Stocker des Mots ne Garantit nullement leur Accès.

Michael Zock[1] Didier Schwab[2]
(1) CNRS, Aix Marseille Université
(2) Univ. Grenoble Alpes
michael.zock@lif.univ-marseille.fr, didier.schwab@imag.fr

Résumé. L'objectif de ce papier est double : (a) montrer que le stockage ou la mémorisation d'une forme lexicale ne garantit nullement son accès ou sa disponibilité, et (b) décrire les étapes nécessaires pour construire une ressource susceptible d'aider les rédacteurs à trouver le mot bloqué sur le bout de leur langue (ou de leur plume).

Pour vérifier le premier point, nous avons réalisé une petite expérience en comparant deux ressources pour voir si elles nous permettaient de trouver le terme recherché (mot cible) et si l'accès était facile. Les ressources en question sont WordNet, ou plutôt une version étendue, eXtended WordNet (xWN) et Wikipedia (WP), converti par nous en une ressource lexicale, nommée WordFinder (WF). Il s'avère que cette dernière ressource permet généralement à trouver assez rapidement le terme recherché, alors que xWN y échoue souvent, ou lorsqu'il y parvient, l'élément en question se trouve assez loin dans la liste des candidats. Ceci paraît surprenant dans la mesure où les deux ressources 'possèdent' le même vocabulaire. Cependant la situation devient vite assez claire lorsqu'on regarde les liens entre les mots (l'index ou l'organisation lexicale) des deux ressources. Contrairement à WN, WF contient beaucoup de liens syntagmatiques (café-noir ; café-Brésil ; café-Starbucks,...), permettant de ce fait d'accéder au mot cible par un bien plus grand nombre de mots source.

Ayant montré que 'stockage' n'implique pas forcément 'accès' ou disponibilité, nous présentons ensuite une feuille de route, esquissant les éléments à élaborer pour construire une ressource susceptible d'aider des rédacteurs à trouver le mot bloqué sur le bout de la langue. La construction de notre future ressource est basée sur les raisonnement suivants. L'accès lexical consiste essentiellement à localiser un élément parmi l'ensemble des formes lexicales stockées dans la ressource lexicale (dictionnaire). Comme il est déraisonnable de chercher le mot cible parmi l'ensemble des formes stockées, nous proposons de décomposer ce processus en deux étapes. Dans un premier temps nous essayons de réduire l'espace initial à un ensemble plus petit. A cette fin on présentera tous les mots directement associés au(x) mot(s) source (l'entrée), mot(s) disponible(s), et mot(s) auquel(s) on pense spontanément en cherchant la cible. Dans un deuxième temps on essayera de guider l'utilisateur en lui présentant une version structurée des mots obtenus lors de la phase précédente. Pour atteindre ce dernier objectif il faut donc structurer la liste des mots, ce qui veut dire, qu'il faut former des groupes (clusters) auxquels on donne des noms (arbre catégoriel). Le défi ici est de nommer ces groupes, parce que c'est sur cette base (le nom de ces catégories) que l'utilisateur décidera dans quelle direction aller pour chercher le mot dans un 'paquet' particulier.

Abstract. Dealing with word access in language production we pursue here two goals: (a) provide evidence that 'storage' does not imply 'access' (or, accessibility) ; (b) describe the steps to be carried out to build a resource allowing for interactive word finding.

In order to show evidence for the first claim we compared two resources, an extended version of WordNet (xWN) and WordFinder (WF), a lexical resource based on Wikipedia (WP). One of the goals was to see their respective performance with respect to word access. It appears that our resource (WF) generally finds quicker and more often the target word than xWN. This seems surprising at first sight as both resources 'have' the same vocabulary. Yet this is not surprising any more if one takes a look at the information on which the organization of the two resources is based. WN lacks syntagmatic links, hence it will not perform well when the relationship between the input and the target is encyclopedic knowledge (coffee-Brazil ; elephant-grey).

In order to build the resource required to support wordfinding we started from the following assumptions. Word access is basically finding a specific item (target word) within the lexicon. Put differently, the task is to reduce the entire set (of words contained in the lexicon) to one, the target. Since it is unreasonable to search in the entire lexicon, we suggest a two-step method. The goal of the first is to reduce the initial search space to a smaller set, while the goal of the second

is to support navigation by presenting the words identified in step-1 in a clustered and labeled form (categorial tree). The challenge here is to name the clusters, as it is on this basis that the user decides on the direction to go in order to search further for a given word.

Mots-clés : Accès lexical, WordNet, Wikipédia, WordFinder, groupement par catégorie, navigation assistée.
Keywords: Lexical access, WordNet, Wikipedia, WordFinder, categorial tree, clustering, navigational aid.

1 Introduction

Tout le monde admettra que posséder un grand vocabulaire est un atout important. Reste à savoir ce qu'il faut entendre par le terme 'posséder'. Pour nous cela signifie trois choses : avoir *stocké* des signes au sens Saussurien (des couples sens/mot-forme), savoir s'en *servir* en effectuant les bons choix entre des alternatives (synonymes) tout en respectant les contraintes de la langue (collocations), et (c) savoir *trouver* (récupérer) à volonté le sens (compréhension) ou la forme des lemmes (production). C'est surtout ce dernier aspect qui nous intéresse ici, la récupération des formes (lemmes) exprimant un certain sens. Concernant l'accès lexical la mémoire humaine semble bien plus fragile que celle des machines. Ce qui a été stocké dans la mémoire d'une machine nous semble accessible ce qui est loin d'être le cas pour le cerveau humain, comme cela a été maintes fois montré via le phénomène du *mot sur le bout de la langue* (Brown et McNeill, 1966, Brown, 1991). Il nous arrive parfois de ne pas trouver un terme, alors que nous l'avions appris et nous en sommes servi il n'y a pas bien longtemps. Le mot en question a donc bel et bien été stocké (donc, mémorisé), mais pour des raisons diverses, pas toujours identifiables, il est (momentanément) inaccessible. Bien qu'une grande partie du mot est accessible (notamment le sens), la forme du mot reste bloquée sur le bout de la langue (Brown, 1991). Ceci dit, contrairement à ce qu'on pourrait croire, un problème d'apparence identique peut également toucher les machines. Ce n'est pas parce qu'une information (par exemple, un mot) a été stockée, qu'elle est toujours accessible et c'est que nous allons montrer par la suite.

Dans la deuxième partie nous allons présenter l'ébauche d'une feuille de route (ou, d'un programme de recherche), précisant la nature du problème et montrant quels éléments doivent être élaborés pour aider les êtres humains à dépasser ce problème, c'est-à-dire, trouver effectivement le mot recherché.

2 L'accès lexical automatique via une ressource externe

2.1 Comparaison de deux ressources

Comme déjà dit, le fait d'avoir stocké des mots ne garantit nullement leur accès. Pour vérifier cette affirmation nous avons réalisé une petite expérience, en comparant deux ressources : une version étendue de WordNet (WN), eXtended WN (Mihalcea et Moldavan, 2001) et Wikipedia (WP), que nous avons converti en une ressource lexicale, nommée WordFinder (voir 2.2). Notre but n'était pas tant de vérifier la qualité de WN ou d'une de ses extensions que de montrer que (a) le stockage ne garantissait pas l'accès, et que (b) l'accès dépendait de plusieurs facteurs qualitatifs, notamment, celui de la *ressource* dans laquelle s'effectue la recherche, de l'*indice*, et du type de la requête. Ayant deux ressources aux caractéristiques différentes, notre objectif était de vérifier leur efficacité relative par rapport à l'accès lexical. Pour des raisons purement pratiques (limitation du temps de traitement), nous avons seulement pris en compte les voisins directs (c'est-à-dire, les mots à une distance de 1). Par conséquent, nous avons défini une fonction nommée voisinage direct (désormais f_{vd}), qui, une fois appliquée à une fenêtre donnée (phrase / paragraphe)[1], produit toutes ses cooccurrences. Bien sûr, ce qui vaut pour les associations directes (notre cas ici), vaut également pour les mots liés indirectement (distance > 1), c'est-à-dire, des associations médiées.

2.1.1 L'usage de WordNet comme un corpus

Un des objectifs de WN était de construire une ressource ressemblant au dictionnaire mental (réseau associatif), permettant un fonctionnement analogue à celui du cerveau humain (propagation d'activation).

La structure de WN est assez différente de celle des dictionnaires conventionnels, qui eux sont organisés par ordre alphabétique. Aussi, plutôt que de multiplier le nombre de dictionnaires, un pour chaque utilisation ou chaque tâche

[1] La taille optimale est une question empirique. Elle peut varier selon le type de texte, encyclopédie texte brut.

(trouver une définition, un synonyme, un antonyme,…), WN a été construit comme une ressource unique, permettant l'accès par des chemins multiples et par le biais de différents type de liens. Comme ce travail est très connu, nous ne le décrirons pas plus en détail ici (Miller, 1990).

Si WN est une ressource lexicale, il peut également être vu comme un corpus. Ceci peut s'avérer très utile, si l'on veut le comparer avec d'autres corpus — comme, par exemple, Wikipedia[2] qui est une encyclopédie multilingue, collaborative et libre — ou si l'on veut faire usage d'une partie spécifique de la base, par exemple, les gloses. Puisque les gloses correspondent schématiquement à la signification d'un mot (définition), leurs éléments (sac de mots) peuvent être utilisés pour accéder au mot dont ils définissent le sens (entrée lexicale, lemme).

WN a eu un grand impact dans la communauté TAL où il est fortement utilisé[3]. Ceci a conduit à la création de nombreuses extensions. Comme déjà mentionné, nous en utilisons l'une d'elles, Extended WN (Mihalcea et Moldovan, 2001), ce qui nous épargne la peine d'avoir à faire face aux problèmes inhérents à l'analyse de textes brut : segmentation, résolution d'ambiguïtés lexicales, lemmatisation,…

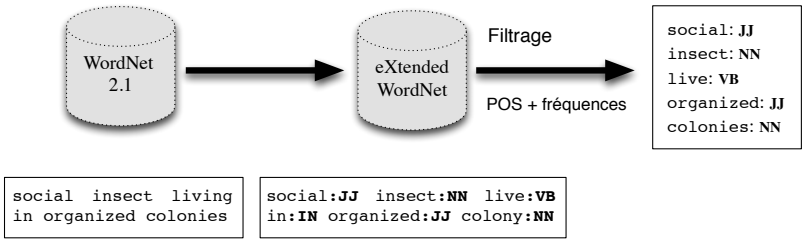

FIGURE 1: WordNet comme corpus (l'exemple étant "ants" (fourmis)).

Deux problèmes demeurent cependant : la taille du corpus (environ 144 000 entrées) et le manque de connaissances encyclopédiques, c'est-à-dire les associations syntagmatiques, faiblesses, qui, pris ensemble, peuvent entraver l'accès lexical. En effet, les concepts fonctionnellement liés comme *dîner-table-repas* ou *pêcher-filet-poisson*, devraient s'évoquer réciproquement, alors que ce n'est souvent pas le cas. Ce problème, connu depuis longtemps par les auteurs de WN est nommé *problème de tennis* (Fellbaum, 1998). Des mots jouant ensemble un rôle dans un domaine ou dans une tâche ne sont pas forcément stockés ensemble. Ainsi, *balle de tennis*, *raquette* et *arbitre* apparaissent dans différentes branches de l'arborescence, alors qu'ils sont tous susceptibles d'être nécessaires lorsqu'on parle du sujet qui les réunit, un match de tennis. De manière analogue, *instrument* et *utilisé_pour*, apparaissent dans différentes parties de la ressource, alors qu'ils sont (quasi-)synonymes. Malgré tout, il faut noter que de réels efforts ont été faits pour surmonter ces problèmes. Par exemple, des informations peuvent être trouvées dans les gloses (dans le cas de *utilisée_pour* et *instrument*), et des mots thématiquement liés peuvent désormais être consultés dans une certaine mesure (Boyd-Graber et al. 2006).

2.1.2 L'usage de Wikipédia comme corpus

Afin de comparer WP et WN, nous avons utilisé la version anglaise, qui, au moment de la rédaction de ce papier (mars 2013) contenait 3.550.567 entrées. WP a exactement des propriétés opposées à WN[4]. Bien qu'il contienne de nombreuses associations syntagmatiques, ce n'est que du texte brut. Ainsi, des problèmes tels que la segmentation du texte ou la lemmatisation doivent être abordés. Pour éviter cela, nous avons utilisé DBpedia (Bizer et al., 2009), une version texte brut de WP. L'utilisation d'un lemmatiseur[5] nous a permis d'annoter les éléments majeurs du paragraphe et de filtrer tous les mots hors de propos, pour ne garder que les plus importants (noms, adjectifs, verbes et adverbes). Ces derniers ont ensuite été utilisés pour la construction de notre base de données.

[2] http://www.wikipedia.org/

[3] Voir Fontenelle (2012) sur l'impact des réseaux sémantiques à la WordNet sur la lexicographie contemporaine.

[4] Ces deux ressources ont été alignées, par exemple, dans BABELNET (Navigli et Ponzetto 2010).

[5] Dans cette expérience, nous avons utilisé notre propre lemmatiseur basé sur le dictionnaire anglais DELA (http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/ telechargement.html)

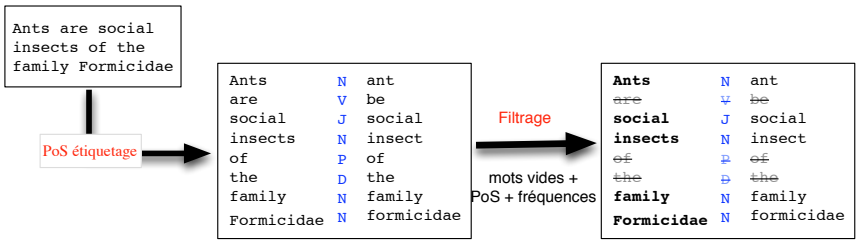

FIGURE 2: Wikipedia comme corpus

2.1.3 Exploitation et comparaison des ressources

Construire la ressource nécessite le traitement d'un corpus et la construction d'une base de données. À cette fin, nous avons utilisé un corpus en appliquant notre fonction de voisinage f_{vd} à une fenêtre prédéterminée : un paragraphe dans le cas des encyclopédies. Le résultat (c'est-à-dire les cooccurrences) est stocké dans la base de données, avec leurs poids, (c'est-à-dire le nombre de fois que deux termes apparaissent ensemble) et le type de lien. Comme mentionné plus haut, ce genre d'information est nécessaire plus tard pour le classement des termes et la navigation.

Les cooccurrences sont stockées sous forme de triplets (M_S, M_{CP}, NB_{occ}), où M_S et M_{CP} désignent respectivement le *mot source* (c'est-à-dire, le mot déclencheur ou *mot requête*) et le *mot cible potentiel*, terme obtenu en réponse à la requête (association directe), tandis que NB_{occ} (nombre d'occurrences), représente le poids, c'est-à-dire le nombre de fois que deux termes apparaissent ensemble dans le corpus, la portée des cooccurrences étant le paragraphe. Bien sûr, il y a d'autres façons de déterminer le poids (par exemple, des informations partagées), et surtout, d'autres facteurs peuvent avoir influer sur l'accessibilité d'un terme, par exemple, la récence. Aussi, les mots produits suite à une requête (M_S), ne sont que des mots cible potentiels. Ils peuvent être la cible, sans l'être nécessairement. Ils peuvent être des termes intermédiaires entre la source et la cible (association indirecte) ou être un terme associé au M_S, sans être le mot recherché pour autant (la cible). Il s'agit simplement d'un terme associé.

2.2 Utilisation

Pour montrer les qualités relatives d'une requête, nous avons développé WordFinder, un site web en Java (bientôt disponible sur nos pages d'accueil respectives). L'utilisateur communique au programme via cette interface les mots source. Le programme calcule alors les mots les plus probables d'être la cible, puis il transmet cette liste après avoir mise à jour la page. L'utilisateur peut alors choisir de rajouter des mots à sa requête en l'ajoutant dans le champ prévu à cet effet ou en cliquant sur les termes de la liste. Par exemple, si les entrées sont *récolte*, *vin*, *raisin*, le système va afficher tous les mots co-occurrents (associations directes, figure 3). Bien sûr, si nous utilisons plusieurs corpus, comme c'est le cas ici, nous devons afficher les résultats pour chacun d'eux.

La sortie est une liste ordonnée de mots, l'ordre étant fonction du score global : le nombre de cooccurrences entre les M_S et le mot associé, appelé le *mot cible potentiel* (M_{CP}). Par exemple, si le M_S *bouquet* apparaissait cinq fois avec *vin* et huit fois avec *récolter*, nous obtiendrions un score ou poids global de 13 : ((vin-5, récolte-8), bouquet, 13). Les poids peuvent être utilisés pour classer les mots en terme d'ordre (de priorité) et pour choisir les mots à présenter. Ceci peut devenir nécessaire pour peu que la liste soit longue.

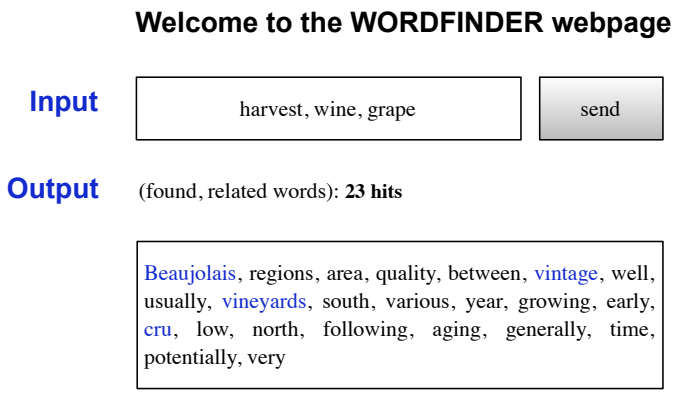

FIGURE 3: Sorties produites en réponse aux entrées 'récolte, vin, raisin'

2.2.1 Exemples de requêtes et comparaison des deux ressources

La figure 4a ci-dessous montre les résultats produits respectivement par WN et par WP pour les entrées *vin*, *récolte* ou leur combinaison : *vin* + *récolte*.

Entrées :	*Sorties* de WordNet	*Sorties* de Wikipedia
wine	**488 candidats :** grape, sweet, serve, france, small, fruit, dry, bottle, produce, red, bread, hold…	**3045 candidats :** name, lord characteristics, christian, grape, France, … vintage (81 ème position), …
harvest	**30 candidats :** month, fish, grape, revolutionary, calendar, festival, butterfish, dollar, person, make, wine, first,…	**4583 candidats :** agriculture, spirituality, liberate, production, producing, …, vintage (112 ème position), …
wine + harvest	**6 candidats :** make, grape, fish, someone, commemorate, person	**353 candidats :** grape, France, vintage (3 ème position), …

FIGURE 4a: Comparaison de deux corpus pour trois entrées différentes

Notre objectif était de trouver le terme *vintage* (vendange). Les résultats montrent que *récolte* est un meilleur terme de requête que *vin* (488 vs 30 candidats) et que leur combinaison est meilleure que chacun des deux termes seul (6 candidats). Ce qui est plus intéressant est le fait qu'aucun de ces termes ne correspond au mot cible, bien que celui-ci soit dans WN, ce qui étaye notre hypothèse que le stockage d'un terme ne garantit nullement son accès (voir également Sinopalnikova & Smrz, 2006, Tulving & Pearlstone, 1966).

Les choses peuvent beaucoup changer lorsque nous construisons notre index sur la base d'autres informations, par exemple, sur la base d'informations encyclopédiques, comme celles contenues dans WP. Dans ce cas, le terme *vin* évoque beaucoup plus de mots que WN (3045 au lieu de 488, avec *vendange* dans la position 81). Pour 'récolte' nous obtenons 4583 réponses au lieu de 30, *vendange* arrivant en position 112. La combinaison des deux produit 353 réponses, propulsant le mot cible à la 3ème position, donc, très proche de la tête de la liste.

Nous espérons que cet exemple suffit pour convaincre le lecteur de l'intérêt qu'il y a à utiliser des corpus équilibré, c'est-à-dire, des textes riches, mais hétérogènes, pour construire l'index grâce auquel l'usager peut naviguer dans la ressource pour trouver le mot qu'il a sur le bout de la langue, mot qu'il connaît sans pouvoir l'activer (complètement) pour autant. On notera, que ce problème n'est pas sans rappeler le déclin progressif d'une fonction cérébrale nommé *dégradation gracieuse*, phénomène pris en compte par des architectures connexionistes (Bechtel et Abrahamsen, 1991).

2.2.2 Analyse de cet échec relatif

On peut se demander pourquoi nous n'avons pas réussi à accéder aux informations dans WN, alors qu'elles y étaient, et pourquoi WP a fait tellement mieux. Nous croyons que l'échec relatif de WN est principalement dû à deux facteurs : la taille du corpus (114000 mots au lieu de 3 550 000 dans le cas de WP), et le nombre de liens syntagmatiques, qui tous les deux sont assez faibles par rapport à WP. Ce dernier point a déjà été souligné par G. Miller, lorsqu'il écrit : "WordNet provides a good account of paradigmatic associations, but contains very few syntagmatic links. .... If we knew how to add to each noun a distinctive representation of the contexts in which it is used... WordNet would be much more useful." (Miller, in Fellbaum, 1998: 33-34). C'est précisément ce que nous comptons faire (voir section 3).

Évidemment, comme WP est une encyclopédie, elle contient beaucoup plus de liens syntagmatiques que WN. Par vocation, WP contient beaucoup plus d'informations générales que WN concernant chacun des mots. Autrement dit, la taille de WN n'est pas un argument affaiblissant notre conclusion. Ceci dit, nous pouvons échouer à trouver l'objet recherché, même dans un très grand corpus. La réussite dépendant de la qualité de la ressource (couverture, adéquation), de la qualité de la requête, ou des deux. De plus, comme déjà mentionné, le point faible ne réside pas tant dans la quantité de données, que dans la qualité de l'index (la rareté relative des liens).

Afin d'être juste envers WN, il faut admettre que, si nous avions construit notre ressource différemment, par exemple, en incluant dans la liste tous les termes liés, non seulement les mots directement évoqués (mots cibles potentielles), mais aussi tous les mots contenant le mot-source (*wine*, i.e. *vin*) dans leur définition (*Bordeaux*, *Retsina*, *Tokai*), nous aurions sûrement obtenu le terme *vendange*, puisque le mot *vin* est contenu dans sa définition (*vintage* : a season's yield

of wine from a vineyard). On peut aussi remarquer que le succès peut varier assez considérablement, en fonction des termes choisis (mots cibles). Comme le montre le tableau ci-dessous, WN obtient des meilleures performances que WP pour les termes *ball*, *racket* et *tennis*. Ceci dit, WP suit de près, tout en contenant beaucoup d'autres mots susceptibles d'induire le mot cible, les termes *player*, *racket*, et *court*, étant classés respectivement 12, 18 et 20. N'étant pas une encyclopédie, WN ne possède pas la plupart d'entre eux. En revanche, ce qui est plus surprenant, et probablement un fait assez local et exceptionnel, il contient des informations très spécifiques et de nature encyclopédique, à savoir, le nom de deux grandes anciennes championnes de tennis : Monica Seles et Steffi Graf.

Entrées :	*Sorties* de WordNet	*Sorties* de Wikipedia
ball	**346 candidats :** game, racket, player, court, volley, wimbledon, championships, inflammation, …, tennis (15 ème), …	**4891 candidats :** sport, league, football, hand, food, foot, win, run, game, …, tennis (position 27), …
racket	**114 candidats :** break, headquarter, gangster, lieutenant, rival, kill, die, ambush, tennis (38 ème), …	**2543 candidats :** death, kill, illegal, business, corrupt, …, tennis (position 72), …
ball + racket	**11 candidats :** game, tennis, (2 ème), …	**528 candidats :** sport, strike, tennis (3 ème position), …

FIGURE 4b : Comparaison de différentes entrées dans deux corpus

Dernier point, contrairement à ce que l'on pourrait croire en apprenant que WN a été conçu en s'appuyant sur des données psycholinguistiques, WN n'a pas été conçu en vue d'une consultation. Voici les mots de son concepteur: "WordNet is an online lexical database designed for use under program control." (Miller, 1995, p. 39).

C'est pour combler cette lacune que nous allons esquisser dans le reste de cet article, une feuille de route afin de construire un dictionnaire destiné aux producteurs de langue (rédacteurs, locuteurs).

3 Une feuille de route pour construire la carte sémantique permettant à l'explorateur de s'orienter dans l'espace lexical

Chercher un mot dans un dictionnaire sans bon index est un peu comme s'orienter sur une île déserte sans carte convenable. Autrement dit, il faut construire une carte permettant à l'utilisateur de s'orienter dans cet espace lexical. Nous allons esquisser ci-dessous la construction de cette ressource, mais d'abord nous allons essayer de clarifier ce qu'il faut entendre par 'accès lexical', terme qui semble *a priori* évident. Et pourtant,...

3.1 Prémisses et fonctionnement de la recherche lexicale

Tous les mots du dictionnaire sont liés entre eux par des associations. Ces liens sont soit directs (associations immédiates, voisins directs), soit plus ou moins indirects : associations médiatisées (les voisins de voisins, des voisins,...). Aussi, si 'jaune' évoque 'canari' ou 'citron' on dira que 'jaune-canari' et 'jaune-citron' sont liés directement, l'un pouvant évoquer l'autre. Ceci dit, cette information serait insuffisante si la cible était le mot exprimant la saveur du fruit mentionné. Mais comme le mot 'citron' évoque entre autre la notion d'acidité, on trouvera le terme recherché à l'étape suivante, puisque 'jaune' (mot source) et 'acide' (mot cible) sont liés indirectement via le mot fruit (citron).

Le dictionnaire est donc un graphe connexe ce qui a pour conséquence que tous les mots sont accessibles à partir de n'importe quel mot. Le nombre d'étapes dépendra de la distance entre le *mot source* (M_s), mot ne vous venant pas à l'esprit, et le *mot cible* (M_c), mot représentant le but de la recherche. Chercher un mot consiste donc à entrer le réseau à un endroit quelconque en fournissant le mot source et de suivre les liens jusqu'au mot cible (mot bloqué sur le bout de la langue). Si ce dernier est un voisin direct, le système l'affiche immédiatement, et le problème est résolu. Dans le cas contraire, l'utilisateur peut continuer en changeant de M_c. Celui-ci peut être un des termes obtenues suite au M_c initial, soit un tout autre mot.

L'association (ou, l'index créée à partir d'associations) est donc l'une des bases de notre méthode de recherche. Elle a pour vocation de révéler le mot cible (voisin direct), soit de nous guider vers un mot plus proche (voisin indirect). Dans

tous les cas, cette méthode nous permettra de réduire l'espace de recherche. L'étape suivante consiste à grouper et à nommer les grappes de mots obtenus suite à l'entrée, le M_s. L'objectif de ce travail est d'aider l'utilisateur à naviguer dans une liste de mots (désormais) structurés.

Pour résumer : comme lancer une recherche dans l'intégralité d'une ressource (dictionnaire) pour trouver un mot (M_c) paraît déraisonnable, nous proposons de diviser ce processus en deux étapes. Lors de la première, on réduit l'espace initial, en ramenant l'ensemble des mots stockés dans la ressource aux voisins directs de l'entrée (M_c), liste qu'on structure ensuite en formant des groupes (clusters) auxquels on donne des noms. Aussi l'utilisateur pourrait-il naviguer dans un arbre catégoriel plutôt que dans une liste plate, ce qui devrait considérablement accélérer la recherche.

La difficulté de cette deuxième étape consiste essentiellement à trouver des noms adéquats aux groupes formés. Idéalement ces noms devraient correspondre à ceux que la majorité des gens donneraient à ces groupes, car, c'est via ces noms ou catégories qu'ils vont décider dans quelle direction orienter leurs efforts pour chercher le mot dans un groupe plutôt que dans un autre. Le figure 5 ci-dessous résume notre objectif, notre raisonnement et notre méthode. Ceci dit, beaucoup de détails restent à être clarifier : quels corpus utiliser, quel algorithme développer pour grouper et nommer ces listes de mots.

3.2 L'accès lexical : un processus en deux étapes

D'abord que faut-il entendre par 'accès lexical' (en mode production) ? Cela peut vouloir dire plusieurs choses. Pour ce qui nous concerne ici cela signifie « trouver *un* élément spécifique (mot cible) parmi l'ensemble des mots stockés dans la ressource (le dictionnaire) ». Ceci peut vouloir dire pour un être humain, trouver un terme parmi, environ 50.000 autres (son dictionnaire mental). La tâche consiste donc à réduire l'ensemble des candidats (ensemble de mots contenus dans le dictionnaire) à un seul, le mot cible. Comme il est hors question d'effectuer une recherche dans l'ensemble du dictionnaire, nous proposons de procéder en plusieurs étapes, plus précisément, deux. L'objectif de la première est de réduire l'espace de recherche initial (50.000) à un ensemble plus petit (par exemple, 100-150 mots), alors que l'objectif de la seconde est d'aider l'explorateur (l'être humain naviguant dans ce sous-ensemble) à naviguer. À cette fin on lui présente les mots identifiés à l'étape 1 dans des ensembles étiquetés (arbre catégoriel). Il y a donc deux aspects importants dans cette deuxième phase : grouper les mots et donner aux groupes des noms utilisables (*meaningful*) par l'utilisateur. À cet égard, utiliser « plus général » paraît plus pertinent qu'utiliser 'hyperonyme', parce que compréhensible par un plus grand nombre d'utilisateurs.

Il convient de noter que les locuteurs dans l'état du mot sur le bout de la langue (MBL) savent toujours quelque chose à propos du mot cible (Brown et McNeill, 1966). C'est précisément de cette information que nous allons nous servir. Ce sera l'entrée, la première prise de contact avec le dictionnaire. Étant donnée une entrée, le système affichera alors tous les mots directement liés (mots à une distance de 1, c'est à dire, toutes les associations directes).[6] Ce genre d'informations peut être glané dans une ressource comme le *Edinburgh Association Thesaurus* (EAT) (Kiss et al. 1973).[7] Comme ceci produira toutefois une liste trop longue pour permettre de trouver rapidement le terme recherché, nous proposons de regrouper les mots par familles, et de donner aux groupes des noms afin de faciliter la navigation. Comme on le voit, cette deuxième étape est cruciale, car sans elle l'utilisateur serait noyé sous une énorme liste de mots non-structurés. La figure 5 résume l'ensemble des opérations.

Notez que pour afficher correctement l'espace de recherche, c'est-à-dire l'ensemble de mots parmi lesquels chercher le mot cible (étape 1), il faut, dans un premier temps, lever toute incertitude sur l'entrée (désambiguïsation) afin d'éviter au maximum le bruit. Ne sachant pas quel sens est celui souhaité par l'utilisateur, le système risque d'afficher l'ensemble des associations possibles : « souris :animal » vs. « souris : dispositif informatique ». Il s'agit d'un désagrément que l'on aimerait éviter.

Notez également, que pour construire le guide en question, deux éléments doivent être construits (ou utilisés) : (1) un réseau lexical basé sur la notion d'association et (2) une méthode permettant de grouper les mots donnés en réponse à l'entrée. Ces groupes se verront attribuer un nom parlant pour que l'utilisateur de cette ressource puisse comprendre ce qui les réunit. Si la première étape commence à poser moins de problèmes à l'heure actuelle, la construction automatique de l'arbre catégoriel en question est loin d'être résolu, et ceci malgré la très grande littérature consacrée au problème de la catégorisation (Zhang et al., 2012, Bieman, 2012 ; Everitt et al. 2011).

[6] Si l'utilisateur fournit plusieurs termes en entrée, le système affichera l'intersection des termes associés.

[7] http://www.eat.rl.ac.uk

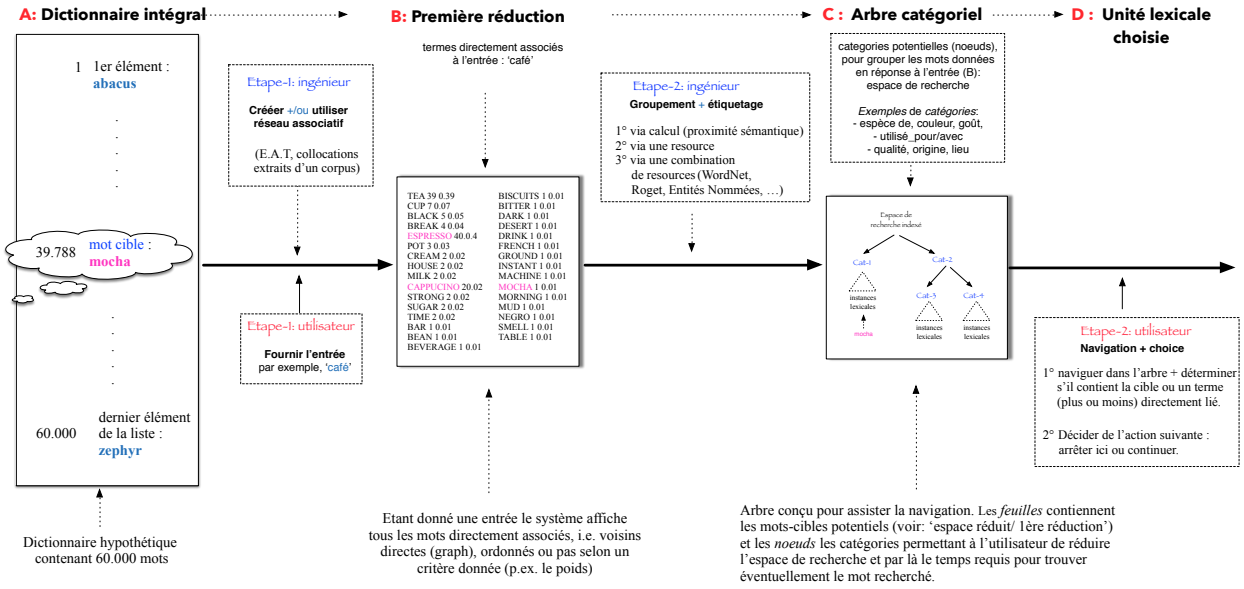

FIGURE 5 : L'accès lexical comme un dialogue en deux étape

4 Conclusion

L'objectif de ce papier était d'attirer l'attention sur le fait que d'avoir *stocké* un mot ne signifiait nullement pouvoir y accéder. Pour le permettre, nous nous avons esquissé une feuille de route précisant (1) la nature du dialogue entre l'utilisateur et la machine et (2) les éléments à mettre en place afin de permettre une navigation par association. La suite consistera donc à mener des expériences concrètes pour voir quel type de ressource (corpus ou autre) nous fournira la meilleure carte (étape-1) et quelle méthode nous permettra de présenter ce résultat sous forme d'un arbre dont les nœuds sont des catégories compréhensibles par l'être humain, tout en nommant de manière compréhensible et non-ambigüe la classe dont les éléments font partie (étape-2).

Références

BECHTEL, W. & ABRAHAMSEN, A. (1991). *Connectionism and the mind: A introduction to parallel processing in networks*. Oxford: Basil Blackwell. Traduction française par J. Proust. *Le connexionnisme et l'esprit: Introduction au traitement parallele par réseaux*, Paris: Editions la Decouverte, 1993.

BIEMANN, C. (2012). Structure Discovery in Natural Language. Theory and Applications of Natural Language Processing. Springer Berlin / Heidelberg.

BIZER, C., LEHMANN J., KOBILAROV G., AUER S., BECKER C., CYGANIAK R. & HELLMANN S. (2009). DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Issue 7, 154–165.

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. pp. 29-35

BROWN, A. (1991). A review of the *tip of the tongue* experience. *Psychological Bulletin*, 10, 204-223

BROWN, R. & MC NEILL, D. (1966). The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, 325-337

EVERITT, B.S., LANDAU, S., LEESE, M. et STAHL, D. (2011). Cluster Analysis: 5th Edition, John Wiley & Sons, Ltd

FELLBAUM, C. (éd.) (1998). WordNet: An Electronic Lexical Database and some of its Applications. Cambridge, MA: MIT Press.

KISS, G., ARMSTRONG, C., MILROY, R. & PIPER, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley and N. Hamilton-Smith (eds.). The Computer and Literary Studies. Edinburgh: University Press.

MIHALCEA, R & MOLDAVAN, D. (2001). Extended WordNet: progress report. In NAACL 2001 - *Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA.

MILLER, G.A. (éd.). (1990). WordNet: An On-Line Lexical Database. *International Journal of Lexicography,* 3(4).

MILLER G. A. (1995). WordNet : A lexical database for english. Communications of the ACM, 38 (11), 39–41.

NAVIGLI, R. & PONZETTO, S.P. (2010). BabelNet: Building a very large multilingual semantic network. *Actes du 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Suède, pages 216-225.

SINOPALNIKOVA, A. & SMRZ, P. (2006). Knowing a word vs. accessing a word: WordNet and word association norms as interfaces to electronic dictionaries. In Proceedings of the Third International WordNet Conference, pages 265–272, Korea.

TULVING, E. & PEARLSTONE, Z. (1966). *Availability* versus *accessibility* of information in memory for words. Journal of Verbal Learning and Verbal Behavior, 5, 381-391

ZHANG, Z., GENTILE A.L. & CIRAVEGNA, F. (2012). Recent Advances in Methods of Lexical Semantic Relatedness – a Survey . In the *Journal of Natural Language Engineering*, 19(4), 411-479, Cambridge Universtiy Press.