PtiClic et PtiClic-kids : jeux avec les mots permettant une double acquisition.

Virginie Zampa, Mathieu Lafourcade

Sciences et Technologies de l´Information et de la Communication pour l´Éducation et la Formation

Volume 18, 2011
Article de recherche

PtiClic et PtiClic-Kids : jeux avec les mots permettant une acquisition lexicale par le joueur et par la machine

Virginie ZAMPA (LIDILEM, Grenoble) et Mathieu LAFOURCADE (LIRMM, Montpellier)

■ RÉSUMÉ : Cet article présente deux jeux lexicaux qui permettent à l'utilisateur d'acquérir ou de consolider des connaissances sur les mots, et à la machine de construire une ontologie généraliste. PtiClic et PtiClic-Kids se fondent sur deux méthodes d'acquisition lexicale, à savoir l'Analyse Sémantique Latente (LSA) et JeuxDeMots (JDM). Nous présenterons d'abord ces deux méthodes qui, même si toutes deux permettent d'obtenir des relations entre termes, diffèrent par de nombreux points : valeur, typage, sens des relations et supports d'où elles sont extraites. Nous exposerons ensuite l'intérêt à combiner ces deux méthodes afin de combler les lacunes de chacune au travers de ces deux jeux. Enfin, nous détaillerons ces jeux, c'est-à-dire le public visé, les différences, etc. Nous expliquerons comment ils permettent une double acquisition : de vocabulaire par les utilisateurs et lexicale par la machine. Ceci a donc un intérêt à la fois en TICE et en TALN. En effet, ces jeux permettent d'acquérir du vocabulaire et de travailler sur certains types de relations. Quant aux données recueillies, elles peuvent conduire à la constitution d'un lexique lié à l'âge d'acquisition des mots et des relations entre mots, dont de multiples applications, comme la correction ou la génération de textes, peuvent tirer profit.

■ MOTS CLÉS : jeux lexicaux, acquisition lexicale, relations lexicales, LSA, JeuxDeMots, âge d'acquisition.

■ ABSTRACT : This paper presents two lexical games, PtiClic and PtiClic-Kids, which are based on two lexical acquisition methods, namely Latent Semantic Analysis (LSA) and JeuxDeMots (JDM). We, first detail those two methods, which even if they both produce relations between terms, differ in several aspects: value, types, and directionality of the relations but also the way they are obtained. We present then the benefits of combining them in order to overcome their respective drawbacks. Secondly, we present how these games allow a lexical learning for the users and lexical relation acquisition for the computer. Finally, the overall architecture of the system and the obtained information are described as well as their benefit for research in NLP. Those games allow to gather data on age level for terms and may help to constitute such a lexicon, that would be very useful for applications like text generation.

■ KEYWORDS : Lexical games, lexical relation, LSA, JeuxDeMots, age-of-acquisition, lexical acquisition

1. Introduction

Le but de cet article est de présenter deux applications logicielles ayant pour objectif l'acquisition de connaissances lexicales conjointement par les utilisateurs et par la machine. L'utilisateur va acquérir ou

consolider des connaissances sur des mots (ou des termes composés ou des expressions) en jouant à des jeux gratuits disponibles sur la toile : PtiClic et PtiClic-Kids.

PtiClic n'a pas de public spécifique. S'il est utilisé par des adultes, c'est essentiellement la machine qui acquiert des connaissances, c'est-à-dire qu'un réseau lexical est construit. Par contre, s'il est utilisé par des enfants ou préadolescents, que ce soit en autonomie ou au sein d'une séquence didactique, le bénéfice est plus pour eux. PtiClic-Kids, en revanche, s'adresse essentiellement à des enfants ou préadolescents et, tout comme PtiClic, il peut être utilisé en autonomie ou avec un enseignant au sein d'une séquence didactique.

A l'aide du couplage de deux méthodes d'acquisition lexicale (JeuxDeMots et LSA), le système va être en mesure d'une part, de créer/compléter des informations au sein d'un réseau lexico-sémantique généraliste (dit a-domaine), ceci via l'activité des joueurs, et d'autre part de réaliser des estimations de l'âge d'acquisition des termes et des relations entre termes (avec PtiClic-Kids). Il s'agit ainsi de savoir, à partir de quel âge en moyenne, chaque sens d'un mot est maîtrisé par les locuteurs. Par exemple, nous pouvons penser que pour le mot *souris* le sens *animal* sera connu en premier, puis la *souris d'ordinateur*, puis la *viande d'agneau*, etc. De plus, il est aussi important au fur et à mesure des acquisitions des termes de savoir quelles sont les relations qui sont tissées entre eux et quels sont les types de ces relations. En reprenant l'exemple de *souris* au sens *animal*, il peut être utile de savoir qu'à partir de certains âges le terme est librement associé (par une association non typée) à *rongeur*, *souricière*, *souris blanche*, et que plus tard ces associations deviennent typées et orientées : que *souris* et *souris blanche* sont des spécifiques de *rongeur*, que *souricière* est un lieu dans lequel une *souris* peut se trouver mais aussi un mot de la même famille que *souris* et dispose d'au moins un autre sens métaphorique, etc.

Cette démarche a pour objectif de créer des lexiques adaptés à des publics d'âges différents dans un but pédagogique, mais également en analyse/génération automatiques de textes de guider la tâche en fonction de l'âge du rédacteur/lecteur. Certaines applications du TALN (Traitement Automatique du Langage Naturel), comme la correction orthographique et l'indexation de texte, mais aussi la désambiguïsation lexicale, peuvent grandement profiter de ce type d'information.

La ressource visée ne se restreint pas à un domaine particulier, mais vise à représenter des connaissances globales à la fois sur le monde et sur la langue. Par exemple, le terme *neige* pourra être associé à *froid* et *blanc* comme caractéristiques relevant du monde, mais également à *lourde* ou *poudreuse* comme phénomène linguistique pour nommer des états possibles. En TALN, et plus précisément en analyse sémantique, ces deux facettes sont primordiales, car intiment liées dans les textes.

Pour cela, nous partons de deux sources d'information : a) un réseau déjà existant, celui de JeuxDeMots (JDM) et b) LSA (Latent Semantic Analysis) qui permet, à partir de textes, de trouver des termes proches d'un terme donné (dans le même champ lexical). L'acquisition des informations visées est réalisée à travers une approche ludique, mais sans que la contribution soit explicitement demandée. Ce genre d'approche semble particulièrement adaptée à de jeunes sujets, car les activités proposées n'ont pas une coloration scolaire et leur paraissent ainsi moins rebutantes.

Pour résumer, l'objectif de ce travail de recherche est double. Il s'agit à la fois de proposer des outils pédagogiques portant sur le lexique de façon adaptée à l'âge des utilisateurs et à leur maîtrise supposée du vocabulaire, mais également de récupérer les informations d'âge d'acquisition pour les termes de ce même lexique. Au fur et à mesure de son utilisation, le système ajuste les données collectées en fonction des performances des joueurs. Une des originalités du travail présenté ici, est ce bouclage entre les données collectées et les termes proposés dans le contexte de l'acquisition lexicale conjointe humain/machine.

Dans un premier temps, nous présentons les deux méthodes d'acquisition lexicale considérées en nous focalisant sur leurs différences et sur l'avantage de les utiliser ensemble. Puis, nous introduisons deux jeux issus de cette combinaison : PtiClic et PtiClic-Kids. PtiClic permet de compléter/affiner le réseau de JDM avec des mots issus de l'analyse fournie par LSA et s'adresse à un public général dans un contexte ouvert. PtiClic-Kids, à visées plus pédagogiques, permet d'obtenir des connaissances sur les âges d'acquisition des termes et des relations entre termes, et ce dans un contexte tutoré.

2. LSA et JEUXDEMOTS : deux méthodes d'acquisition

lexicale

2.1. Différences

Les différences entre LSA et JDM concernent essentiellement la méthode d'acquisition des connaissances et les relations entre les mots (valeur, typage, sens). Les deux approches se distinguent par la méthode (les algorithmes) mais aussi par le support d'où sont extraites les informations (respectivement, corpus et joueurs).

2.1.1. Construction des associations entre termes

LSA et JDM sont deux méthodes d'acquisition lexicale. Pour les deux, il s'agit de construire une ontologie a-domaine (c'est-à-dire à la fois généraliste et incluant des informations de spécialités), comportant des connaissances du monde (relations ontologiques) et sur la langue (relations lexicales). Mais LSA et JDM diffèrent par leur méthode d'acquisition. Avec LSA il s'agit d'une analyse et compilation de corpus automatique, alors que JDM est un jeu contributif s'appuyant sur des utilisateurs.

L'objet de LSA (Landauer & Dumais,1997) est de représenter dans un espace multidimensionnel (environ 300 dimensions) les mots de la langue. Grâce à une analyse statistique, chaque mot est caractérisé par un vecteur. La similarité de sens entre deux mots est approximée par le cosinus de l'angle entre leurs vecteurs. Pour construire cet espace, LSA prend un ensemble de textes en entrée et construit une matrice d'occurrences, qui est réduite par le biais d'une analyse statistique afin d'obtenir l'espace multidimensionnel. Ceci permet ainsi de faire ressortir les relations sémantiques saillantes entre mots ou entre textes. Avec LSA, la construction de l'espace est donc automatique, rapide et est très efficace dans les domaines de spécialité (moins polysémiques que le domaine général). Par contre, l'approche dépend totalement de l'ensemble de textes utilisés au départ. De plus, la nature des relations entre mots est non spécifiée. Par exemple, « apprenant » est proche de « correction », « didactique », « rétroaction », etc. mais rien n'indique quelle est la relation qui les unie. Les catégories morpho-syntaxiques ne sont pas renseignées non plus. Enfin, les mots composés, expressions, etc. ne sont pas pris en compte : un vecteur correspond à une graphie.

JeuxDeMots (Lafourcade & Joubert, 2009) est un jeu en ligne disponible depuis septembre 2007 (http://www.jeuxdemots.org) qui a pour but la construction d'un réseau lexical associatif entre termes. Il est demandé au joueur de donner des mots correspondants à la relation avec le mot cible comme le montre l'exemple suivant.

Figure 1 : Exemple de partie de JDM. Le mot cible est *naissance*, et la relation en jeu celle de la simple association d'idées.

Dans la copie d'écran précédente, par exemple, le mot cible est *naissance* et le type de la relation à jouer est *idées associées* (c'est-à-dire les termes évoqués par le mot cible). Les mots donnés par le joueur sont à gauche de l'écran.

Une fois que le joueur a terminé, c'est-à-dire quand le temps alloué est écoulé, sa partie est mise en relation avec celle d'un autre joueur. Par exemple dans la figure suivante, pour la consigne « idées associées au terme naissance » les deux joueurs ont trois mots communs qui sont *sage-femme, maternité* et *bébé*.

Figure 2 : Exemple de résultat de partie

En pratique, les validations des propositions sont faites par concordance entre paires de joueurs, c'est-à-dire en prenant en compte les « mots » communs aux propositions des deux joueurs. Toutefois, le joueur est en présence de deux contraintes contradictoires : a) tenter de maximiser le nombre de mots qu'il peut avoir avec l'autre joueur, et b) essayer d'être le plus original possible. En effet, les relations qui sont très fortement activées dans le réseau ne rapportent plus de points. Il faut donc pour « gagner » être aussi original que possible mais en partageant les propositions avec les autres joueurs.

Ce processus de validation par accord rappelle celui utilisé par (von Ahn et Dabbish, 2004) pour l'indexation d'images, ou plus récemment par (Lieberman et al., 2007) pour la collecte de « connaissances de bon sens ». L'avantage avec JDM est que l'approche par consensus populaire permet d'obtenir des connaissances sur des choses évidentes pour tout un chacun, mais peu décrites dans les textes ou trop implicites pour être extraites automatiquement à partir de corpus.

Contrairement à LSA, il s'agit d'un vocabulaire actif, forcé par la consigne, les joueurs étant sollicités selon un mécanisme de production (à opposer à une approche par reconnaissance). Un biais sur les connaissances effectives des utilisateurs peut être lié à l'utilisation de ressources externes (comme des dictionnaires, des encyclopédies, etc.), cependant le filtrage positif de ce qui est pertinent comme information n'en demeure pas moins l'activité essentielle de l'utilisateur. Un autre biais peut être lié aux joueurs qui ne représentent pas l'ensemble de la population : 60% des joueurs sont des joueuses et l'âge moyen se situe autour de la trentaine.

L'évolution du réseau, aussi bien qualitative que quantitative, est intimement liée à la participation des joueurs. Quantitativement, l'évolution est directement corrélée à la somme des activités des joueurs. Qualitativement, il est possible de diviser l'évolution du réseau en trois tiers : une part du réseau suit l'actualité (ajout de termes tels que *Révolution de Jasmin*, *Place Tahrir*, *iPad*, etc. et rajout de nouvelles relations : *Ben Ali* a été récemment associé à *fuite*, *révolution*), une part représente les centres d'intérêt liés aux joueurs, et le dernier tiers le vocabulaire général.

Enfin, comme nous allons le voir, les relations entre mots sont typées, orientées et pondérées.

2.1.2. Liens entre termes

Comme nous l'avons indiqué, la similarité entre deux mots avec LSA correspond au cosinus de l'angle que forme les vecteurs des deux mots dans l'espace à 300 dimensions. Il s'agit donc d'une valeur unique pour une relation entre deux termes. Cette relation implicite n'est pas typée et peut être calculée pour tout couple de termes. Si nous prenons, par exemple, les termes *trompette*, *instrument*, *champignon*, *musique* et *instrument de musique* nous obtenons le réseau implicite suivant (notons que *instrument de musique* est un terme composé et est donc traité comme des mots distincts).

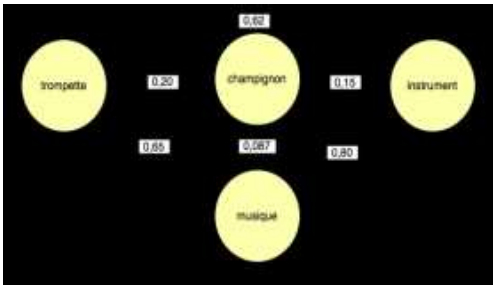

Figure 3 . Réseau entre les mots « trompette », « musique », « instrument » et « champignon » dans LSA. Les valeurs correspondent à la similarité entre termes.

Dans JDM, les relations entre mots sont typées, orientées et pondérées (cf. Figure 4). En reprenant les mêmes mots qu'avec LSA, le réseau (RezoJDM) obtenu avec JDM est donc plus complet et a une structure plus complexe. Nous pouvons ainsi constater que même si certaines relations sont à peu près réciproques (*trompette* est lié à *champignon* par la relation hyperonyme avec un poids de 70 et *champignon* est relié à *trompette* par une relation hyponyme avec un poids de 50) pour d'autres ce n'est pas le cas (*trompette* est lié à *musique* par deux relations *domaine* (poids de 25) et *idée associée* (poids de 310) alors que *musique* n'a aucun lien vers *trompette*).

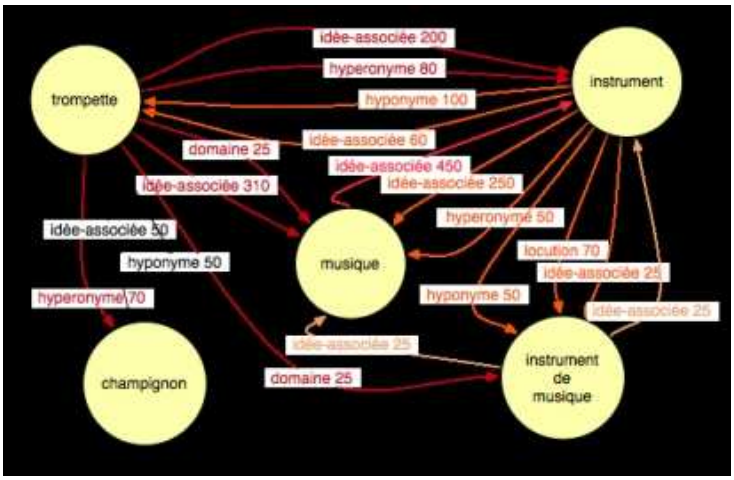

Figure 4 . Exemple de réseau RezoJDM dans JDM. Les valeurs des arcs correspondent au niveau d'activation pour la relation (résultat de l'activité de jeu des joueurs)

Dans JDM, il existe 39 types de relations (Clas et al., 1995) présentés en annexe de ce document. Grâce à ces relations, il est possible « d'expliquer » (partiellement), par exemple, les phrases ci-dessous de la façon suivante :

« Après leur dispute, ils ont remplacé la vaisselle. »
dispute =carac=> violent
dispute =conséquence=> destruction
détruire =patient=> vaisselle
vaisselle =carac=> brisée

« Il s'est brûlé la main sur la plaque et a beaucoup pleuré. »
plaque =hypo=> plaque de cuisson
plaque de cuisson =carac=> chaud, brûlant
chaud, brûlant =conséquence=> se brûler
se brûler =conséquence => douleur

L'extraction automatique de ce genre d'information à partir de corpus est au-delà des capacités des systèmes en TALN. Par contre, si les relations en question sont répertoriées, il est possible, à partir de textes, d'en faire l'analyse (c'est-à-dire de calculer quelles relations s'appliquent en contexte). Il n'existe pas à notre connaissance de réseau lexical représentant ce type d'information. Usuellement, les réseaux

lexicaux ou les ontologies (comme Wordnet (Miller et al., 1990), par exemple) se limitent aux relations ontologiques (hyperonyme, hyponyme, partie/tout) et à quelques relations lexicales (synonymes, contraires, termes de la même famille, etc.). La principale raison de cette situation provient du fait que la construction de ce type de ressources est manuelle, et donc coûteuse et longue.

2.2. Ressemblances

Il est aussi possible de calculer une valeur de similarité entre deux termes avec JDM et cette dernière semble très proche de celle renvoyée par LSA.

Le calcul de la similarité entre deux termes dans le RezoJDM passe par la construction de signatures lexicales. La signature d'un mot dans JDM est calculée de la façon suivante. Pour un mot M (par exemple *musique* cf. Figure 2) en relation avec le terme T (par exemple *instrument*), toutes les relations entrantes et sortantes sont additionnées afin d'obtenir le « poids » de la relation (dans notre exemple, nous avons deux relations entrantes : *idées-associées* à 250 et *hyperonymes* à 50, et une relation sortante : *idées-associées* à 450). Le poids est calculé pour tous les termes en relation avec le mot M.

Relation M-T	Poids
musique-instrument	450+250+50=750
musique-trompette	310+25=335
musique-instrument de musique	25

Tableau 1 . Calcul des poids à partir de l'exemple donné en figure2

Enfin, cet ensemble pondéré est ensuite normalisé, c'est-à-dire que les pondérations sont divisées par la norme de l'ensemble (la norme étant égale à la racine carrée de la somme des poids au carré). La signature d'un mot M est donc

$S(M) = <T_1 :poids_1/N\ T_2 : poids_2/N \ldots T_x : poids_x/N>$ avec N pour norme.

Norme (N)= $\sqrt{}$((poids (musique-trompette))^2 + (poids (musique-instrument))^2 + (poids (musique-instrument de musique))^2

= $\sqrt{}$(335² + 25² +750) = $\sqrt{}$(112225 + 625 + 562500)

= 822

La signature est ainsi donc un vecteur normé des différents termes reliés à M dans le réseau.

Signature(Musique)
= <instrument : 750/822 trompette : 335/822 instrument de musique 25/822>
= <instrument : 0.91 trompette : 0.40 instrument de musique : 0.03>

Nous constatons ainsi que les valeurs de similarité fournies par LSA sont très proches des valeurs de similarités entre signatures lexicales dans JDM (cf. Tableau 2).

	trompette		musique		instrument		champignon	
	LSA	JDM	LSA	JDM	LSA	JDM	LSA	JDM
Trompette	1							
Musique	0,65	0,52	1					

Instrument	0,62	0,62	0,80	0,63	1		
champignon	0,20	0,22	0,087	0,084	0,15	0,13	1

Tableau 2 . Similarités fournies par LSA et similarités entre signatures lexicales dans JDM

2.3. Complémentarité : type de vocabulaire et relations entre termes

L'analyse avec LSA étant issue de ressources écrites, les similarités renvoient à un vocabulaire plus précis et dans un registre plus soutenu. À l'inverse, JDM fournit des associations plus générales, issues de ressources spontanées. Ainsi, par exemple, avec le mot *sida* les mots les plus proches fournis par les deux méthodes sont les suivants :

	LSA		JDM	
	mot	proximité	mot	proximité
1	sida	1.000	sida	1.000
2	virus	0.948	mst	0.952
3	infection	0.900	vih	0.925
4	contamination	0.888	hiv	0.915
5	immunodéficience	0.880	maladie	0.861
6	sexuellement	0.877	contagion	0.854
7	transmissibles	0.872	maladie infantile	0.850
8	dépistage	0.856	maladie grave	0.850
9	infections	0.855	infantile	0.850
10	vih	0.849	varicelle	0.850
11	infectées	0.844	maladie mortelle	0.849
12	infectieuses	0.841	maladie bénigne	0.849
13	vaccin	0.837	hépatite	0.849
14	anti-vih	0.836	choléra	0.848
15	contaminée	0.831	quarantaine	0.847
16	immuno	0.828	infection	0.844
17	contaminé	0.825	rougeole	0.844
18	opportunistes	0.822	incurable	0.842

19	séropositifs	0.813	peste	0.842
20	contaminés	0.803	bénigne	0.841

Tableau 3 . Mots proches de « sida » avec LSA et avec JDM

D'une façon générale, le vocabulaire obtenu par LSA semble plus riche que celui acquis via JDM. Par exemple, un terme comme *immunodéficience* apparaît en 5e position avec LSA mais n'apparaît pas dans le réseau lexical de JDM comme particulièrement lié à *sida*. En effet, via le jeu, les connaissances acquises sont des associations plus immédiates, ceci est lié au fait que les joueurs ne sont pas experts du domaine et que le temps de jeu est limité. Certaines associations, évidentes, apparaissent très rapidement dans le réseau lexical du jeu. C'est le cas par exemple de l'association *sida-maladie* (5e position). Cette association est beaucoup plus faible avec LSA (proximité de 0,71) car elle est trop évidente pour être donnée explicitement dans les textes. Ce phénomène semble d'autant plus marqué que le terme cible est spécifique (relevant d'un domaine de spécialité). D'autres associations, plus difficiles parce que techniques sont plus lentes à émerger dans JDM.

Dans LSA la segmentation des textes est faite à partir des caractères de séparation (blancs, virgules, etc.) ce qui interdit l'apparition de termes composés. Ce n'est évidement pas le cas dans JDM où les joueurs ont toute liberté dans le choix des termes qu'ils suggèrent.

Il nous semble donc intéressant d'augmenter le réseau de JDM en intégrant des mots issus de LSA correspondant à du vocabulaire passif (tel immunodéficience), c'est-à-dire des mots soit inexistants, soit très peu reliés dans le réseau de JDM. En faisant jouer de tels mots, cela permet de typer et pondérer les relations entre eux et les autres termes de la base. Nous constatons que pour un terme donné, les mots proches fournis par LSA couvrent l'ensemble des relations issues de JDM mais que les relations pertinentes restent à identifier.

3. PtiClic et PtiClic-Kids

Comme nous l'avons montré, LSA tout comme JDM présentent des biais et ne proposent qu'une couverture partielle du vocabulaire. De plus, ces méthodes ont chacune des parts distinctes de bruit (associations qui devraient être plus faibles) et de silence (associations qui devraient exister ou être plus fortes). En partant de la complémentarité de ces approches, nous avons donc créé, dans un premier temps, le jeu PtiClic (http://www.lirmm.fr/pticlic/pticlic.php) afin de les combiner et qu'ainsi chacune compense les défauts de l'autre, puis PtiClic-Kids qui correspond à une approche plus pédagogique pour des enfants ou des adolescents.

3.1. PtiClic

La consigne donnée au joueur est la suivante :

« Tu prends chaque mot et tu le déposes gentiment sur une des zones visibles, en fonction de son rapport au mot cible (celui qui se trouve au milieu du nuage de mots). Certains mots n'ont rien à voir et ne doivent pas être déposés. Pour d'autres, il y a plusieurs possibilités, en choisir une bonne suffit. Quand tu as fini, clique sur le bouton en bas. »

Dans l'exemple suivant, le mot cible est *veste*, les quatre relations sont « un lieu pour veste est ... », « un spécifique de veste est ... », « ...est une partie de veste » et « veste fait partie de ... ». Nous pouvons donc avoir des réponses telles que « un lieu pour veste est ... » *placard* ou *penderie*, etc. Par contre des mots tels que *matelas* ou *portefeuille* ne correspondent à aucune des quatre relations.

Figure 5 . Partie de PtiClic en cours. Le joueur doit glisser-déposer les termes sur les zones adéquates

Le principe est le suivant, un mot source est sélectionné aléatoirement dans la base de JeuxDeMots. Si le mot n'est pas connu par LSA, un autre mot est choisi et ainsi de suite jusqu'à obtenir un mot connu. De ce fait un mot inconnu de LSA ne sera jamais proposé dans PtiClic. À partir de ce mot source, LSA sélectionne les cinq à vingt mots les plus proches (après un léger traitement ; par exemple le mot au pluriel n'est pas retenu) que le joueur doit placer dans une à quatre relations.

Seulement 10 des relations de JDM sont gardées afin que le jeu ne soit pas trop compliqué. Les dix relations maintenues sont : idées associées, synonyme, agent, patient, instrument, lieu, partie de, tout, antonyme, et hyponyme.

Nous avons bien entendu vérifié que les dix relations se retrouvent souvent dans les vingt mots fournis par LSA. PtiClic permet ainsi de typer les relations sur des mots qui sont dans un vocabulaire « moins actif », plus précis, que celui de JeuxDeMots.

Comme nous l'avons déjà montré, le joueur doit placer les mots cibles qui conviennent dans les catégories proposées. Lorsque deux joueurs ont eu la même partie, le résultat est affiché (cf. Figure 6) et les points gagnés sont calculés par comparaison entre les accords, différences et oublis. Dans la copie d'écran suivante, les mots en vert correspondent à l'accord entre les joueurs, ils apportent chacun 1 point. Les mots en gris sont ceux qui ont été mis par le premier joueur mais pas par le second. Chacun fait baisser le score de 0,5. Enfin les mots en rouge sont ceux mis par le second joueur mais pas par le premier, chacun faisant reculer le score de 0,5.

Figure 6 . Résultat de la partie PtiClic

Le principe est le même que dans JeuxDeMots, à savoir que la relation n'est validée dans la base qu'après accord entre paires d'utilisateurs. PtiClic est ainsi une composante de JDM agissant sur le même réseau

lexical. Contrairement à ce dernier, c'est un jeu en monde clos pour les utilisateurs (le joueur sélectionne une proposition parmi celles fournies mais ne peut pas en faire une nouvelle). Ce choix de conception permet d'obtenir des associations sur des termes relevant du vocabulaire passif sélectionnés par LSA, termes qui n'auraient pas spontanément été proposés par les joueurs. L'ajout de PtiClic dans JeuxDeMots réduit le bruit (des termes mal orthographiés ou des confusions de sens) ainsi que le silence (de nouveaux termes sont introduits via LSA). PtiClic permet ainsi de consolider les relations de la base et de densifier le réseau lexical.

Concernant les aspects éducatifs, PtiClic peut être utilisé aussi bien par des adultes que par des enfants. Il permet de travailler de façon ludique sur les différents types de liens entre les mots, mais également d'introduire en douceur avec des jeunes apprenants la notion de polysémie, par exemple. La comparaison des réponses et la confrontation des idées peut être source d'apprentissage.

3.2. PtiClic-kids

À partir de PtiClic, l'idée est venue de faire un jeu plus ciblé sur les enfants-adolescents, et en particulier visant à évaluer l'âge à partir duquel les mots, mais surtout les relations entre mots, étaient acquis. Il s'agit ici d'une version « pédagogique » de PtiClic, s'adressant essentiellement aux enfants entre 6 et 15 ans. Les parties sont créées par des enseignants (professeurs des écoles ou professeurs de collèges). Ces derniers choisissent un mot cible. Le système renvoie quarante mots : les vingt mots les plus proches fournis par JDM ainsi que les vingt mots les plus proches fournis par LSA. L'enseignant peut, à ce moment là, sélectionner les mots qu'il veut garder, supprimer, mais il peut aussi en ajouter d'autres. Cette liste de propositions est faite pour l'assister dans la création de la partie. Il sélectionne en plus entre une et quatre relations. Contrairement à PtiClic, avec cette version, toutes les relations présentes dans JeuxDeMots sont proposées. En effet cela permet par exemple de travailler sur les rôles grammaticaux avec des relations telles que *agent* ou *patient* pour un verbe. Ensuite, l'enseignant donne pour chacune des propositions la/les relation(s) entre chacun des mots cibles et le mot source. En effet, un mot cible peut être lié au mot source par plusieurs relations (par exemple la relation *mot associé* peut toujours être donnée). De plus, l'enseignant indique l'âge des enfants pour lesquels il a conçu la partie. Ceci correspond à l'étape 1 dans la figure 5. La partie ainsi créée représente une référence à laquelle celles des élèves seront comparées. Elle est stockée dans une base de façon permanente.

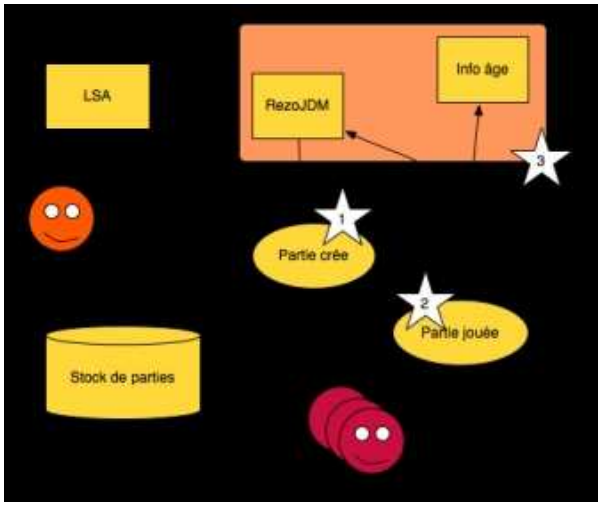

Figure 7 . Architecture et séquence d'évènements pour PtiClic-Kids

L'approche est collaborative, chaque partie créée par un enseignant étant disponible pour les autres. Un enseignant peut ainsi prendre des parties faites par un autre ayant des classes de même niveau, de niveau plus faible (pour des révisions) ou de niveau un peu plus élevé (pour mettre ses élèves en contact avec des termes/concepts qu'ils maîtrisent moins/pas afin de les aborder en jouant). C'est dans ce but que l'enseignant indique l'âge des élèves auxquels la partie est adressée. Le système pioche aléatoirement les parties pour les élèves en fonction de leur âge parmi celles créées par leur enseignant (étape 2). Le mécanisme est identique à celui du PtiClic original, à ceci près que conjointement au renforcement du RezoJDM, l'information d'âge est stockée dans une base annexe sous la forme d'un quadruplet : <numéro

de relation, âge, nombre de parties jouées, score>. À l'issue de chaque partie, le quadruplet est mis à jour. Il est ainsi possible d'obtenir pour chaque relation et chaque âge un pourcentage de réussite (étape 3).

PtiClic-Kids permet l'acquisition directe d'une information sur la connaissance, selon l'âge, d'une occurrence de relation. Plus précisément, pour une relation <A R B> (de type R entre deux termes A et B) nous obtenons un ensemble de couples âge-pourcentage. Le pourcentage étant une évaluation de la maîtrise de la relation pour l'âge donné. Par exemple, pour <chat syn félin>, nous pourrions ainsi obtenir <8:0,5 9:0,6 12:0.7>. Nous notons :

$M_{r,â}$ (t1,t2), la « maîtrise » de la relation « r » à l'âge « â » entre les termes t1 et t2. Dans l'exemple ci-dessous, nous aurions ainsi : $M_{syn,8}$ (chat,félin) = 0,5.

$M_{r,â}$ (t), la « maîtrise » de la relation « r » pour le terme « t » sur l'ensemble des termes auxquels il est associé : $M_{r,â}$ (t) = moyenne($M_{r,â}$ (t,ti)). Cette formule est une version simplifiée de ce qui serait une moyenne pondérée en fonction de l'importance du terme associé (en utilisant le réseau de jeudemots pour connaître l'importance que nous faisons correspondre à la fréquence d'utilisation).

$M_{â}$ (t), la « maîtrise » du terme « t ». De la même façon il s'agit de la moyenne pondérée des $M_{r,â}$ (t) pour l'ensemble des relations « r ».

À partir de ces différentes mesures, il est donc possible de calculer une mesure de « lisibilité lexicale » d'un texte pour un âge donné, notée $L_{â}$(T), correspondant à la moyenne des maîtrises des mots le composant pour cet âge.

Cette information n'est pas forcément disponible pour tous les âges, tous les mots, toutes les relations, cela est fonction des parties réalisées. Il est donc possible de créer un lexique avec l'âge d'acquisition moyen de chacun des termes. Ces informations peuvent être utilisées dans différentes applications. Tout d'abord, elles peuvent servir à affiner le calculer de lisibilité d'un texte pour des élèves d'un âge donné en fonction de la maîtrise lexicale de chaque terme le composant. Dans le cas où des données sont manquantes, la valeur est neutralisée en mettant la moyenne.

Par exemple dans le texte T : « les oiseaux gazouillent », si M_{8} (oiseau) = x et M_{8}(gazouiller) = y alors L_{8}(T) = moyenne(x,y).

En effet, comme le souligne Mesnager (2002), la difficulté des textes est un problème auquel les maîtres sont quotidiennement confrontés. En parlant de l'ignorance du sens de certains mots par de jeunes enfants il précise que « un seul terme vous manque et tout est incompris ». Or il n'existe pas de lexique donnant une approximation de l'âge d'acquisition des mots. Certaines méthodes utilisent la fréquence des mots comme indicateur de difficulté, or il ne s'agit pas d'un indice fiable pour de multiples raisons : par exemple les termes utilisés dans les contes de fées sont connus mais ne sont pas forcément fréquents dans le vocabulaire courant, de plus le calcul automatique de la fréquence ne prend pas en compte la synonymie (la tour (dans un contexte de château) est certainement connue plus tôt que le tour (machine outil) et que le tour de magie), etc. De plus, afin de pouvoir comprendre certaines phrases il faut être capable de comprendre les inférences.

De façon plus fine, il est possible de confronter les relations entre les termes du texte aux données recueillies. Par exemple, en prenant la phrase suivante, les relations issues de JDM, ainsi que les données de PticlicKids : « Après leur dispute, ils ont remplacé la vaisselle. ». Dans un premier temps, nous énumérons les relations pertinentes pour ce texte dans JDM (actuellement cette recherche est manuelle).

dispute =carac=> violent
dispute =conséquence=> destruction
détruire =patient=> vaisselle
vaisselle =carac=> brisée
Puis, nous regardons, dans le réseau PtiClic-Kids, pour chacun des mots du texte, s'il est acquis à l'âge donné, c'est-à-dire si M_{8}(t)> seuil donné. Après nous cherchons si les relations pertinentes issues de JDM sont acquises. Dans l'exemple, nous cherchons donc si $M_{carac,â}$ (dispute,violent), $M_{cons,â}$ (dispute,destruction), $M_{patient,â}$ (détruire,vaisselle) et $M_{carac,â}$ (vaisselle,brisée) existent et ont une

valeur supérieure au seuil.

Savoir à partir de quel âge chacun de ces mots est acquis et à partir de quel âge chacune de ces relations est connue, permet à la fois de pouvoir déterminer si un enfant comprendra, mais aussi de lui donner automatiquement les explications dont il a besoin en spécifiant chacune des relations. Il est tout à fait possible d'envisager des exercices dans lesquels l'enfant aurait à chercher parmi une liste de relations celles qui lui permettent de soulever les sous-entendus ou de comprendre une expression idiomatique.

De plus, savoir à partir de quel âge quel terme est relié à quel autre par quelle relation peut offrir un support concret à une étude sur l'acquisition des relations entre termes. Par exemple à quel âge la relation de synonymie pour le terme *chat* renvoie à *minou* ou *minet*, etc. et à partir de quel âge elle donnera *félin*. Mais cela permet aussi pour les relations fortes chez les adultes de savoir à partir de quand elles sont ancrées (ceci pour des relations comme « idées associées » chat-chien). Il est aisé d'étendre cette étude aux termes eux-mêmes.

Enfin, en analyse de textes, il est ainsi possible d'estimer automatiquement l'âge de l'auteur ou l'âge du lectorat visé. Cette information peut guider une analyse automatique (de désambiguïsation lexicale) en formulant des préférences, en particulier sur le type de relations à privilégier. Pour la génération de textes, un outil d'assistance du vocabulaire en fonction de l'âge visé est une application directe de notre expérience.

L'approche que nous proposons prend clairement la structure d'une boucle. Le jeu permet l'acquisition d'information, offrant des usages multiples, parmi lesquels celui de pouvoir affiner le jeu. Ces données étant libres d'accès, elles peuvent être exploitées par des chercheurs en didactique, linguistique, psychologie, etc. pour des études portant sur l'acquisition du vocabulaire.

4. Conclusion

Combiner deux méthodes d'acquisition lexicale (LSA et JeuxDeMots) dans un même jeu, qu'il s'agisse de PtiClic ou de PtiClic-Kids, permet d'obtenir des résultats plus intéressants qu'avec chaque approche prise isolément, et ceci qu'il s'agisse de l'acquisition par la machine ou par le joueur. Pour la machine, cela permet d'introduire ou de renforcer des relations sur du vocabulaire passif (via LSA) tout en ayant les relations typées, orientées et pondérées (via JDM). Pour le joueur/enseignant, cela permet de voir plus de vocabulaire, de se focaliser sur certains types de relations lexicales (par exemple, tous les mots de la même famille que ...), ontologiques (génériques, spécifiques), d'usage (magn, antimagn), ou sémantique/de typicalité (agent, patient, lieu, ...), etc.

Concernant les avantages pour l'enseignant et/ou l'enfant, cela permet, entre autre, de travailler sur du vocabulaire « passif » qui est présenté à l'utilisateur. De plus, l'ajout des données concernant l'âge d'acquisition pour les relations entre mots permet, de façon incrémentale, de sélectionner des mots réellement appropriés à l'apprenant en fonction de son âge et de son niveau supposé. Il est donc facilement envisageable de sélectionner, dans le nuage de mots, des termes qui auraient pour la plupart des relations ayant un niveau légèrement plus élevé que celui de l'apprenant (ni trop proche, ni trop éloigné – sorte de zone proximale de développement) ceci afin d'optimiser son apprentissage. Ceci permettrait aussi, comme nous l'avons déjà signalé, d'obtenir l'âge d'acquisition des termes et des relations, ce qui faciliterait le calcul de la lisibilité des textes pour les enfants. Un enseignant peut utiliser le réseau ainsi obtenu afin de créer de nombreuses activités : en prenant par exemple toutes les relations *agent* d'un verbe donné accessible à des enfants de l'âge désiré ; en faisant sélectionner les relations une à une afin d'expliquer une phrase ; en extrayant les différents sens pour un mot polysémique acquis à un âge donné ; etc.

Concernant les apports pour le TALN, l'information sur l'âge peut guider les processus d'analyse automatique de textes. En effet, si l'hypothèse de l'homogénéité du vocabulaire est admise, des préférences peuvent être émises par l'analyseur en fonction du vocabulaire rencontré et des sens possibles de chaque terme. À l'inverse, en génération automatique de textes, à l'aide des informations relatives à l'âge, il est concevable de produire des versions multiples d'un même texte avec un vocabulaire adapté à différentes tranches d'âge.

RÉFÉRENCES BIBLIOGRAPHIQUES

VON AHN L., DABBISH L. (2004). Labelling Images with a Computer Game. In ACM *Conference on Human Factors in Computing Systems* (CHI), p.319-326.

CLAS A., MEL'CUK I.A., POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot AUPELF-UREF.

LAFOURCADE M., JOUBERT A. (2009). Similitude entre les sens d'usage d'un terme dans un réseau lexical. Dans *Traitement Automatique des Langues* (TAL), Vol. 50 Numéro 1. Varia, p.179-200, 2009.

LANDAUER T. K., DUMAIS S. T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge . *Psychological Review*, vol.104, p.211-240.

LIEBERMAN H., SMITH D.A., TEETERS A. (2007). Common Consensus: a web-based game for collecting commonsense goals, *International Conference on Intelligent User Interfaces* (IUI'07), Hawaii, USA.

MESNAGER J. (2002). Pour une étude de la difficulté des textes ou la lisibilité revisitée. *Le Français aujourd'hui*, Numéro 137.

MILLER G.A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K.J. (1990). Introduction to WordNet: an on-line lexical database, *International Journal of Lexicography* 3 (4), pp. 235-244.

RÉFÉRENCES COMPLÉMENTAIRES NON CITÉES DANS L'ARTICLE

DENHIÈRE G., LEMAIRE B. (2004) A Computational Model of a Child Semantic Memory, *Proc. 26th Annual Meeting of the Cognitive Science Society* (CogSci'2004), p. 297-302.

KINTSCH W. (2001). Predication . *Cognitive Science*, vol.25-2, p.173-202.

KIPFER B.A. (2001). *Roget's International Thesaurus*, sixth edition, Harper Resource (1rst Ed:1852)

LAPATA M., Keller F. (2005) Web-based Models for Natural Language Processing. In ACM *Transactions on Speech and Language Processing*, vol.2, n°1, pp. 1-30.

ANNEXE : liste des relations de JDM

Toutes les relations sont présentées dans la liste suivante avec leur nom, le nombre actuel d'occurrences de chaque relation ainsi que la consigne donnée au joueur.

- idée associée : 627797 ; Donner des IDEES ASSOCIEES au terme qui suit ;
- domaine : 83029 ; Donner des THEMES/DOMAINES pour le terme qui suit (par exemple, 'sports', 'médecine', 'cinéma', 'cuisine,' etc.) ;
- synonyme : 152215 ; Donner des SYNONYMES du terme qui suit (par exemple, 'chat' pour 'matou') ;
- générique : 40918 ; Donner des GENERIQUES pour le terme qui suit (par exemple, 'véhicule' pour 'voiture', 'félin', 'animal' pour 'chat') ;
- antonyme : 12025 ; Donner des CONTRAIRES pour le terme qui suit (par exemple, 'froid' pour 'chaud', 'haut' pour 'bas') ;
- hyponyme : 9683 ; Donner des SPECIFIQUES pour le terme qui suit (par exemple, 'chat', 'chien', 'animal de compagnie', etc. pour 'animal' - ou encore 'voiture', 'train', 'véhicule spatial', etc. pour 'véhicule') ;
- partie de : 11761 ; Donner des PARTIES du terme suivant : (une partie est une composante de l'objet, par exemple : 'moteur', 'roue', etc. pour 'voiture' - ou encore 'couverture', 'pages', 'chapitre' etc. pour 'livre') ;
- holonyme : 9020 ; Donner des TOUT du terme suivant : (le tout est ce qui englobe/contient/possède la

partie, par exemple : 'corps', 'bras', etc. pour 'coude' - ou encore 'banque' pour 'guiche') ;

- locution : 9244 ; Donner des LOCUTIONS pour le terme qui suit (par exemple, 'langue au chat', 'chat de gouttière', 'chat à neuf queues', ... pour 'chat');

- agent : 11213 ; Donner des SUJETS typiques pour le VERBE qui suit (le sujet est celui qui effectue l'action, par exemple 'chat', 'animal', 'personne', ... pour 'manger') ;

- patient : 8751 ; Donner des OBJETS typiques pour le VERBE qui suit (l'objet est ce qui subit l'action, par exemple : 'viande', 'fruit', 'bonbon', ... pour 'manger' ou encore 'personne', 'homme politique', 'otage', ... pour 'assassiner') ;

- lieu : 15578 ; Donner des LIEUX typiques pour le terme qui suit (par exemple : 'pré', 'écurie', 'champs de courses', ... pour 'cheval') ;

- instrument : 7906 ; Donner des INSTRUMENTS typiques pour le VERBE qui suit (l'instrument est quelque'chose avec lequel on peut effectuer l'action, par exemple : 'pelle', 'pioche', 'main', ... pour 'creuser');

- caractéristique : 10052 ; Donner des CARACTERISTIQUES typiques pour le terme qui suit. Par exemple, 'liquide', 'blanc', 'buvable', ... pour 'lait', ou 'coupant', 'tranchant', ... pour 'lame');

- magn : 2928 ; Qu'est ce qui est PLUS INTENSE que le terme qui suit (par exemple, 'forte fièvre', 'fièvre de cheval', ... pour 'fièvre' - ou encore 'vivre intensément' pour 'vivre';

- anti-magn : 2834 : Qu'est ce qui est MOINS INTENSE que le terme qui suit (par exemple, 'maisonette', ... pour 'maison' - ou encore 'marcher lentement', 'trainer' pour 'marcher');

- famille : 10174 ; Donner des mots de la MEME FAMILLE pour le terme qui suit (par exemple, 'chanter', 'chanteur'... pour 'chant' - ou encore 'vente', 'vendeur', 'vendu' pour 'vendre');

- caractéristique-1 : 1118 ; Qu'est-ce qui possède la CARACTERISTIQUE qui suit (par exemple, 'eau', 'vin', 'lait' pour 'liquide');

- agent-1 : 1400 ; Que peut faire le SUJET qui suit (par exemple, 'manger', 'dormir', 'chasser' pour 'lion');

- instrument -1 : 2903 : Que peut-on faire avec l'INSTRUMENT qui suit (par exemple, 'écrire', 'dessiner', 'gribouiller' pour 'stylo');

- patient-1 : Que peut subir l'OBJET qui suit (par exemple, 'gratin' peut 'cuire', 'être mangé', ...);

- domaine -1 : 14 ; Donner des termes du DOMAINE qui suit (par exemple, 'touche', 'penalty', 'but' pour 'Football') ;

- lieu -1 : 4527 ; Que trouve-t-on dans le LIEU qui suit (par exemple, 'poisson', 'coquillage', 'algue' pour 'mer');

- lieu-action : 2948 ; Que peut-on faire dans le LIEU qui suit (par exemple, 'manger', 'boire', 'commander' pour 'restaurant' -- des verbes sont demandés) ;

- action-lieu : 2919 : Dans quels LIEUX peut-on faire l'action qui suit (par exemple, 'restaurant', 'cuisine', 'fast-food' pour 'manger' -- des lieux sont demandés);

- sentiment : 1573 ; A quels SENTIMENTS/EMOTIONS peut être associé le terme qui suit;

- manière : 4023 : De quelles MANIERES peut être effectuée l'action qui suit (il s'agira d'un adverbe ou d'un équivalent, par exemple : 'rapidement', 'sur le pouce', 'goulûment', 'salement' ... pour 'manger');

- sens/signification : 9176 ; Quels SENS/SIGNIFICATIONS pouvez vous donner au terme qui suit (il s'agira de termes évoquant chacun des sens possibles, par exemple : 'forces de l'ordre', 'contrat d'assurance', 'police typographique', ... pour 'police');

- rôle télique : 1239 ; Quels BUT/FONCTION (rôle télique) pouvez vous donner au terme qui suit (il s'agira d'un verbe, par exemple : 'couper' pour 'couteau', 'lire' pour 'livre', ...);

- rôle agentif : 987 ; Quels MODES DE CREATION (rôle agentif) pouvez vous donner au terme qui suit (il s'agira d'un verbe, par exemple : 'construire' pour 'maison', 'rédiger'/'imprimer' pour 'livre', ...);
- rôle causatif : 1192 ; Quelles CONSEQUENCES (A entraine B) pouvez vous donner au terme qui suit (il s'agira d'un verbe ou d'un nom : 'tomber' -> 'se blesser', 'faim' -> 'voler'/'dérober', ...);
- conséquence : 1317 ; Quelles CAUSES (A a pour cause B) pouvez vous donner au terme qui suit (il s'agira d'un verbe ou d'un nom : 'se blesser' -> 'tomber', 'voler'' -> 'faim', 'pauvreté', ...);
- action-temps : 85 ; Quel MOMENT/PERIODES/TEMPS peut-on associer au terme qui suit (par exemple : 'dormir' -> 'nuit', 'manger' -> 'midi', 'soir', 'fatigue' -> 'soir') ;
- objet - matière/substance : 532 ; De quelles MATIERES/SUBSTANCE est fait le terme qui suit (par exemple, 'acier' pour 'épée' ou 'bois' pour 'chaise' ...)
- matière/substance – objet : 112 ; Quel est la ou les CHOSES qui sont composées de la MATIERE/SUBSTANCE qui suit (par exemple pour 'marbre' -> 'statue', 'table', 'escalier', ...));
- succession : 111 ; Qu'est ce qui peut SUIVRE le terme suivant : (par exemple Noêl -> jour de l'an, guerre -> paix, jour -> nuit, pluie -> beau temps, repas -> sieste, etc);
- fabrication/producteur: 11 ; Que peut PRODUIRE le terme ? (par exemple abeille -> miel)
- produit de : 26 ; Le terme est le RESULTAT/PRODUIT de quoi ?
- opposition : 32 ; A quoi le terme suivant S'OPPOSE / COMBAT /EMPECHE ? (par exemple : médicament -> maladie)

Référence de l´article :

Virginie ZAMPA et Mathieu LAFOURCADE, PtiClic et PtiClic-Kids : jeux avec les motspermettant une acquisition lexicale par le joueur et par la machine, *Revue STICEF*, Volume 18, 2011, ISSN : 1764-7223, mis en ligne le 9/01/2012, http://sticef.org