Combinaison de ressources linguistiques pour l'aide à l'accès lexical : étude de faisabilité

Laurianne SITBON
Laboratoire d'Informatique d'Avignon - Université d'Avignon
Laboratoire Parole et Langage - Université de Provence
laurianne.sitbon@univ-avignon.fr

Résumé. Cet article propose une évaluation combinée et comparative de 5 ressources (descriptive, paradigmatique et syntagmatiques) pour l'aide à l'accès lexical en situation de "mot sur le bout de la langue", en vue de la création d'un outil utilisant la combinaison de ces ressources. En situation de "mot sur le bout de la langue", l'utilisateur n'accède plus au mot qu'il veut dire ou écrire mais est capable d'en produire d'autres sémantiquement associés. L'évaluation se base sur un corpus de 20 mots "sur le bout de la langue" pour lesquels on dispose de 50 groupes de 5 associations sémantiques effectuées par des utilisateurs. Les résultats montrent que les ressources sont complémentaires et peu redondantes. De plus au moins une association proposée parmi les 5 permettrait de retrouver le mot "sur le bout de la langue" dans 79% des cas, à condition de le sélectionner parmi les 2500 mot potentiels. Enfin, les résultats montrent des disparités entre les utilisateurs, ce qui permettrait de définir des profils d'utilisateur pour une amélioration des performances.

Abstract. This paper describes a joint and comparative evaluation of 5 lexical resources (descriptive, paradigmatic and syntagmatic) from a lexical access angle, with the further perspective of constructing a tool based on a combination of these resources to avoid the "tip of the tongue" (TOT) phenomenon. This phenomenon characterises a person who has difficulty in saying or writing an intended word but one who is able to produce semantically associated words. The evaluation corpus is composed of 20 TOT examples each linked to 50 users' association sets of 5 semantically associated words. The results highlight that all the tested resources are complementary. Moreover, 79% of proposed association sets contain at least one association leading to the TOT through its relative words in at least one resource (in the worse case the TOT has to be found among 2500 words). Finally the results show variations between users which could increase performance thanks to user profiles.

Mots-clés : réseaux sémantiques, accès lexical, profil d'utilisateur.

Keywords: semantic networks, lexical access, user profiling.

1 Introduction

Les outils du traitement automatique du langage ont démontré à plusieurs reprises leur capacité à remédier ou compenser des handicaps de langage. Le problème du *mot sur le bout de la langue*, plus généralement connu comme étant un déficit d'accès lexical en production de phrases (Tip

Of the Tongue en anglais (TOT)), est l'une des manifestation de tels handicaps. Si le phénomène a été étudié dans le cadre général (des personnes sans déficit particulier peuvent en souffrir par moments), les observations diffèrent en situation de handicap. En effet les travaux de (Brown & McNeill, 1966) ont montré que en situation de TOT les personnes connaissent des informations sur le mot recherché, d'ordre phonologique aussi bien que sémantique. Les modèles actuels de la production du langage proposent une vision connexioniste (Dell, 1986), où l'information verbale serait stockée dans trois systèmes interconnectés (le système sémantique, le système phonologique et le système orthographique). D'après les travaux de (Burke & Shafto, 2004) sur le déficit d'accès lexical chez les personnes âgées, c'est essentiellement un déficit d'accès phonologique qui est en cause, ce qui signifie que l'accès sémantique est intact. Dans les modèles connexionistes, l'interconnexion entre les unités sémantiques est beaucoup plus importante que la connexion unique entre un mot et sa phonologie ou un mot et sa graphie (qui sont en réalité des connexions des phonèmes vers les graphèmes, plus une connexion du sens vers la graphie qui porte les exceptions), ce qui explique que lorsque les connexions sont "fragilisées" elles sont maintenues au sein du réseau sémantique mais pas pour l'accès phonologique. Les travaux de (Faust & Sharfstein-Friedman, 2003) montrent un comportement similaire chez des adolescents dyslexiques, qui s'explique dans ce cas par un déficit de la conscience phonologique fréquent chez les personnes souffrant de dyslexie.

Un système qui reproduirait le réseau sémantique de chaque individu (et doté d'un convertisseur phonologique robuste entre les sens et les mots) permettrait ainsi de proposer à l'utilisateur en situation de handicap et face à un mot recherché une liste d'hypothèses, obtenue à partir de mots proches dans son réseau sémantique. Ainsi automatiser l'aide à l'accès lexical nécessiterait une représentation exhaustive et individuelle du lexique mental. Dans cet article nous faisons l'hypothèse que la construction d'une telle ressource peut être approchée à l'aide d'une combinaison de ressources déjà disponibles ou réalisables automatiquement, et qui présentent chacune des aspects différents du cheminement de la pensée pour passer de l'idée au mot. Les ressources lexicales auxquelles nous nous sommes intéressé ont déjà fait indépendamment l'objet d'études dans le cadre de l'aide à l'accès lexical (Reuer, 2004). Cependant tous les individus étant différents nous pensons que les ressources lexicales peuvent être utiles à des individus différents à différents degrés. On peut aussi penser que les voies empruntées par le cerveau pour accéder à un mot dépendent de caractéristiques linguistiques ou sémantiques de ce mot. Encore une fois, la ressource la plus appropriée dans un cas précis pourrait dépendre de la nature sémantique ou syntaxique du mot recherché.

L'idée à long terme est d'implémenter ce système avec des associations audio, ce qui permettrait une utilisation en situation quotidienne. Les bonnes performances des systèmes de reconnaissance automatique de la parole nous laissent penser que la prise en compte de peu d'hypothèses de reconnaissance serait suffisante pour converger vers le sens recherché. L'aide à la formulation ou l'enrichissement semi-automatique de requêtes en recherche d'information est également un domaine d'application qui pourra bénéficier de ce système.

La conception d'un tel système soulève plusieurs questions auxquelles nous tenterons de répondre dans cet article. Dans un premier temps le choix des systèmes utilisés est justifié dans la première section. Il doit permettre un recouvrement maximal des types de relations possibles (paradigmatiques, syntagmatiques ou descriptives) ainsi que des domaines sémantiques (journalistique, littéraire, générique). Ensuite l'intérêt d'une telle combinaison doit être attesté à travers une évaluation combinée des systèmes unitaires sélectionnés. Pour cela la seconde section présente l'élaboration d'un corpus d'évaluation avec l'aide de 50 participants qui ont produit des associations sémantiques sur des simulations de TOT sur 20 mots.

2 Les ressources linguistiques testées

Les approches existantes dans le domaine du "mot sur le bout de la langue" ont souvent évoqué l'usage de ressources de natures différentes. (Lortal *et al.*, 2004) proposent d'étendre les classes sémantiques du réseau de relation sémantico-pragmatiques SVELTAN à l'aide d'EuroWordNet[1] et montrent l'intérêt de les utiliser ensemble pour une tâche consistant à retrouver 10 mots supprimés dans des documents. (Zock, 2002) évoque trois accès aux TOT *via* des ressources différentes sans les évaluer : un par la forme du mot, graphique ou phonologique, un autre par son sens en termes de collocations dans le domaine sémantique auquel il appartient, et un accès par la fonction grammaticale. (Reuer, 2004) argumente en faveur de quatre types de ressources linguistiques différentes : des réseaux sémantiques construits manuellement , des réseaux paradigmatiques et syntagmatiques construits automatiquement, et des réseaux phonologiques et morphologiques.

Etant donné que l'accès phonologique est justement la cause du TOT dans les situations de handicap, il n'est pas pertinent dans notre cas d'utiliser des ressources pour l'accès phonologique, et nous nous sommes donc concentré sur la représentativité des 3 premiers types de ressources. Les relations paradigmatiques sont des liens sémantiques de type synonymiques ou hiérarchiques qui peuvent dépendre du contexte. Les relations syntagmatiques sont des associations mnésiques, issues d'un contexte stocké dans la mémoire à long terme. Elles peuvent être en parties extraites à l'aide de cooccurrences. Nous avons ajouté à ces ressources des relations descriptives issues d'un dictionnaire.

2.1 Les relations descriptives

La manière la plus naturelle d'expliquer à une personne ou à un système le mot que l'on recherche est de le décrire avec d'autres mots, même si il s'avère que cette description est parfois impossible car elle implique le même amorçage lexical que celui du mot recherché. Les dictionnaires constituent de bons référentiels en ce qui concerne la description des mots et disposent généralement d'une bonne couverture. Nous avons utilisé un dictionnaire en ligne (`http://fr.answers.com` propose gratuitement un dictionnaire du français) pour nos expériences. Tous les mots d'une définition n'étant pas nécessairement reliés sémantiquement au mot défini (on rencontre beaucoup de mots outils), des mots clés sont extraits en fonction de leur fréquence relative (*tf/idf*) dans un corpus généraliste (extraits du journal *Le Monde*). Les relations sémantiques entre les mots induites dynamiquement par cette technique ne sont pas des relations symétriques (par exemple le mot *pied* peut être dans la définition de *chaussette* mais l'inverse est peu probable).

2.2 Les relations paradigmatiques

EuroWordNet est une base de données multilingue construite sur le modèle de Wordnet [2]. A chaque mot est associé un certain nombre de sens, qui sont hiérarchisés selon des relations paradigmatiques : synonymie, méronymie, hyponymie. La base de données EuroWordNet française dont nous disposons contient ces informations pour les noms, les verbes et les adjectifs.

[1] `http://www.illc.uva.nl/EuroWordNet/`
[2] `http://wordnet.princeton.edu/`

Le principal inconvénient de cette ressource pour l'accès lexical est que les relations ne sont pas quantifiables, on ne peut pas leur associer de score. Le niveau dans la hiérarchie d'un méronyme ou d'un hyponyme n'est pas non plus utilisable car selon les mots, le niveau de détails hiérarchique peut être très variable. Par exemple une *pomme* sera un *fruit* alors qu'un *citron* sera un *agrume* qui est aussi un *fruit*. Pourtant la relation *pomme/fruit* est aussi forte que la relation *citron/fruit*.

2.3 Les relations syntagmatiques apprises automatiquement

Les relations syntagmatiques sont les moins délimitées, et il est impensable d'assurer leur représentativité. Cependant nous nous sommes concentrés sur trois types de corpus de natures différentes : des articles de presse, des romans, et le web. Des méthodes d'extraction de relations différentes sont appliquées sur chacun des corpus, mais elles sont toutes basées sur l'idée de cooccurrences entre les mots associés et le mot recherché.

Un premier réseau de relations syntagmatiques appris automatiquement est une carte sémantique apprise à l'aide de l'outil Infomap[3] sur le corpus littéraire corpatext[4]. Infomap crée des cartes sémantiques en se fondant sur le principe de LSA (Latent Semantic Analysis) (Deerwester *et al.*, 1990), c'est à dire qu'il effectue une réduction de l'espace lexical composé des mots du corpus afin de les regrouper en classes sémantiques, en se basant sur les cooccurrences des mots. (Landauer *et al.*, 1998) a montré la représentativité du lexique mental à travers LSA, qui est utilisé pour cette propriété par des psycholinguistes. Cependant cette approche ne permet pas d'explorer tous les sens d'un mot, au contraire elle le classe dans sa classe sémantique la plus représentée. La création de *contextonymes* proposée par (Ji *et al.*, 2003) permet de distinguer les différents sens en créant des classes de termes associés, mais nous ne les avons pas expérimentés ici.

Un second réseau de relations syntagmatiques est un réseau de cooccurrences. Il est utilisé par (Ferret & Zock, 2006) pour des travaux similaires. Il est construit selon la méthode proposée par (Church & Hanks, 1990) qui calcule l'information mutuelle entre les termes d'une fenêtre glissante de 20 mots sur le corpus journalistique (extraits du journal *Le Monde*). On obtient ainsi des scores de cohésion entre les paires de mots cooccurrents du corpus. Cependant, comme pour les relations paradigmatiques apprises par LSA, les cooccurrences ne résolvent généralement pas les ambiguïtés de sens lorsqu'un sens est plus fréquent qu'un autre.

La troisième ressource créée n'est pas un réseau de relations mais plutôt un générateur dynamique de relations syntagmatiques, selon le même principe que la génération de relations descriptives. Elle permet d'obtenir les mots clés des 10 premières réponses du moteur de recherche Google [5] à la requête correspondant au mot que l'on cherche à mettre en relation. On considère que les mots de plus grande fréquence relative (*tf/idf*) dans les 10 sites les plus pertinents selon Google sont des cooccurents importants et sont relié au mot par une relation syntagmatique.

(Joyce, 2005) propose la construction d'un réseau de relations syntagmatiques à partir d'un réseau d'associations qui reposerait sur l'expérience personnelle des individus, en demandant à beaucoup de personnes de participer à la création de ce réseau. Il s'agit d'une partie d'un programme de création de ressources à très grande échelle qui est pratiqué sur la langue japonaise.

[3] http://infomap-nlp.sourceforge.net/
[4] http://www.lexique.org/public/corpatext.php
[5] http://www.google.fr

3 Constitution d'un corpus d'associations sémantiques

Jusqu'à ce jour les réseaux proposés ont été évalués dans des cadres d'autres applications, comme la traduction ou l'aide à la rédaction. Nous proposons un cadre évaluatif reflétant la tâche d'un système d'aide à l'accès lexical en situation quotidienne pour des personnes âgées ou dyslexiques, en composant un réseau d'associations réelles à l'aide de participants.

Pour 20 mots sélectionnés (TOT), 50 personnes ont donné les 5 premiers mots qui leur venaient à l'esprit en association avec le mot proposé, en évitant toute forme de digression. Cette expérience bénévole a été réalisée à l'aide d'une interface web, de manière à laisser les utilisateurs libres de leurs conditions d'expérimentation. La consigne exacte était : *"Cette expérience s'intéresse aux associations d'idées. Il va vous être proposé une série de 20 mots auxquels vous devrez associer à chaque fois les 5 premier mots qui vous viennent à l'esprit. par exemple : OREILLER : plumes, taie, bataille, dormir, moelleux. Ces 5 mots ne doivent pas nécessairement être reliés entre eux, mais doivent tous être associés au mot de départ. Ainsi, on n'attend PAS pour oreiller : plumes, oiseau, arbre, feuille, papier ... (ici chaque mot est relié au précedemment cité)."*

Les 20 TOT ont été sélectionnés de manière à ce qu'ils fassent tous partie des ressources linguistiques statiques (EuroWordNet, la carte LSA sur Corpatext et le réseau de cooccurrences). La liste des mots ainsi que leurs fréquences dans Lexique 3 [6] (fréquence dans des dialogues issus de films, fréquence dans un corpus de livres) est dans le tableau 1. La plupart sont des noms communs, mais il y a aussi des adjectifs et des noms propres, ainsi que des noms d'origine étrangère. Plusieurs de ces mots appartiennent à plusieurs catégories syntaxiques (java est à la fois un nom propre et un nom commun, massif est à la fois un adjectif et un nom commun), certains sont des noms communs repris par des marques commerciales (Casino, Lion), et d'autres sont aussi fortement polysémiques (facteur, baril, brioche, quiche). Tous les mots sont de fréquence inférieure à 20, donc peu fréquents voir rares.

TOT	f. film	f. livre	mot	f. film	f. livre	mot	f. film	f. livre
veau	6,20	16,96	baril	4,22	3,04	casino	17,41	10,81
entretien	17,71	27,77	cambriolage	9,34	2,77	java	0,72	2,30
menhir	0,18	0,68	quiche	0,66	0,68	kleenex	2,11	2,91
lion	17,05	33,04	brioche	7,29	7,09	facteur	11,27	14,32
festin	5,12	5,68	chaussette	14,58	22,84	massif	8,13	22,30
virtuel	2,53	2,16	rugby	1,87	3,11	landau	1,20	4,59
snowboard	N/C	N/C	oie	5,90	9,32			

TAB. 1 – Mots sélectionnés pour la constitution du corpus : chaque utilisateur devait proposer les 5 premiers mots ou expressions qui leur venaient à l'esprit. Les fréquences de chaque mot dans des dialogues de films (f. film) ou dans un corpus de livres (f. livre) montrent qu'il s'agit de mots peu fréquents dans le langage oral comme dans le langage écrit.

Les 50 personnes sont des adultes d'âges, de professions et de niveaux d'études variés, on part ainsi sans a priori social sur l'échantillon de population qu'elles représentent. Elles ont toutes proposé les 5 associations demandées. On note que des participants (non retenus dans le corpus) n'ont pas pu terminer l'expérience pour qui il était trop difficile de trouver plus d'un mot associé.

Le corpus ainsi créé contient pour chacun des 20 TOT au total 250 associations. En moyenne, on dénombre 76,8 associations différentes par TOT. [7]

[6] `http://www.lexique.org`

[7] Le corpus est disponible sur demande par mail à l'auteur.

4 Etude numérique

On cherche à répondre à l'aide du corpus à trois grandes questions, à savoir : est-il utile de combiner plusieurs systèmes ? Y-a-t-il des mots pour lesquels certaines ressources sont plus représentatives du lexique mental que d'autres ? Est-il possible de déterminer des profils d'utilisateurs dans la prise en compte des ressources ?

Pour tenter de répondre à ces questions, nous avons recherché le TOT dans les 5 environnements sémantiques (listes de mots relatives à chaque ressource) de chaque association proposée pour ce TOT (soit 250 par TOT). La figure 1 illustre le principe de cette évaluation. Pour chaque association, on recherche les mots en relation dans chacune des 5 ressources, en limitant aux 100 premiers mots (au sens de scores lorsqu'ils sont disponibles). Le TOT est considéré retrouvé si fait partie d'une relation retrouvée par une ressource. En effet les relations sont des termes ou des expressions nominales. Ainsi si le mot recherché est *tarte*, l'expression *tarte à la fraise* sera acceptable.

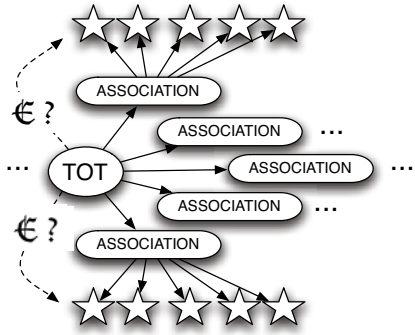

FIG. 1 – Evaluation des ressources : on recherche le TOT dans les listes de mots issues des relations de chacun des mots associé à ce TOT par un utilisateur dans chacune des 5 ressources disponibles.

Ce mode d'évaluation recherche le TOT dans 5×5 listes de 100 mots, soit une liste de 2500 mots différents au maximum. Sachant qu'au final on souhaite retrouver le TOT dans une liste de 10 mots proposés à l'utilisateur en fonction des associations qu'il propose, il ne parait pas pertinent d'augmenter le nombre de mots contenus dans les listes retournées par chaque ressource. On définit les associations "utiles" (AU) comme les associations proposées par les utilisateurs et pour lesquelles la liste de mots (obtenue à partir de relations dans au moins l'une des 5 ressources (500 mots) ou bien dans une ressource en particulier (100 mots) contient le TOT.

4.1 Couverture entre les ressources

La couverture de l'ensemble des ressources est estimée à partir du nombre d'associations comprenant le TOT dans les relations d'une seule des 5 ressources disponibles. On calcule le pourcentage de ces associations par rapport au nombre total d'associations comprenant la cible dans au moins une des 5 ressources.

relations	descriptives	paradigmatiques	syntagmatiques		
	def	ewn	lsa	coocc	web
%ressource	40%	33%	23%	47%	70%
%global	6%	6%	2%	16%	44%

TAB. 2 – Pourcentages de rappel unique (nombres d'associations pour lesquelles la ressource est la seule à retrouver le TOT dans les mots reliés), soit par rapport au rappel global pour cette ressource (nombre d'associations pour lesquelles la ressource retrouve le TOT dans les mots reliés) (%ressource) , soit par rapport au nombre total d'AU (associations "utiles") (%global).

Les pourcentages de rappel unique de chaque ressource par rapport à son rappel total, consignés dans le tableau 2, sont compris entre 23% et 70%, ce qui signifie que chacune des ressources est essentielle pour une bonne part des informations auxquelles elle permet d'accéder, étant donné que dans ces cas là elle est la seule à pouvoir proposer le TOT recherché à l'aide de l'AU. Les pourcentages de rappel unique par rapport à la globalité des réponses montrent néanmoins que l'impact des ressources descriptives ou paradigmatiques est beaucoup moins important dans l'ensemble que les relations syntagmatiques.

4.2 Résultats par mots

Nous nous sommes penchés sur la répartition des relations apportées par chacune des ressources pour chaque TOT du corpus. Pour cela nous avons observé les cas où les associations proposées par les utilisateurs contenaient le TOT dans chacune des ressources. Les résultats dans le tableau 3 montrent tout d'abord que certains mots, comme *veau*, *entretien* ou *rugby* sont plus enclins à être retrouvés par l'ensemble des ressources, alors que d'autres comme *cambriolage* ou *menhir* nécessite la complémentarité de deux ressources (en effet dans ces deux cas la somme des associations reliant au TOT pour chaque ressource est pratiquement égale au total d'associations reliant au TOT, ce qui signifie que ce sont des associations différentes qui ont permis de retrouver le TOT pour chaque ressource). Par ailleurs certains mot ne sont atteignables que par une seule ressource, et pas la même dans tous les cas. Ainsi, *snowboard* n'a été retrouvé que par des associations de relations dans la définition, facteur n'a été atteint pratiquement que à l'aide du réseau de cooccurrences, *festin* majoritairement par la carte sémantique LSA sur le corpus littéraire.

Cependant la couverture globale reste assez faible (33 % des associations permettent de retrouver le TOT pour lequel elles ont été citées). Ainsi nous nous sommes interrogés sur les cas où au moins une ou deux, où toutes les associations, et où la première association donnée par un utilisateur est une AU. On dénombre ainsi les groupes d'associations (5 associations proposées par un utilisateurs pour un TOT) où N associations sont "utiles". Ainsi le tableau 3 contient les pourcentages de ces cas pour chaque TOT. Le pourcentage de cas où les 5 associations sont des AU est de 0% dans la plupart des cas, ce qui exclut toute possibilité d'intersection entre les différents réseaux de l'ensemble des associations proposées pour une combinaison des ressources. Par ailleurs pour 80 % des TOT plus de 70% des groupes d'associations contiennent au moins une AU. Les pourcentages de cas où la première association est "utile" sont à considérer en regard des cas où au moins une association est "utile". On constate alors que si pour des mots comme *landau* ou *veau* la première association est presque toujours une AU, pour des mots comme *lion* ou *java* ou *entretien* l'écart est important ce qui signifie que les AU viennent plutôt après la première.

Mot cible	% d'AU						% cibles : N AU			% cibles :
	def	ewn	lsa	coocc	web	total	N > 0	N > 1	N = 5	1ere AU
veau	23,6	43,2	27,2	46,8	54	80,8	100	98	32	96
baril	0,8	0	0	26,8	24	36	86	62	0	60
casino	0	0	0	32	13,6	44,8	100	76	0	70
entretien	10	10,4	10	0	23,2	23,6	84	30	0	36
cambriolage	0,8	12	0	0,4	11,6	24	72	38	0	40
java	0,4	0,8	0	0,4	4	4,8	24	0	0	0
menhir	13,6	0	0	0	33,2	46,8	96	82	4	62
quiche	0	0	0	0	34,4	34,4	90	64	0	74
kleenex	0	0	0	0	15,2	15,2	72	2	0	50
lion	14,8	4	3,2	0,8	16	30	88	50	0	36
brioche	0,4	0,4	0	0	39,6	40	94	70	0	52
facteur	1,2	1,2	0	13,6	2,4	17,6	74	12	0	28
festin	2,8	0,8	12,4	0	5,6	21,6	68	34	0	48
chaussette	1,6	4,8	0	0	6,8	10	40	8	0	4
massif	1,2	15,6	0	19,2	4,8	38,8	92	68	2	52
virtuel	0	0	0	17,2	24	28,8	82	46	0	30
rugby	17,2	10,4	0	46,8	20,8	63,6	98	90	14	84
landau	0,4	10,4	0	0	29,6	30	98	50	0	84
snowboard	5,2	0	0	0	0	5,2	26	0	0	6
oie	10,4	16,4	18	31,6	48,4	68,4	100	98	8	78
Moyenne	5,22	6,52	3,54	11,78	20,56	33,22	79,2	48,9	3	49,5

TAB. 3 – Pourcentages d'associations "utiles" (AU), parmi les 250 proposées pour chaque mot. Les ressources sont les mots clés du web (web), les mots clés des définitions (def), les liens dans EuroWordNet (ewn), les liens dans la carte sémantique (lsa), et les liens dans le réseau de cooccurrences (coocc). Pourcentages de TOT avec au moins une AU, au moins 2 AU, ou exactement 5 AU, et pour lesquels la première association proposée est "utile".

4.3 Profils d'utilisateurs

Nous faisons l'hypothèse que chaque personne utilise des voies d'accès privilégiées en fonction d'un profil cognitif, et que chacune de ces voies correspond à une des cinq ressources proposées. Nous supposons que s'il est possible de catégoriser automatiquement les données de chaque personne, c'est que d'une part il existe des profils reconnaissables et d'autre part on doit les prendre en compte.

Afin de déterminer si l'on pouvait dégager des profils d'utilisateurs, nous avons effectué une catégorisation automatique des répartitions d'AU pour chaque ressource par utilisateurs en un nombre de classes variable à l'aide de l'algorithme K-Means implémenté dans la plate-forme WEKA. On a ainsi pu obtenir dans la meilleure configuration trois classes de répartitions, dont deux majoritaires qui se différencient essentiellement par la capacité de la ressource de mots clés du web à retourner le TOT à partir des associations proposées. La répartition des AU pour les 20 TOT de chaque utilisateur montre qu'on peut les séparer en deux classes. En effet pour chaque utilisateur une majorité de ses associations appartient à une des deux classes (33 utilisateurs dans la classe favorisant les mots clés web, et 17 dans l'autre).

Cependant, la variété des associations et ressources "utiles" est également fortement liée aux TOT eux mêmes, comme le montrent les résultats du tableau 3. Nous avons donc exploré une autre méthode afin de définir s'il existe ou non des profils d'utilisateurs. Pour cela, nous avons appliqué l'algorithme EM (Expectation-Maximization) sur les données comprenant pour chaque utilisateur et chaque ressource le nombre total d'associations "utiles" parmi les 100 proposées pour les 20 mots réunis. On définit 5 paramètres, qui correspondent aux 5 ressources. Le tableau 4 décrit la composition des 3 clusters proposés par l'algorithme, à travers les moyennes

ressource :	Cluster 1		Cluster 2		Cluster 3	
attribut	Moy.	E.C.	Moy.	E.C.	Moy.	E.C.
web	20,6	3,1	25,2	2,3	13,6	2,6
def	4,6	1,4	6,9	2,2	4,3	1,6
ewn	5,9	1,9	9,4	1,6	3,7	1,3
lsa	3,7	1,1	4,5	1,4	1,8	1
coocc	11,6	1,8	13,4	2,9	9,7	2
instances	48%		32%		20%	

TAB. 4 – Clusters proposées par l'algorithme Expectation-Maximization : valeurs moyennes et écart type pour chaque attribut (pourcentage d'associations utiles par individu pour l'ensemble des 20 mots, pour chaque ressource), et nombre d'instances attribuées à chaque cluster.

pour chaque attribut. L'écart type indique la consistance des données autour de ces moyennes.

Les clusters décrits dans le tableau 4 montrent trois aspects de la prise en compte des ressources pour un profil d'utilisateur. Tout d'abord le cluster 2 se distingue par une place plus importante à apporter aux ressources descriptives et paradigmatiques telles que les définitions et EuroWordNet, alors que les individus du cluster 1 seront majoritairement aidés par les mots clés du web et les cooccurrences. Les individus du cluster 3 se caractérisent essentiellement par leur incompatibilité avec l'ensemble des ressources sélectionnées, c'est à dire que les scores sont faibles pour toutes les ressources. Cependant, il faut noter que de ce fait les ressources sont à prendre en compte pratiquement sur un pied d'égalité. Dans tous les cas cependant, la part accordée aux associations reliées par les mots clés sur le web sont à prendre en compte en large majorité.

5 Conclusions et perspectives

L'évaluation combinée des ressources proposées pour l'aide à l'accès lexical dans le cadre de la recherche du TOT dans des listes de mots obtenues à partir d'associations montre qu'elles sont fortement complémentaires. Cependant le taux de rappel global (nombre d'AU par rapport au nombre total d'associations proposées) n'est que de 33%, ce qui implique qu'une combinaison des ressources devra être capable de sélectionner les associations utiles, puis en extraire le TOT. Dans 79% des exemples du corpus les TOT sont accessibles via au moins une AU, ce qui est la limite maximum du rappel que pourra avoir un système combinant les ressources sélectionnées.

Les relations syntagmatiques obtenues à l'aide des premières pages retournées par Google sont les plus riches en termes de rappel global (21%), ce qui ne signifie pas nécessairement que la technique d'extraction est performante mais plutôt que le corpus représenté par les pages du Web est peut être plus pertinent. De la même manière, le faible rappel de la carte sémantique apprise à l'aide de LSA (3,54%) ne remet absolument pas en cause la qualité de la méthodologie, mais doit être imputé à la nature du corpus à partir duquel la ressource est construite. Nous n'avons pas encore évalué les méthodologies d'extraction de relations syntagmatiques, et cela pourra être intéressant afin de sélectionner la représentation la plus adaptée à chaque nature de corpus, et ainsi améliorer le rappel global.

D'autre part, les associations proposées n'ont pas été caractérisées d'un point de vue psychologique, mais intuitivement on remarque des tendances. Par exemple, beaucoup d'associations sont des noms de couleurs, ce qui suggère des associations sémantiques visuelles plus que verbales. De la même manière 18 personnes ont associé le mot chien au mot facteur, ce qui correspond plus à une image de carte postale ou de film (où le facteur se fait mordre) qu'à un

réel lien sémantique.

Le principal problème qui reste en suspens est comment peut-on pondérer les listes de mots issues de chaque réseau pour parvenir à une liste unique ordonnée, sachant qu'on ne peut présenter qu'une dizaine de mots ou expressions à l'utilisateur pour rester raisonnable en termes d'ergonomie. Au delà de scores de pertinence ou de caractérisation des relations disponibles dans les ressources, les mots concernés par le TOT étant essentiellement des mots peu fréquents, on peut pondérer les mots proposés en fonction de leur fréquence. Etant donné la possibilité de définir des profils utilisateurs, il est peut être également possible de pondérer les mots proposés en fonction de la ressource dont ils sont issus, et d'apprendre ces paramètre de manière empirique. La pondération doit aussi prendre en compte l'appartenance du mot proposé à la sémantique globale dégagée par les associations proposées, ce qui reviendrait à désambiguïser les associations.

Références

BROWN R. & MCNEILL D. (1966). The tip of the tongue phenomenon. *Journal of verbal learning and verbal behaviour*, **5**, 325–337.

BURKE D. M. & SHAFTO M. A. (2004). Aging and language production. *American psuchological society*, **13**(1), 21–24.

CHURCH K. & HANKS P. (1990). Word association norms, mutual information , and lexicography. *Computational Linguistics*, **16**(1), 177–210.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6), 391–407.

DELL G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, **93**, 283–321.

FAUST M. & SHARFSTEIN-FRIEDMAN S. (2003). Naming difficulties in adolescents with dyslexia : application of the tip-of-the-tongue paradigm. *Brain and cognition*, **53**(2), 211–217.

FERRET O. & ZOCK M. (2006). Enhancing electronic dictionaries with an index based on associations. In *Coling/ACL joint conference*, Sydney, Australia.

JI H., PLOUX S. & WEHRLI . E. (2003). Lexical knowledge representation with contextonyms. In *MT Summit IX*, New Orleans, USA.

JOYCE T. (2005). Constructing a large-scale database of japanese word associations. *Glottometrics*, **10**, 82–98.

LANDAUER T., FOLTZ P. W. & LAHAM . D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, **25**, 259–284.

LORTAL G., GRAU B. & ZOCK M. (2004). Système d'aide à l'accès lexical : trouver le mot qu'on a sur le bout de la langue. In *TALN'04*, p. 259–268.

REUER V. (2004). Language resources for a network-based dictionary. In *Workshop on Enhancing and Using Electronic Dictionaries ; following COLING 2004*, p. 81–84.

ZOCK M. (2002). Sorry, but what was your name again, or, how to overcome the tip-of-the tongue problem with the help of a computer ? In *COLING-Workshop on building and using semantic networks*, Taipei, Taiwan.