul pour le terme « le » donne un  $tf-idf$  de  $\frac{4}{\frac{86}{\frac{861}{996}}} = 0.05380$ , donc presque 3 fois moins.

Le coefficient de Jaccard peut être utilisé pour identifier la continuité thématique dans un texte, en le calculant sur une fenêtre glissante.

Deux textes consécutifs (deux segments de 500 caractères consécutifs) ont plus de chance de partager des termes en communs que deux segments pris au hasard.

Par exemple, `xeq` et `xer` sont plus similaires (pour les termes en commun) qu'en mesure un coefficient

$Jaccard = \frac{16}{111} = 0.14414$ ) que deux textes distants de 5000 caractères, par `xeq` et `xfa` dont le coefficient

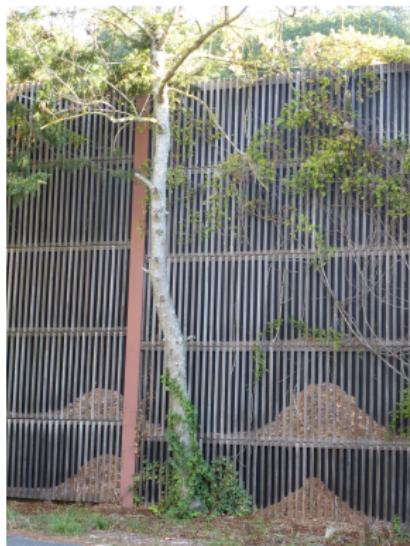
$Jaccard = \frac{13}{114} = 0.11403$

On utilise beaucoup la fréquence comme mesure :

$$\frac{\text{nombre d'evenements cibles}}{\text{nombre total d'evenements}}$$

car c'est le meilleur estimateur de la probabilité des événements cibles lorsqu'ils ont tous les mêmes chances de se produire (événements équiprobables). Nous sommes alors confrontés au hasard « pur » .

Mais le hasard « pur » suit une loi de Gauss définie par deux paramètres, la moyenne et l'écart-type (écart moyen par rapport à la moyenne). Par exemple lorsque l'on mesure la taille d'un population, la distribution des tailles mesurées donne une courbe « en cloche » caractéristique de la loi de Gauss, cela est vrai pour n'importe quel phénomène naturel equiprobable comme on peut s'en rendre compte avec la courbe dessinée par les feuilles qui sont tombées au hasard sur la photo ci-dessous.



# Wordnet

## Présentation du réseau lexical WordNet, <http://wordnet.princeton.edu/>

The screenshot displays two Firefox browser windows side-by-side.

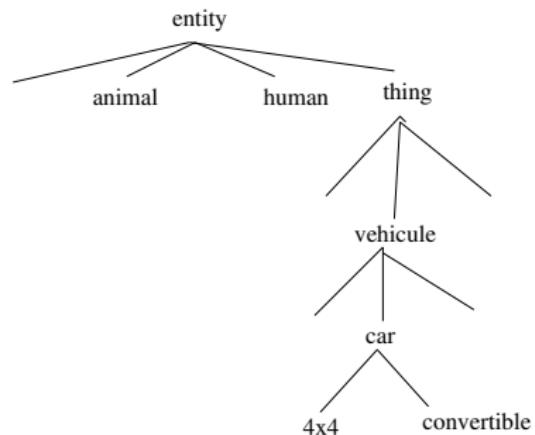
**Left Window (Princeton University WordNet Homepage):**

- Title: About WordNet - WordNet
- URL: https://wordnet.princeton.edu
- Content: Features a large image of a tree on a campus path, with text below it: "WordNet A lexical database for English". Below the image is a section titled "What is WordNet?" containing links like "What is WordNet?", "People", "News", "Use WordNet online", and "Download".

**Right Window (WordNet Search - 3.1):**

- Title: WordNet Search - 3.1 - Mozilla Firefox
- URL: wordnetweb.princeton.edu/peri/webm... Rechercher
- Content: A search interface with a "Word to search for: Foot" input field and a "Search WordNet" button. Below it are "Display Options" (dropdown menu), "Key: "S" - Show Synset (semantic) relations, "W" - Show Word (lexical) relations, and "Display options for sense: (gloss)" "an example sentence". A list of definitions for "Foot" follows:
  - S: (n) **foot, human foot, pes** (the part of the leg of a human being below the ankle joint) "his bare feet projected from his trousers"; "armored from head to foot"
  - S: (n) **foot, ft** (a linear unit of length equal to 12 inches or a third of a yard) "he is six feet tall"
  - S: (n) **foot** (the lower part of anything) "curled up on the foot of the bed"; "the foot of the page"; "the foot of the list"; "the foot of the mountain"
  - S: (n) **animal foot** foot (the pedal extremity of vertebrates other than human beings)
  - S: (n) **foundation, base, pedestal, foot, groundwork, substructure, understructure** (lowest support of a structure) "it was built on a base of solid rock"; "he stood at the foot of the tower"
  - S: (n) **foot, invertebrate foot** (any of various organs of locomotion or attachment in invertebrates)
  - S: (n) **foot, travel by walking** "he followed no foot"; "The earliest reflex"

On peut utiliser la topologie du réseau sémantique en parcourant les chemins étiquetés par les différentes relations sémantiques (synonymie, antonymie, méronomie etc.) pour calculer des distances sémantiques en comptant le nombre d'arcs parcourus. Cependant une telle distance ne prend pas en compte la plus ou moins grande spécificités des concepts. Par exemple, il y a la même distance de 2 (arcs) entre « animal » et « humain » qu'entre « 4x4 » et « convertible », alors qu'intuitivement la première est beaucoup plus grande. Une solution à ce problème est la mesure de Wu & Palmer qui prend en compte le niveau de spécificité.



# Exemple

La mesure de Wu & Palmer de distance conceptuelle dans un réseau sémantique. Zhibiao Wu and Martha Palmer, « Verbs semantics and lexical selection », Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94), pages 133-138, doi :10.3115/981732.981751.

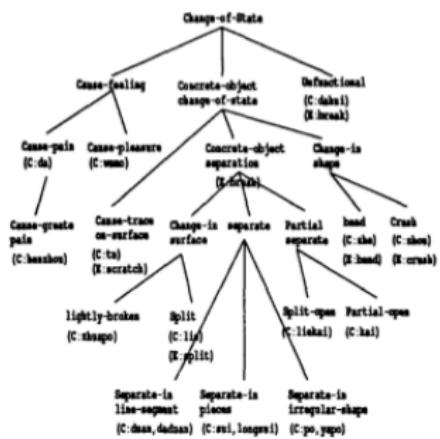


Figure 3. Change-of-state domain for English and Chinese

Within one conceptual domain, the similarity of two concepts is defined by how closely they are related in the hierarchy, i.e., their structural relations.

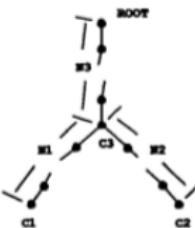


Figure 4. The concept similarity measure

The conceptual similarity between C1 and C2

is:

$$\text{ConSim}(C1, C2) = \frac{2 \cdot N3}{N1 + N2 + 2 \cdot N3}$$

C3 is the least common superconcept of C1 and C2. N1 is the number of nodes on the path from C1 to C3. N2 is the number of nodes on the path from C2 to C3. N3 is the number of nodes on the path from C3 to root.