

Thésaurus pour la Sécurité Radiologique à partir d'un corpus de textes et d'outils linguistiques en ligne

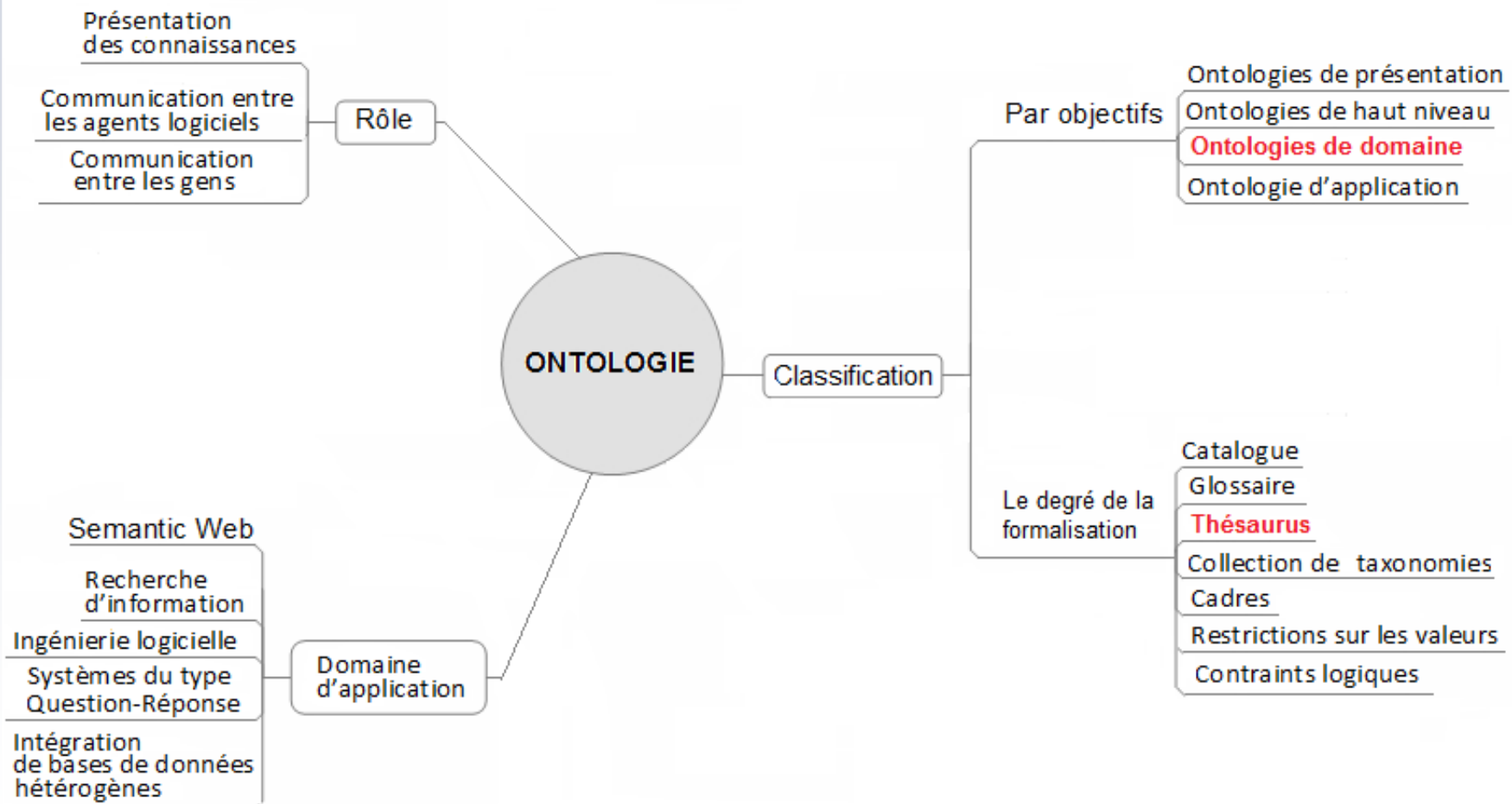
Olena OROBINSKA – ERIC, Université de Lyon & KhPI Kharkov

Jean-Hugues CHAUCHAT – ERIC, Université de Lyon,

Natalie SHARONOVA – KhPI, Université Polytechnique de Kharkov, Ukraine

Objectifs

- Créer le thésaurus d'un domaine spécialisé (la Sécurité Radiologique)
- à partir d'un corpus original de textes,
- ... et d'outils linguistiques en ligne (Parseurs & Dictionnaire de synonymes) ...
- ... pour le français (et le russe)
- en allégeant le travail des experts



Les corpus construits

Période : 1999 à nos jours

- Les séries des Normes de Sécurité ;
- Les recommandations de l'AIEA (Agence Internationale pour l'Energie Atomique) et des Commissions nationales ;
- Les rapports sur les accidents liés aux risques radiologiques ;
- Des articles scientifiques.

corpus français : 1 600 000 mots

corpus russe : 600 000 mots

le tout à partir de documents de format PDF,
en supprimant tout ce qui n'était pas le texte principal

Isoler les mots avec TreeTagger : étiquetage morphologiques pour trouver les lemmes du corpus

Exemple

«...*détriment causé au personnel médical est pris en compte de façon subsidiaire.*»

on obtient la structure étiquetée :

détriment /NOM/***détriment***

causé/PPP/***causer***

au PRP/***au***

personnel NOM/***personnel***

médical ADJ/***médical***

est /VER/***être***

pris/PPP/***prendre***

en/PRP/***en***

compte/NOM/***compte***

de/PRP/***de***

façon/NOM/***façon***

subsidiaire/ADJ/***subsidiaire***

./SENT/.

Construire la liste initiale des mots du thésaurus

Chercher les noms *spécifiques de certains textes du corpus* avec l'indice Tf/Idf

⇒ Une liste provisoire de 100 noms

Confronter ces noms aux classes proposées par le projet
« RISQUE »

Validation de cette liste initiale par l'expert du domaine

Construire la liste initiale des mots : le noyau d'ontologie

cadre du RISQUE

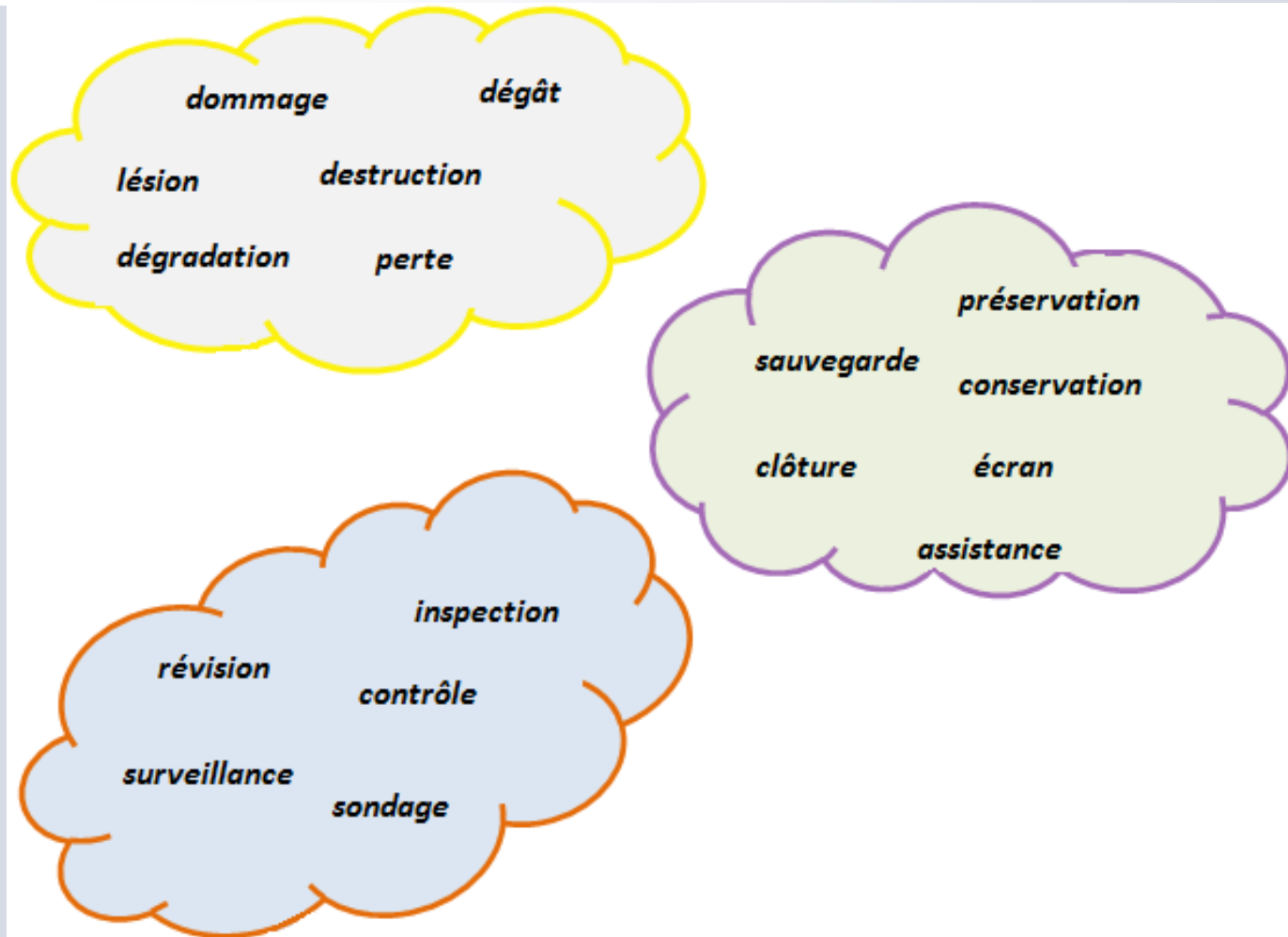
<i>Dans FrameNet</i>	<i>Dans notre modèle</i>
Possibilité (Chance)	-
Dommage (Harm)	Dommage
Victime (Victim)	Population, Personnel
L'acteur (Actor)	Personnel
Motivation	Contrôle, Protection
L'intention	-
Gain	Sécurité, Sûreté
L'objet évalué (Valued Object)	Source
Bénéficiaire	-
Situation à risque (Risky situation)	Facteurs de risque
L'acte, action (Deed)	Accident

Liste initiale des termes

dommage
exposition
contrôle
personnel
population
rayonnement
risque
sûreté
source

« Peupler » le noyau d'ontologie :

Pour chaque classe pour notre domaine, comment trouver les termes synonymes, équivalents pour designer un concept ?



L'idée de notre approche

La fréquence des entrées lexicales dans le corpus est souvent **trop faible** pour une détection avec les méthodes purement statistiques

=> il faut utiliser les outils linguistiques disponibles

Les noms cherchés sont des synonymes qui partagent les mêmes propriétés.

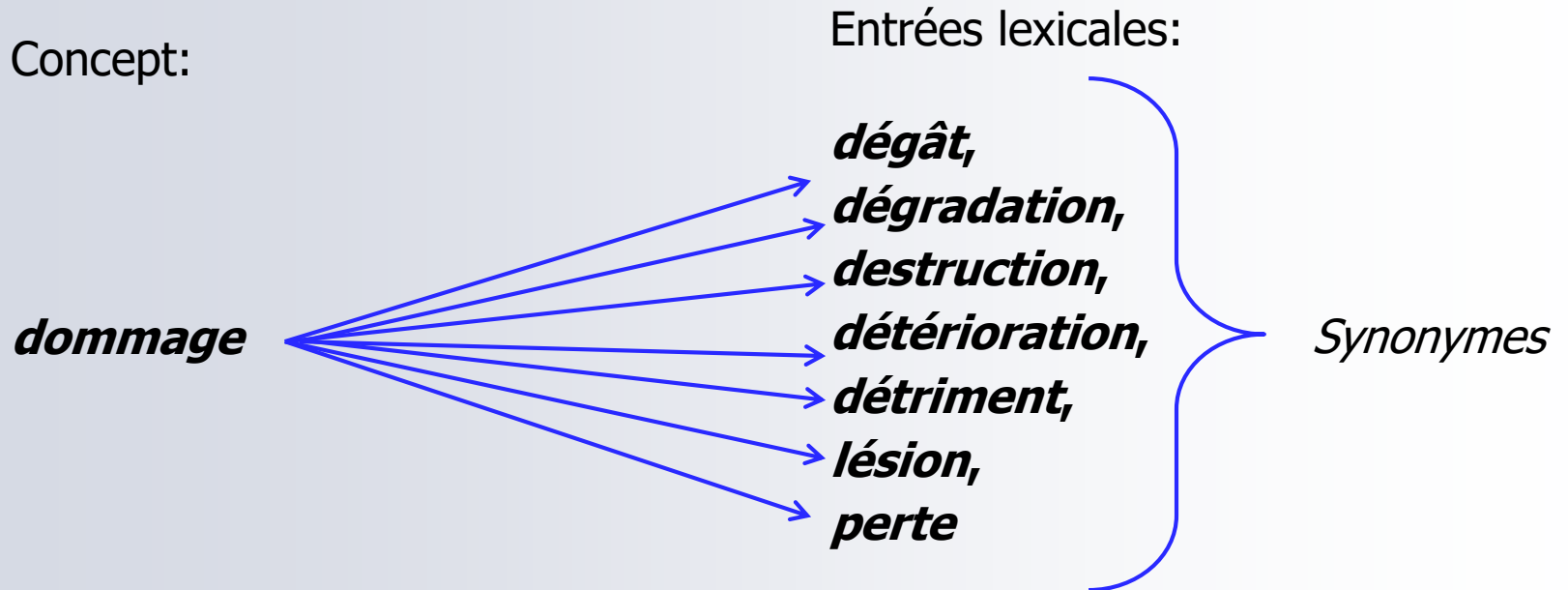
Concept:

dommage

Entrées lexicales:

***dégât,
dégradation,
destruction,
détérioration,
détriment,
lésion,
perte***

Synonymes



Pour nous,
les noms équivalents pour designer un concept du domaine
sont **synonymes** et associés aux mêmes verbes

Nous avons testé les dictionnaires de synonymes disponibles en ligne :

- CRISCO (Université de Caen)
- Alexandria (TV5 Monde)
- BabelNet
- Wolf (analogue français de WordNet)

Les résultats d'interrogation de CRISCO

dommage	exposition	personnel	population	protection	rayonnement	source	sûreté	contrôle
atteinte	abandon		démographie	abri	chaleur	agent	abri	blocage
avarie	argument		foule	affermisssement	développement	amont	acuité	critique
bête	ban		gens	aide	éclat	base	assurance	encadrement
brèche	carcan		habitant	aile	expansion	canal	cachette	examen
calamité	concours		habitants	appui	fluide	cause	cadenas	expérience
casse	délaissement		nation	armure	fraîcheur	centre	caution	expérimentation
coup	démonstration		peuple	asile	gaieté	commencement	certitude	gouvernement
dam	détail		peuplement	assistance	gloire	endroit	clairvoyance	inspection
dégât	éclairage		plèbe	auspice	illustration	étymologie	clause pénale	limitation
dégradation	étalage			baderne	influence	ferment	cran	maîtrise
déprédation	exhibition			bastion	irradiation	filon	dédit	possession
désavantage	exorde			bénédiction	jet	fondement	dépôt	régulation
destruction	explication			blindage	joie	fontaine	fermeté	sondage
détérioration	vernissage			bouclier	lumière	foyer	fidélité	supervision
détriment	exposé			boulevard	lustre	générateur	force	surveillance
ennui	foire			capuchon	nitescence	germe	gage	vérification
entorse	forum			carapace	parure	geyser	garant	visite
grabuge	galerie			caution	prestige	griffon	garantie	
grief	introduction			chaperon	propagation	idée	hypothèque	
injure	montre			cloche	radiation	mère	infaillibilité	
injustice	narration			clôture	rayon	mine	justesse	
intérêt	orientation			conservation	réflexion	naissance	lucidité	
lésion	ostension			couvercle	réverbération	origine	nantissement	
mal	pilori			couverture	satisfaction	pépinière	ordre	
méfait	position			cuirasse	splendeur	point	police	
nuisance	préambule			dédain	transmission	point d'eau	précaution	
outrage	prélude			défense		point de départ	précision	
perte	présentation			écran		principe	sécurité	
préjudice	protase			égide		provenance	serrure	
ravage	récit			encouragement		puits	sincérité	
regrettable	rétrospective			enduit		racine	solidité	
ribordage	salon			enveloppe		résurgence		
sévice	situation			faveur		seuil		
tort				gaine				
vimaire				garantie				

Pour nous,
les noms équivalents pour designer un concept du domaine
sont synonymes et associés aux mêmes verbes

Notre coefficient d'association

$$K_{ij} = MI(c_i, v_j) \times JaccardCoefficient(c_i, v_j)$$

Information Mutuelle

$$MI(c_i, v_j) = |W| \times \log \frac{TF(c_i, v_j)}{TF(c_i) \times TF(v_j)}$$

L'information mutuelle favorise l'association des mots rares et spécifiques.

Coefficient de Jaccard

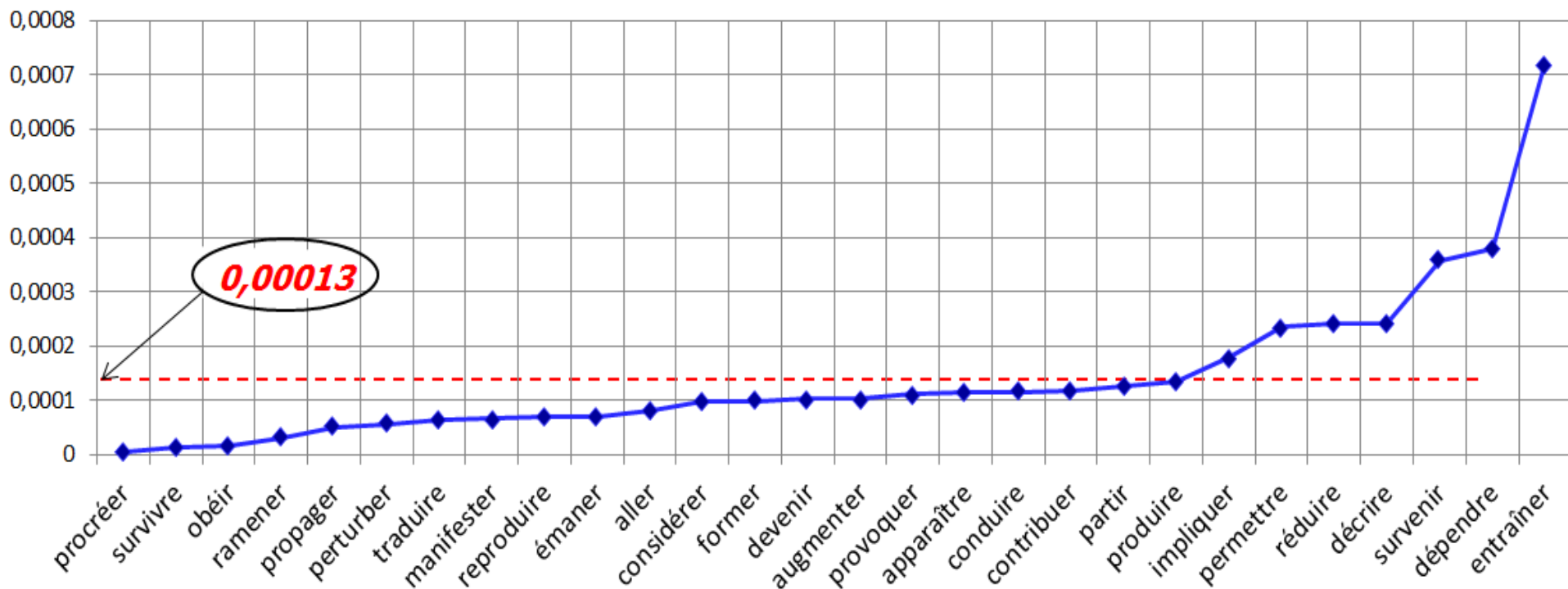
$$JaccardCoefficient(c_i, v_j) = \frac{TF(c_i, v_j)}{TF(c_i, \bar{v}_j) + TF(\bar{c}_i, v_j)}$$

Le coefficient de Jaccard favorise l'association des mots fréquents.

Les noms équivalents pour designer un concept du domaine sont les synonymes (CRISCO) associés aux mêmes verbes dans le corpus.

Exemple : les verbes associés au terme « dommage »

Taux d'association
concept-verbe



Pour« dommage » - 1

Dans le dictionnaire des synonymes CRISCO on trouve 38 synonymes de « dommage » ; dans le corpus il n'y a que 11 mots de cette liste.

Concept		<i>dommage</i>
CRISCO	38	atteinte, avarie, bête, brèche, calamité, casse, coup, dam, dégât, dégradation, déprédation, désavantage, destruction, détérioration, détriment, ennui, entorse, grabuge, grief, injure, injustice, intérêt, lésion, mal, méfait, nuisance, outrage, perte, préjudice, ravage, regrettable, ribordage, sévices, sinistre, tant pis, tort, vimaire
Présents dans le Corpus	11	atteinte, brèche, dégât, dégradation, destruction, détérioration, détriment, ennui, intérêt, lésion, perte

Pour « dommage » -2 noms associés aux mêmes verbes puis retour à l'expert

**Parmi les 11 candidats-termes issus de *dommage*,
7 sont associés, dans le corpus, aux mêmes verbes :**
*{brèche, dégradation, destruction, détérioration, détriment,
lésion, perte}*

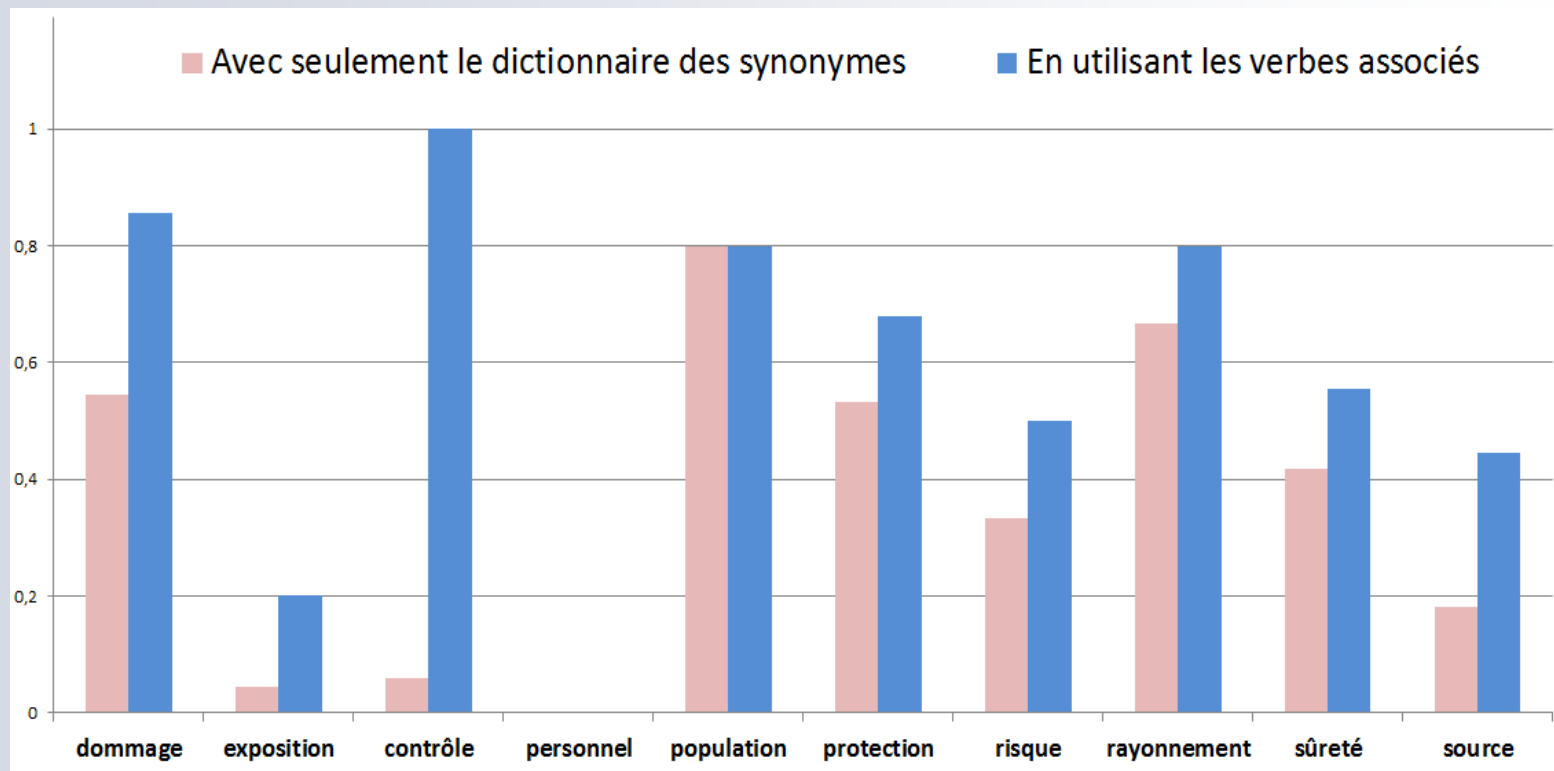
Le mot «*brèche*» a été exclu par l'expert du domaine.

La valeur moyenne de la précision est égale à $7/11=68\%$

Bilan les termes peuplant les classes sémantiques

Classe	Membres
Domage	<i>dommage, dégradation, destruction, détérioration, détriment, lésion, perte</i>
Exposition	<i>exposition, exposé</i>
Contrôle	<i>contrôle, inspection, révision, sondage, surveillance</i>
Personnel	—
Population	<i>population, habitant, gens, démographie</i>
Rayonnement	<i>rayonnement, chaleur, lumière, irradiation, radiation, rayon</i>
Protection	<i>protection, assistance, abri, secours, sauvegarde, préservation, défense, clôture, recommandation</i>
Risque	<i>risque, danger, menace</i>
Sûreté	<i>sûreté, garantie, sécurité, précaution</i>
Source	<i>source, point, mine, endroit, cause</i>

Amélioration du taux d'acceptation par l'expert quand on utilise les verbes associés aux noms



Résumé de notre approche

Quatre étapes :

- Mise en forme du corpus et modélisation primaire avec des experts du domaine.
À la sortie de cette étape on doit obtenir une liste des concepts génériques du domaine.
- Récupération de toutes les paires (c,v) où c est l'un des concepts sujets, et v est son verbe associé (prédicat).
- Interrogation **en-ligne** de ressources lexicales, récupération les synonymes potentiels dans le corpus, et comparaison de ces deux listes.
- Validation des résultats et peuplement de l'ontologie.

Conclusion et perspectives

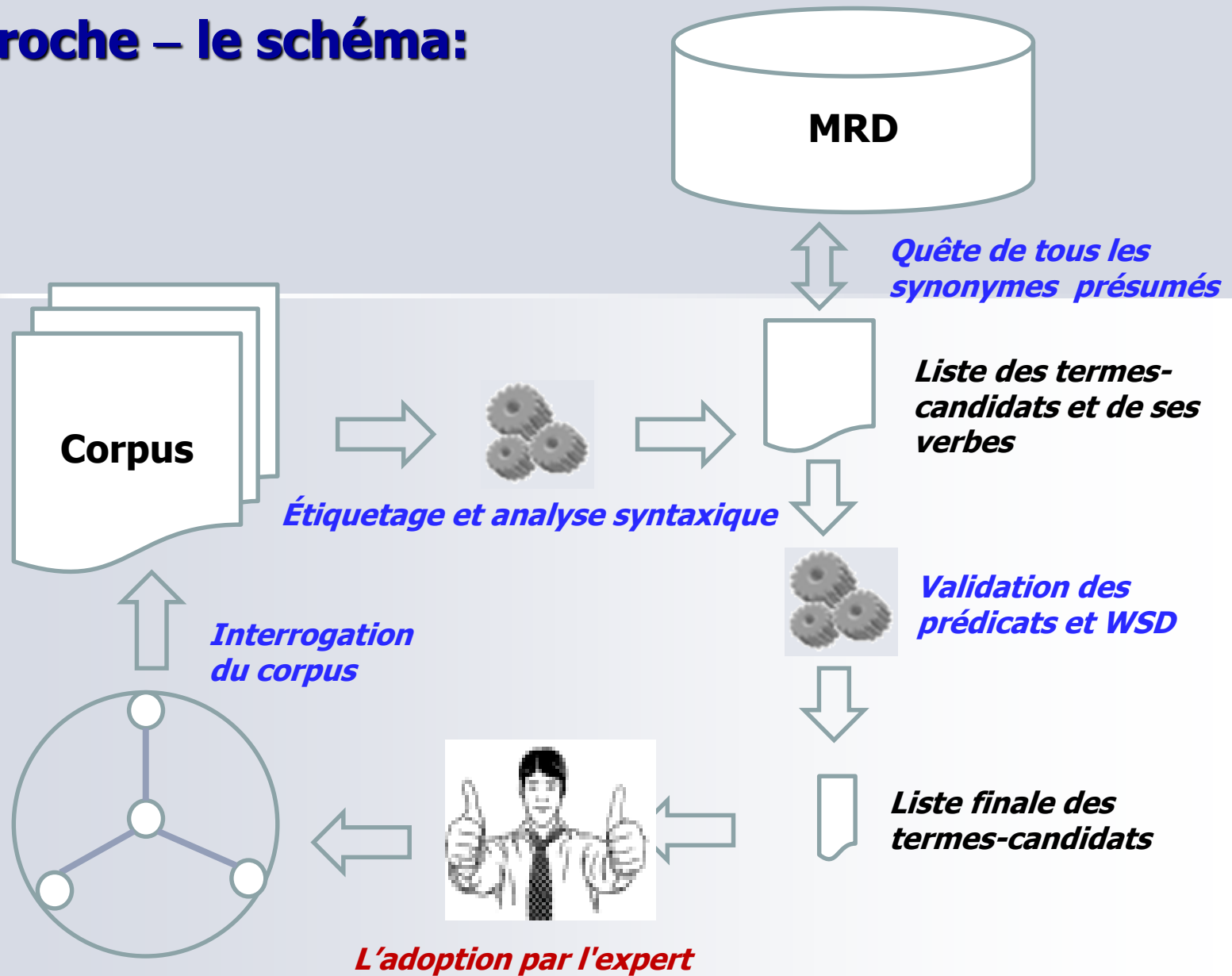
- Dans cette approche la qualité du dictionnaire des synonymes joue un rôle crucial. Ces ressources s'améliorent progressivement.
- D'un côté les verbes peuvent servir à indiquer des propriétés significatives des concepts. Comme la plupart des verbes sont polysémiques, il faut utiliser un indice quantitatif de sélection.
- Afin d'améliorer le rappel pour les termes détectés dans le corpus il faut créer des classes de verbes ayant le même sens – c'est possible.
- Les valeurs de précision, relativement basses, s'expliquent du fait que les concepts génériques (comme *dommage*, par exemple) sont polysémiques.
- Les vrais termes du domaine sont des locutions composées de plusieurs mots (groupes nominaux). C'est en cours.

Merci pour votre attention !

Et pour vos questions...



Approche – le schéma:



L'idée de notre approche

La fréquence des entrées lexicales dans le corpus est souvent **trop faible** pour une détection avec les méthodes statistiques.

Mais ils sont des synonymes qui partagent les mêmes propriétés.

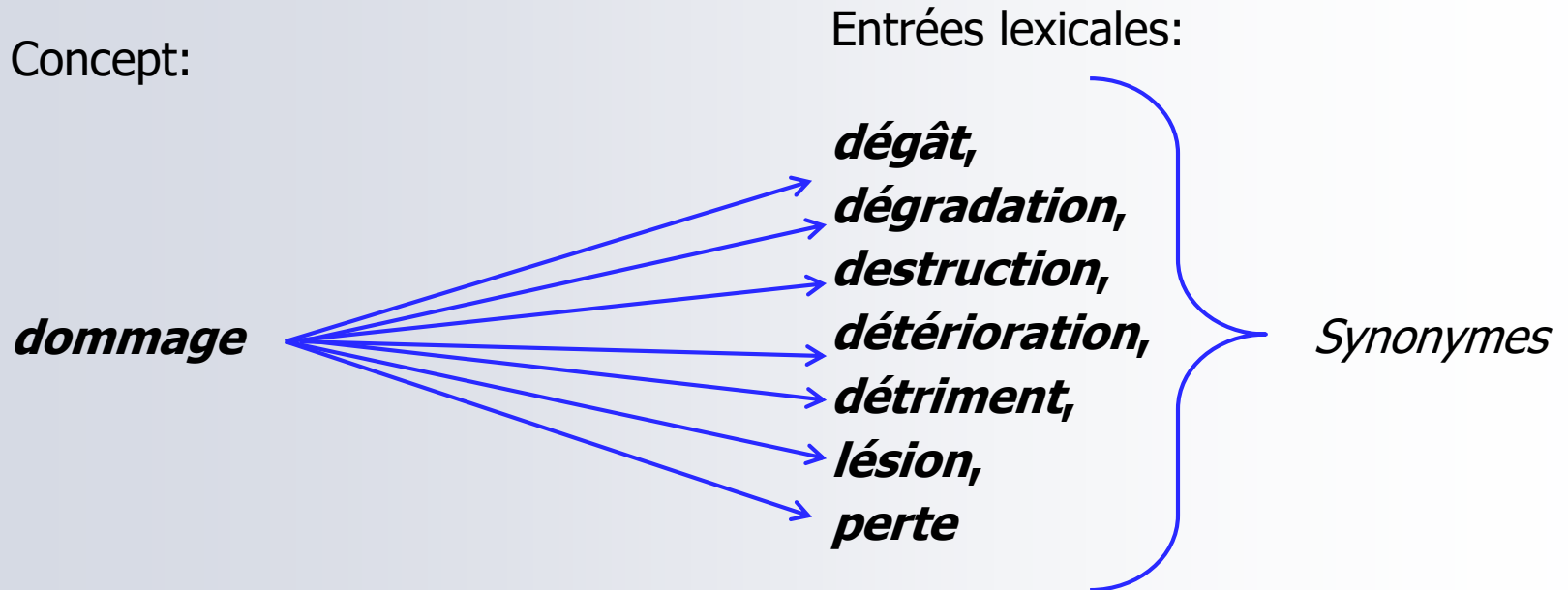
Concept:

dommage

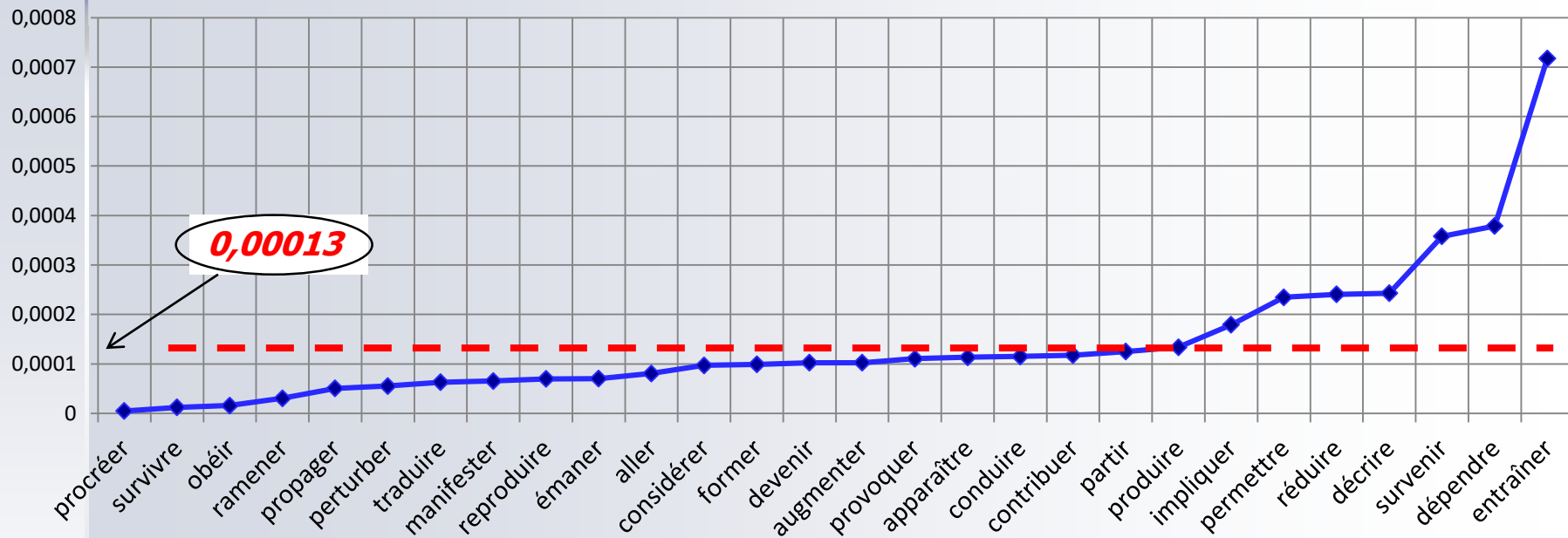
Entrées lexicales:

***dégât,
dégradation,
destruction,
détérioration,
détriment,
lésion,
perte***

Synonymes



Les verbes associés à concept "*dommage*", rangés selon notre indice COMB



Le seuil a été établie heuristiquement au niveau 0,00013.

Les verbes pour lesquels le valeur de notre indice est inférieur à cette grandeur ont été conservés pour être soumis à l'expert.

Modèle mathématique - 1

- $C = \{c_i \mid i=1 \dots n, n \in \mathbb{N}\}$ la liste des concepts représentant le modèle de domaine ;
- $W = \{w\}$ l'ensemble des substantifs (noms) apparaissant au moins une fois dans le corpus ;
- $V = \{v\}$ l'ensemble des verbes apparaissant au moins une fois dans le corpus ;
- $CV = \{c_i, v_j\}$ l'ensemble des paires « nom – verbe » où le nom correspond à un des concepts c_i en tant que sujet dans la phrase, et où v_j est son prédicat.

Modèle mathématique - 2

$DL = \{DL_i \mid i = 1 \dots n, n \in N\}$ – est l'ensemble des synonymes de tous les concepts qui ont été extraits de dictionnaire CRISCO ;

DL_i – est l'ensemble des synonymes du dictionnaire pour le concept i ;

$CL_i = W \cup DL_i$ – est la liste de tous les synonymes potentiels pour le concept i qu'on trouve dans le corpus ;

Et $CL_i := \{l \subseteq W \mid \forall w : (w, c_i) \in I_{syn}\}$

et $I_{syn} \subseteq DL \times W$ – est la relation binaire de synonymie pour l'ensemble des mots.

Exemple complet pour « dommage »

- 2 - retour au corpus

Concept		<i>dommage</i>
verbes du corpus liés à <i>dommage</i>	28	aller, apparaître, augmenter, conduire, considérer, contribuer, décrire, dépendre, devenir, émaner, entraîner, former, impliquer, manifester, obéir, partir, permettre, perturber, procréer, produire, propager, provoquer, ramener, réduire, reproduire, survenir, survivre, traduire
prédicats – les verbes qui n’ont pas dépassé le seuil ($\leq 0,00013$)	20	procréer, survivre, obéir, ramener, propager, perturber, traduire, manifester, reproduire, émaner, aller, considérer, former, devenir, augmenter, provoquer, apparaître, conduire, contribuer, partir

Critères de validation des entrées lexicales pour les concepts donnés

Rôle crucial pour résoudre le problème de la désambiguïsation des candidats-termes ou différents entrées lexicales récupérées dans le dictionnaire des synonymes.

Nous avons choisi un **coefficient combinant deux critères**, capables de caractériser le niveau de la « compatibilité » entre le concept et certain verbe:

- *L'information mutuelle* qui permet de calculer la dépendance statistique de deux variables
- *L'indice de Jaccard modifié*, pour évaluer leurs similarité.

Coefficient combiné de valorisation des prédicats : l'association « prédicat + verbe » dans le corpus est-elle fortuite ?

Information mutuelle :

$$I(c_i, v_j) = \sum f_{ij} * \log \frac{f_{ij}}{f_{ci} * f_{vj}}$$

L'indice de Jaccard modifiée :

$$k_{ij} = \frac{f_{ij}}{N_i + N_j}$$

où:

- f_{ij} est la fréquence du concept ci en tant que le sujet du prédicat vj
- $f(ci)$ et $f(vj)$ sont les fréquences du concept ci et du prédicat vj

Le coefficient est : $COMB = I(c_i, v_j) * k_{ij}$

Rappel de la méthode AFC (analyse formelle des concepts)

Définition 1. Le triple $K=(G; M; I)$ est un *contexte formel* si G et M sont des ensembles disjoints et I est une relation binaire entre G et M . Les éléments de G sont nommés **les objets**, ceux de M sont nommés **les attributs** et I est **l'incidence** du contexte.

Définition 2. Le *concept formel* (A,B) est la paire des ensembles disjoints des objets (A) et des attributs (B) respectivement, si l'affirmation suivante est vrai : chaque objet partage tous les attributs de l'ensemble d'attributs et tous les attributs sont partagés par chaque objet de l'ensemble des objet.

Soit (G, M, I) un contexte formel.

Nous définirons A' pour $A \subseteq G$: $A' := \{ m \subseteq M \mid (\forall g \in A : (g, m) \in I) \}$

comme l'ensemble des attributs communs pour tous les objets A ;

et B' pour $B \subseteq M$ $B' := \{ g \subseteq G \mid \forall m \in B : (g, m) \in I \}$

comme l'ensemble des objets qui possèdent tous les attributs B .

Étant donné la définition 2, la paire (A, B) est un *concept formel* $(G; M; I)$ si et seulement si $A \subseteq G, B \subseteq M$ $A' = B, B' = A$

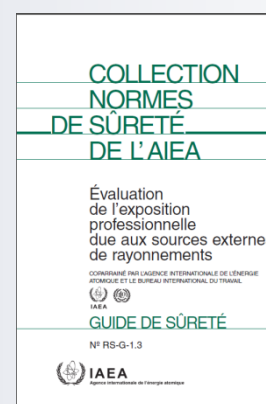
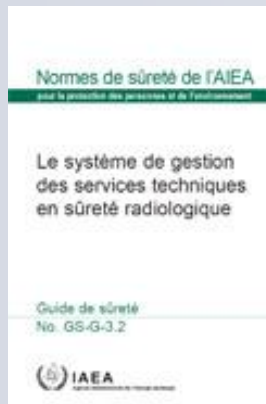
Description des «conditions initiales»

« Conditions initiales » : notre corpus

Le corpus de textes français ; environ 1 500 000 mots :

- les normes de sûreté radiologique,
- les rapports de l'IAEA
- les rapports des Commissions internationales et nationales de protection radiologique,
- des articles de revues spécialisées publiés depuis 1999.

Ces textes étaient en PDF, on a dû les convertir en texte ou HTML



« Conditions initiales » étiquetage morphologiques et lemmes

Avec TreeTagger. Ex. pour le texte initial :

«...détriment causé au personnel médical est pris en compte de façon subsidiaire.»

on obtient la structure étiquetée :

*détriment /NOM/détriment
causé/PPP/causer
au PRP/au
personnel NOM/personnel
médical ADJ/médical
est /VER/être
pris/PPP/prendre
en/PRP/en
compte/NOM/compte
de/PRP/de
façon/NOM/façon
subsidiaire/ADJ/subsidiaire
./SENT/.*

« Conditions initiales »

9 concepts de niveau supérieur

Après modélisation primaire, à l'aide d'expert : 9 concepts de niveau supérieur : $C = \{\text{dommage, exposition, contrôle, personnel, population, protection, rayonnement, sûreté, source}\}$.

Fragment de
l'ontologie de noyau
en format OWL

```
<owl:Class>
  <label xml:lang="fr">Sources</label>
  <comment xml:lang="fr">Les sources artificielles de rayonnement
ionisant</comment>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#RadionuclidesArtificiels"/>
    <owl:Thing rdf:about="#RéacteurNucléaire"/>
    <owl:Thing rdf:about="#AccélérateurDeParticules"/>
  </owl:oneOf>
</owl:Class>

<owl:Class rdf:about="&protection;RayonnementIonisant">
  <label xml:lang="en">Ionising radiation</label>
  <label xml:lang="fr">Rayonnement_Ionisant</label>
  <comment xml:lang="en">is radiation composed of particles that individually
carry enough kinetic energy to liberate an electron from an atom or molecule,
ionizing it</comment>
  <comment xml:lang="fr">un rayonnement capable de déposer assez
d'énergie dans la matière qu'il traverse pour créer une ionisation</comment>
  <comment xml:lang="ru"></comment>
</owl:Class>
```

Comparaison des ressources lexiques : CRISKO

Résultats renvoyés par CRISKO pour le concept « *dommage* »
(38 synonymes potentiels)



université de Caen
Basse-Normandie

CRISKO KALI DICTIONNAIRE DES SYNONYMES

Dictionnaire Electronique des Synonymes (DES)

Tapez l'unité lexicale recherchée puis cliquez sur *Valider* ou tapez *Entrée*

Rechercher :

dommage *définition*

38 synonymes
atteinte, avarie, bête, brèche, calamité, casse, coup, dam, dégât, dégradation, déprédation, désavantage, destruction, détérioration, détriment, endommagement, ennui, entorse, grabuge, grief, injure, injustice, intérêt, lésion, mal, méfait, nuisance, outrage, perte, préjudice, ravage, regrettable, ribordage, sévice, sinistre, tant pis, tort, vimaire

12 antonymes
avantage, bénéfice, bien, bienfait, bonheur, dédommagement, gain, heureux, indemnité, produit, profit, réparation

Classement des premiers synonymes

tort	
dégât	
ravage	
dégradation	
préjudice	
atteinte	
casse	
détérioration	



Comparaison des ressources lexiques : ALEXANDIA

Résultats du dictionnaire ALEXANDIA TV5.org pour le concept « *dommage* »
deux groupes de synonymes potentiels, regroupés a priori selon de deux sens différents :
les 12 synonymes potentiels)

TV5MONDE TV5MONDE+ AFRIQUE JEUNESSE CINEMA DOCUMENTAIRE

TV5MONDE

► Météo
► Dictionnaire
► Les JT

PROGRAMMES INFORMATION LANGUE FRANÇAISE DIVERTISSEMENT PRATIQUE TV5

Accueil /

Le dictionnaire multifonctions

avec



Le mot du jour **exhaure**

SAISISSEZ UN TERME

OK

lexique

DÉFINITIONS

SYNONYMES

CONJUGAISONS

STYLE

FRANÇAIS/ANGLAIS

ANGLAIS/FRANÇAIS

> **dommage**

1/ préjudice: dégât, dégradation, détérioration, atteinte, tort, détriment.

2/ ennuyeux: fâcheux, regrettable, triste, bête.

Pour le concept «dommage», BabelNet propose 11 sens différents, dont deux absents dans d'autres dictionnaires :
pour le concept « dommage »

BabelNet propose, en plus, « prix » et « dommage en droit civil »

Meaning: damage ² • ID: bn:00025081n • Type: Concept		  [details] [explore]
Senses:   damage ² , equipment casualty ¹   كَسَارَةٌ, كَسَارَةٌ مُعَدَّات,  préjudice, dommage   معدات الضحايا,  设备伤亡,  équipements victime,  geräte unfallversicherung,  εξοπλισμός θύμα,  ציוד הרוג,  उपकरण हताहत,  attrezzature ferito,  機器死傷者,  оборудование жертв,  equipo víctima		
Glosses:   Loss of military equipment		

Meaning: damage ¹ • ID: bn:00025080n • Type: Concept		  [details] [explore]
Senses:   damage ¹ , harm ² , impairment ¹   أذى, خَسَارٌ, فَتَادٌ, كَلَفٌ,  損失,  préjudice, dommage , détérioration, dégât,  εξασθένιση, ζημιά,  avaria, danno,  ダメージ, 不都合, 傷害, 危害, 厄, 厄難, 害, 害悪, 弊害, 損傷, 損壊, 損害, 毀傷, 毀損, 禍害,  daño, deterioro, perjuicio [w]  damage,  dégât, dommage ,  Schaden,  danno,  損失,  ущерб, убыток,  daño   dommages,  נזק,  नुकसान,  danno,  daño		
Glosses:   The occurrence of a change for the worse [w]   Cost or expense.		

Meaning: Dommage (échecs) • ID: bn:04980424n • Type: Concept		  [details] [explore]
Senses:   Dommage (échecs)		
Glosses:   Dommage est un terme du problèmes d'échecs.		
Categories:   Problème d'échecs		

Meaning: damages ¹ • ID: bn:00003333n • Type: Concept		  [details] [explore]
--	--	---

Comparaison des ressources lexiques : WOLF

Dans WOLF, analogue français de WordNet, on trouve 6 variations du sens « **dommage** ».

WOLF n'a pas de l'interface graphique.

The screenshot shows a text editor window titled "Syntext Serna Free 4.4.0-20111114 - wolf-0.1.5.xml". The window contains XML data representing a lexical resource. The XML structure is as follows:

```
<SYNSET><ID>ENG20-00384902-n</ID><POS>n</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00359080-n</ILR><ILR><TYPE>eng_derivative</TYPE>ENG20-00439460-
<SYNSET><ID>ENG20-00384937-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>similar_to</TYPE>ENG20-00384332-a</ILR><DEF>as if dulled in color with a sprinkling of powder</DI
<SYNSET><ID>ENG20-00385010-v</ID><POS>v</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00431897-v</ILR><ILR><TYPE>verb_group</TYPE>ENG20-00385187-v<
<SYNSET><ID>ENG20-00385038-n</ID><POS>n</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00384902-n</ILR><ILR><TYPE>eng_derivative</TYPE>ENG20-00443273-
<SYNSET><ID>ENG20-00385055-b</ID><POS>b</POS><SYNONYM></SYNONYM><ILR><TYPE>near_antonym</TYPE>ENG20-00384894-b</ILR><ILR><TYPE>derived</TYPE>ENG20-02348204-a<
<SYNSET><ID>ENG20-00385058-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>near_antonym</TYPE>ENG20-00384332-a</ILR><ILR><TYPE>be_in_state</TYPE>ENG20-0471968
<SYNSET><ID>ENG20-00385151-b</ID><POS>b</POS><SYNONYM><LITERAL>transitivement</SENSE>ManValOK:0/1:derivation</SENSE></LITERAL></SYNONYM><ILR><TYPE>near_antonym</T
<SYNSET><ID>ENG20-00385156-n</ID><POS>n</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00359080-n</ILR><ILR><TYPE>eng_derivative</TYPE>ENG20-00430741-
<SYNSET><ID>ENG20-00385187-v</ID><POS>v</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00431323-v</ILR><ILR><TYPE>subevent</TYPE>ENG20-00361270-v</IL
<SYNSET><ID>ENG20-00385268-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>similar_to</TYPE>ENG20-00385058-a</ILR><DEF>slightly dark</DEF><USAGE>darkish red</USAGI
<SYNSET><ID>ENG20-00385311-b</ID><POS>b</POS><SYNONYM><LITERAL>absolument</SENSE>ManValOK:0/1:clg08</SENSE></LITERAL><LITERAL>intransitivement</SENSE>ManValOK:0/2:der
<SYNSET><ID>ENG20-00385318-n</ID><POS>n</POS><SYNONYM><LITERAL>hybridation</SENSE>0/1:en.wikipedia</SENSE></LITERAL></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-003:
<SYNSET><ID>ENG20-00385350-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>near_antonym</TYPE>ENG20-00385489-a</ILR><ILR><TYPE>category_domain</TYPE>ENG20-06:
<SYNSET><ID>ENG20-00385382-v</ID><POS>v</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00385553-v</ILR><DEF>break or fall apart into fragments</DEF><USAGE:
<SYNSET><ID>ENG20-00385459-n</ID><POS>n</POS><SYNONYM><LITERAL>dommage</SENSE>2/1:fr.roen</SENSE></LITERAL></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00359080-
<SYNSET><ID>ENG20-00385487-b</ID><POS>b</POS><SYNONYM></SYNONYM><ILR><TYPE>derived</TYPE>ENG20-02754247-a</ILR><DEF>in an intravenous manner</DEF><USAGE>the patier
<SYNSET><ID>ENG20-00385489-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>near_antonym</TYPE>ENG20-00385350-a</ILR><ILR><TYPE>category_domain</TYPE>ENG20-06:
<SYNSET><ID>ENG20-00385553-v</ID><POS>v</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00135037-v</ILR><ILR><TYPE>near_antonym</TYPE>ENG20-00453592-
<SYNSET><ID>ENG20-00385614-b</ID><POS>b</POS><SYNONYM></SYNONYM><ILR><TYPE>derived</TYPE>ENG20-01382373-a</ILR><DEF>in an intuitive manner</DEF><USAGE>inventors see
<SYNSET><ID>ENG20-00385678-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>near_antonym</TYPE>ENG20-00386049-a</ILR><ILR><TYPE>similar_to</TYPE>ENG20-00385868-
<SYNSET><ID>ENG20-00385701-n</ID><POS>n</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00385459-n</ILR><ILR><TYPE>eng_derivative</TYPE>ENG20-00464634-
<SYNSET><ID>ENG20-00385833-b</ID><POS>b</POS><SYNONYM><LITERAL>imaginativement</SENSE>ManValOK:0/1:derivation</SENSE></LITERAL></SYNONYM><ILR><TYPE>derived</TYPE>
<SYNSET><ID>ENG20-00385833-n</ID><POS>n</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00385459-n</ILR><ILR><TYPE>eng_derivative</TYPE>ENG20-00283700-
<SYNSET><ID>ENG20-00385868-a</ID><POS>a</POS><SYNONYM></SYNONYM><ILR><TYPE>similar_to</TYPE>ENG20-00385678-a</ILR><DEF>on the Italian or Roman side of the Alps</DEF><U
<SYNSET><ID>ENG20-00385946-v</ID><POS>v</POS><SYNONYM></SYNONYM><ILR><TYPE>hypernym</TYPE>ENG20-00385553-v</ILR><ILR><TYPE>verb_group</TYPE>ENG20-00202062-v<
<SYNSET><ID>ENG20-00385989-b</ID><POS>b</POS><SYNONYM></SYNONYM><ILR><TYPE>derived</TYPE>ENG20-00953701-a</ILR><DEF>in a manner causing resentment</DEF><DOMAIN>
```

The word "dommage" is highlighted in red in the XML data. The window also shows a "Message" pane at the bottom with the text "XML" and a "PLAIN TEXT MODE" indicator.

Les résultats d'interrogation de CRISCO

dommage	exposition	personnel	population	protection	rayonnement	source	sûreté	contrôle
atteinte	abandon		démographie	abri	chaleur	agent	abri	blocage
avarie	argument		foule	affermisssement	développement	amont	acuité	critique
bête	ban		gens	aide	éclat	base	assurance	encadrement
brèche	carcan		habitant	aile	expansion	canal	cachette	examen
calamité	concours		habitants	appui	fluide	cause	cadenas	expérience
casse	délaissement		nation	armure	fraîcheur	centre	caution	expérimentation
coup	démonstration		peuple	asile	gaieté	commencement	certitude	gouvernement
dam	détail		peuplement	assistance	gloire	endroit	clairvoyance	inspection
dégât	éclairage		plèbe	auspice	illustration	étymologie	clause pénale	limitation
dégradation	étalage			baderne	influence	ferment	cran	maîtrise
déprédation	exhibition			bastion	irradiation	filon	dédit	possession
désavantage	exorde			bénédiction	jet	fondement	dépôt	régulation
destruction	explication			blindage	joie	fontaine	fermeté	sondage
détérioration	vernissage			bouclier	lumière	foyer	fidélité	supervision
détriment	exposé			boulevard	lustre	générateur	force	surveillance
ennui	foire			capuchon	nitescence	germe	gage	vérification
entorse	forum			carapace	parure	geyser	garant	visite
grabuge	galerie			caution	prestige	griffon	garantie	
grief	introduction			chaperon	propagation	idée	hypothèque	
injure	montre			cloche	radiation	mère	infaillibilité	
injustice	narration			clôture	rayon	mine	justesse	
intérêt	orientation			conservation	réflexion	naissance	lucidité	
lésion	ostension			couvercle	réverbération	origine	nantissement	
mal	pilori			couverture	satisfaction	pépinière	ordre	
méfait	position			cuirasse	splendeur	point	police	
nuisance	préambule			dédain	transmission	point d'eau	précaution	
outrage	prélude			défense		point de départ	précision	
perte	présentation			écran		principe	sécurité	
préjudice	protase			égide		provenance	serrure	
ravage	récit			encouragement		puits	sincérité	
regrettable	rétrospective			enduit		racine	solidité	
ribordage	salon			enveloppe		résurgence		
sévice	situation			faveur		seuil		
tort				gaine				
vimaire				garantie				

« Conditions initiales » :

les verbes associés aux concepts génériques

les fragments de paires (C,V) pour certains concepts génériques :

dommage	aller apparaître augmenter conduire considérer contribuer décrire dépendre devenir émaner entraîner former impliquer manifester obéir partir permettre perturber procréer produire propager provoquer ramener réduire reproduire survenir survivre traduire	population	atteindre choisir compenser concerner conférer coudre créer décrire démontrer différer entrevoir éprouver équilibrer estimer expliquer fabriquer fournir gérer identifier intervenir introduire jouer manifester manquer miser partir persister peser	rayonnement	aboutir actualiser affecter aller apparaître appartenir attaquer atteindre attendre atténuer augmenter autoriser avertir calculer cesser changer cibler compenser compliquer comporter composer compter concerner conduire considérer
----------------	--	-------------------	--	--------------------	---

Résumé des résultats sur le corpus français

<i>Nb de substantifs différents dans le corpus (sous forme de lemmes)</i>	<i>Nb de tous les verbes différents dans le corpus</i>	<i>Nb des verbes associés aux concepts génériques</i>	<i>Nb des prédicats</i>
2464	1094	423	211

Les résultats pour les 9 concepts génériques

<i>Concept</i>	<i>Nb de verbes associés</i>	<i>Nb de prédicats</i>	<i>Nb de candidats- synonymes</i>	<i>Nb de candidats- synonymes dans le corpus</i>	<i>Nb de candidats- termes après AFC</i>	<i>Nb des entités adoptés</i>
dommage	28	20	37	11	7	6
exposition	92	69	34	22	5	1
contrôle	111	92	17	17	1	1
personnel	69	54	-	-	-	-
population	42	31	9	5	5	4
protection	148	127	77	32	25	17
rayonnement	144	103	26	12	10	8
sûreté	86	51	31	12	9	5
source	159	117	35	22	9	4

Exemple complet pour « dommage » - 1

Dans le dictionnaire des synonymes CRISCO on trouve 38 synonymes pour « dommage » ; dans le corpus il n'y a que 11 mots de cette liste.

Concept		<i>dommage</i>
CRISCO	38	atteinte, avarie, bête, brèche, calamité, casse, coup, dam, dégât, dégradation, déprédation, désavantage, destruction, détérioration, détriment, ennui, entorse, grabuge, grief, injure, injustice, intérêt, lésion, mal, méfait, nuisance, outrage, perte, préjudice, ravage, regrettable, ribordage, sévices, sinistre, tant pis, tort, vimaire
Présents dans le Corpus	11	atteinte, brèche, dégât, dégradation, destruction, détérioration, détriment, ennui, intérêt, lésion, perte

Exemple complet pour « dommage » fin : retour à l'expert

**Parmi les 11 candidats-termes issus de *dommage*
7 se trouvent, dans le corpus, avec les mêmes
verbes :**

*{brèche, dégradation, destruction, détérioration, détriment,
lésion, perte}*

Le mot «*brèche*» a été exclu par l'expert du domaine.

La valeur moyenne de la précision est égale à 68%

Conclusion et perspectives

- Dans cette approche la qualité du dictionnaire des synonymes joue un rôle crucial. Ces ressources sont en développement progressif.
- D'un côté les verbes peuvent servir à indiquer des propriétés fiables et significatives des concepts. Comme la plupart des verbes sont polysémiques, il faut utiliser un indice quantitatif de sélection.
- Afin d'améliorer le rappel pour les termes détectés dans le corpus il faudrait créer des classes de verbes ayant le même sens – c'est un travail à faire.
- Les valeurs de précision, relativement basses, s'expliquent du fait que les concepts génériques (comme *dommage*, par exemple) sont polysémiques.
- Les vrais termes du domaine sont des locutions composées de plusieurs mots (groupes nominaux). Pour réduire l'ambiguïté, il faudrait d'affiner l'algorithme d'extraction des nouveaux termes par des patrons syntactique et morphologique.

Merci pour votre attention !

Et pour vos questions...