Similitude entre les sens d'usage d'un terme dans un réseau lexical

Mathieu Lafourcade – Alain Joubert

LIRMM – Université Montpellier 2 - CNRS
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
161, rue Ada, F-34392 Montpellier cedex 5

{lafourcade, joubert}@lirmm.fr

RÉSUMÉ : *Les relations lexicales typées entre termes sont indispensables pour les tâches réalisées en TALN, mais leur collecte peut s'avérer difficile. En effet, effectuée manuellement, cette tâche nécessite la compétence d'experts et la durée nécessaire peut être prohibitive, 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. À partir d'une base de termes préexistante, ce sont ainsi les joueurs qui vont construire le réseau lexical, en fournissant des associations qui ne sont validées que si elles sont proposées par au moins une paire d'utilisateurs. De plus, ces relations typées sont pondérées en fonction du nombre de paires d'utilisateurs qui les ont proposées. Nous abordons ensuite la question de la détermination des différents sens d'usage d'un terme, en analysant les relations entre ce terme et ses voisins immédiats dans le réseau lexical, avant d'introduire la notion de similitude entre ces différents sens d'usage. Nous pouvons ainsi construire pour un terme son arbre des usages. Enfin, nous présentons brièvement la réalisation et les premiers résultats obtenus.*

ABSTRACT : *Typed lexical relations between terms are indispensable for the tasks realized in NLP, but collecting lexical information is a difficult process. Indeed, when done manually, it requires the competence of experts and the duration can be prohibitive. When done automatically, the results can be biased by the chosen 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. From an already existing base of terms, the players themselves thus build the lexical network, by supplying associations which are validated only by an agreeing pair of users. Furthermore, these typed relations are weighted according to the number of pairs of users who provide them. We then approach the question of word usage determination for a term, by searching relations between this term and its neighbours in the network, before introducing the notion of similarity between these different word usages. We are thus able to build the tree of word usages for a term. Finally, we briefly present the realization and the first obtained results.*

MOTS-CLÉS : *traitement automatique du langage naturel, réseau lexical, relations typées pondérées, sens d'usage d'un terme, jeu en ligne.*

KEYWORDS : *Natural Language Processing, lexical network, typed and weighted relations, word usage, web-based game.*

1. Introduction

Un très grand nombre d'applications en traitement automatique des langues (TAL), en particulier celles requérant un processus de désambiguïsation, nécessite la connaissance de relations lexicales ou fonctionnelles entre termes. Ces relations, que l'on trouve généralement dans des thésaurus ou des ontologies, peuvent être mises en évidence de façon manuelle ; il est possible de citer ici, par exemple, le thésaurus (Roget, 1852), l'un des plus anciens, le thésaurus Larousse (Pechoin, 1991) pour la langue française ou l'un des plus célèbres réseaux lexicaux, WordNet (Miller *et al.*, 1990). De telles relations peuvent aussi être déterminées automatiquement à partir de corpus de textes, par exemple (Robertson et Spark Jones, 1976), (Wettler et Rapp, 1992) ou (Lapata et Keller, 2005), dans lesquels sont effectuées des études statistiques sur les distributions de mots. De nombreux travaux ont également porté sur la détection de collocations comme ceux de (Spence et Owens, 1990), (Smadja, 1993) ou plus récemment (Ferret, 2002). La méthode *Latent Semantic Analysis* (LSA), présentée par (Dumais, 1994) ou (Landauer et Dumais, 1997), s'appuie également sur des ensembles de textes ; elle permet de calculer la proximité sémantique entre mots, produisant ainsi des nuages de termes appartenant à un même champ sémantique. LSA s'appuie sur le contexte dans lequel les mots sont utilisés : deux mots sont considérés comme sémantiquement proches si les contextes dans lesquels on les rencontre dans le corpus sont similaires. On en trouvera une réflexion récente dans (Wandmacher *et al.*, 2008). En outre, certaines applications du TAL requièrent 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, nécessite des ressources (en durée et en personnel) qui peuvent être prohibitives, alors que leur extraction automatique sur un corpus de textes semble beaucoup trop dépendante du domaine des textes choisis.

La méthode développée ici s'appuie sur un système contributif, où ce sont les utilisateurs qui font évoluer la base, au travers d'une interface présentée sous forme d'un jeu en ligne. Cette méthode s'apparente à celle utilisée par (Zock et Quint, 2004) pour l'apprentissage de la langue japonaise. De plus, le prototype introduit ici réalise l'acquisition d'informations lexicales évolutives, contrairement à la plupart des méthodes classiques qui permettent d'acquérir des informations lexicales généralement statiques, même si actuellement bon nombre de bases de connaissance évoluent de façon incrémentale.

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

1. JeuxDeMots est accessible à l'adresse http://jeuxdemots.org. Depuis peu, il existe aussi une version anglaise, ainsi qu'une version thaï et une version japonaise (toutes deux en cours de développement), à l'adresse http://www.lirmm.fr/jeuxdemots/world-of-jeuxdemots.php

des éléments clés étant qu'une relation ne peut être validée que si elle est proposée par au moins deux utilisateurs. Dans une deuxième partie, utilisant le réseau ainsi obtenu, nous abordons la problématique de la détermination de la polysémie d'usage, puis de la similarité entre les différents usages d'un même terme qui nous permet d'introduire la notion d'arbre des usages. Enfin, nous présentons brièvement la réalisation et les premiers résultats obtenus.

2. Construction du réseau lexical

2.1. *Structure du réseau lexical*

La structure du réseau lexical que nous cherchons à obtenir s'appuie sur les notions de nœuds et de relations entre nœuds, selon un modèle initialement présenté à la fin des années 1960 par (Collins et Quillian, 1969), développé dans (Sowa, 1992), utilisé dans les petits mondes par (Gaume, 2006) et (Gaume *et al.*, 2007), et plus récemment explicité par (Polguère 2006). Chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies. Les relations entre nœuds sont typées. Certaines de ces relations correspondent à des fonctions lexicales, telles qu'explicitées par (Mel'čuk, 1988), (Mel'čuk *et al.*, 1995) et (Polguère, 2003). Nous aurions souhaité que notre réseau comporte toutes les fonctions lexicales définies dans (Mel'čuk, 1988), mais, compte tenu du principe de notre logiciel JeuxDeMots explicité en section 2.2, cela n'est pas raisonnablement possible. En effet, certaines de ces fonctions lexicales sont trop spécialisées ; par exemple, (Mel'čuk, 1988) fait la distinction entre les fonctions *Conversif*, *Antonyme* et *Contrastif*. Il considère également des raffinements, avec des fonctions lexicales caractérisées de « *plus larges* » ou « *plus étroites* ». JeuxDeMots s'adressant à des utilisateurs qui sont de « simples internautes » et non pas nécessairement des experts en linguistique, de telles fonctions auraient pu être mal interprétées par eux. De plus, certaines de ces fonctions sont trop peu lexicalisées, c'est-à-dire que trop peu de termes possèdent des occurrences de telles relations ; c'est par exemple le cas des fonctions de *Métaphore* ou de *Fonctionnement avec difficulté*.

JeuxDeMots possède une liste prédéterminée de types de relations ; les joueurs ne peuvent pas introduire de nouveaux types. Ces types de relations sont de plusieurs catégories (une liste détaillée est fournie en annexe) :

– **relations lexicales** : synonymie, antonymie, locution, famille. Il s'agit de type de relations portant sur le vocabulaire ;

– **relations ontologiques** : générique (hyperonymie), spécifique (hyponymie), partie (méronymie), tout (holonymie)… Il s'agit de relations portant sur des connaissances liées à des objets du monde ;

– **relations associatives** : association libre, sentiment associé, signification. Il s'agit plutôt de connaissances subjectives et globales ; certaines d'entre elles peuvent être considérées comme des associations syntagmatiques ;

– relations prédicatives : agent typique, patient typique… Il s'agit de relations associées à un verbe et les valeurs de ses arguments (au sens très large).

Les types de relations mis en œuvre dans JeuxDeMots sont donc de plusieurs natures, en partie selon une distinction faite par (Schwab et Lafourcade, 2007) : certains relèvent de connaissances du monde (hyperonymie/hyponymie, par exemple), d'autres concernent des connaissances linguistiques (synonymie, antonymie, locution ou famille, par exemple). La plupart des joueurs ne sont pas sensibles à cette distinction qui souvent est floue. La liste des types de relations utilisés dans JeuxDeMots est donnée en annexe.

Dans la suite de cet article, lorsque nous parlons d'une relation, il faut entendre une occurrence de relation, et non un type de relation. De telles relations peuvent être considérées comme des quadruplets : terme source, type de relation, terme cible, poids de cette relation. La figure 1 montre quelques exemples de relations acquises dans JeuxDeMots.

2.2. *Principe du logiciel*

Afin d'éviter les écueils d'un système où un utilisateur pourrait, volontairement ou non, donner des informations erronées, et donc pour assurer la qualité et l'intégrité de la base lexicale, il a été décidé que les validations des relations proposées anonymement par un joueur seraient effectuées par d'autres joueurs, tout autant anonymement. En pratique, les validations sont 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 ou plus récemment par (Lieberman *et al.*, 2007) pour la collecte de « connaissances de bon sens ». À notre connaissance, il n'a jamais été mis en œuvre dans le domaine des réseaux lexicaux.

Une partie se déroule entre deux joueurs, en asynchrone, fondée sur la concordance de leurs propositions. Lorsqu'un premier joueur que nous appellerons (A) débute une partie, une consigne correspondant à un type de relation (synonymies, contraires, parties de…) est affichée, ainsi qu'un terme[2] T tiré aléatoirement dans une base de mots. Ce joueur (A) a alors un temps limité pour répondre en donnant des propositions correspondant, selon lui, à la consigne appliquée au terme T. Le nombre de propositions qu'il peut faire est limité : nous souhaitons que les joueurs réfléchissent « un minimum ». Ce même terme, avec cette même consigne, est proposé par la suite à un autre joueur que nous appellerons (B) ; le processus est identique. Afin d'accroître l'aspect ludique, 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.

2. Un terme peut être constitué de plusieurs mots (exemple : *feu de bois* ou *énergie renouvelable*).

La démarche présentée ici est complémentaire de celle développée par (Zock et Bilac, 2004) et (Zock et Schwab, 2008) qui cherchent à créer un index fondé sur la notion d'association afin d'assister à la navigation, pour aider un être humain à trouver un mot qu'il a « sur le bout de la langue ». Leur démarche est de type *bottom-up :* on connaît les termes (co-occurrences trouvées dans un corpus), mais pas la « nature » du lien ; celle-ci doit être induite, ce qui est loin d'être trivial. Dans notre cas, nous connaissons l'un des deux termes (le terme source) ainsi que le type de la relation (imposé par la consigne donnée aux joueurs). Nous cherchons plusieurs deuxièmes termes (les termes cibles). Notre démarche est plutôt de type *top-down*.

Pour le terme source T, nous mémorisons les réponses communes aux joueurs (A) et (B). Nous ne mémorisons 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 paire 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. Initialement, les nœuds sont constitués des termes de notre base de départ, mais celle-ci peut s'accroître ; effectivement, si les deux joueurs (A) et (B) d'une même partie proposent un terme initialement inconnu, alors ce terme est ajouté à notre base. La figure 1 présente une partie de l'ensemble des termes que les joueurs ont associés au terme *bois* en fonction de la consigne qui leur était imposée. La relation « idée associée » (correspondant à une association libre) est la plus forte, parce que c'est celle qui est proposée en premier aux joueurs, afin de les familiariser avec JeuxDeMots. Remarquons que ce lien « idée associée » correspond à un type de relation beaucoup moins spécifique que la plupart des autres : on peut le considérer comme un lien sous-spécifié.

'bois'

125 relations ==>

bois ---r_associated:340--> arbre
bois ---r_associated:250--> forêt
bois ---r_has_part:250--> arbre
bois ---r_has_part:160--> fibre
bois ---r_associated:140--> meuble
bois ---r_holo:120--> arbre
bois ---r_holo:120--> meuble
bois ---r_lieu:100--> forêt
bois ---r_carac:80--> dur
bois ---r_lieu-1:80--> arbre
bois ---r_carac:70--> marron
bois ---r_familly:70--> boisé
bois ---r_locution:60--> petit bois
bois ---r_lieu:60--> cheminée
bois ---r_lieu:60--> meuble
bois ---r_carac:60--> sec
bois ---r_carac:60--> tendre
bois ---r_carac:60--> vieux
bois ---r_lieu_action:60--> couper
bois ---r_lieu_action:60--> scier
…

280 relations <==

feu ---r_associated:300--> bois
champignon ---r_associated:210--> bois
forêt ---r_associated:200--> bois
arbre ---r_associated:190--> bois
sabots ---r_has_part:140--> bois
charpente ---r_has_part:130--> bois
entailler ---r_patient:110--> bois
parquet ---r_associated:110--> bois
travailler ---r_patient:110--> bois
baguette ---r_isa:100--> bois
caisse ---r_carac:100--> bois
campagne ---r_lieu-1:100--> bois
acacia ---r_holo:90--> bois
arbre ---r_magn:90--> bois
brûler ---r_agent:90--> bois

menuiserie ---r_hypo:90--> bois
sculpture ---r_carac:90--> bois
bûcher ---r_associated:80--> bois
bûcher ---r_has_part:80--> bois
orée ---r_holo:80--> bois
…

Figure 1. *Ensemble (partiel) des termes associés par les joueurs au terme bois. Sont présentées tout d'abord les relations*[3] *dont le terme bois est source, puis celles pour lesquelles le terme bois est cible. Pour chacune de ces relations, on a son type (« idée associée », « a pour partie »…) correspondant à la consigne, ainsi que son poids. Le calcul de cette pondération est expliqué à la section 2.3*

Nous n'avons pas souhaité compléter notre réseau par l'ajout automatique de relations, alors que dans certains cas cela est manifestement possible. En effet, par exemple dans le cas de relations ontologiques, si la relation *animal générique de chat* est donnée par les joueurs, la relation *chat spécifique de animal* aurait pu être créée automatiquement. Nous nous y sommes refusés ; nous souhaitons que notre réseau reflète les idées qui viennent spontanément à l'esprit des joueurs. Dans l'exemple du couple *chat – animal*, les deux relations *générique* et *spécifique* existent, ce n'est pas le cas pour le couple *antilope – animal*.

Il aurait pu être envisagé de mémoriser toutes les réponses, depuis le début du jeu, avec leurs fréquences. Notre base se serait accrue beaucoup plus rapidement, mais cela aurait été au détriment de sa qualité, en particulier le bruit aurait été très important. L'intérêt de la solution retenue (validation par couple de joueurs) est de limiter de façon beaucoup plus drastique les réponses « fantaisistes » ou les erreurs dues à une mauvaise compréhension de la consigne, voire du terme T lui-même. L'émergence des solutions « originales » sera plus lente, mais elle se fera tout de même, après saturation des solutions les plus courantes, grâce au processus des termes « tabous ». En effet, lorsqu'une relation *terme T → terme proposé* a été établie par un grand nombre de couples de joueurs, elle devient banale ou taboue ; elle est affichée en même temps que le terme T, afin que les joueurs ne la proposent plus. Ainsi, les joueurs sont amenés à faire d'autres propositions, généralement plus originales. Ceci favorise l'émergence de relations plus rares, mais non l'émergence d'erreurs. En fait, notre méthode de validation par paire d'utilisateurs élimine les erreurs occasionnelles, mais les erreurs courantes, comme certaines fautes d'orthographe, pourront toutefois émerger ; on peut en voir un exemple en annexe avec le terme *théâtre* écrit sans accent circonflexe par certains utilisateurs.

2.3. *Émergence et pondération des relations entre termes*

Côté jeu, il s'agit de définir le nombre de points gagnés par les joueurs (A) et (B), avec la même consigne sur un même terme T. Côté réseau lexical, il s'agit

3. Ces relations sont accessibles à l'adresse http://jeuxdemots.org/rezo.php.

d'établir des relations entre termes, grâce aux propositions faites par (A) et (B). 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 : T → x_i.

L'un des points forts de notre méthode réside dans la 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. À partir d'un certain nombre d'occurrences, une relation R est bien établie : elle devient alors banale, ou « taboue ». 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 : (1000 – poids(R)) / 10. Plus la relation est récente, plus elle a de valeur : cela revient à « récompenser la primauté ». Cette fonction est décroissante : une relation rapporte de moins en moins de points.À partir d'un certain seuil de poids, fixé actuellement à 300, la relation devient taboue ; elle est alors indiquée en même temps que la consigne et le terme T (c'est une solution donnée, exemple : *bois* → *arbre* ou *poutre* → *bois* pour le type « idée associée »). Cette proposition n'est plus intéressante pour les joueurs qui sont donc invités à faire d'autres propositions. 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 : *élection* → *Obama*). À chaque partie sur un terme T, le poids des relations existantes dans notre base à partir de ce terme T, qui ne sont proposées par aucun des deux joueurs est très légèrement diminué

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.

(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].

3. Détermination des sens d'usage

3.1. *Principe général*

En première approximation, il est possible de considérer que si un terme T est polysémique, les termes qui lui sont reliés forment plusieurs groupes distincts, chacun de ces groupes constituant un sens d'usage de T. Nous faisons ici la distinction entre les notions de sens d'usage et de sens. La notion de sens d'usage (appelée plus communément usage) est beaucoup plus fine que celle de sens qui, comme l'a montré (Véronis, 2001), est relativement pauvre lorsqu'on se réfère aux dictionnaires traditionnels ou à des ressources comme WordNet. L'usage est donc en TALN une notion plus importante que le sens. Pour citer l'exemple que nous prendrons à la section 3.3, *bois-arbre-tronc (bois : constituant du tronc d'un arbre)* et *bois-arbre-forêt-chêne (bois : constituant des arbres d'une forêt de chênes)* constituent deux usages distincts du terme *bois* dans notre réseau, alors qu'il s'agit manifestement du même sens de ce terme.

3.2. *Détermination des cliques*

Comment déterminer une clique ? C'est un ensemble de termes « fortement » reliés entre eux, c'est-à-dire deux à deux adjacents, constituant un sous-graphe induit complet (ou clique) dans le réseau lexical. Les liens considérés ici sont uniquement des relations de type « *idée associée* » symétrique. Nous faisons abstraction des autres types de relations (« *tout de* », *« partie de* »…) en raison de leur caractère non symétrique. L'objectif est la construction d'un réseau lexical dans lequel chaque nœud est constitué par un usage d'un terme, et non plus par un terme regroupant ses éventuels différents usages.

Dans le réseau lexical, un terme T est directement relié à n termes : $T_1,\ldots,T_i,\ldots,T_j,\ldots,T_n$. Les termes T_i et T_j appartiennent à deux usages différents de T si :

– $\text{poids}(T_i\text{-}T_j) \leq k * \min(\text{poids}(T\text{-}T_i), \text{poids}(T\text{-}T_j))$.

où k est un coefficient de seuil et $\text{poids}(T_i\text{-}T_j)$ désigne le poids de la relation « idée associée » entre les termes T_i et T_j.

5. Ce processus d'érosion est encore au stade expérimental : nous n'avons pas assez de recul pour voir ces effets dans la durée.

Pour simplifier, nous avons actuellement pris k = 0, c'est-à-dire que les termes T_i et T_j appartiennent à deux usages distincts de T s'ils ne sont pas directement reliés.

Nous nous attachons ici aux différents usages du terme T ; le but est donc de regrouper ces n termes en un ou plusieurs groupes, chacun constituant un usage de T. Les termes T_{i1}, … ,T_{im} constituent le i-ième usage de T si les (m + 1) * m / 2 relations entre ces (m + 1) termes existent. Ainsi, un terme T_j n'appartiendra pas à ce i-ième usage de T si au moins une des relations entre ce terme T_j et l'un des termes T_{i1}, … ,T_{im} n'existe pas.

3.3. *Exemple du terme* bois

Les figures 2 et 3 illustrent, sur un exemple simple, les résultats obtenus permettant de déterminer les différents usages d'un même terme, ainsi que la pertinence de chacun de ces usages (le principe de calcul de cette pondération est explicité à la section 3.4). Il est possible de remarquer sur cet exemple qu'un usage étant constitué d'une clique de termes, et non simplement d'une composante connexe, un même terme T_i relié à T peut appartenir à plusieurs cliques et donc un même terme T_i peut servir à définir plusieurs usages de T.

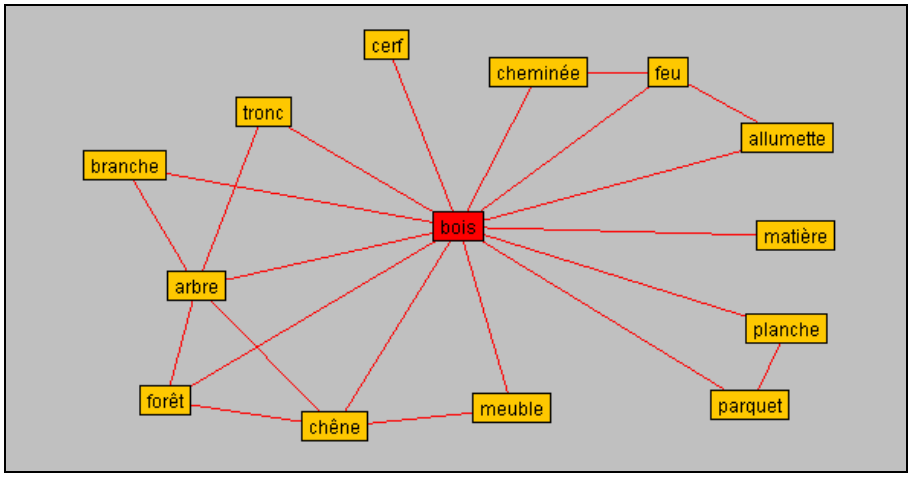


Figure 2. *Cette figure montre, autour du terme bois, le réseau lexical en ne considérant que les relations de type « idées associées » symétrique. Ainsi reliés au terme bois se trouvent treize termes qui correspondent selon les critères définis dans cet article à neuf usages*

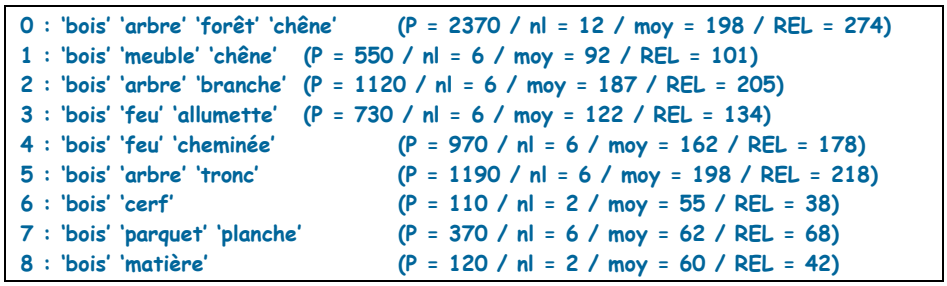


Figure 3. *Cette figure montre les neuf usages du terme bois effectivement décelés. Pour chacun de ces usages du terme bois figure également une évaluation de sa pertinence (valeur REL)*

Le terme *bois* est polysémique, les neuf usages (voir figures 2 et 3) correspondent à plusieurs sens distincts : *forêt*, *matière* et *cornes des cervidés*. On remarquera que notre réseau n'est pas encore complet, le sens *instrument de musique* n'y figure pas. Dans notre réseau, la relation *bois* → *flûte* existe, mais elle n'est pas symétrique, la relation *flûte* → *bois* n'existant pas actuellement.

3.4. *Pertinence d'un usage*

Évaluer la pertinence d'un usage consiste à obtenir une mesure de son importance à la fois en termes de fréquence d'utilisation mais aussi en termes de couverture lexicale. On émettra l'hypothèse que pour un terme donné et en dehors de tout contexte spécifique, l'usage le plus pertinent est celui auquel on pense généralement en premier. Ainsi donc, lors d'une analyse sémantique de texte, les usages peuvent être pondérés par défaut en fonction de leur pertinence *a priori*. Compte tenu du principe de la pondération des relations dans notre réseau lexical, le poids d'un usage est corrélé aux poids des relations entre les termes de la clique qui caractérise cet usage. Ainsi, pour une clique C de m termes et comprenant le terme T, le poids de l'usage correspondant sera égal à :

– $P(C) = \sum_{i,j} poids(T_i\text{-}T_j)$

– $Rel(C) = Ln(m) * P(C) / [m*(m -1)]$

Le terme $poids(T_i\text{-}T_j)$ est le poids de la relation entre Ti et Tj. La pertinence est la moyenne des poids des relations existant entre les termes, valeur qui exprime la cohérence de la clique, que multiplie le logarithme du nombre de termes impliqués dans la clique.

4. Similarité entre cliques

4.1. *Définition*

Il s'agit de définir un coefficient, compris entre 0 et 1, qui traduira la proximité sémantique de deux usages, réalisés dans le réseau sous forme de deux cliques. Plus l'intersection entre deux cliques sera importante, plus le coefficient de similarité sera proche de 1. De nombreux travaux ont été réalisés sur la similarité. (Tversky, 1977) définit la similarité de deux objets comme étant fonction de leurs caractéristiques communes par rapport à l'ensemble de leurs caractéristiques. En TALN, on trouve plusieurs définitions de la similarité, par exemple (Manning et Schütze, 1999), ou plus récemment (Fairon et Ho, 2004). En particulier, dans un certain nombre de modélisations une représentation vectorielle ou matricielle est utilisée ; la fonction cosinus est alors couramment employée pour exprimer la similarité entre deux vecteurs. Dans notre cas, elle correspond au rapport du poids de l'intersection de deux cliques sur la moyenne géométrique des poids de ces deux cliques. En considérant chaque clique comme un ensemble de relations pondérées, la similarité entre deux cliques C1 et C2 s'écrira :

$$Sim(C1,C2) = \sum_{C1 \cap C2} Poids(relations) \; / \; [\sum C1 \; Poids(relations) * \sum C2 \; Poids(relations)]^{1/2}$$

ou, en utilisant les notations ci-dessus :

$$Sim(C1,C2) = P(C1 \cap C2) \; / \; [P(C1) * P(C2)]^{1/2}$$

où P(C) désigne la somme des poids des relations composant la clique C.

4.2. *Raffinement des usages*

Comme vu à la section 3, chaque clique correspond à un usage d'un terme. Pour un terme, deux cliques voisines correspondent à deux usages voisins de ce terme : plus le coefficient de similarité est proche de 1, plus proches sémantiquement sont les cliques. En utilisant le coefficient de similarité, il est possible de regrouper les différents usages d'un même terme : il est possible d'estimer qu'au-delà d'un certain seuil du coefficient de similarité, proche de 1, deux cliques d'un même terme correspondent en fait à deux raffinements d'un même usage de ce terme. Il est d'ailleurs probable qu'à terme deux cliques très voisines puissent fusionner, dans une évolution ultérieure du réseau, en fonction de l'activité des joueurs.

Par exemple, dans l'état actuel de notre réseau, le terme *soleil* possède 27 cliques. La figure 4 présente un extrait de cet ensemble de cliques.

```
...
Clique 4 : 'soleil' 'étoile' 'lune' 'planète' 'ciel'              (P = 2440 / nl = 20 / moy = 122 / REL = 196)
Clique 5 : 'soleil' 'étoile' 'astre' 'planète' 'galaxie' (P = 2400 / nl = 20 / moy = 120 / REL = 193)
Clique 6 : 'soleil' 'étoile' 'astre' 'lune' 'planète'             (P = 2210 / nl = 20 / moy = 111 / REL = 178)
...
Clique 9 : 'soleil' 'astre' 'astronomie' 'galaxie'                (P = 1060 / nl = 12 / moy = 88 / REL = 122)
...
Clique 11 : 'soleil' 'planète' 'galaxie' 'système solaire' (P = 1100 / nl = 12 / moy = 92 / REL = 127)
...
Clique 22 : 'soleil' 'ciel' 'nuage' 'pluie'                       (P = 2680 / nl = 12 / moy = 223 / REL = 310)
...
```

Figure 4. *Extrait de la liste des cliques du terme* soleil.

– *Clique 4* => soleil *:* étoile *gravitant dans le* ciel *avec les* planètes *et la* lune
– *Clique 5* => soleil *:* étoile *d'un système* planétaire *de la* galaxie
– *Clique 6* => soleil *: dans le sens* astre*, comme la* lune *et les* planètes
– *Clique 9* => *aspect plus scientifique (*astronomie*) de l'*astre soleil
– *Clique 11* => soleil *parmi les objets célestes, avec échelle des tailles :* planète – soleil – système solaire – galaxie
– *Clique 22* => soleil *en tant qu'élément de météorologie*

Les cinq premières cliques reproduites ici correspondent à cinq usages du terme *soleil* dans le sens *étoile*. Les autres cliques du terme *soleil* correspondent à des usages centrés sur des sens de *chaleur*, *lumière*, *vacances* ou *météorologie*, comme par exemple la clique 22. Nous avons calculé les similarités entre ces cliques ; en voici quelques valeurs :

Sim(C4,C5) = 0,294 Sim(C4,C6) = 0.514 Sim(C4,C22) = 0,150
Sim(C5,C6) = 0,669 Sim(C5,C22) = 0 Sim(C6,C22) = 0

Les cliques C5-C6 d'une part, et C22 d'autre part, correspondent manifestement à deux sens différents du terme *soleil*. Remarquons toutefois que notre réseau est évolutif : il est possible que certaines cliques existantes fusionnent, et de nouvelles cliques peuvent apparaître.

Il est également possible d'envisager qu'il existe un second seuil du coefficient de similarité, relativement faible, en deçà duquel deux cliques d'un même terme correspondent à deux sens distincts de ce terme. Comme mentionné à la section 3.1, la notion de sens est beaucoup plus large que celle d'usage, un sens regroupant généralement plusieurs usages. En particulier, deux cliques d'un même terme ne possédant que ce terme en commun correspondront fort probablement à deux sens distincts de ce terme. Par exemple, le terme *barreau* possède trois cliques (voir la figure 5). Ces trois cliques correspondent manifestement à trois sens distincts du terme *barreau*.

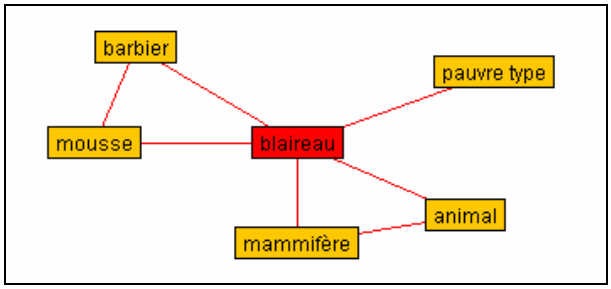
```
Clique 0 : 'barreau' 'avocat' 'justice'              (P = 760 / nl = 6 / moy = 127 / REL = 139)
Clique 1 : 'barreau' 'prison' 'prisonnier'  (P = 900 / nl = 6 / moy = 150 / REL = 165)
Clique 2 : 'barreau' 'chaise'                        (P = 150 / nl = 2 / moy = 75 / REL = 52)
```

Figure 5. *Les trois cliques du terme* barreau
- *Clique 0* => barreau *: ordre professionnel des* avocats
- *Clique 1* => barreau *: barre métallique dans une* prison
- *Clique 2* => barreau *: élément d'une* chaise

Ces trois cliques n'ayant que le terme barreau *en commun, leurs similarités seront nulles*

4.3. *Arbre des usages*

Notre objectif est d'obtenir une représentation des différents usages d'un terme T sous forme d'un arbre, la racine regroupant tous les sens de T, les branches correspondant à ses différents usages. La figure 6 illustre la construction de cet arbre sur un exemple simple : le terme *blaireau*. Ce terme possède actuellement trois cliques disjointes, leurs similarités sont nulles ; le terme *blaireau* présente donc trois usages. Son arbre des usages, présenté également à la figure 6, est constitué d'une racine regroupant tous ses usages et de trois branches simples correspondant chacune à un usage.



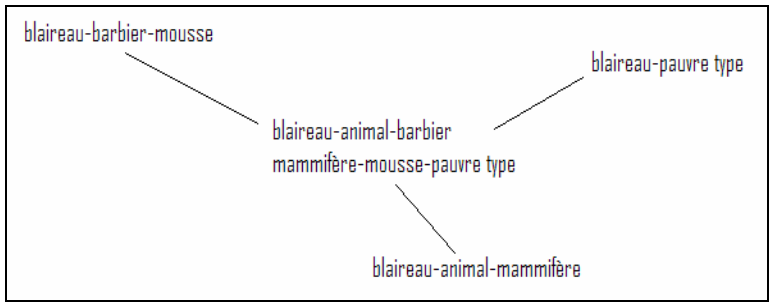
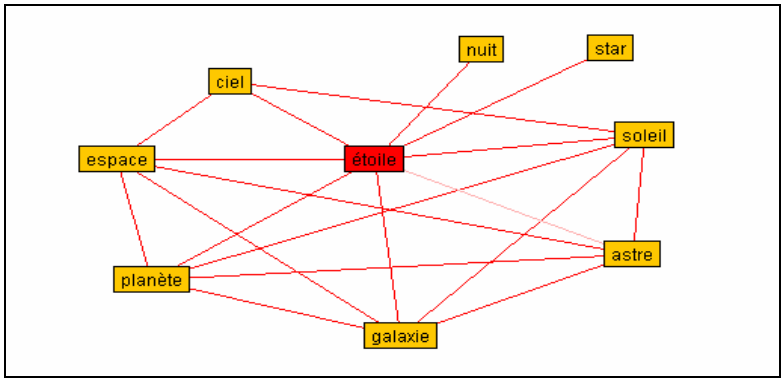


Figure 6. *Réseau lexical réduit aux cliques du terme* blaireau *et l'arbre des usages correspondant*

De manière générale, la plupart des termes possèdent plusieurs cliques non disjointes. Dans ce cas, l'arbre des usages est plus complexe que celui de l'exemple ci-dessus : plus on s'éloigne de la racine, plus on rencontre des distinctions fines d'usages. En réalité, nous construisons l'arbre des usages d'un terme T depuis l'ensemble de ses cliques, c'est-à-dire depuis ses feuilles en remontant jusqu'à sa racine qui regroupe tous les sens de T. Pour cela, nous fusionnons ces cliques, deux à deux, en commençant par celles dont le coefficient de similarité est le plus élevé et qui donc correspondent aux usages les plus proches : nous constituons ainsi des quasi-cliques représentant des regroupements d'usages. L'algorithme de fusion s'arrête lorsque tous les coefficients de similarité sont nuls ; nous faisons l'hypothèse que les quasi-cliques obtenues alors pourraient correspondre aux différents sens de T, tels qu'on pourrait les trouver dans un dictionnaire. La figure 7 montre, pour le terme *étoile*, l'ensemble des termes constituant ses six cliques. Pour ce terme, nous avons calculé son arbre des usages. Celui-ci est présenté à la figure 8.

L'arbre des usages d'un terme est une structure exprimant les raffinements de ses différents sens. Il constitue donc un arbre de décision, structure de données qui pourra être exploitable en désambiguïsation.



Clique 0 : 388 'étoile' 'ciel' 'espace' (P = 940 / nl = 6 / moy = 157 / REL = 172)
Clique 1 : 448 'étoile' 'ciel' 'soleil' (P = 1400 / nl = 6 / moy = 233 / REL = 256)
Clique 2 : 331 'étoile' 'soleil' 'astre' 'galaxie' 'planète' (P = 2670 / nl = 20 / moy = 134 / REL = 215)
Clique 3 : 288 'étoile' 'star' (P = 270 / nl = 2 / moy = 135 / REL = 94)
Clique 4 : 271 'étoile' 'astre' 'espace' 'galaxie' 'planète'(P = 2650 / nl = 20 / moy = 133 / REL = 213)
Clique 5 : 272 'étoile' 'nuit' (P = 310 / nl = 2 / moy = 155 / REL = 107)

Figure 7. *Réseau lexical réduit aux cliques du terme* étoile*, ainsi que la liste de ces cliques (précédées de leur numéro)*

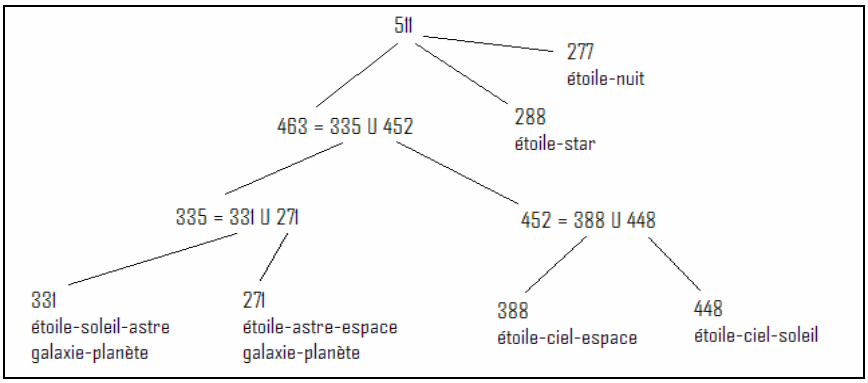


Figure 8. *Arbre des usages du terme* étoile*. Les valeurs numériques correspondent aux numéros des différentes cliques ou quasi-cliques. Les termes les constituant n'ont été indiqués qu'aux feuilles de l'arbre, afin de ne pas surcharger la figure*

5. Mise en œuvre et résultats

5.1. *Réalisation*

Le logiciel JeuxDeMots a été développé en PHP/MySQL; et certains programmes annexes ont été réalisés en langage JAVA et C++. L'interface, la comptabilisation de points, mais également des notions de niveau, honneur, captures de mots, procès entre joueurs… ainsi que l'affichage du classement des joueurs, ont été mis en œuvre afin d'accroître l'aspect attrayant du jeu. Le but recherché est d'inciter les joueurs à revenir régulièrement sur le site, et donc d'augmenter d'autant le nombre de relations acquises : c'est l'intérêt majeur de cette approche 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.

5.2. *Déroulement d'une partie*

Chaque fois qu'un joueur se connecte au site et démarre une partie, une consigne est affichée pendant quelques secondes (par exemple : *« 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 donner ses réponses. La figure 9 illustre un exemple de partie. Si le joueur est (B), il est procédé à l'affichage immédiat du résultat de la partie (voir figure 10) : propositions qu'avait faites le joueur (A) et nombre de points gagnés. 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.

Figure 9. *Exemple d'une partie en cours : le terme affiché est* stade*, l'utilisateur a proposé 11 termes qui lui semblent associés*

Figure 10. *Affichage des résultats de la partie précédente : les relations* stade → pelouse, stade → sport, stade → spectateur *et* stade → athlétisme *seront créées ou renforcées*

Si, à l'affichage d'un mot et de la consigne, un joueur estime n'avoir aucune idée, il a la possibilité de « passer » : la partie se termine alors prématurément. L'absence de réponse du joueur peut avoir trois causes principales : soit le terme n'est pas un terme courant (par exemple : *« forfanterie* »), soit la consigne appliquée à ce terme n'a pas une grande signification (par exemple : *« contraires de écureuil ?* »), soit – et c'est le cas le plus intéressant – tous les termes auxquels le joueur pensent sont tabous : le terme source est alors « saturé », toutes les relations qui le concernent sont bien établies. 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 ; cela est plus motivant pour le joueur qui ainsi voit plus souvent le résultat immédiat de ses propositions.

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.

5.3. *Résultats et comparaison avec Euro WordNet*

Cette première version de JeuxDeMots est relativement récente : son lancement a eu lieu en juillet 2007. En un peu plus d'une année et demie, plus de 1 200 joueurs se sont enregistrés et une bonne proportion d'entre eux se connectent plusieurs fois par semaine. Plus de 120 000 parties ont été jouées : elles ont fait émerger près de 180 000 relations, dont 70 000 de type « idées associées ». Actuellement, plus de 2 500 relations sont taboues, soit plus de 1 % du nombre total de relations. Il y a une émergence rapide des relations et on constate que les plus fortes (celles qui viennent le plus spontanément à l'esprit des joueurs) sont statistiquement créées en premier. Sans que cela soit une surprise, nous avons constaté qu'il y a une corrélation très forte entre le poids d'une relation et son rang de création parmi les autres relations pour un même terme source. L'évolution de la base de termes est nécessairement plus lente : elle compte à ce jour environ 164 000 termes ; les joueurs y ont déjà ajouté près de 10 000 nouveaux termes, principalement conjoncturels ou liés à l'actualité. La figure 11 présente un exemple partiel du réseau lexical obtenu.

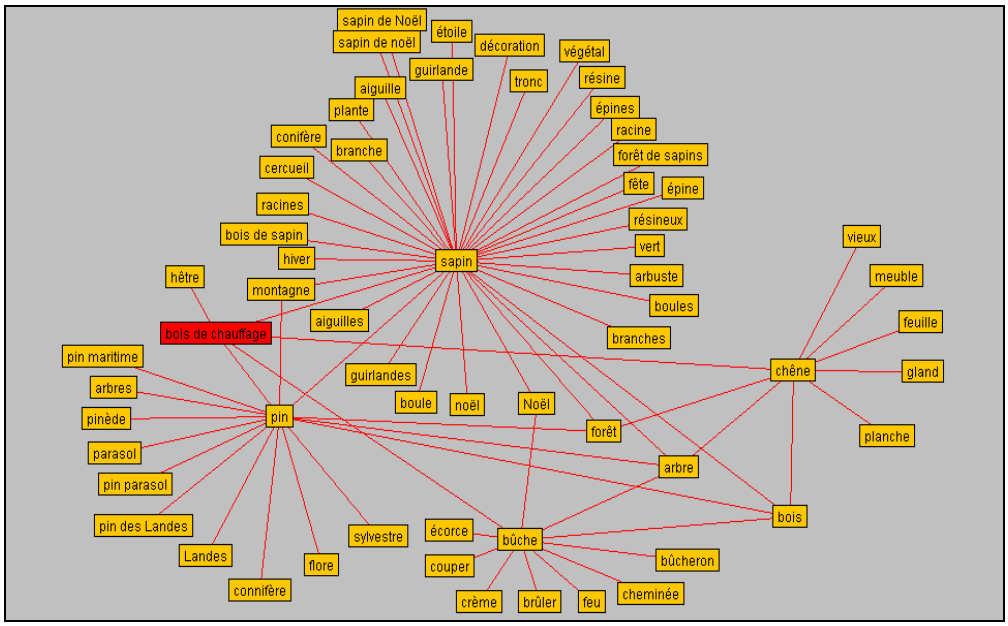


Figure 11. *Exemple partiel du réseau lexical du terme* bois de chauffage *avec une distance de 2. On constate que l'une des relations principales est celle liant* bois de chauffage *à* sapin*, d'où l'affichage des relations au départ de* sapin*. Les nœuds correspondant à des termes généraux (ici :* pin *ou* sapin*) sont des «* hubs *» dans le réseau, sources ou cibles d'un grand nombre de relations. Le graphe présenté ici ne tient pas compte de l'orientation des relations*

Nous avons comparé les résultats obtenus à ce jour grâce à JeuxDeMots (JDM) avec les données d'Euro WordNet Français (EWF). JDM possède environ 7 fois plus de termes qu'EWF qui en comporte environ 23 000. EWF compte actuellement un peu plus de 100 000 relations ; JDM en possède déjà 180 000 et ce nombre est en progression quasi constante d'environ 10 000 relations supplémentaires par mois.

Dans EWF, tous les termes sont reliés avec une moyenne de 4 relations par terme. Dans JDM, approximativement 18 000 termes ont des relations sortantes et 30 000 termes possèdent au moins une relation entrante ; de très nombreux termes ne sont reliés à aucun autre, ce qui montre que notre réseau est loin d'être « complet ». Le réseau lexical actuellement obtenu possède une volumineuse composante connexe regroupant près de 30 000 termes, plusieurs centaines de composantes connexes ne comportant que quelques termes, ainsi que plus de 100 000 termes isolés.

En ce qui concerne la détermination des usages d'un terme, plusieurs figures, ainsi que les annexes ci-après, montrent des exemples de cliques détectées dans le

réseau lexical, avec leurs poids respectifs (poids total de la clique et moyenne des poids de ses relations) et leurs pertinences.

Nous avons effectué des mesures sur les 1 000 termes les plus fréquemment proposés par les joueurs. Nous avons divisé les termes en quatre quartiles en fonction de leur fréquence d'utilisation dans le jeu par les joueurs. Nous avons obtenu les résultats reproduits sur le tableau de la figure 12 (extraits à fin novembre 2008) :

Quartile	**NB termes**	**NB moy cliques**	**< 100**	**100-199**	**200-299**	**> 300**
Q1	70	13,04	4,03	4,21	3,06	1,74
Q2	167	7,52	3,19	1,21	1,58	0,75
Q3	298	5,22	2,35	1,33	1,08	0,58
Q4	465	2,76	1,68	0,66	0,36	0,29

Figure 12. *Les données de ce tableau correspondent aux 1 000 termes les plus fréquemment rencontrés dans JDM. Leurs fréquences d'utilisation ont été divisées en quatre quartiles sur la fréquence, Q1 correspondant aux termes dont la fréquence d'utilisation est la plus élevée. Pour chaque quartile, nous avons indiqué le nombre de termes, le nombre moyen de cliques (donc d'usages) de chaque terme, ainsi que la répartition des pertinences de ces cliques*

On observe donc qu'en moyenne plus un terme est utilisé fréquemment, plus il a d'usages différents. Ce n'est certes pas une découverte lexicale, mais que cela soit observé dans les données issues de JeuxDeMots est un indice positif sur la validité linguistique de celles-ci. Ce résultat s'explique aussi par le taux plus faible de parties jouées sur les termes rares que pour les termes fréquents. Les cliques dont la pertinence est très faible (< 100) relèvent de sens peu souvent indiqués par les joueurs, voire d'hapax. Seules les cliques raisonnablement formées (> 200) peuvent être considérées comme correspondant à des usages pertinents.

Nous avons également constaté que les usages peu fréquents (ex : *bois instrument de musique*) n'émergent pas, ou n'émergent que tardivement lorsque les usages plus courants sont devenus tabous. Ceci peut être considéré comme un défaut de JDM ; les approches fondées sur des méthodes manuelles, ou de type dictionnairique, semblent dans ce cas plus performantes que JDM.

6. Conclusion

JeuxDeMots est un jeu en ligne sur le Web dont l'objectif est la construction d'un réseau lexical. L'émergence de relations typées et pondérées entre termes s'effectue grâce au concours d'un grand nombre d'utilisateurs dont l'activité a pour effet de bord la construction de ce réseau. Ces utilisateurs ne sont certes pas des linguistes, mais nous pensons que leur nombre permettra d'obtenir un réseau évolutif de bonne qualité, avec une couverture satisfaisante de l'ensemble des connaissances générales. Notre but n'est pas la constitution d'une base d'experts, mais d'une base de connaissances « moyennes », représentant une culture générale commune quelque peu biaisée, nous le reconnaissons, par le fait que JDM soit un jeu en ligne impliquant donc que les utilisateurs ont un profil d'internaute.

De plus, au vu des résultats actuels, bien que récents et nécessairement partiels, nous pensons arriver à séparer les différents sens d'usage parmi ceux représentés pour chaque terme du réseau. Ce dernier travail n'en est qu'à ses débuts : il serait possible de considérer en plus de la relation « *idée associée* » d'autres relations symétriques dans la détermination des cliques[6]. De même, pourquoi ne pas aussi considérer les quasi-cliques (sous-graphes induits presque complets) dans la détermination des usages d'un terme. Pour un même terme, deux cliques qui ont une forte proportion de termes en commun correspondent-elles réellement à deux usages distincts de ce terme ? Cette question reste actuellement ouverte : il est possible que, malgré l'évolution de notre réseau, des cliques actuellement séparées ne fusionneront pas, alors qu'on peut manifestement les considérer comme d'un même usage. Par exemple, pour le terme *ciel*, parmi les cliques existantes on trouve *ciel-nuage-pluie-gris* et *ciel-nuage-pluie-soleil* qui ne fusionneront probablement pas, car la relation *soleil-gris* a peu de chances d'émerger. Ce biais est-il induit par notre méthodologie ou est-il dû à la représentation en réseau lexical ? Affaire à suivre…

7. Bibliographie

von Ahn L. et Dabbish L., « Labelling Images with a Computer Game », *ACM Conference on Human Factors in Computing Systems (CHI)*, p. 319-326, 2004.

Collins A. et Quillian M.R., « Retrieval time from semantic memory », *Journal of verbal learning and verbal behaviour* 8 (2), p. 240-248, 1969.

Dumais S.T., « Latent Semantic Indexing (LSI) and TREC-2 », *The Second Text REtrieval Conference*, National Institute of Standards and Technology Special Publication, vol 500, n°215, p. 105-116, 1994.

6 Nous avons récemment implémenté une version de notre logiciel dans laquelle nous considérons les relations de tout type pour déterminer les cliques, mais nous n'avons pu encore faire d'évaluation à ce sujet.

Fairon C. et Ho N.D., « Quantité d'information échangée : une nouvelle mesure de la similarité des mots », *Journées internationales d'Analyse statistiques des Données Textuelles (JADT)*, Louvain-la-Neuve (Belgique), 2004.

Ferret O., « Using Collocations for Topic Segmentation and Link Detection », *Proc. of the Coling Conference on Computational Linguistics*, Taipei, p. 261-266, 2002.

Gaume B., « Cartographier la forme du sens dans les petits mondes lexicaux », *Journées internationales d'Analyse statistiques des Données Textuelles (JADT)*, Besançon, France, p. 451-465, 2006.

Gaume B., Duvignau K. et Vanhove M., « Semantic associations and confluences in paradigmatic networks », in : *Typologie des rapprochements sémantiques*, M. Vanhove éd., 2007.

Landauer T.K. et Dumais S.T., « A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge », *Psychological Review*, 104, p. 211-240, 1997.

Lapata M. et Keller F., « Web-based Models for Natural Language Processing », *ACM Transactions on Speech and Language Processing*, vol.2, n°1, p. 1-30, 2005.

Lieberman H., Smith D.A. and Teeters A., « Common Consensus: a web-based game for collecting commonsense goals », *International Conference on Intelligent User Interfaces (IUI'07)*, Hawaii, USA, 2007.

Manning C.D. et Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, 1999.

Mel'čuk I.A., *Dictionnaire explicatif et combinatoire du français contemporain*, Les Presses de l'Université de Montréal, 1988.

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

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

Pechoin D., *Thésaurus: Des idées aux mots, des mots aux idées*, Larousse, Paris, 1991.

Polguère A., *Lexicologie et Sémantique lexicale*, Les Presses de l'Université de Montréal, 2003.

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

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

Roget P.M., *Thesaurus of English words and phrases*, Longman, London, 1852.

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

Schwab D. et Lafourcade M., « Modelling, Detection and Exploitation of Lexical Functions for Analysis », *ECTI Journal*, 2007, vol.2, ISSN 1905-050X, p. 97-108.

Smadja F., « Retrieving collocations from text: Xtract », *Computational Linguistics*, 19, p. 143-177, 1993.

Spence D.P. et Owens K.C., « Lexical co-occurrence and association strength », *Journal of Psycholinguistic Research*, 19 (5), 1990.

Sowa J., *Semantic networks*, Encyclopedia of Artificial Intelligence, edited by S.C. Shapiro, Wiley, New York, 1992.

Tversky A., *Features of similarity*, Psychological Review, 84, p. 327-352, 1977.

Véronis J., « Sense tagging: does it make sense ? », *Corpus linguistics' 2001 Conference*, Lancaster, U.K, 2001.

Wandmacher T., Ovchinnikova E. and Alexandrov T., « Does Latent Semantic Analysis reflect Human Associations ? », *Proc. of the Lexical Semantics workshop at ESSLLI'08*, Hamburg, Germany, 2008.

Wettler M. et Rapp R., « Computation of Word associations based on the co-occurrences of words in large corpora », *Proc. of the 1st Workshop on Very Large Corpora*, Academic and Industrial Perspectives, p. 84-93, 1992.

Zock M. et Bilac S., « Word lookup on the basis of associations: from an idea to a roadmap », *Proc. of Coling workshop: Enhancing and using dictionaries*, Geneva, p. 29-34, 2004.

Zock M. et Quint J., « Why have them work for peanuts, when it is so easy to provide reward ? Motivations for converting a dictionary into a drill tutor », Papillon-2004, 5th workshop on Multilingual Lexical Databases, Grenoble, 2004.

Zock M. et Schwab D., « Lexical access based on underspecified input », COGALEX workshop, Coling, Manchester, p. 9-17, 2008.

Annexes

Liste des types de relations rencontrés dans JeuxDeMots, avec leur signification, et quelques exemples caractéristiques.

Types de relations	Significations	Exemples (cible → valeur)
Idée associée	Association libre - relation associative	*chat → chien*
Synonyme	Terme ayant un sens identique ou proche - relation lexicale	*chat → matou* *chien → cabot*
Contraire	Terme ayant un sens contraire - relation lexicale	*chaud → froid* *haut → bas*
Spécifique	Terme associé à un spécifique de la cible – relation ontologique	*chat → chat persan*
Générique	Terme associé à un spécifique de la cible – relation ontologique	*chat →félin* *chat → animal*
Chose > domaine	Domaine auquel peut appartenir la cible – relation ontologique	*chat → zoologie* *tennis → sport*
Domaine > chose	Terme typique pouvant appartenir au domaine – relation ontologique	*zoologie → taxonomie*
Partie	Terme désignant des parties de la cible – relation ontologique	*chat → queue*

Tout	Terme désignant un tout dont fait partie la cible – relation ontologique	*moustache → chat* *racine → arbre*
Chose > lieu	Lieu typique où peut se trouver la cible – relation ontologique	*chat → maison*
Lieu > chose	Objet typique que l'on peut trouver dans ce lieu – relation ontologique	*canapé → chat*
Chose > caractéristique	Caractéristique typique associée à la cible – relation ontologique	*chat → agile*
Caractéristique > chose	Terme typique associé à la caractéristique – relation ontologique	*brûlant → feu* *rapide → TGV*
Locution	Locution contenant le mot cible – relation lexicale	*chat → chat perché*
Famille	Terme de la même famille – relation lexicale	*chat → chatière*
Magn	Terme typique désignant l'intensification de la cible	*chat → tigre* *averse → déluge*
Antimagn	Terme typique désignant l'amoindrissement de la cible	*chat → chaton* *forêt → bois*
Action > manière	Manière typique liée à l'action – relation prédicative	*miauler → bruyamment*
Action > agent	Agent typique pouvant réaliser l'action – relation prédicative	*vacciner → médecin*
Agent > action	*Inverse de la précédente*	*médecin → soigner*
Action > patient	Patient typique subissant l'action – relation prédicative	*vacciner → enfant*
Patient > action	*Inverse de la précédente*	*enfant → gronder*
Action > instrument	Instrument typique permettant de réaliser l'action – relation prédicative	*vacciner → seringue*
Instrument > action	*Inverse de la précédente*	*pistolet → tirer*
Sentiment	Sentiment associé à la cible -relation associative	*chat → affection*
Sens/signification	Terme typique pouvant désigner un sens possible de la cible - relation associative	*bois → matière* *bois → lieu* *bois → cornes*

Ci-après, les différents usages actuellement décelés dans notre réseau lexical pour le terme *ciel*. Il paraît peu probable que les cliques *1 : ciel-nuage-pluie-gris* et *2 : ciel-nuage-soleil-pluie* puissent fusionner, alors que les cliques *5 : ciel-étoile-*

lune-espace et *8 : ciel-lune-espace-astronomie* fusionneront probablement prochainement, dès que la relation *étoile-astronomie* sera créée.

0: 'ciel' 'bleu'	(P = 1220 / nl = 2 / moy = 610 / REL = 423)
1: 'ciel' 'nuage' 'pluie' 'gris'	(P = 2300 / nl = 12 / moy = 192 / REL = 266)
2: 'ciel' 'nuage' 'soleil' 'pluie'	(P = 2660 / nl = 12 / moy = 222 / REL = 307)
3: 'ciel' 'avion' 'air'	(P = 530 / nl = 6 / moy = 88 / REL = 97)
4: 'ciel' 'soleil' 'étoile' 'lune'	(P = 1890 / nl = 12 / moy = 158 / REL = 218)
5: 'ciel' 'étoile' 'lune' 'espace'	(P = 1360 / nl = 12 / moy = 113 / REL = 157)
6: 'ciel' 'soleil' 'pluie' 'arc-en-ciel'	(P = 2010 / nl = 12 / moy = 168 / REL = 232)
7: 'ciel' 'soleil' 'lune' 'astronomie'	(P = 1360 / nl = 12 / moy = 113 / REL = 157)
8: 'ciel' 'lune' 'espace' 'astronomie'	(P = 970 / nl = 12 / moy = 81 / REL = 112)
9: 'ciel' 'oiseau' 'air'	(P = 470 / nl = 6 / moy = 78 / REL = 86)
10: 'ciel' 'pluie' 'nuages' 'gris'	(P = 1050 / nl = 12 / moy = 88 / REL = 121)
11: 'ciel' 'étoiles' 'espace' 'astronomie'	(P = 960 / nl = 12 / moy = 80 / REL = 111)
12: 'ciel' 'espace' 'astronomie' 'univers'	(P = 920 / nl = 12 / moy = 77 / REL = 106)
13: 'ciel' 'gris' 'noir'	(P = 530 / nl = 6 / moy = 88 / REL = 97)
14: 'ciel' 'paradis'	(P = 120 / nl = 2 / moy = 60 / REL = 42)

Ci-dessous, les huit usages actuellement décelés dans notre réseau lexical pour le terme *pièce*. Ces usages correspondent aux quatre sens suivants :

- *pièce d'un logement* : cliques 0 et 2 ;
- *pièce de monnaie* : cliques 1 et 4 ;
- *pièce de théâtre* : cliques 5, 6 et 7 ;
- *partie d'un tout* : clique 3.

La distinction entre les cliques 6 et 7 provient d'une faute d'orthographe relativement courante : le mot *théâtre* étant écrit sans l'accent circonflexe sur la lettre *a*. Nous remarquons également que notre réseau est incomplet : les sens pièce d'un jeu d'échecs ou pièce de tissu n'y figurent pas encore.

0 : 'pièce' 'chambre' 'maison' 'appartement'	(P = 1450 / nl = 12 / moy = 121 / REL = 168)
1 : 'pièce' 'monnaie' 'argent' 'euro' 'billet'	(P = 3630 / nl = 20 / moy = 182 / REL = 292)
2 : 'pièce' 'maison' 'salon' 'appartement'	(P = 1350 / nl = 12 / moy = 113 / REL = 156)
3 : 'pièce' 'morceau' 'partie'	(P = 510 / nl = 6 / moy = 85 / REL = 93)
4 : 'pièce' 'monnaie' 'argent' 'or'	(P = 2510 / nl = 12 / moy = 209 / REL = 290)
5 : 'pièce' 'théâtre' 'salle'	(P = 730 / nl = 6 / moy = 122 / REL = 134)
6 : 'pièce' 'théâtre' 'acteur'	(P = 1020 / nl = 6 / moy = 170 / REL = 187)
7 : 'pièce' 'acteur' 'théatre'	(P = 330 / nl = 6 / moy = 55 / REL = 60)