

Proximité synonymique Badauds : flâneurs ou spectateurs ?

Dominique Laurent (Synapse Développement)
Patrick Séguéla (Synapse Développement)

Résumé – Abstract

Afin d'améliorer un dictionnaire de synonymes existant, pour l'enrichir et affecter les synonymes d'un indice de proximité, nous avons mis en œuvre des méthodes de comparaison vectorielle des contextes d'utilisation des entrées et des synonymes sur de très gros corpus et sur le Web. Le dictionnaire résultant fournit ainsi un indice de proximité synonymique pour chacun des synonymes des différents sens d'un mot, indice qui s'est révélé très pertinent pour l'extension de requête en extraction d'information.

In order to improve an existing dictionary of synonyms, to enrich it and affect the synonyms of a proximity coefficient, we have implemented methods using vectorial comparison of the contexts for entries and synonyms on very large corpora and on Web. The resulting dictionary thus provides a proximity coefficient for each synonym of the various senses of each word, coefficient which appeared very relevant for the request extension in information extraction.

Keywords – Mots Clés

synonyme, synonymie, proximité, similitude, proximité synonymique, indice de proximité synonymique, analogies, équivalents, dictionnaire de synonymes.

synonym, synonymy, proximity, similarity, synonymic proximity, synonymic proximity index, analogies, equivalents, thesaurus, dictionary of synonyms.

1 Synonymie et extraction d'information

Les dictionnaires courants définissent les synonymes comme des mots de sens voisins¹. En lexicographie, les synonymes ont généralement une partie de définition commune et des traits les différenciant². En traitement de la langue, on considère souvent que deux mots sont synonymes s'ils ont des contextes similaires ou très proches, s'ils peuvent être employés l'un

¹ « Se dit de mots ou d'expressions qui ont une signification très voisine et, à la limite, le même sens » (Le Petit Robert 2003). « Se dit de deux ou plusieurs mots de même fonction grammaticale, qui ont un sens analogue ou très voisin » (Petit Larousse 2003).

² On appelle synonymes les termes dont le sens a de grands rapports, et des différences légères, mais réelles. (préface au Dictionnaire universel des synonymes de la langue française, François Guizot, Maradan, 1809)

à la place de l'autre. Selon Jacques Poitou « des termes synonymes sont des termes interchangeables dans tous les contextes (c'est-à-dire qui ont la même distribution) »³.

Pour son analyseur de la langue française, Cordial, notre société a développé un dictionnaire de synonymes il y a maintenant une dizaine d'années. Conçu comme les dictionnaires papier à partir des dictionnaires existants, puis enrichi au fil des années, ce dictionnaire était destiné à l'aide à la rédaction et utilisé comme ses homologues papier pour retrouver « le mot juste ». Il offrait une liste à plat avec une différenciation par type grammatical et par sens. L'utilisation de ce dictionnaire dans le cadre d'un moteur d'extraction d'information (« Chercheur ») a montré les insuffisances de ce format : synonymes mal différenciés sémantiquement et absence de coefficient permettant de classer les synonymes selon la proximité. Ce qui nous a incité à revoir le format et les caractéristiques de ce dictionnaire. Cet article décrit ce travail d'amélioration, les méthodes utilisées et mesure les résultats obtenus.

2 Synonymie : état de l'art

Comme la polysémie, la synonymie est d'abord contextuelle parce que le sens est contextuel. L'usage des mots dans la langue accentue les différences synonymiques ou les réduit. Ainsi le mot « papotage » pourra le plus souvent remplacer le mot « bavardage » même s'il a une connotation plus familière, mais on imagine mal dire « cet élève, je lui ferai passer le goût du papotage ». De sorte que les cooccurrences marquent la synonymie et ses limites.

Les dictionnaires de synonymes ont habituellement été conçus pour un utilisateur humain. Pour réduire la taille des ouvrages, les renvois y sont fréquents. Les dictionnaires informatiques évitent généralement ces renvois, mais ils apportent rarement plus. Nous prendrons ici pour exemple le mot « badaud » dont voici les synonymes fournis par quelques dictionnaires :

flâneur, spectateur, niais, nigaud, gobe-mouche, oisif, crédule, curieux [*ancien dictionnaire de synonymes intégré à Cordial*⁴]

I. Non favorable : crédule, gobe-mouches (fam.), niais, nigaud, oisif, sot. Arg. : cave, pingouin. => bête. II. neutre : curieux, flâneur, lèche-vitrine (fam.), promeneur [*Henri Bertaud du Chazaud*]

benêt – niais – nigaud (cf : nigaud) [*Georges Younes*]

oisif, promeneur, passant, curieux [*Paul Rouaix*]

curieux, flâneur, passant [*Thomas Decker*]

v. flâneur (promeneur⁵), curieux (indiscret) [*É. Genouvrier*]

v. flâneur, spectateur (témoin, observateur) et niais. [*Roger Boussinot*]

v. curieux (badaud, flâneur), flâneur (badaud, curieux, passant, promeneur), promeneur (badaud, flâneur, passant) [*Marc Baratin*]

v. flâneur (badaud, bayeur) [*Henri Bénac*]

³ Sémantique lexicale (http://nte.univ-lyon2.fr/~poitou/Morpho_Lexico/7_rel-sem.html)

⁴ Dictionnaire intégré dans les versions 1 à 8 de Cordial.

⁵ Les mots entre parenthèses sont les synonymes fournis pour les renvois.

Ces listes font apparaître :

- Une absence de catégorisation grammaticale (certes l'usage de « badaud » en adjectif est beaucoup plus rare qu'en substantif mais « badaud » appartient bien aux deux catégories);
- Aucune distinction sémantique. Seul le dictionnaire de Bertaud du Chazaud introduit des niveaux de langage;
- Certaines incohérences : « lèche-vitrine » ne saurait désigner un être humain ; si à « cave » est associée la naïveté d'un des sens de « badaud », « pingouin » est argotiquement un « type », un « zigue », et n'a que peu à voir avec un « badaud ».

Le dictionnaire de synonymes de l'Université de Caen, élaboré à partir de certains des dictionnaires cités ci-dessus, offre, grâce à l'analyse par cliques, une vision plus élaborée puisque les synonymes sont regroupés sous 4 composantes :

1. badin.
2. bête, crédule, gobe-mouche, niais, nigaud, sot, stupide.
3. curieux, flâneur, oisif, promeneur, rôdeur.
4. lèche-vitrine.

Les composantes 1 et 4 correspondent à de fausses pistes des dictionnaires compilés (souvent anciens et parfois peu fiables) mais les composantes 2 et 3 mettent clairement en valeur deux sens réels de « badaud ». On sait que « badaud » vient du provençal « bader » (regarder bouche bée), lui-même issu du latin « batere » (bâiller) et avait au XVI^e siècle le sens de « sot », « niais ». Ce sens s'est presque complètement perdu et la valeur péjorative n'a cessé de se réduire au fil des siècles, sauf pour l'adjectif encore associé à la notion de crédulité, de niaise curiosité⁶. C'est à cet ancien sens que réfère la composante 2 qui, toutefois, associe des adjectifs comme « bête » (dont le substantif n'a pas ici le sens mis en valeur) ou « stupide » à un substantif pur comme « gobe-mouche » et à des adjectifs/substantifs. Le second sens, décrit par la composante 3, correspond bien au sens moderne du substantif « badaud ».

3 Mesure de la proximité sémantique

Les synonymes de notre dictionnaire (présentés plus haut) offraient une plus grande homogénéité grammaticale (tous les synonymes sont des noms) mais aucun indice ne nous permettait de différencier les synonymes les plus proches des plus éloignés. Cette proximité synonymique nous paraissant capitale pour l'extension de requête en extraction d'information, nous avons développé des méthodes permettant de mesurer cette proximité synonymique.

Dans l'exemple de « badaud », notre dictionnaire ne couvrait que les synonymes du nom. Or, en analysant les synonymes fournis, comme « flâneur », nous avons constaté que certains de ces synonymes offraient eux-mêmes des synonymes pour le type adjectival, dont le mot « badaud » en tant qu'adjectif. D'où une première extension possible du dictionnaire.

⁶ « le peuple de Paris tant sot, tant badaud et tant inepte de nature », Anatole France ; « l'immense et badaude majorité », Stendhal in TLF (article « badaud », tome 3, page 1203), accessible sur <http://atilf.inalfr.fr>.

3.1 Extensions du dictionnaire

Dans une première phase, nous avons appliqué de simples règles de commutativité. Ainsi, le mot « badaud » n'avait pas de synonymes pour le type adjectival mais « flâneur » et « niais » avaient « badaud » comme synonyme du type adjectival. La simple application de la règle de commutativité a augmenté le dictionnaire d'environ 15% et les listes de synonymes de « badaud » sont devenues⁷ :

1. adjectif

flâneur (sens 1) | niais (sens 2)

2. nom

crédule (sens 2)		gobe-mouche (sens 2)		nigaud (sens 2)		promeneur
curieux (sens 5)		gobe-mouches (sens2)		oisif (sens 2)		spectateur (sens 1)
flâneur (sens 3)		niais (sens 3)		passant (sens 2)		

Dans une deuxième phase, nous avons recherché les mots ayant pour synonyme un des mots de cette nouvelle liste. Enfin, dans une troisième phase, nous avons recherché les mots ayant pour synonyme un des mots issus de la seconde phase. Chacun des synonymes pouvant être présent dans la liste issue de la première phase, une ou plusieurs fois comme synonyme d'un des mots de la liste de la deuxième phase, et une ou plusieurs fois comme synonyme d'un des mots de la liste de la troisième phase, nous avons affecté les synonymes de la liste finale d'un premier indice de proximité lié au type et au nombre d'occurrences. Dans cet indice i_k entrent deux coefficients dont la valeur a été définitivement fixée lors de la seconde étape (cf 3.2) :

$$i_k = n_1 + 0,5.(n_1=0) + 0,32.(n_2 - (n_1=0)) + 0,17.(n_3 - ((n_1=0)\&(n_2=0))) ;$$

où n_1 est à 1 si le synonyme figure dans la liste issue de la première phase, à 0 sinon (ainsi l'équation logique $n_1=0$ a pour valeur 1 si n_1 est nul, la valeur 0 sinon), et n_2 est le nombre d'occurrences du mot dans la liste issue de la deuxième phase, et n_3 le nombre d'occurrences du mot dans la liste issue de la troisième phase.

Ainsi, l'adjectif « flâneur » a le coefficient $1 + (0,17 \cdot 2) = 1,34$ car il figure dans la liste de départ et il est synonyme de « musard » et « oisif » qui ont eux-mêmes pour synonyme l'adjectif « flâneur ». Il a le coefficient $1 + (0,32 \cdot 2) + (0,17 \cdot 1) = 1,81$ car il figure dans la liste de départ, il est synonyme de « passant » et « promeneur » qui sont dans la liste de départ, et il est synonyme de « marcheur » qui a été ajouté dans la seconde phase.

Voici pour le mot « badaud » les synonymes avec leurs coefficients :

1. adjectif

⁷ Les mentions de sens entre parenthèses (« sens 1 », « sens 2 », etc.) correspondent à notre découpe de sens et de type grammatical. La numérotation utilisée dans nos bases grammaticales et nos dictionnaires s'applique en effet aux types grammaticaux et aux sens. Ainsi « badaud » est classé « sens1 » pour l'adjectif, « sens2 » pour le nom ; « curieux » est classé « sens1 » comme adjectif (« ayant envie de connaître, envie de comprendre »), « sens2 » comme adjectif (« désirant tout découvrir, y compris ce qui est caché »), « sens3 » comme adjectif (« surprenant, étonnant »), « sens4 » comme nom (« celui qui désire connaître, apprendre »), « sens5 » comme nom (« celui qui se rapproche dès qu'un événement insolite se déroule »), « sens6 » comme nom (« étrange ») et « sens7 » comme nom (« argotiquement, juge d'instruction »).

benêt (sens 1) 0,17	flâneur (sens 1)1,34	naïf (sens 1).....0,64	piéton (sens 1)..... 0,32
candide..... 0,17	godiche0,34	niais (sens 2)..... 1,32	sot (sens 1)..... 0,34
crédule (sens 1) 0,34	jobard (sens 1)0,17	nigaud (sens 1)0,49	
curieux (sens 1)..... 0,49	musard (sens 1).....0,49	oisif (sens 1)0,51	

2. nom

benêt (sens 2) 0,34	gobe-mouches (2) ..1,17	niais (sens 3)..... 1,17	piéton (sens 3)..... 1,34
crédule (sens 2) 1,00	gogo.....0,51	nicodème0,17	promeneur..... 1,64
curieux (sens 5)..... 1,32	jobard (sens 2)0,34	nigaud (sens 2) 1,00	sot (sens 2)..... 0,34
flâneur (sens 3)..... 1,81	marcheur.....0,64	oisif (sens 2)..... 1,17	spectateur (sens 1). 1,17
gobe-mouche (2) ... 1,17	musard (sens 2).....0,49	passant (sens 2) 1,49	

Pour un nombre d'entrées peu accru (37 014 entrées), le volume du dictionnaire a nettement augmenté puisqu'il est passé de 151 681 synonymes et antonymes à plus de 580 000. Nous disposons ainsi d'une base d'environ 16 synonymes par entrée (tous sens confondus) mais cette liste indicée reflétait les choix subjectifs des lexicographes ayant constitué le dictionnaire primitif, elle n'offrait pas une mesure objective de la proximité synonymique.

3.2 Vectorisation contextuelle sur corpus et via le Web

Notre société a constitué un ensemble important de ressources linguistiques (analyseur syntaxique, désambiguïseur sémantique, taxonomie...) et dispose également d'un corpus de plus de 500 millions de mots dans lequel chacune des occurrences a été lemmatisée, toutes les phrases comportant un lemme donné étant groupées dans plus de 80 000 fichiers correspondant à chacun de ces lemmes. Grâce à ces fichiers, il nous est possible d'extraire les contextes principaux de chaque entrée du dictionnaire de synonymes. Ainsi pour le mot « badaud », figurant dans 1493 phrases de notre corpus, nous disposons du contexte suivant :

adjectifs précédant «badaud(s)» : quelques (5,2%), nombreux (3,6%), rares (2,3%), simple (2,3%), autres (2,0%)

adjectifs suivant «badauds» : présents (1,4%), massés (1,2%), habituels (1,0%)

substantif suivi de «de badauds» : foule (5,2%), milliers (1,9%), centaine(s) (1,7%), dizaine(s) (1,2%), groupe (1,2%), cercle (1,0%)

substantif suivi de «des badauds» ou «du badaud» : foule (2,1%), yeux (1,0%)

substantif dans les 5 mots suivant «badaud(s)» : touristes (1,3%), rue (1,0%)

verbe + article + «badaud(s)» : attirer (1,1%)

verbe suivant (sauf «avoir» et «être») : venir (2,1%), regarder (1,6%), applaudir (1,2%)

Le contexte de « badaud » se révèle assez peu figé, même si l'ensemble des contextes relevés ci-dessus couvre 34,7 % des occurrences de « badaud » (le total est de 42,6% mais certains contextes sont communs comme dans « la foule des badauds applaudit »). D'après nos tests, ce pourcentage correspond à peu près à la moyenne de contextes obtenus pour l'ensemble des entrées en ne tenant compte que des occurrences supérieures à 1% (ou 2 en nombre absolu).

Pour chacun des synonymes de la liste étendue (33 termes pour « badaud »), les contextes sont comparés aux contextes ci-dessus. Ainsi pour « flâneur », les pourcentage de phrases dans lesquelles on trouvera « quelques flâneurs », « nombreux flâneurs »... « flâneurs présents », « foule de flâneurs », « milliers de flâneurs », « flâneurs regarder », etc. sont relevés pour comparaison avec chacun des contextes de « badaud ».

Pour l'ensemble des contextes pouvant faire l'objet d'une requête sur un moteur Web, la même analyse est effectuée. Les moteurs de recherche sur le Web ne permettant qu'une utilisation réduite des jokers, il est très lourd de rechercher l'ensemble des contextes où « badaud » est, par exemple, suivi de « touriste(s) » ou « rue » dans les 5 mots suivants, sauf à analyser l'ensemble des pages contenant « badaud » (10 090 sur Google en juillet 2002). Afin de contourner cette limitation et afin d'accélérer les recherches, seuls ont été relevés le nombre d'occurrences total du mot-vedette et des synonymes puis le nombre d'occurrences de chacun des contextes pour le mot-vedette et les synonymes.

Ainsi, Google est appelé avec le mot « badaud » et l'on relève le nombre d'occurrences (2040) puis Google est à nouveau appelé avec le pluriel « badauds » et l'on relève le nombre d'occurrences (8050). Ensuite Google est appelé avec chacun des contextes possibles. Le tableau ci-dessous donne la grille des résultats pour le substantif « badaud » et ses synonymes (pour des raisons de présentation, le tableau n'est que partiel) :

	<i>badaud</i>	<i>flâneur</i>	<i>spectateur</i>	<i>promeneur</i>	<i>passant</i>	<i>oisif</i>	<i>curieux</i>	<i>marcheur</i>
nbre occurrences	10090	3810	186600	30300	355700	7180	113000	23650
quelques Xs	7,06%	2,21%	1,14%	1,80%	2,53%	1,95%	2,65%	0,76%
nombreux Xs	5,14%	0,81%	2,77%	2,90%	1,08%	0,29%	2,40%	0,81%
simple X	0,82%	0,93%	1,74%	2,07%	0,97%	0,00%	4,52%	0,11%
foule de Xs	4,51%	0,00%	0,37%	0,40%	0,16%	0,78%	2,09%	0,13%
foule des Xs	3,52%	0,12%	0,36%	0,37%	0,18%	0,00%	0,72%	0,10%
dizaines de Xs	0,78%	0,00%	0,16%	0,53%	0,32%	0,00%	0,32%	0,13%
centaines de Xs	1,75%	0,00%	0,80%	0,36%	0,16%	0,00%	0,66%	0,62%
milliers de Xs	1,95%	1,05%	4,93%	0,46%	0,19%	0,10%	1,14%	1,14%
groupe de Xs	0,78%	0,23%	0,24%	0,50%	0,08%	0,10%	0,15%	2,25%
regard des Xs	0,74%	0,12%	0,19%	0,32%	0,03%	0,00%	0,50%	0,04%
curiosité des Xs	0,82%	0,47%	0,10%	0,22%	3,21%	0,20%	0,00%	0,00%
yeux des Xs	1,03%	0,58%	0,85%	0,64%	0,33%	0,00%	0,28%	0,07%
attirer les Xs	0,82%	0,12%	0,18%	0,13%	0,22%	0,00%	0,63%	0,02%
attire les Xs	0,51%	0,23%	0,06%	0,20%	0,11%	0,00%	0,39%	0,05%
	30,25%	6,86%	13,90%	10,90%	9,57%	3,41%	16,46%	6,23%

Les données numériques extraites des moteurs de recherche sont parfois douteuses. Ainsi il n'est pas possible de distinguer les types grammaticaux, on obtient parfois des pages en langue étrangère lorsque le mot existe dans d'autres langues et même lorsque l'on a pris soin de paramétrer la recherche sur les seuls pages en français, certains mots sont également des formes verbales (« passant » par exemple) ou des noms propres (la plupart des moteurs ne différencient pas les majuscules des minuscules). Ceci fausse en particulier le nombre total d'occurrences pour chacun des mots, les contextes étant moins sensibles à ces erreurs car constitués d'un ensemble de mots. Pour réduire ces effets de bord, les pourcentages sont calculés par rapport au total des pluriels pour les contextes contenant un pluriel, par rapport au singulier pour les autres.

Une fois effectuées ces deux analyses, sur le corpus et sur le Web, les résultats sont associés mais, compte tenu des effets de bord notés ci-dessus, l'analyse sur le Web a été pondérée d'un facteur 1/3 par rapport à l'analyse contextuelle sur corpus, cette valeur étant apparue la plus pertinente lors des tests itératifs effectués. La valeur de proximité contextuelle, obéit elle-même à une équation complexe dont les coefficients ont été déterminés de façon itérative.

Cette valeur de proximité contextuelle est basée sur les carrés des différences entre les pourcentages de fréquence. Cependant il a semblé qu'un contexte plus élevé devait être favorisé par rapport à un contexte moins élevé, ce que les tests itératifs ont confirmé. Ainsi, dans le tableau des données recueillies sur le Web, l'expression « milliers de badauds » a une fréquence de 0,0195 alors que l'expression « milliers de flâneurs » a une fréquence de 0,0105 et que l'expression « milliers de spectateurs » a une fréquence de 0,0493. Dans ce cas, les carrés des écarts seront de 0,000081 et 0,000888, le mot « flâneur » semblant donc beaucoup plus proche de « badaud » que « spectateur » pour ce contexte précis. Après itérations sur un ensemble de valeurs et comparaison des résultats obtenus avec les coefficients lors d'une extension de requête, l'équation suivante a été retenue :

$$\text{Si } fr_{vedette} \geq fr_{synonyme} \quad v_k = (fr_{vedette} - fr_{synonyme})^2$$

$$\text{si } fr_{vedette} < fr_{synonyme} \quad v_k = ((fr_{synonyme} - fr_{vedette}) / 2,6)^2$$

Dans le cas ci-dessus, le coefficient sera toujours de 0,00081 pour flâneur mais sera de 0,000131 pour spectateur, soit une valeur 60% supérieure et non 11 fois supérieure. Un vecteur résultante est associé à l'indice i_k calculé précédemment. On obtient ainsi l'indice de proximité ip_k :

$$ip_k = 4,7 \cdot \left(\frac{3}{4} \cdot \left(1 - \sum_{i=nc}^{i=1} v_{kc} \right) + \frac{1}{4} \cdot \left(1 - \sum_{i=nc}^{i=1} v_{kw} \right) \right) + (n_1 + 0,5 \cdot (n_1 == 0) + 0,32 \cdot (n_2 - (n_1 == 0))) + 0,17 \cdot (n_3 - ((n_1 == 0) \& (n_2 == 0)))$$

où nc désigne le nombre de contextes étudiés (13 pour « badaud »), où v_{kc} désigne le carré des écarts de fréquence sur le corpus pondéré selon la formule décrite plus haut, et où v_{kw} désigne le carré des écarts de fréquence pour le Web. Les coefficients (4,7; 3/4; 1/4; 0,5; 0,32; 0,17), ainsi que le coefficient 2,6 utilisé pour pondérer l'écart positif, résultent d'une multitude d'itérations effectuées en étendant les requêtes sur un jeu de 200 questions et en comparant les résultats obtenus en fonction des différents coefficients.

Le coefficient résultat est ramené sur une échelle de 0 à 255, fournissant ainsi un indice de proximité associé à chacun des synonymes du mot-vedette. Pour le mot « badaud », dans ses deux types grammaticaux, les listes deviennent :

1. adjectif

flâneur (sens 1)	255
oisif (sens 1)	236
crédule (sens 1)	71
niais (sens 2)	32
nigaud (sens 1)	32
curieux (sens 1)	31
piéton (sens 1)	30
musard (sens 1)	21
sot (sens 1)	16
jobard (sens 1)	10
godiche	9
benêt (sens 1)	5
naïf (sens 1)	3
candide	2

2. substantif

flâneur (sens 3)	255	jobard (sens 2)	7
musard (sens 2)	123	benêt (sens 2)	4
gobe-mouches (sens 2)	110	marcheur (sens 2)	3
oisif (sens 2)	86	crédule (sens 2)	1
gobe-mouche (sens 2)	85	gogo	1
spectateur (sens 1)	72		
promeneur (sens 1)	64		
piéton (sens 3)	24		
passant (sens 2)	23		
curieux (sens 5)	21		
nigaud (sens 2)	21		
niais (sens 3)	20		
sot (sens 2)	10		
nicodème	10		

Cette double liste indicée appelle plusieurs remarques :

1. les synonymes de l'adjectif « badaud » divergent sensiblement des synonymes du substantif. Le concept de « niais », « crédule » s'applique bien à l'adjectif, beaucoup moins au nom (« nigaud » vient en 11^e position, suivi de « niais » et « sot »).
2. certains synonymes inconnus des dictionnaires de synonymes usuels apparaissent avec de bons taux de proximité. C'est par exemple le cas du mot « musard », sous ses deux formes, adjectivale et nominale.
3. La présence du mot « musard » ou encore des deux écritures de « gobe-mouche(s) » met en valeur une spécificité de notre méthode : les mots rares ne sont pas laminés par l'algorithmique, pour autant que leur présence soit effective dans le corpus et surtout sur le Web : « musard » figure 71 fois dans le corpus (sur le Web : 2830), « gobe-mouche » 9 fois (sur le Web : 502) et « gobe-mouches » 48 fois (Web : 680).

4 Dictionnaire résultat.

Une fois effectués les traitements décrits ci-dessus pour les 37 014 mots vedettes et leurs 580 000 synonymes, le dictionnaire résultat a été réduit par suppression des synonymes obtenant un taux trop réduit de proximité. Le taux actuel de sélection a été fixé à 21, ce qui laisse tout de même 273 179 synonymes et 50 881 antonymes. Comme notre dictionnaire gère les formes fléchies et les formes conjuguées en fournissant les synonymes aux formes correspondantes, l'ensemble représente 283 624 mots vedettes pour un total de 3 583 146 synonymes et 618 591 antonymes.

Dans notre moteur, l'indexation porte sur des blocs (actuellement de 8 ko). Après analyse de la requête de l'utilisateur, les blocs correspondant aux mots pivots et aux concepts de la requête sont analysés afin d'en extraire les phrases censées répondre à la requête. Comme l'indique le tableau ci-dessous, l'intégration du nouveau dictionnaire de synonymes a nettement amélioré les résultats. Nous avons testé notre moteur avec l'ancien dictionnaire puis avec le nouveau dictionnaire sur un même ensemble de 200 questions appliquées à deux corpus indexés d'environ 10 Mo et 100 Mo. Ces corpus sont constitués de :

corpus 10 Mo : 4 091 dépêches AFP pour une taille totale de 6,83 Mo
 223 notices de l'Encyclopaedia Universalis pour une taille de 3,34 Mo

corpus 100 Mo : 26 503 dépêches AFP pour une taille totale de 44,0 Mo
 3 593 notices de l'Encyclopaedia Universalis pour une taille de 55,2 Mo

Voici les résultats sur ces deux corpus avec les deux dictionnaires :

	corpus 10 Mo ancien dictionnaire de synonymes	corpus 10 Mo nouveau dictionnaire de synonymes	corpus 100 Mo ancien dictionnaire de synonymes	corpus 100 Mo nouveau dictionnaire de synonymes
1 ^e réponse juste	69,0 %	78,0 %	61,0 %	67,0 %
1 ^e ou 2 ^e réponse juste	76,0 %	86,5 %	72,0 %	77,5 %
1 ^e à 5 ^e réponse juste	86,0 %	94,5 %	78,5 %	86,5 %
1 ^e à 10 ^e réponse juste	88,5 %	96,0 %	83,0%	93,0 %

pas de réponse juste dans les 10 premières	11,5 %	4,0 %	17,0 %	7,0 %
pas de réponse juste dans les 64 premières	10,0 %	3,5 %	14,5 %	4,5 %

L'amélioration est donc importante, tout particulièrement pour ce qui est des questions laissées sans réponse, soit dans les 10 premières phrases fournies, soit dans les 64 phrases fournies (maximum paramétrable dans le logiciel), d'autant que 4 questions (2%) n'ont aucune réponse acceptable dans les deux corpus de textes ! Si l'on enlève des résultats ces 4 questions sans réponse possible, le nombre de questions sans réponse est en gros divisé par cinq. Cet écart considérable résulte directement de la diminution du bruit engendré antérieurement par l'extension des pivots de la requête aux synonymes. En augmentant nettement le nombre de blocs pouvant faire l'objet d'une réanalyse, sans possibilité de pondérer la probabilité des blocs autrement que par des comparaisons des cumuls de mots pivots et de synonymes pour chacun des blocs, la méthode antérieure noyait les blocs corrects dans un océan de blocs incorrects, éliminant de la réanalyse certains blocs contenant les réponses.

Ces résultats intéressants mériteraient d'être validés avec d'autres corpus et d'autres jeux de questions. Notre jeu de 200 questions avait été constitué afin de tester l'échantillon le plus complet possible des types de questions, tant sur les caractéristiques physiques (taille, volume, vitesse, prix, poids, aspect, matériau, etc.) que sur des caractéristiques abstraites (adresse, fonction, dénomination, définition, etc.) sans oublier les catégories de questions les plus délicates : agent manière, appréciation, causalité, but, comparaison, supposition, etc. Par rapport aux corpus du type TREC, il fait probablement moins appel au dictionnaire de noms propres et sollicite plus souvent le dictionnaire de synonymes.

Une extension prévue de ce travail consistera à réaliser les mêmes opérations sur le dictionnaire de synonymes d'expressions verbales (plus de 10 000 synonymes portant sur 7 500 expressions) et le dictionnaire de synonymes d'expressions non verbales (plus de 41 000 synonymes portant sur environ 28 000 expressions). Les améliorations à attendre de ce traitement complémentaire seront cependant plus réduites, la proportion d'expressions dans les requêtes étant faible et une expression sur deux seulement ayant un ou plusieurs synonymes. Pourrait également être envisagé un travail similaire sur le dictionnaire de synonymes des noms propres (qui associe par exemple à «Jacques Chirac» les dénominations «Chirac», «le président de la République Française», «le président de la République», «le président français» ou «le président», dont il apparaît clairement que le degré de correspondance fluctue tant dans l'espace que dans le temps (ces deux notions sont prises en compte par le dictionnaire mais n'apparaissent pas toujours d'évidence dans le texte analysé). Toutefois il est difficile d'imaginer une automatisation de ce travail d'indexation !

En conclusion, le badaud est plus flâneur que spectateur, mais il associe les deux composantes. L'affectation d'un coefficient de proximité, en dehors de l'amélioration apportée à notre moteur d'extraction d'informations, se révèle également fructueuse pour l'aide à la rédaction elle-même. Des synonymes classés par ordre de proximité se dégage souvent une définition du mot : le badaud est un flâneur (ou un promeneur, en tout cas un piéton) qui a du temps pour regarder (un oisif et un spectateur) et qui est doté de curiosité sinon d'une certaine naïveté (curieux plus que naïf).

Références

Dictionnaires :

- Baratin M., Baratin-Lorenzi M. (1996), *Dictionnaire des synonymes*, Paris, Hachette.
- Bénac H. (1982), *Dictionnaire des synonymes*, Paris, Hachette.
- Bertaud du Chazaud H. (1988), *Dictionnaire de synonymes*, Paris, Robert.
- Boinwilliers J.-F. (1826), *Dictionnaire universel des synonymes de la langue française*, Paris.
- Boussinot R. (1988), *Dictionnaire des synonymes, analogies, antonymes*, Paris, Bordas.
- Decker T. (1996), *Dictionnaire des synonymes pour trouver vite le mot juste*, Paris, éd. Moréna.
- Genouvrier É., Désirat C., Hordé T. (1996), *Dictionnaire des synonymes*, Paris, Larousse.
- Guizot F. (1809), *Dictionnaire universel des synonymes de la langue française*, Paris, Maradan.
- Lafaye B. (1878), *Dictionnaire des synonymes de la langue française avec une introduction sur la théorie des synonymes*, Paris.
- Larousse P. (1960-1964), *Grand Larousse encyclopédique*, Paris, Librairie Larousse.
- Lecointe J. (1993), *Dictionnaire des synonymes et des équivalences*, Paris, Librairie Générale Française.
- Rey A. (2001), *Le Grand Robert de la langue française*, Paris, VUEF.
- Rouaix P. (1897), *Trouver le mot juste, dictionnaire des idées suggérées par les mots*, Paris, Armand Colin.
- Younes G. (1981), *Dictionnaire de synonymes*, Allier, éditions Marabout.

Articles :

- Besançon R., Rozenkop A., Chappelier J.-C., Rajman M. (2001), Intégration probabiliste de sens dans la représentation de textes, Actes de *TALN 2001*, 83-91.
- Dessus P. (1999), Vérification sémantique de liens hypertextes avec LSA, in J.-P. Balme, A. Lelu, S. Natkin, I. Saleh, *Hypertextes, hypermédiats et internet (H2PTM'99)*, Paris, Hermès, 112-129.
- De Loupy C. (2002), Évaluation des taux de synonymie et de polysémie dans un texte, Actes de *TALN 2002*, 225-234.
- Lafourcade M., Prince V. (2001), Synonymie et vecteurs conceptuels, Actes de *TALN 2001*, 233-242.
- Manguin J.-L., Victorri B. (1999), Représentation géométrique d'un paradigme lexical, Actes de *TALN 1999*, 363-368.
- Schwab D., Lafourcade M., Prince V. (2002), Amélioration de la représentation sémantique lexicale par les vecteurs conceptuels : le rôle de l'antonymie, Actes de *JADT 2002*.