Acquisition de synonymes à partir du TLFi : analyse de données et expérimentation

Nabil Hathout & Philippe Muller

30 novembre 2007

Introduction

- objectif : extraire des liens lexicaux d'un dictionnaire
- moyen : analyse des définitions
- exemple : expérience sur la recherche de synonymes
- techniquement : recherche de traits pertinents par apprentissage automatique
- évaluation : manuelle, référent Dicosyn

SÉPARER, verbe

1re Section. *Empl. trans.* Diviser; éloigner; isoler.

I. – [Le compl. d'obj. désigne une globalité] **Séparer qqc.[1] en qqc.[2].** Diviser, partager.

A. – 1. Couper, casser, déchirer (un objet, un matériau) en deux, en plusieurs morceaux. *cartonnage de la momie* (Gautier, *Rom. momie*, 1858, p. 182). *Le général (...) coupa* [*à ses miettes. Ils buvaient souvent, par gouttes, pour ne pas étouffer* (D'Esparbès, *Tumulte*, 1905

▸ [Le compl. second. est s.-ent.] *Tout en séparant le beefsteak, Hussonnet apprit à son con sent.*, t. 1, 1869, p. 41).

– *P. métaph.* [*Proust*] *a eu beau s'acharner à séparer en parcelles infimes la matière impa le lecteur referme-t-il son livre que par un irrésistible mouvement d'attraction toutes ces cohérent...* (Sarraute, *Ère soupçon*, 1956, p. 83).

2. Partager en deux une masse composée d'éléments libres les uns par rapport aux autres. *tête, tombaient en tresses indépendantes sur des épaules marmoréennes* (Lautréam., *Chant*

B. – 1. *CHIM., PHYS.* Dissocier, décomposer une substance, une particule de matière en *l'énergie nécessaire à la séparer en 2 ou plusieurs fragments* (Daudel, *Fond. chim. théor.*,

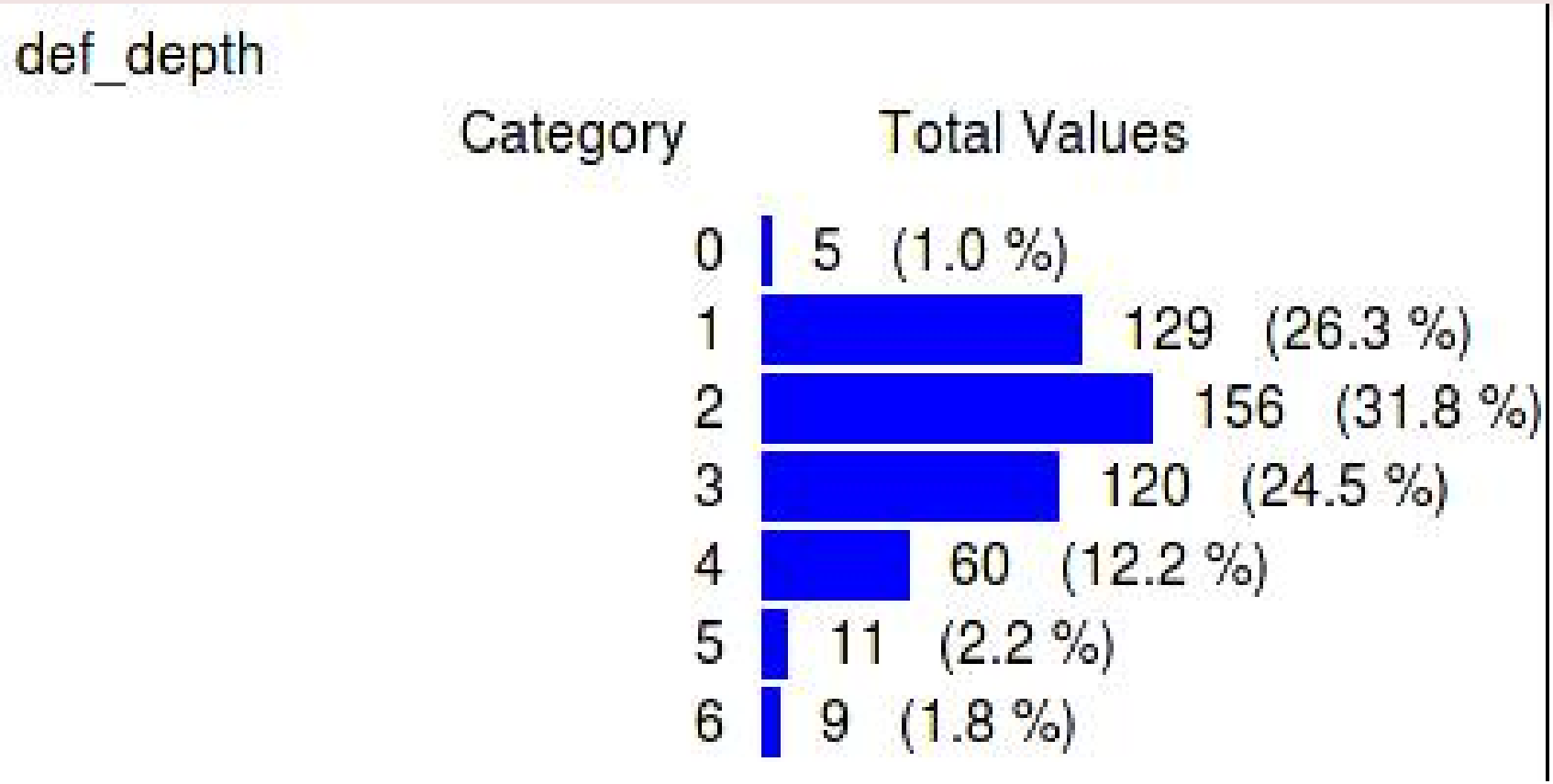
▸ [Souvent à la forme passive] *Les distillats sont séparés en deux fractions: huile paraffi* (Chartrou, *Pétroles natur. et artif.*, 1931, p. 93). *Le bois est complètement séparé en ses*

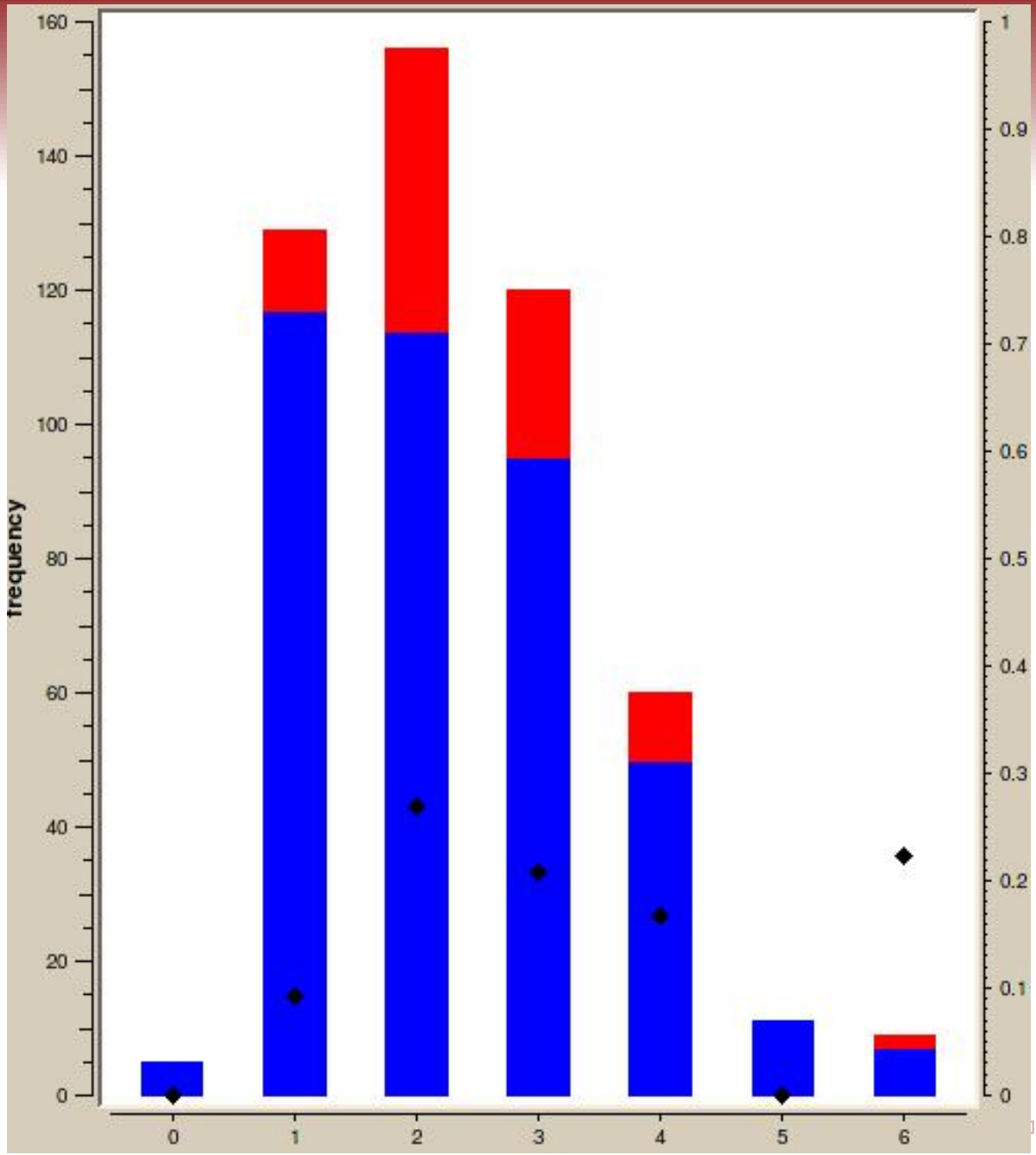
Analyse de données

facteurs pertinents pour l'apparition de synonymes ? étant donné une entrée et un mot apparaissant dans une de ses définitions

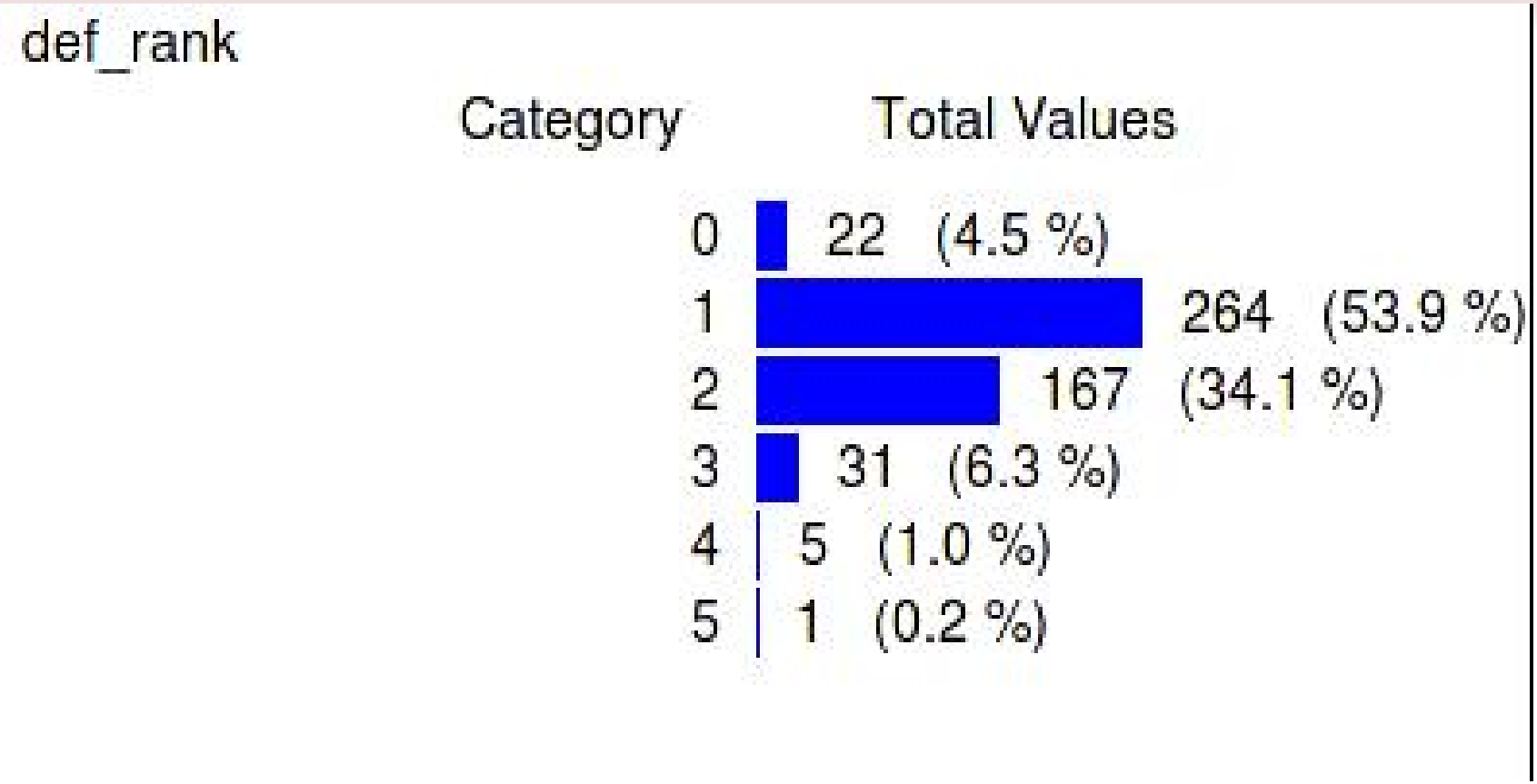
- position de la définition dans l'article
  - premières définitions plus importantes
  - définition plus haute dans la hiérarchie plus importante
- position dans la définition
  - premiers mots plus importants
- taille des définitions, nb de définitions
- le premier mot de la définition, pour les substantifs (type)
- rôle syntaxique dans la définition

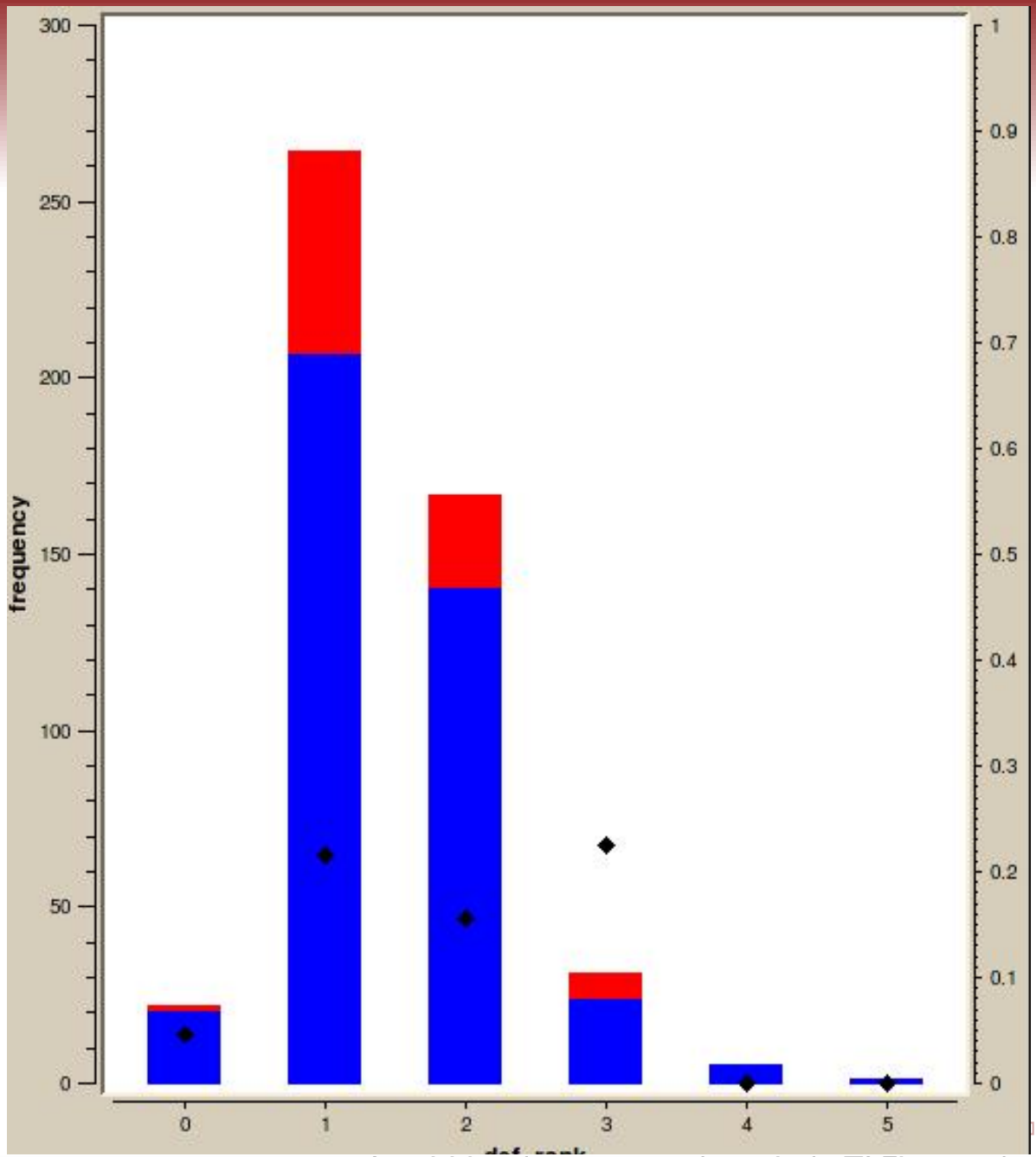
def_depth


Category
Total Values
0 5 (1.0 %)
1 129 (26.3 %)
2 156 (31.8 %)
3 120 (24.5 %)
4 60 (12.2 %)
5 11 (2.2 %)
6 9 (1.8 %)


frequency
160
140
120
100
80
60
40
20
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6

def_rank


Category
Total Values
0 22 (4.5 %)
1 264 (53.9 %)
2 167 (34.1 %)
3 31 (6.3 %)
4 5 (1.0 %)
5 1 (0.2 %)


frequency
300
250
200
150
100
50
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5

Autres facteurs

- "Proxémie" entre lemmes cibles, calculée sur le TLF
- morphologie : suffixes, préfixes partagés (abandonné)
- recouvrement des lexiques des définitions des deux mots candidats (ex. Jaccard)
- information mutuelle :

$$\mu(w_1, w_2) = -\log\left[\frac{freq(w_1, w_2)}{freq(w_1) \times freq(w_2)}\right]$$

calculée ici sur 10 ans du monde, contexte= fenêtre de trois mots

Proxémie

DONNER DES IDEES ASSOCIEES AU TERME SUIVANT :

ovale

Réponse données par vous (philippe m) pour cette partie (21165) : tourner - dessin - courbe - traçer - géométrie - traver - cercle - conique - ellipse - rugby - ballon -

Réponse données par le créateur ($ebos$) de cette partie (21165) : ballon - ovalie - rugby - tete - forme - corps - oeuf - ellipse - visage - géométrie -

Intersection : ballon - rugby - ellipse - géométrie -

DONNER DES TOUT (PAR EXEMPLE : 'CORPS', 'BRAS',
- OU ENCORE 'BANQUE' POUR 'GUICHET') DU TERME SU

mélanine

Réponse données par vous (philippe m) pour cette partie (27474) : peau - épiderme - derme - bronzage - corps - surface - organisme - animal - humain -

Réponse données par le créateur (gabzou) de cette partie (27474) : épiderme - substance - pigment - couleur - brune - brun - bronzage - cellule - protéine - peau -

Intersection : épiderme - bronzage - peau -

Information mutuelle

sample_mi_1000V (Examples)

	2 mutual_info	syno	inst_id
1	9.944281578063965	False	V.rehausser%V.affaisser
2	9.846164703369141	True	V.refondre%V.remanier
3	9.822524070739746	False	V.papillonner%V.agiter
4	9.770794868469238	True	V.ployer%V.tordre
5	9.770023345947266	True	V.retoucher%V.rectifier
6	9.760146141052246	False	V.imbriquer%V.chevaucher
7	9.712531089782715	True	V.amadouer%V.flatter
8	9.591949462890625	False	V.refouler%V.inhiber
9	9.576280593872070	False	V.effilocher%V.étirer
10	9.502298355102539	True	V.vocaliser%V.chanter
11	9.434070587158203	True	V.copier%V.imiter
12	9.424160957336426	True	V.desceller%V.briser
13	9.378643035888672	False	V.halluciner%V.obséder
14	9.319133758544922	False	V.mutiler%V.déformer
15	9.244208335876465	True	V.électriser%V.exalter
16	9.235912322998047	False	V.refondre%V.moderniser
17	9.210223197937012	False	V.assigner%V.comparaître
18	9.204236984252930	False	V.amaigrir%V.affaiblir

Recouvrement de définition

V.remonter	V.monter	0.36	TRUE
V.immatriculer	V.matriculer	0.36	TRUE
V.chatter	V.chatonner	0.27	TRUE
V.prouver	V.démontrer	0.26	FALSE
V.sursauter	V.tressauter	0.25	TRUE
V.natter	V.tresser	0.24	TRUE
V.pâlir	V.affaiblir	0.23	FALSE
V.terminer	V.finir	0.22	TRUE
V.fraîchir	V.refroidir	0.2	FALSE
V.panteler	V.haleter	0.2	TRUE
V.croître	V.augmenter	0.19	TRUE
V.reboiser	V.boiser	0.19	TRUE
V.assouvir	V.rassasier	0.19	TRUE
V.fleurer	V.flairer	0.19	TRUE

Une expérience d'apprentissage automatique

- problème de catégorisation binaire : synonyme (Vrai)/ non-synonyme (Faux)
- apprentissage = induction sur une base d'exemples + prédictions sur de nouveaux cas
- en pratique, à partir d'un ensembles de cas (les instances), on sépare en un ensemble d'apprentissage utilisé pour faire l'induction, et un ensemble de tests pour évaluer le prédicteur.
- diverses méthodes applicables pour l'induction

Apprentissage simple bayésien

probabilité d'une classe sachant les traits de l'exemple =

$$\log(p(c|e)) = \sum_{t_i \in traits} \log \frac{p(c, t_i(e)) * p(c)}{p(t_i(e))}$$

Particularité de la tâche

- gros biais vers le "non-synonyme" ($\approx$ 80% sur les verbes)
- trouver une paire de synonyme nous intéresse plus que trouver une paire non-synonyme !
- ... mais les algorithmes d'induction partent tous des probabilités de départ et vont se contenter de trouver à peu près les non-synonymes
- $\longrightarrow$ on déforme la distribution initiale pour apprendre sur 50% d'exemple de chaque (Vrai ou Faux)
- on teste sur la distribution écologique, et on regarde les valeurs suivantes

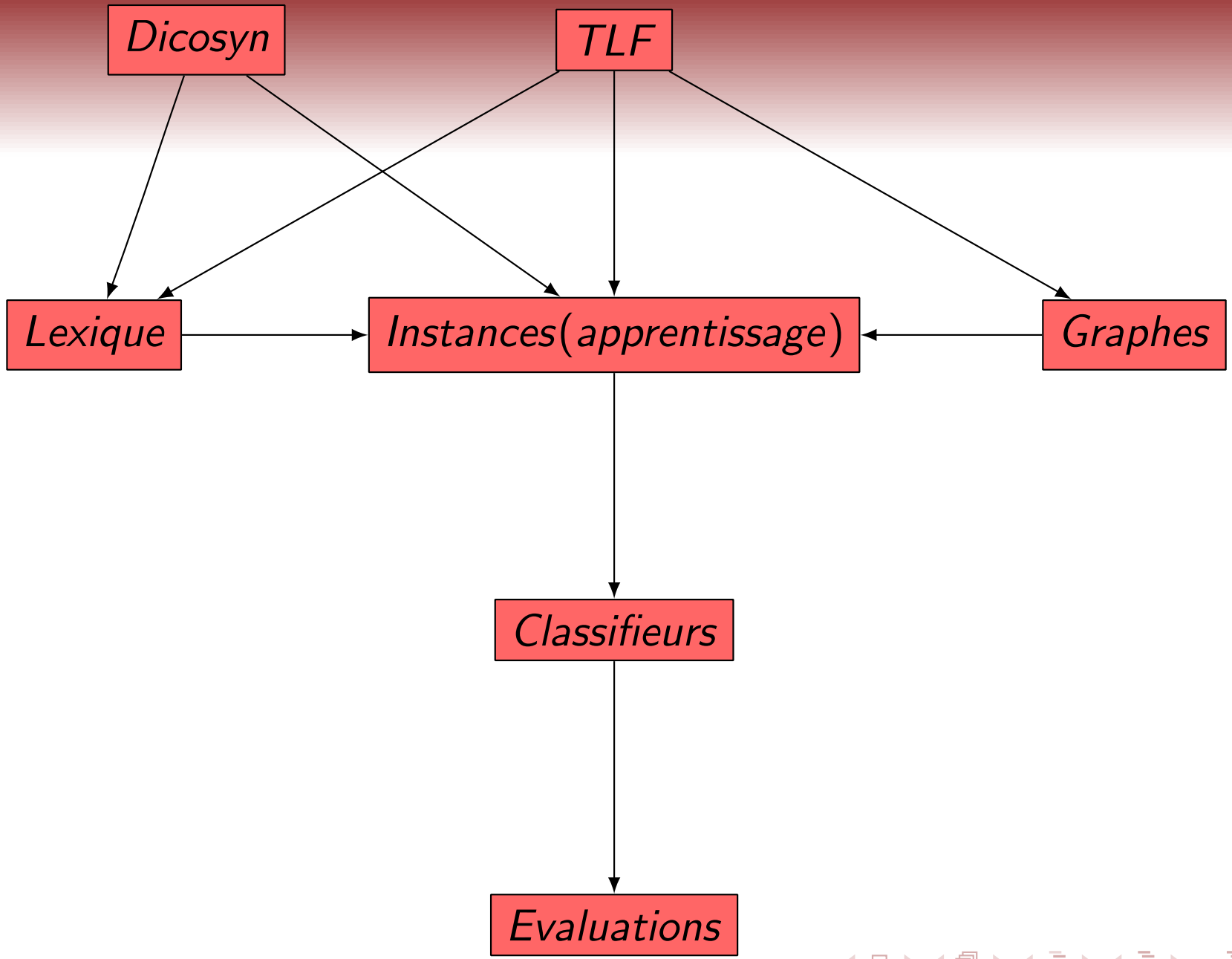
Evaluation

- "précision" de la classe "synonyme" =

$$\frac{\text{nb de couples marqués synonymes}}{\text{nb réel de couples synonymes}}$$

- "rappel" des synonymes =

$$\frac{\text{nb de couples marqués synonymes correctement}}{\text{nb total de couples synonymes à trouver}}$$


Dicosyn
TLF
Lexique
Instances(apprentissage)
Graphes
Classifieurs
Evaluations

Points de comparaison

- bornes inférieures : "baselines"
  1. tout est synonyme dans une définition
  2. seulement le premier mot de la même catégorie que l'entrée
- bornes supérieures : la même tâche réalisée par l'humain
  - dictionnaire de synonymes de référence
  - expérience d'annotation par lexicographes

Résultats

environ 5600 instances au total

Méthode	Precision	Rappel	F-score
bayes	0.385	0.632	0.478
arbre de décision	0.481	0.419	0.448
premier mot	0.347	0.545	0.420
règles (cn2)	0.403	0.403	0.403
meilleur voisin (knn)	0.280	0.532	0.367
tous syno.	0.211	1.000	0.348

grande variabilité de ce genre de tests (trop peu d'exemples)

L'humain ...

Cohérence de Dicosyn :

	1	2	3	4	5	6	7
1		0.591	0.303	0.341	0.441	0.243	0.274
2			0.415	0.262	0.362	0.283	0.304
3				0.115	0.156	0.420	0.555
4					0.814	0.261	0.254
5						0.249	0.248
6							0.532
7							

TAB.: F-score des dictionnaires de Dicosyn entre eux, restreints à leur vocabulaire commun

L'humain ...

	Rappel sur l'ensemble des verbes de dicosyn
Bailly	0.090
Benac	0.144
Bertaud	0.541
Guizot	0.043
Lafaye	0.060
Larousse	0.499
Robert	0.567

Conclusion

- la synonymie est une notion mal définie (surprise...)
- référentiel sujet à caution
- accord inter-humain faible ($\kappa \approx 0.5$)
- sélection des traits cruciale face au manque d'exemple ;
  les plus informatifs : premier mot de la définition, profondeur de la définition, nombre de définition d'une entrée, rank d'une définition, recouvrement, info mutuelle, proxémie.
- succès mitigé, travail en cours