# JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes

Mathieu Lafourcade, Alain Joubert

# JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes

Mathieu Lafourcade, Alain Joubert
*{lafourcade, joubert}@lirmm.fr*

LIRMM – Univ. Montpellier 2 - CNRS
Laboratoire d'Informatique, de Robotique
et de Microélectronique de Montpellier
161, rue Ada – 34392 Montpellier Cédex 5 – France

## Abstract

Collecting lexical information is a difficult task. Indeed, when done manually, it requires the competence of experts and the duration and cost implied can be prohibitive. When done automatically, the results can be biased by the corpus of texts. The approach we present here consists in having people take part in a collective project by offering them a playful application accessible on the web. The players themselves thus stock the database, by supplying associations which will be validated only if they are suggested by a pair of users. Furthermore, these typed relations will be weighted according to the number of pairs of users who will provide them.

**Keywords :** Natural Language Processing, lexical network, typed and weighted relations, web-based game

## Résumé

La collecte des informations lexicales est une tâche difficile. En effet, effectuée manuellement, elle nécessite la compétence d'experts et la durée et le coût nécessaires peuvent être prohibitifs, alors que réalisée automatiquement, les résultats peuvent être biaisés par les corpus de textes retenus. L'approche présentée ici consiste à faire participer un grand nombre de personnes à un projet contributif en leur proposant une application ludique accessible sur le web. Ce sont ainsi les joueurs qui vont alimenter la base, en fournissant des associations qui ne seront validées que si elles sont proposées par une paire d'utilisateurs. De plus, ces relations typées seront pondérées en fonction du nombre de paires d'utilisateurs qui les auront proposées.

**Mots-clés :** Traitement Automatique du Langage Naturel, réseau lexical, relations typées pondérées, jeu sur internet

## 1. Introduction

Un très grand nombre de tâches réalisées en Traitement Automatique des Langues (TAL) nécessite la connaissance de relations lexicales ou fonctionnelles entre termes, relations que l'on trouve généralement dans des thésaurus ou des ontologies. Ces relations peuvent être mises en évidence de façon manuelle, par exemple Roget (Kipfer 2001) ou Wordnet (Miller 1990), ou bien automatiquement à partir de corpus de textes, par exemple (Robertson et Spark Jones 1976) ou (Lapata et Keller 2005), dans lesquels sont effectuées des études statistiques sur les distributions de mots. En outre, le TAL requiert des informations de différentes natures, comme la synonymie ou l'antonymie, mais également des relations d'hyperonymie/hyponymie, holonymie/méronymie, … L'établissement de telles relations, s'il est effectué manuellement par un ensemble d'experts, requiert des ressources (en durée et en personnel) qui peuvent être prohibitives, alors que leur extraction automatique sur un corpus de textes est beaucoup trop dépendante des textes choisis.

La méthode développée ici s'appuie sur un système contributif, tel Wikipédia, où ce sont les utilisateurs qui font évoluer la base. Pour inciter les utilisateurs à participer, l'interface est présentée sous forme d'un jeu. De plus, contrairement aux méthodes classiques qui permettent d'acquérir des informations lexicales généralement statiques, le prototype introduit ici réalise l'acquisition d'informations lexicales évolutives.

Dans cet article, nous présentons les principes d'un jeu (JeuxDeMots[1]) visant à construire la base de relations. L'objectif poursuivi ici concerne avant tout la fiabilité et la qualité des informations recueillies auprès des utilisateurs. Les données recueillies peuvent évoluer quantitativement et qualitativement.

## 2. Objectifs et méthode

### *2.1. Principe de base*

Afin d'éviter les écueils d'un système où n'importe quel utilisateur pourrait écrire n'importe quoi, la solution aurait pu reposer sur l'existence d'un modérateur-expert humain qui aurait validé (ou invalidé) les propositions faites par les utilisateurs. Mais dans ce cas, toute la base de relations aurait dépendu de la compétence de cet expert dont le travail aurait été fastidieux. De plus, des joueurs mal intentionnés auraient pu faire un nombre de propositions tel que les validations par l'expert auraient été matériellement irréalisables. Il a donc été décidé que les validations des relations proposées anonymement par un joueur seraient effectuées par d'autres joueurs, tout autant anonymement. Pratiquement, les validations seront faites par concordance des propositions entre paires de joueurs. Ce processus de validation rappelle celui utilisé par (von Ahn et Dabbish, 2004) pour l'indexation d'images. A notre connaissance, il n'a jamais été mis en œuvre dans le domaine des réseaux lexicaux.

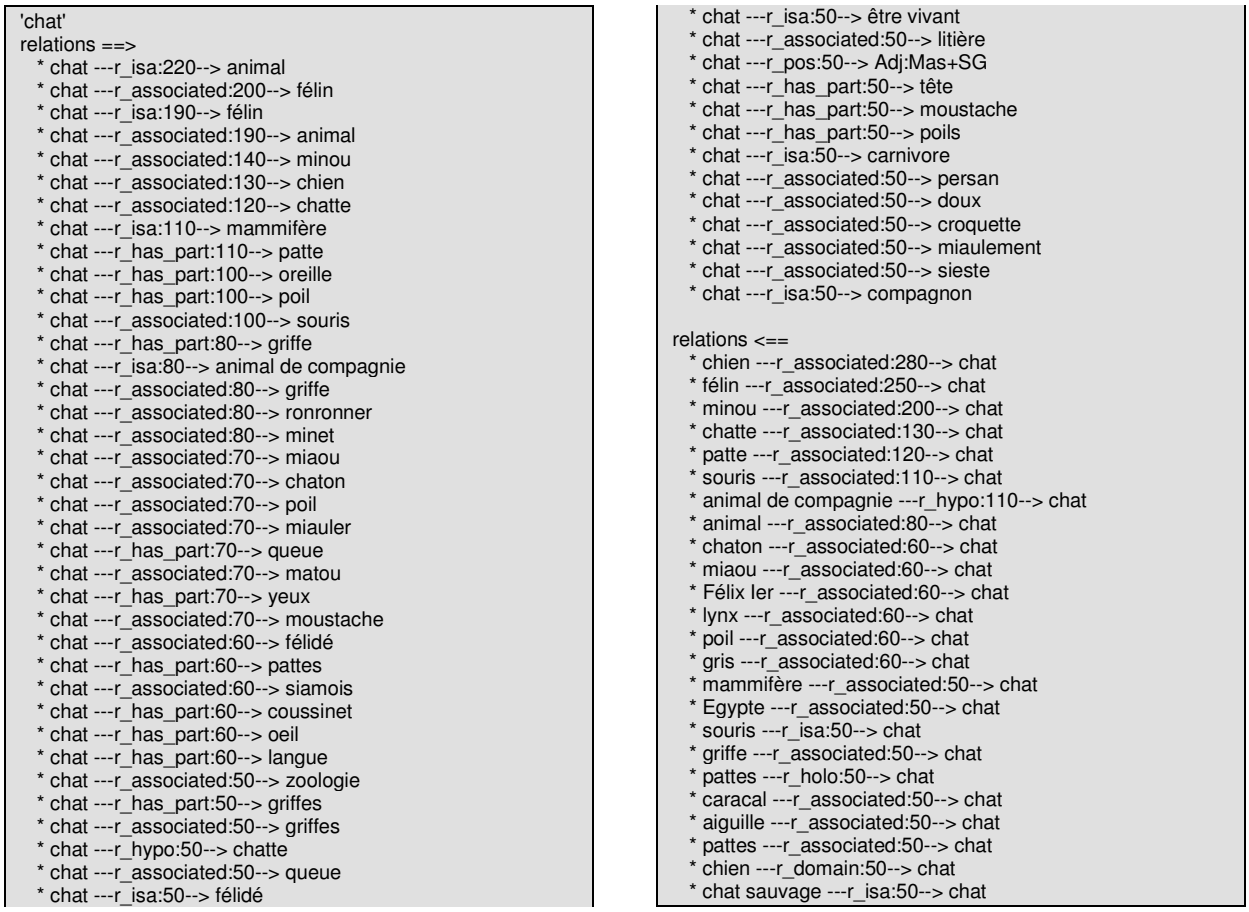
### *2.2. Déroulement du jeu*

Une partie se déroule entre deux joueurs, en double aveugle, basée sur la concordance de leurs propositions. Lorsqu'un joueur (A) débute une partie, une consigne concernant un type de compétence (synonymes, contraires, domaines …) est affichée, ainsi qu'un mot[2] M tiré aléatoirement dans une base de mots. Le joueur (A) a alors un temps limité pour répondre en donnant des propositions répondant, selon lui, à la consigne appliquée au mot M. Le nombre de propositions qu'il peut faire est limité pour éviter que des joueurs ne frappent n'importe quoi le plus vite possible. Ce même mot, avec cette même consigne, est proposé à un autre joueur (B) ; le processus est identique. Les deux demi-parties, celle du joueur (A) et celle du joueur (B), ne sont pas simultanées, mais asynchrones. Pour toute réponse commune dans les propositions de (A) et (B), ces deux joueurs gagnent un certain nombre de points. Le calcul de ce nombre de points est explicité en section 2.3.

Pour le mot cible M, on mémorise les réponses communes aux joueurs (A) et (B). On ne mémorise pas les réponses proposées uniquement par l'un des deux joueurs. Cela permet la construction d'un réseau lexical reliant les termes par des relations typées et pondérées, validées par paires de joueurs. Ces relations sont typées par la consigne imposée aux joueurs ; elles sont pondérées en fonction du nombre de paires de joueurs qui les ont proposées, comme explicité en section 2.3. La structure du réseau lexical que nous cherchons ainsi à obtenir s'appuie sur les notions de nœuds et de relations entre nœuds, telles que rappelées par (Polguère, 2006). Chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies et les relations entre nœuds traduisent des fonctions

[1] JeuxDeMots est accessible à l'adresse http://www.lirmm.fr/jeuxdemots
[2] Pour la suite de cet article, *mot* et *terme* sont considérés comme synonymes.

lexicales, telles que présentées par (Mel'čuk et al., 1995). Initialement, les nœuds sont constitués des termes de notre base de départ, mais celle-ci peut s'accroître ; effectivement, si les deux joueurs d'une même partie proposent un terme initialement inconnu, alors ce terme est ajouté à notre base. La figure 1 présente les relations acquises pour le terme *chat*.



'chat'
relations ==>
* chat ---r_isa:220--> animal
* chat ---r_associated:200--> félin
* chat ---r_isa:190--> félin
* chat ---r_associated:190--> animal
* chat ---r_associated:140--> minou
* chat ---r_associated:130--> chien
* chat ---r_associated:120--> chatte
* chat ---r_isa:110--> mammifère
* chat ---r_has_part:110--> patte
* chat ---r_has_part:100--> oreille
* chat ---r_has_part:100--> poil
* chat ---r_associated:100--> souris
* chat ---r_has_part:80--> griffe
* chat ---r_isa:80--> animal de compagnie
* chat ---r_associated:80--> griffe
* chat ---r_associated:80--> ronronner
* chat ---r_associated:80--> minet
* chat ---r_associated:70--> miaou
* chat ---r_associated:70--> chaton
* chat ---r_associated:70--> poil
* chat ---r_associated:70--> miauler
* chat ---r_has_part:70--> queue
* chat ---r_associated:70--> matou
* chat ---r_has_part:70--> yeux
* chat ---r_associated:70--> moustache
* chat ---r_associated:60--> félidé
* chat ---r_has_part:60--> pattes
* chat ---r_associated:60--> siamois
* chat ---r_has_part:60--> coussinet
* chat ---r_has_part:60--> oeil
* chat ---r_has_part:60--> langue
* chat ---r_associated:50--> zoologie
* chat ---r_has_part:50--> griffes
* chat ---r_associated:50--> griffes
* chat ---r_hypo:50--> chatte
* chat ---r_associated:50--> queue
* chat ---r_isa:50--> félidé
* chat ---r_isa:50--> être vivant
* chat ---r_associated:50--> litière
* chat ---r_pos:50--> Adj:Mas+SG
* chat ---r_has_part:50--> tête
* chat ---r_has_part:50--> moustache
* chat ---r_has_part:50--> poils
* chat ---r_isa:50--> carnivore
* chat ---r_associated:50--> persan
* chat ---r_associated:50--> doux
* chat ---r_associated:50--> croquette
* chat ---r_associated:50--> miaulement
* chat ---r_associated:50--> sieste
* chat ---r_isa:50--> compagnon

relations <==
* chien ---r_associated:280--> chat
* félin ---r_associated:250--> chat
* minou ---r_associated:200--> chat
* chatte ---r_associated:130--> chat
* patte ---r_associated:120--> chat
* souris ---r_associated:110--> chat
* animal de compagnie ---r_hypo:110--> chat
* animal ---r_associated:80--> chat
* chaton ---r_associated:60--> chat
* miaou ---r_associated:60--> chat
* Félix Ier ---r_associated:60--> chat
* lynx ---r_associated:60--> chat
* poil ---r_associated:60--> chat
* gris ---r_associated:60--> chat
* mammifère ---r_associated:50--> chat
* Egypte ---r_associated:50--> chat
* souris ---r_isa:50--> chat
* griffe ---r_associated:50--> chat
* pattes ---r_holo:50--> chat
* caracal ---r_associated:50--> chat
* aiguille ---r_associated:50--> chat
* pattes ---r_associated:50--> chat
* chien ---r_domain:50--> chat
* chat sauvage ---r_isa:50--> chat

*Figure 1 : Ensemble des relations acquises pour le terme* chat. *Cette figure présente tout d'abord les relations dont le terme* chat *est origine, puis celles pour lesquelles le terme* chat *est destinataire. Pour chacune de ces relations, on a en outre son type (est un, idée associée, a pour partie …) ainsi que son poids. Le calcul de cette pondération est expliqué à la section 2.3.*

Il aurait pu être envisagé de mémoriser toutes les réponses, depuis le début du jeu, avec leurs fréquences. L'intérêt de la solution retenue est de limiter de façon beaucoup plus drastique les réponses « fantaisistes » ou les erreurs dues à une mauvaise compréhension de la consigne ou du mot M lui-même, même si cela ralentit fortement l'émergence de propositions plus rares. Effectivement, si pour un mot M, un même terme est proposé plusieurs fois, mais dans des parties à deux joueurs différentes, on n'en tiendra pas compte (comme si ce terme n'avait jamais été proposé !). L'émergence des solutions « originales » sera plus lente, mais elle se fera tout de même, après élimination des solutions les plus courantes, grâce au processus des termes « tabous ». Effectivement, lorsqu'une relation mot M → terme proposé a été faite par un grand nombre de couples de joueurs, elle devient taboue ; elle est affichée en même temps que le mot M, afin que les joueurs ne la proposent plus. Ainsi, les joueurs sont amenés à faire d'autres propositions. Ceci favorisera l'émergence de relations plus rares, mais non l'émergence d'erreurs.

Par contre, il peut y avoir émergence de relations erronées, non pas volontaires de la part des joueurs, mais parce qu'il existe des confusions dans leur esprit. Il s'agit d'informations

erronées d'usage ; l'un des exemples que nous avons rencontré est la confusion entre *Dalila* et *Dalida*. La figure 2 illustre cette confusion d'usage. Actuellement, notre système ne permet pas de détecter ces erreurs. Nous souhaiterions arriver, sans intervention d'un expert humain, à repérer ce genre d'erreur[3] ; nous pourrions alors conserver ces relations dans le réseau, comme actuellement, mais en les typant en tant que relations erronées.

```
'Dalila'
relations ==>
   * Dalila ---r_associated:50--> Bible
   * Dalila ---r_associated:60--> suicide
   * Dalila ---r_associated:80--> Samson
   * Dalila ---r_associated:60--> star
   * Dalila ---r_associated:50--> musique
   * Dalila ---r_associated:130--> chanteuse
   * Dalila ---r_associated:60--> chanson
   * Dalila ---r_associated:50--> bible

'Dalida'
relations ==>
   * Dalida ---r_associated:50--> suicide
   * Dalida ---r_associated:60--> blonde
   * Dalida ---r_associated:70--> chanteuse
   * Dalida ---r_associated:50--> chanson
   * Dalida ---r_associated:60--> Egypte
```

*Figure 2 : Relations « idées associées » relatives aux mots* Dalila *et* Dalida*. On constate en particulier que la relation la plus forte est* Dalila → chanteuse *(avec un poids égal à 130), bien que ce ne soit pas sa caractéristique principale selon la Bible. C'est manifestement une confusion d'usage.*

### *2.3. Pondération des relations et décompte des points*

Il s'agit de définir le nombre de points gagnés par les joueurs (A) et (B), avec la même consigne sur un même mot M. Pour cette partie, notons :

propositions de (A) : $x_1, x_2, \ldots x_i \ldots, x_n$

propositions de (B) : $y_1, y_2, \ldots y_j \ldots, y_m$

Pour tous les couples (i,j) tels que $x_i = y_j$ , nous mémorisons la relation R : M → $x_i$.

L'un des avantages de notre méthode réside dans gestion de la pondération des relations entre termes. En effet, il est possible d'affecter un poids à la relation R : plus elle a été proposée de fois, plus son poids est important. Dans cette première version de notre prototype, nous avons envisagé un poids de 50 pour sa première occurrence, puis nous augmentons ce poids de 10 pour chaque occurrence suivante. Rappelons qu'une occurrence de R correspond à une proposition de R par le joueur (A) ainsi que le joueur (B) lors d'une même partie. On peut envisager qu'à partir d'un certain nombre d'occurrences, une relation R est bien établie : elle devient alors banale, ou « taboue », et elle est indiquée en même temps que le mot M, afin que les joueurs ne la proposent plus. Ce processus permet de faire émerger plus facilement de nouvelles relations ; sa conséquence sur la base est donc une augmentation du taux de rappel[4].

Le nombre de points obtenus par (A) et (B) dépend du poids de la relation R. Ce nombre de points vaut actuellement : 10% (1000 - poids(R)). Plus la relation est récente, plus elle a de valeur : cela revient à « payer la primauté ». Cette fonction est décroissante : une relation rapporte de moins en moins de points. A partir d'un certain seuil, fixé actuellement à 300 pour la valeur du poids, la relation devient taboue ; elle est alors indiquée en même temps que la

[3] Notre réflexion sur ce point n'est pas encore suffisamment avancée pour la mentionner dans cet article.

[4] D'après (Salton 1968), le taux de rappel peut être défini par le rapport du nombre de relations pertinentes trouvées sur le nombre de relations pertinentes, la précision correspondant au rapport du nombre de relations pertinentes trouvées sur le nombre de relations proposées.

consigne et le mot M (c'est une solution donnée, exemple : *Wimbledon → tennis*). Elle continue à rapporter des points, mais beaucoup moins : elle n'est plus intéressante pour les joueurs. Avec les valeurs indiquées ci-dessus, une relation devient taboue quand elle a été proposée par 25 couples de joueurs.

Même lorsqu'une relation devient taboue, son poids n'est pas figé, mais il évolue beaucoup moins vite car cette relation est proposée moins souvent par les joueurs. Il est tout de même intéressant que le poids de la relation continue d'évoluer. En effet, au bout d'un certain temps, pour un même terme plusieurs relations peuvent être taboues. Si elles avaient le même poids, on ne saurait pas laquelle a atteint cet état en premier et donc on ignorerait celle qui est la plus « forte ».

Il a également été prévu un phénomène d'érosion des relations. Effectivement, une relation a pu être créée à la suite d'une erreur commune à deux joueurs, ou bien une relation a pu être conjoncturelle et être beaucoup moins forte, donc moins proposée, par la suite (exemple : *Paris → Jeux olympiques*). A chaque partie sur un mot M, le poids des relations existantes dans notre base à partir de ce mot M qui ne sont proposées par aucun des deux joueurs est très légèrement diminué (actuellement -1). Cela diminuera inexorablement le poids des relations accidentelles, mais nous espérons que cette érosion n'aura qu'un effet négligeable sur les relations fortes[5].

### *2.4. Niveau d'un terme – niveau d'un joueur*

A chaque terme est associée une estimation de son niveau de difficulté (par exemple, il est plus difficile de trouver des synonymes au mot *bijection* qu'au mot *maison*). Initialement, le niveau d'un terme est relié à sa fréquence.

A chaque joueur est également associé un niveau ; on pourrait parler de niveau de compétences (un enfant de dix ans aura probablement un niveau inférieur à celui d'un étudiant préparant une thèse). Initialement, le niveau de chaque joueur est identique.

Ces niveaux peuvent évoluer : un terme peut être relativement fréquent mais évoquer peu d'associations, le niveau d'un joueur augmentera au fur et à mesure de l'expérience qu'il va acquérir sur ce type de jeu. Si un joueur gagne des points sur un terme de niveau supérieur au sien, le niveau du joueur augmentera et le niveau du terme diminuera. Par contre, si un joueur ne gagne pas de points, alors le niveau du terme augmentera et le niveau du joueur sera diminué.

### *2.5. Honneur et efficacité d'un joueur*

L'honneur d'un joueur représente sa contribution à la base, c'est-à-dire sa capacité à établir des relations R : honneur est ici synonyme d'expertise. C'est un artéfact du jeu qui permet d'établir un classement entre les joueurs et donc concourt à leur stimulation.

Contrairement à l'honneur, le nombre de points, correspondant à de l'argent virtuel gagné par le joueur, peut servir à donner à celui-ci un certain contrôle sur la partie : achat de temps supplémentaire, achat de parties, … comme mentionné en section 3 ; c'est la raison pour laquelle le nombre de points est aussi appelé « crédits ». L'honneur ne permet pas ce genre de transaction : on ne vend pas son honneur ! Un joueur peut donc avoir un certain honneur, sans pour autant avoir beaucoup de crédits. Les mécanismes d'augmentation ou de diminution d'honneur, similaires au gain ou à la perte de crédits, sont toutefois plus simples. Pour chaque proposition concordante, les deux joueurs d'une même partie gagnent un point d'honneur

[5] Ce processus d'érosion est encore au stade expérimental : nous attendons de voir ces effets dans la durée.

chacun. Chaque fois qu'un joueur s'abstient sur un terme, il perd un point d'honneur. L'honneur d'un joueur diminue également s'il passe un certain temps sans jouer.

Il a également été défini l'efficacité d'un joueur : il s'agit d'avoir une mesure de sa contribution à la construction de la base par partie jouée. L'efficacité d'un joueur est égale au rapport de son honneur par le nombre de parties auxquelles il a participé. Il est prévu que cette notion d'efficacité puisse avoir un impact sur l'évolution de la base : plus l'efficacité d'un joueur sera forte, et plus le renforcement du poids des relations qu'il propose sera important. L'efficacité d'un joueur sera alors assimilable à la notion de confiance que le système pourra lui accorder.

## 3. Réalisation

Le logiciel a été développé en PHP/MySQL sous forme d'un site web; les programmes annexes ont été réalisés en langage JAVA et C++. Les notions de niveau, honneur, crédits, captures de mots, … ainsi que l'affichage du classement des joueurs, ont été mis en œuvre afin d'accroître l'aspect attrayant du jeu. En incitant les joueurs à revenir régulièrement sur le site, on augmente d'autant le nombre de relations acquises : c'est l'intérêt majeur de cette dimension jeu par rapport à un logiciel qui se contenterait de demander des relations à des utilisateurs qui, certes, auraient plus conscience de leur rôle d'«experts », mais qui, probablement, y consacreraient moins de temps.

Chaque fois qu'un joueur accède au site, il lui est demandé de se connecter, sinon il joue en tant qu'invité. Une consigne est alors affichée pendant 3 secondes (par ex : « Donner des idées associées au terme suivant »), avant que le terme sur lequel il doit appliquer cette consigne n'apparaisse à l'écran. Ce terme est tiré aléatoirement dans une base d'environ 150.000 termes. Il a alors une minute pour répondre à la consigne, éventuellement plus s'il acquiert du temps supplémentaire en échange de points. Si le joueur est (B), il est procédé à l'affichage immédiat du résultat de la partie : nombre de points gagnés (appelés crédits), évolution éventuelle de ses points d'honneur … S'il est joueur (A), ces informations lui seront envoyées par mail après que (B) ait joué. Les parties proposées au joueur sont soit des parties en création où il est joueur (A), soit des parties à finir pour lesquelles il est joueur (B). Il y a donc en permanence un ensemble de parties à finir.

Lorsqu'un joueur vient de terminer une partie en création sur un mot M, s'il est satisfait de ses propositions, il a la possibilité d'acheter des parties supplémentaires (en échange de crédits), afin que ce même mot M soit proposé à un plus grand nombre de joueurs. Par contre, si à l'affichage du mot et de la consigne le joueur estime n'avoir aucune idée, il a la possibilité de « passer » : la partie se termine alors prématurément, le joueur perd un point d'honneur et son niveau est légèrement diminué. L'absence de réponse du joueur peut avoir deux causes principales : soit le terme n'est pas un terme courant (par exemple : *sycophante*), soit la consigne appliquée à ce terme n'a pas une grande signification (par exemple : *contraires de Star Trek ?*). Le système mémorise alors le fait que ce terme est peu productif, en particulier par rapport à cette consigne ; ce terme appliqué à cette consigne sera moins souvent proposé.

Toute partie créée avec un joueur (A) génère deux parties à finir avec des joueurs (B). En effet, si tel n'était pas le cas, il suffirait que le joueur (B) passe le mot sans faire de proposition pour initier un sentiment de frustration chez le joueur (A), ce qui le démotiverait à revenir participer. Il est donc proposé à tout joueur qui se connecte un plus grand nombre de parties à finir que de parties en création, et ce d'autant plus qu'un joueur (A) satisfait de ses propositions peut acheter des parties, comme nous venons de le voir. Les proportions sont telles qu'il est actuellement proposé environ trois parties à finir pour une seule partie en création.

Lorsque deux joueurs obtiennent sur un mot M un nombre de points supérieur à un certain seuil (initialement 250 points), ils capturent ce mot. Mais si un couple de joueurs obtient sur ce mot M un nombre de points supérieurs, ils capturent ce mot qui est donc perdu par les deux joueurs qui le possédaient. Les joueurs vont donc tenter de se « voler des mots », ce qui augmente l'activité du site, et donc le nombre et le poids des relations.

Afin de permettre à tout joueur de se comparer aux autres membres de la communauté, il est possible d'afficher un tableau récapitulatif des joueurs enregistrés, avec leurs performances, par ordre de classement selon les points d'honneur, ainsi que les meilleurs scores obtenus sur une partie.

## 4. Résultats

Cette première version de JeuxDeMots est relativement récente : son lancement a eu lieu en juillet 2007. En trois mois, plus de 300 joueurs sont enregistrés et la plupart d'entre eux se connectent plusieurs fois par semaine. Les quelques 18.000 parties jouées ont fait émerger plus de 22.000 relations, dont 13.000 de type « idées associées », 2000 de type « synonymes », 1000 « contraires », … Actuellement, seules une trentaine de relations sont taboues. La figure 3 illustre l'évolution du nombre de relations que comporte notre réseau. On constate une émergence rapide des relations, même si pour l'instant ce sont les plus fortes qui sont créées. L'évolution de la base de termes est nécessairement plus lente : elle compte à ce jour environ 152.000 termes ; les joueurs y ont déjà ajouté plus de 500 nouveaux termes, principalement conjoncturels ou liés à l'actualité. La figure 4 présente un exemple partiel du réseau lexical obtenu.

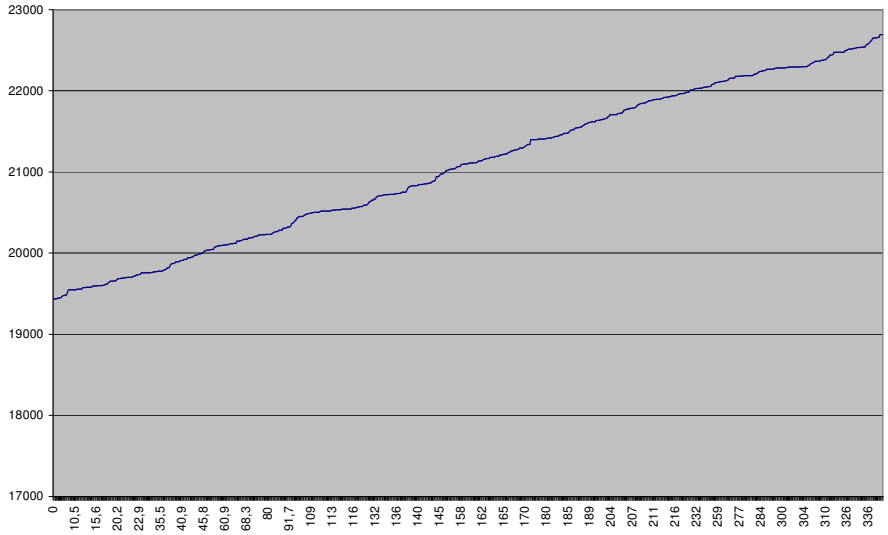



*Figure 3 : Evolution du nombre de relations sur les deux premières semaines d'octobre 2007 (en abscisse : l'échelle du temps exprimée en heures)*

Nous avons comparé les résultats obtenus à ce jour grâce à JeuxDeMots (JDM) avec les données d'Euro Wordnet Français (EWF). JDM possède environ 6 fois plus de termes qu'EWF qui en comporte environ 23.000. Par contre, EWF est beaucoup plus riche au niveau des relations : il en compte actuellement un peu plus de 100.000 ; mais le nombre de relations dans JDM est en progression quasi constante d'environ 6000 relations supplémentaires par mois.

Sur un échantillon des 100 termes les plus couramment proposés par les utilisateurs de JDM (voir Annexe), nous avons étudiés les termes auxquels ils sont associés. Dans 97% des cas,

ces associations sont au moins correctes. Les 3% restants correspondent à des erreurs (fautes d'orthographe, par exemple *Mongolfière* pour *Montgolfière*) ou des confusions (par exemple : *Dalila* → *chanteuse*). Les données collectées grâce à JDM apportent beaucoup d'originalité, mais le taux de précision est bien moins important que celui des données obtenues manuellement dans EWF. Ce manque de précision semble dû à la relative nouveauté de JDM, les relations les plus précises augmentant régulièrement en poids.

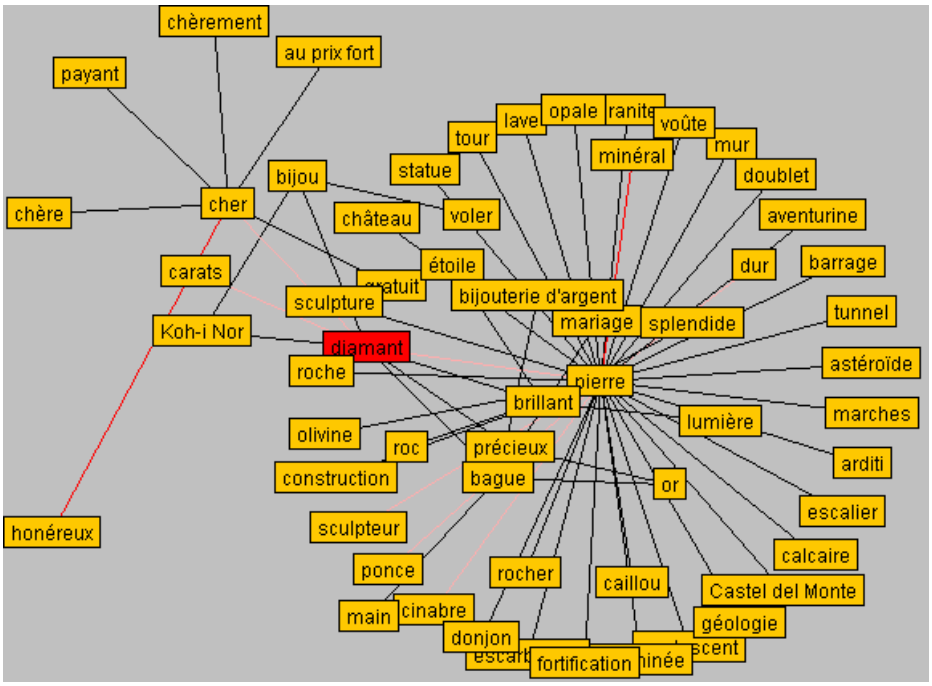



*Figure 4 : Exemple partiel du réseau lexical avec une distance de 2, c'est-à-dire que l'on présente un terme central (ici :* diamant*), tous les termes directement reliés à ce terme central, ainsi que tous les termes qui leur sont reliés. On constate que l'une des relations principales est* diamant → pierre*, d'où l'affichage des relations au départ de* pierre*. Les nœuds correspondant à des termes très généraux (ici :* pierre*) sont des « hubs » dans le réseau, destinataires d'un grand nombre de relations. Le graphe présenté ici ne tient pas compte de la direction des relations. On remarquera également un effet de bord dans la présence de la relation* pierre → arditi*, dû à l'homonymie entre le prénom* Pierre *et le nom commun* pierre.

## 5. Conclusion

Le prototype JeuxDeMots présenté ici a pour objectif la construction d'un réseau lexical via une activité ludique proposée sur le web. L'émergence de relations typées et pondérées entre termes s'effectue grâce au concours d'un grand nombre d'utilisateurs. Au vu des résultats actuels, bien que récents, nous espérons obtenir un tel réseau, évolutif et de bonne qualité, avec une couverture satisfaisante de l'ensemble des connaissances[6].

## Références

vonAhn L. et Dabbish L. (2004) Labelling Images with a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 319-326.

Kipfer B.A. (2001). *Roget's International Thesaurus*, sixth edition, Harper Resource (First Edition : 1852)

[6] De plus, une fois le projet développé, cette acquisition de ressources lexicales est totalement gratuite !

Lapata M. et 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.

Mel'čuk I.A., Clas A., Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot AUPELF-UREF

Miller G.A., Beckwith R., Fellbaum C., Gross D. and Miller K.J. (1990) Introduction to WordNet: an on-line lexical database. In: *International Journal of Lexicography* 3 (4), pp. 235-244.

Polguère A. (2006) Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability* (COLING/ACL 2006), Sydney, pp. 50-59.

Robertson S. et Spark Jones K. (1976) Relevance weighting of search terms, *Journal of the American Society for Information Science*, n° 27, pp. 129-146.

Salton G. (1968) *Automatic Information Organization and Retrieval*, Mac Graw Hill, NY.

## Annexe

Les 100 termes actuellement les plus courants dans JeuxDeMots :

voiture -- 552
maladie -- 513
mer -- 512
animal -- 509
musique -- 466
eau -- 457
livre -- 448
guerre -- 445
Harry Potter -- 404
sport -- 394
magie -- 388
argent -- 388
avion -- 381
ville -- 370
pays -- 364
train -- 352
sexe -- 346
maison -- 340
bateau -- 336
femme -- 332
sorcier -- 330
amour -- 326
couleur -- 322
président -- 320
chat -- 308
film -- 303
soleil -- 300
homme -- 300
blanc -- 294
arbre -- 292
oiseau -- 290
politique -- 286
médecin -- 286
froid -- 281
école -- 278
peinture -- 278
cinéma -- 272
jeu -- 264
noir -- 259
yeux -- 258
plante -- 258
rouge -- 255
tête -- 254
fleur -- 252
ordinateur -- 250
langue -- 250
bois -- 248
roman -- 245
tennis -- 242
manger -- 238
feu -- 236
religion -- 229
métal -- 226
médecine -- 226
pluie -- 224
malade -- 224
art -- 224
neige -- 222
fruit -- 221
lit -- 218
corps -- 216
vélo -- 216
dormir -- 214
enfant -- 212
vin -- 209
informatique -- 208
porte -- 208
chanteur -- 204
peintre -- 203
chien -- 202
préservatif -- 202
internet -- 200
mort -- 198
écrivain -- 198
rugby -- 198
ballon -- 198
course -- 195
chaud -- 194
pied -- 194
médicament -- 192
France -- 192
bleu -- 190
baguette -- 190
cheval -- 186
temps -- 184
os -- 184
prénom -- 184
félin -- 184
famille -- 180
père -- 180
acteur -- 180
ciel -- 180
espace -- 178
moteur -- 176
température -- 175
plage -- 174
montagne -- 174
roi -- 172
nuit -- 170