Sens et usages d'un terme dans un réseau lexical évolutif

Mathieu Lafourcade, Alain Joubert, Stéphane Riou

LIRMM – Université Montpellier 2 - CNRS
Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier
161, rue Ada
F-34392 Montpellier cedex 5
{lafourcade, joubert}@lirmm.fr

Résumé L'obtention d'informations lexicales fiables est un enjeu primordial en TALN, mais cette collecte peut s'avérer difficile. L'approche présentée ici vise à pallier les écueils de cette difficulté en faisant participer un grand nombre de personnes à un projet contributif via des jeux accessibles sur le web. Ainsi, les joueurs vont construire le réseau lexical, en fournissant de plusieurs manières possibles des associations de termes à partir d'un terme cible et d'une consigne correspondant à une relation typée. Le réseau lexical ainsi produit est de grande taille et comporte une trentaine de types de relations. A partir de cette ressource, nous abordons la question de la détermination des différents sens et usages d'un terme. Ceci est réalisé en analysant les relations entre ce terme et ses voisins immédiats dans le réseau et en calculant des cliques ou des quasi-cliques. Ceci nous amène naturellement à introduire la notion de similarité entre cliques, que nous interprétons comme une mesure de similarité entre ces différents sens et usages. Nous pouvons ainsi construire pour un terme son arbre des usages, qui est une structure de données exploitable en désambiguïsation de sens. Nous présentons quelques résultats obtenus en soulignant leur caractère évolutif.

Abstract Obtaining reliable lexical information is an essential task in NLP, but it can prove a difficult task. The approach we present here aims at lessening the difficulty: it consists in having people take part in a collective project by offering them playful applications accessible on the web. The players themselves thus build the lexical network, by supplying (in various possible ways) associations between terms from a target term and an instruction concerning a typed relation. The lexical network thus obtained is large and it includes about thirty types of relations. From this network, we then discuss the question of meaning and word usage determination for a term, by searching relations between this term and its neighbours in the network and by computing cliques or quasi-cliques. This leads us to introduce the notion of similarity between cliques, which can be interpreted as a measure of similarity between these different meanings and word usages. We are thus able to build the tree of word usages for a term: it is a data structure that can be used to disambiguate meaning. Finally, we briefly present some of the results obtained, putting the emphasis on their evolutionary aspect.

Mots-clés : Traitement Automatique du Langage Naturel, réseau lexical évolutif, relations typées pondérées, similarité entre sens et usages, arbre des usages.

Keywords: Natural Language Processing, evolutionary lexical network, typed and weighted relations, meaning and word usage similarity, tree of word usages.

1. Introduction

L'obtention de relations lexicales ou fonctionnelles fiables entre termes est un enjeu primordial en Traitement Automatique des Langues (TAL). En effet, de telles relations sont nécessaires dans un grand nombre d'applications. 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, l'un des plus anciens thésaurus, le (Roget, 1852) ou l'un des plus célèbres réseaux lexicaux, WordNet (Miller et al., 1990) auquel (Boyd-Graber et al., 2006) ont récemment ajouté des liens associatifs pondérés. Il est également possible de déterminer automatiquement de telles relations 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. La méthode Latent Semantic Analysis (Dumais, 1994) ou (Landauer et Dumais, 1997) s'appuie également sur des ensembles de textes ; elle permet de calculer une proximité sémantique entre mots et ainsi produire des nuages de termes appartenant à un même champ sémantique. 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. Contrairement à ces méthodes classiques qui permettent d'acquérir des informations lexicales généralement statiques, le prototype introduit ici, où ce sont les utilisateurs qui font évoluer la base, réalise l'acquisition d'informations lexicales évolutives.

Dans cet article, nous rappelons brièvement les principes de deux jeux visant à construire une base de relations entre termes. Le premier de ces deux jeux (JeuxDeMots[1]) permet la construction d'un réseau lexical : c'est un jeu ouvert grâce auquel les utilisateurs fournissent des associations entre termes, en proposant des termes destination à partir d'un terme origine et d'une consigne correspondant à une relation typée. Le deuxième jeu (PtiClic[1]) permet aux utilisateurs de consolider les associations acquises grâce à JeuxDeMots en leur proposant de valider, ou non, des associations potentielles entre termes : c'est un jeu fermé, les utilisateurs ne proposant pas de terme. L'objectif poursuivi concerne avant tout la fiabilité et la qualité des informations recueillies auprès des utilisateurs, l'un 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. La similarité entre les différents usages d'un même terme est abordée dans une troisième partie ; elle nous permet de construire l'arbre des usages d'un terme constituant un raffinement dans la différentiation de ses usages. Enfin, nous présentons quelques résultats, en soulignant le caractère évolutif du réseau lexical obtenu.

2. Réseau lexical : principe et construction

2.1. Principe du logiciel initial : JeuxDeMots

Le principe de base du logiciel, le déroulement d'une partie, ainsi que la construction progressive du réseau lexical, ont déjà été décrit par (Lafourcade et Joubert, 2008). Une partie se déroule entre deux joueurs, en asynchrone. Lorsqu'un premier joueur (A) débute une partie, un terme[2] T tiré aléatoirement dans une base de mots est affiché, ainsi qu'une consigne concernant un type de compétence (synonymes, domaines, association libre…). Ce joueur (A) a alors un temps limité pour répondre en proposant un nombre limité de termes correspondant, selon lui, à la consigne appliquée au terme T. Ce même terme, avec cette même consigne, est proposé par la suite à un joueur (B) ; le processus est identique. Pour ce terme cible T, nous mémorisons les réponses communes aux joueurs (A) et (B). Les validations sont donc faites par concordance des propositions entre paires de joueurs. Ce

[1] JeuxDeMots et PtiClic sont accessibles à l'adresse http://jeuxdemots.org et http://pticlic.org. Il existe également 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

[2] Un terme peut être constitué de plusieurs mots (par exemple : *pomme de terre*).

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″. A notre connaissance, il n'a jamais été mis en œuvre dans le domaine des réseaux lexicaux.

Les propositions faites par les joueurs (A) et (B) vont permettre d'établir des relations entre termes. Pour une même partie, notons :

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

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

Pour tous les couples (i,j) tels que $x_i = y_j$, la relation $R : T \rightarrow x_i$ est validée.

Le logiciel développé 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. La structure du réseau lexical que nous cherchons ainsi à obtenir s'appuie sur les notions de nœuds et de relations entre nœuds, selon un modèle initialement présenté vers la fin des années 1960 par (Collins et Quillian, 1969) et plus récemment explicité par (Polguère, 2006). Cette structure s'apparente à celle du réseau obtenu pour la relation de synonymie par (Ploux et Victorri, 1998) pour la constitution du Dictionnaire électronique des synonymes. 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 lexicales, telles que présentées par (Mel'čuk et al., 1995). La figure 1 présente une partie des relations acquises pour le terme *ciel*.

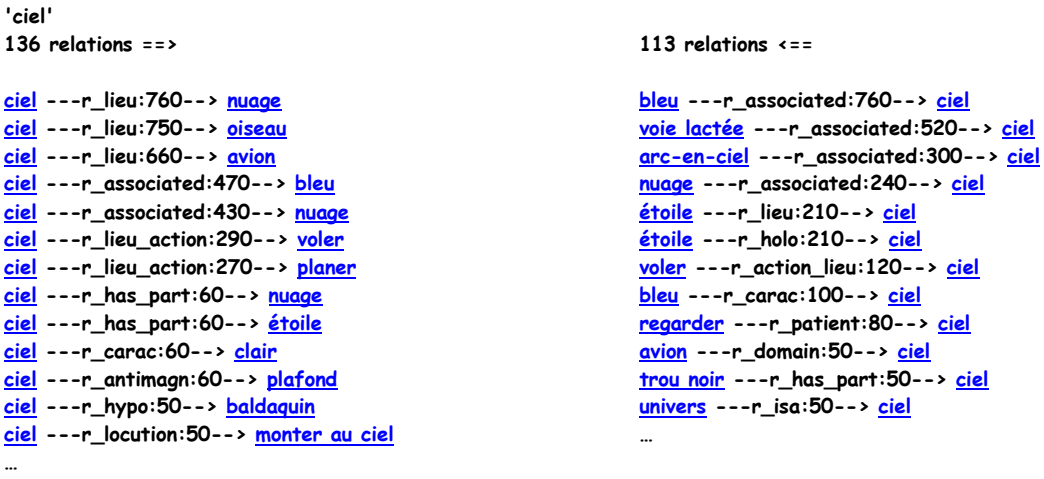


Figure 1 : Ensemble (partiel) des relations acquises pour le terme ciel. *Sont présentées tout d'abord les relations dont le terme* ciel *est origine, puis celles pour lesquelles le terme* ciel *est destinataire. Pour chacune de ces relations, on a en outre son type ("lieu caractéristique", "idée associée"…) ainsi que son poids. Les relations sont ici présentées par ordre de poids décroissant.*[3]

2.2. Consolidation du réseau lexical : logiciel PtiClic

Tout comme pour JeuxDeMots (JDM), une partie de PtiClic se déroule en asynchrone entre deux joueurs. Un premier joueur (A) se voit proposer un terme cible T, ainsi qu'un nuage de mots provenant de l'ensemble des termes reliés à T dans le réseau lexical produit par JDM. Plusieurs consignes correspondant à des types de relations sont également affichées. Le joueur associe, par cliquer-glisser, des mots du nuage aux consignes auxquelles il pense qu'ils correspondent. Ce même terme T, ainsi que le même nuage de mots et les mêmes consignes, sont également proposées à un joueur (B). Selon un principe analogue à celui mis en place

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

pour JDM, les propositions de (B) sont comparées à celles faites par (A) et seules leurs propositions communes sont prises en compte, renforçant ainsi les relations du réseau lexical.

Contrairement à JDM, PtiClic est un jeu fermé où les utilisateurs ne peuvent pas proposer de nouveaux termes, mais sont contraints de choisir parmi ceux affichés. Ce choix de conception doit permettre de réduire le bruit dû aux termes mal orthographiés ou aux confusions de sens. PtiClic réalise donc une consolidation des relations produites par JDM et permet de densifier le réseau lexical.

Afin de réduire le silence correspondant aux termes non proposés par les utilisateurs de JDM, (Zampa et Lafourcade, 2009) ont suggéré de générer le nuage de mots à l'aide de LSA. Cette solution permet d'augmenter le réseau lexical par l'ajout de nouvelles relations, en particulier de relations non symétriques dans JDM (exemple : *ciel* → *aigle*, mais la relation inverse n'existant pas actuellement, aucune clique du terme *ciel* ne comporte le terme *aigle*).

3. Détermination des sens d'usage

3.1. Principe général

Lorsqu'un terme T est polysémique, les termes qui lui sont directement 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 ne citer qu'un exemple, *ciel-nuage-pluie-gris* et *ciel-nuage-soleil-pluie* constituent deux usages distincts du terme *ciel*, alors qu'il s'agit manifestement du même sens de ce terme.

3.2. Détermination des cliques et calcul de leur pertinence

Comment déterminer une clique ? C'est un ensemble de termes « fortement » reliés entre eux constituant un sous-graphe induit complet (ou clique) dans le réseau lexical. Les liens qui sont considérés ici sont uniquement des relations de type "*idée associée*" symétrique. Nous faisons abstraction des autres types de relation ("*tout de*", "*partie de*", …) en raison de leur caractère non symétrique.

Dans le réseau lexical, un terme T est directement relié à n termes : $T_1,\ldots,T_i,\ldots,T_j,\ldots,T_n$. Nous nous attachons ici aux différents usages de ce terme T ; le but est donc de regrouper ces n termes en un ou plusieurs groupes, chacun constituant un usage de T. Les termes $T_{i1}, \ldots, T_{im}$, reliés à T, constituent le $i^{ème}$ usage de T si les (m+1)*m/2 relations entre ces (m+1) termes existent. Ainsi, un terme T_j , relié à T, n'appartiendra pas à ce $i^{ème}$ usage de T si au moins une des relations entre ce terme T_j et l'un des termes $T_{i1}, \ldots, T_{im}$ n'existe pas.

Evaluer la pertinence d'un usage consiste à obtenir une mesure de son importance à la fois en fréquence d'utilisation mais aussi en 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 en premier en général. 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, la pertinence (Rel) de l'usage correspondant sera égale à :

$$Rel(C) = Ln(m) * P(C) / [m*(m-1)] \qquad \text{où } P(C) = \sum_{i,j} poids(T_i\text{-}T_j)$$

Le terme *poids(T_i-T_j)* est le poids de la relation entre T_i etT_j. 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.

Il est à noter que la pertinence est une valeur relative entre cliques d'un même terme, et non une valeur absolue permettant la comparaison entre cliques de termes différents, car les poids des relations à partir d'un terme cible dépendent du nombre de fois où ce terme a été joué.

La figure 2 présente les différents usages obtenus pour le terme *ciel*, ainsi que la pertinence de chacun de ces usages.

0: 'ciel' 'bleu'	(P = 1230 / nl = 2 / moy = 615 / REL = 426)
1: 'ciel' 'nuage' 'pluie' 'gris'	(P = 2410 / nl = 12 / moy = 201 / REL = 278)
2: 'ciel' 'nuage' 'soleil' 'pluie'	(P = 2720 / nl = 12 / moy = 227 / REL = 314)
3: 'ciel' 'soleil' 'étoile' 'lune'	(P = 1980 / nl = 12 / moy = 165 / REL = 229)
4: 'ciel' 'étoile' 'lune' 'espace'	(P = 1370 / nl = 12 / moy = 114 / REL = 158)
5: 'ciel' 'soleil' 'pluie' 'arc-en-ciel'	(P = 2070 / nl = 12 / moy = 173 / REL = 239)
6: 'ciel' 'soleil' 'lune' 'astronomie'	(P = 1420 / nl = 12 / moy = 118 / REL = 164)
7: 'ciel' 'lune' 'espace' 'astronomie'	(P = 980 / nl = 12 / moy = 82 / REL = 113)
8: 'ciel' 'oiseau' 'air'	(P = 470 / nl = 6 / moy = 78 / REL = 86)
9: 'ciel' 'avion' 'air'	(P = 590 / nl = 6 / moy = 98 / REL = 108)
10: 'ciel' 'avion' 'nuages'	(P = 680 / nl = 6 / moy = 113 / REL = 125)
11: 'ciel' 'pluie' 'nuages' 'gris'	(P = 1100 / nl = 12 / moy = 92 / REL = 127)
12: 'ciel' 'étoiles' 'espace' 'astronomie'	(P = 960 / nl = 12 / moy = 80 / REL = 111)
13: 'ciel' 'espace' 'astronomie' 'univers'	(P = 930 / nl = 12 / moy = 78 / REL = 107)
14: 'ciel' 'gris' 'noir'	(P = 580 / nl = 6 / moy = 97 / REL = 106)
15: 'ciel' 'paradis'	(P = 120 / nl = 2 / moy = 60 / REL = 42)

Figure 2 : Cette figure montre les 16 usages du terme ciel *actuellement décelés.*
Pour chacun de ces usages du terme ciel *figure également une évaluation de sa pertinence (valeur* REL*). Le terme* ciel *est polysémique, les 16 usages ci-dessus correspondent à plusieurs sens distincts des domaines :* météorologie, astronomie, aéronautique *et* religion*.*

4. Similarité entre cliques : arbre des usages

4.1. Définition

Nous venons de définir les usages d'un terme par la notion de cliques dans le réseau lexical. L'exemple ci-dessus montre clairement que, pour un même terme, certaines cliques correspondent à des usages sémantiquement plus proches que d'autres. Il s'agit donc de définir un coefficient, compris entre 0 et 1, qui traduira la proximité sémantique de deux usages. 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). PROX, développé par (Gaume, 2004), établit une distance entre termes à partir d'un graphe de dictionnaire, en calculant des similarités entre ses sommets (les entrées du dictionnaire en question). 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. Regroupement des usages : détermination des sens

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 28 cliques. La figure 3 présente un extrait de cet ensemble de cliques.

```
            …
Clique 4: 'soleil' 'étoile' 'lune' 'planète' 'ciel'          (P = 2650 / nl = 20 / moy = 133 / REL = 213)
Clique 5: 'soleil' 'étoile' 'astre' 'planète' 'galaxie'      (P = 2660 / nl = 20 / moy = 133 / REL = 214)
Clique 6: 'soleil' 'étoile' 'astre' 'lune' 'planète'         (P = 2470 / nl = 20 / moy = 124 / REL = 199)
            …
Clique 9: 'soleil' 'astre' 'astronomie' 'galaxie'            (P = 1120 / nl = 12 / moy = 93 / REL = 129)
            …
Clique11: 'soleil' 'planète' 'galaxie' 'système solaire'     (P = 1190 / nl = 12 / moy = 99 / REL = 137)
            …
Clique 23: 'soleil' 'ciel' 'nuage' 'pluie'                   (P = 2720 / nl = 12 / moy = 227 / REL = 314)
            …
```

Figure 3 : 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 23* => 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 23. Nous avons calculé les similarités entre ces cliques ; en voici quelques valeurs :*

Sim(C4,C5) = 0,339 *Sim(C4,C6) = 0.583* *Sim(C4,C23) = 0,171*
Sim(C5,C6) = 0,664 *Sim(C5,C23) = 0* *Sim(C6,C23) = 0*

Les cliques C4-C5-C6 d'une part, et C23 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 puissent fusionner, et de nouvelles cliques peuvent apparaitre.*

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 4). 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 4 : 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 5 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 à la figure 5, est constitué d'une racine regroupant tous ses usages et de trois branches simples correspondant chacune à un usage.

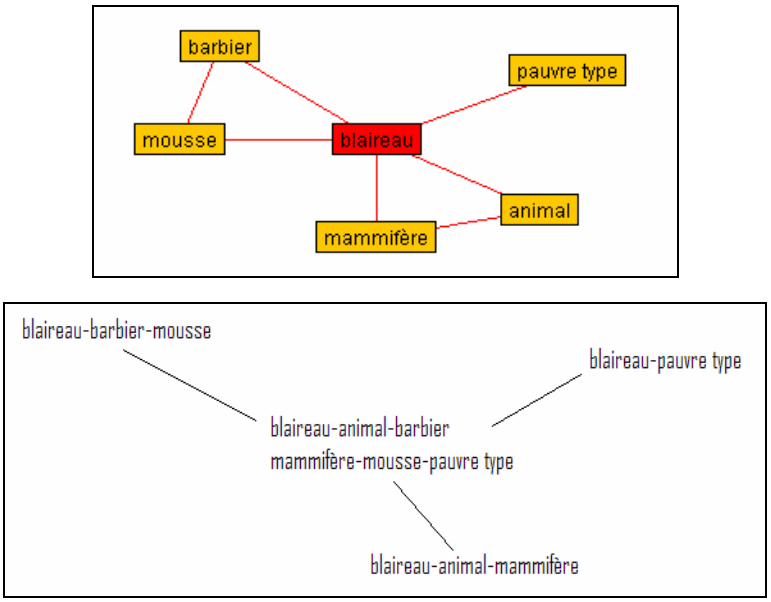
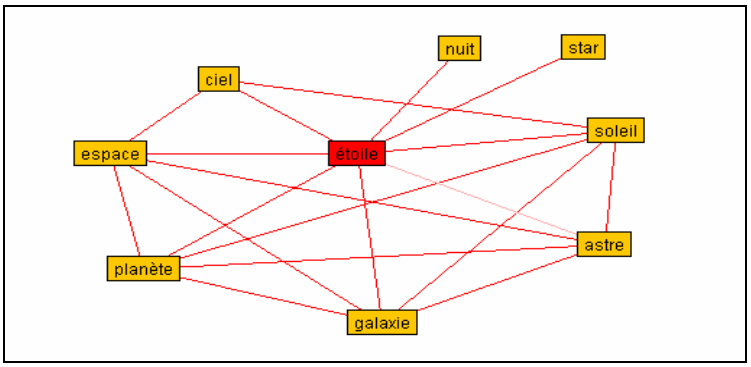


Figure 5 : Réseau lexical réduit aux cliques du terme blaireau *et l'arbre des usages correspondant.*

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 : 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 6 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 7.

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

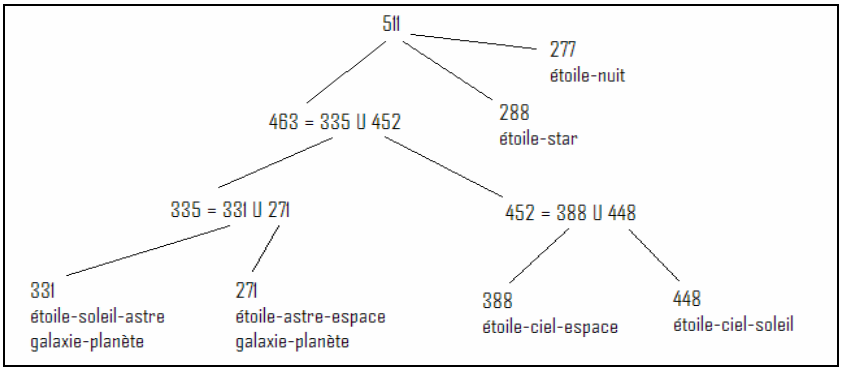


Figure 7 : Arbre des usages du terme étoile*. Les valeurs numériques correspondent aux numéros des différentes cliques ou quasi-cliques. Les termes les constituants n'ont été indiqués qu'aux feuilles de l'arbre, afin de ne pas surcharger la figure.*

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.

5. Résultats et évolution du réseau lexical obtenu

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.

Le lancement de JeuxDeMots a eu lieu en juillet 2007[4]. Depuis environ un an et demi, plus de 1600 joueurs se sont enregistrés et une bonne proportion d'entre eux se connectent plusieurs fois par semaine. Plus de 130.000 parties ont été jouées : elles ont fait émerger plus de 180.000 relations, en progression quasi constante d'environ 10.000 relations supplémentaires par mois. Il y a une émergence rapide des relations et on constate que les

[4] Le lancement de PtiClic est beaucoup trop récent (janvier 2009) pour que son influence sur le réseau lexical ait pu être objectivement mesurée.

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 cible. L'évolution de la base de termes est nécessairement plus lente : elle compte à ce jour environ 165.000 termes ; les joueurs y ont déjà ajouté près de 10.000 nouveaux termes, principalement conjoncturels ou liés à l'actualité.

Dans JeuxDeMots, approximativement 20.000 termes ont des relations sortantes et plus de 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"[5]. Le réseau lexical actuellement obtenu possède une volumineuse composante connexe regroupant près de 40.000 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, les figures précédentes montrent des exemples de cliques détectées dans le réseau lexical, avec leurs poids respectifs (poids total de chaque clique, ainsi que sa pertinence, telle que définie en section 3.2). La figure 8 montre l'évolution des cliques du terme *ciel* sur plusieurs mois, depuis la création du réseau lexical. On y constate l'augmentation du nombre de cliques, ainsi que celle de leurs pertinences.

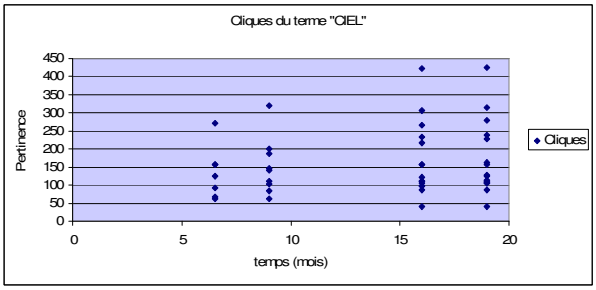


Figure 8 : Evolution des cliques du terme ciel. *Sur ce graphique, l'origine de l'axe du temps est fixée à juillet 2007, mois de lancement du prototype de JeuxDeMots. Le nombre et la pertinence des cliques du terme* ciel *ont été déterminés à quatre dates différentes, entre février 2008 et février 2009.*

6. Conclusion

A partir d'une base de termes préexistante nécessaire pour amorcer le processus, les jeux en ligne JeuxDeMots et PtiClic permettent la construction d'un réseau lexical évolutif. L'émergence de relations typées et pondérées entre termes s'effectue grâce au concours d'un grand nombre d'utilisateurs dont l'activité ludique 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 JeuxDeMots et PtiClic étant des jeux en ligne, les utilisateurs ont donc essentiellement un profil d'internaute.

De plus, au vu des résultats actuels, bien que nécessairement partiels, nous pensons arriver à identifier les différents sens d'usage parmi ceux représentés pour chaque terme du réseau. Ce dernier travail n'en est qu'à ses débuts : pourquoi ne pas considérer également les quasi-cliques dans la détermination des usages d'un terme ? Faut-il descendre jusqu'aux feuilles de l'arbre des usages d'un terme pour en déterminer les usages ? Pour un même terme, deux cliques pour lesquelles le coefficient de similarité est proche de 1 correspondent-elles réellement à deux usages distincts de ce terme ? Ces questions restent actuellement ouvertes :

[5] Compte tenu de l'évolution de notre réseau depuis sa création il y a 18 mois, et en supposant que cette évolution se poursuive au même rythme, notre réseau pourrait être "complet" au bout … d'une dizaine d'années, mais il compterait alors près de 1,5 million de relations.

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 un même usage. Par exemple, pour le terme *ciel*, parmi les cliques existantes on trouve *ciel-nuage-pluie-gris* et *ciel-nuage-soleil-pluie* qui, bien que très proches, 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 ?

Références

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

BOYD-GRABER J., FELLBAUM C., OSHERSON D., SCHAPIRE R. (2006) Adding dense, weighted connections to WordNet, Proceedings of the third Global WordNet Meeting, Jeju Island, Korea

COLLINS A., QUILLIAN M.R. (1969) Retrieval time from semantic memory, *Journal of verbal learning and verbal behaviour*, 8 (2), pp. 240-248

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

FAIRON C., HO N.D. (2004) Quantité d'information échangée : une nouvelle mesure de la similarité des mots, *Journées internationales d'Analyse statistiques des Données Textuelles (JADT'04)*, Louvain-la-Neuve (Belgique)

GAUME B. (2004) Balades aléatoires dans les Petits Mondes Lexicaux, *I3 Information Interaction Intelligence*, vol. 4, n°2, Cepadues Ed.

LAFOURCADE M., JOUBERT A. (2008) Détermination des sens d'usage dans un réseau lexical construit grâce à un jeu en ligne, *Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, Avignon, pp.189-199

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*, 104, pp. 211-240

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

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

MANNING C.D., SCHÜTZE H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge

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

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.

PLOUX S., VICTORRI B. (1998) Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement Automatique des Langues*, 39, n°1, pp. 161-182

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.

ROGET (1852) *Thesaurus of English Words and Phrases*, Longman, London

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

TVERSKY A. (1977) *Features of similarity*, Psychological Review, 84, pp.327-352

VÉRONIS J. (2001) Sense tagging: does it make sense?, *Corpus linguistics' 2001 Conference*, Lancaster, U.K.

ZAMPA V. et LAFOURCADE M. (2009) Evaluations comparées de deux méthodes d'acquisitions lexicale et ontologique : JeuxDeMots vs Latent Semantic Analysis, *XVIèmes rencontres de Rochebrune : ontologie et dynamique des systèmes complexes, perspectives interdisciplinaires.*