Caractérisation des mots abstraits et concrets

Il est considéré dans la littérature qu'il existe deux systèmes du codage de l'information verbale et visuelle. Ces codages sont réalisés dans des régions du cerveau différentes, ce qui est prouvé par des expériences avec l'utilisation de l'imagerie par résonance magnétique fonctionnelle (IRMf) et l'électroencéphalographie (EEG).

M. Just et ses collègues (Just et al., 2004) ont observé que les mots abstraits sont souvent associés aux régions du cerveau touchées chez les enfants dyslexiques ; cependant, il est nécessaire de distinguer les types de dyslexie. Les erreurs sémantiques ne sont pas uniquement présentes dans les cas de dyslexie profonde, acquise et non observée chez les enfants. Une expérience avec des phrases de haute et basse iconicité montre que cela prend plus de temps de répondre 'vrai' ou 'faux' pour les phrases avec un haut degré d'iconicité.

Cependant, Paivio a constaté que l'effet de concrétude, caractérisé par le temps de réponse plus rapide a des mots concrets, ne se produit pas seulement face à des données provenant

[1] ANR ALECTOR. Consulté le 20 mai 2020, à l'adresse https://alectorsite.wordpress.com/.

d'individus souffrant de dyslexie profonde, mais également chez les normo-lecteurs (Paivio, 1991). Les principales théories expliquant l'effet de concrétude chez les normo lecteurs incluent la théorie du double codage (Paivio, 1990) qui soutient que les mots concrets ont un avantage en termes de traitement car ils activent le système verbal (linguistique) et le système non-verbal (d'imagerie), tandis que les mots abstraits activent seulement le système verbal (Figure 1).

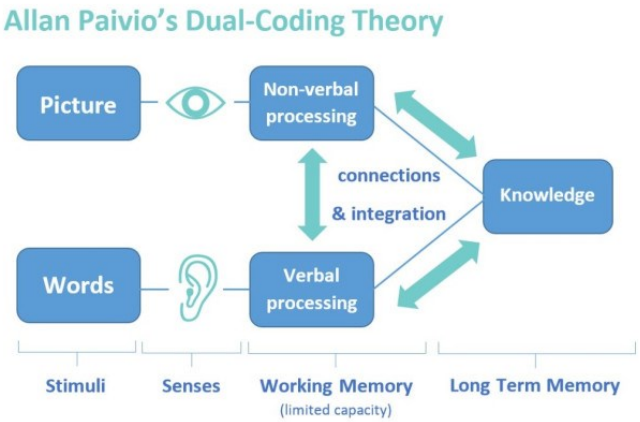

Figure 1. Système de double codage

Une autre explication de l'effet de concrétude est la théorie de la disponibilité du contexte (Schwanenflugel et al., 1988; Schwanenflugel & Shoben, 1983; Schwanenflugel & Stowe, 1989) qui soutient que les mots concrets sont fortement associés à quelques contextes, tandis que les mots abstraits sont faiblement associés à de nombreux contextes.

L'effet de concrétude est toutefois exagéré chez les personnes souffrant de dyslexie profonde, de sorte qu'il peut être impossible de lire des mots abstraits à cause d'un déficit sémantique de ces mots. Certaines preuves suggèrent que cet effet de concrétude exagéré se reflète également dans les différences d'activation neuronale chez les normo-lecteurs et les personnes aphasiques (Sandberg & Kiran, 2014). Diverses théories expliquant l'effet de concrétude dans la dyslexie profonde ont été proposées. Selon l'hypothèse de Coltheart (Coltheart et al., 1988) l'hémisphère gauche permet une lecture abstraite des mots. Les lecteurs souffrant de dyslexie profonde ayant des dommages à l'hémisphère gauche utilisent fortement l'hémisphère droit, ce qui entraîne des difficultés avec les mots abstraits. Morton et Patterson (1980) proposent un modèle à double voie dans lequel la dyslexie profonde résulte de lésions multiples. Dans ce modèle, la lecture s'effectue via la voie sémantique ; cependant, la sémantique des mots abstraits est altérée. De même, Plaut et Shallice (1993), dans leur modèle connexionniste, proposent un avantage pour la lecture de mots concrets, car les mots concrets sont plus simples à caractériser que les mots

abstraits. En outre, le modèle des différents cadres de représentation de Crutch et Warrington (2005) propose que les mots concrets sont représentés dans un cadre catégorique (basé sur la similarité sémantique) et les mots abstraits sont principalement représentés par une association sémantique (contextes linguistiques). Cette théorie soutient que les mots concrets partagent davantage de représentations avec d'autres mots similaires (par exemple, *vache – mouton*) qu'avec d'autres mots associés (par exemple, *vache – étable*), tandis que les mots abstraits partagent davantage de représentations avec d'autres mots associés (par exemple, *vol-punition*) qu'avec d'autres mots similaires (par exemple, *vol - crime*). En conséquence, les lecteurs souffrant de dyslexie profonde produisent plus d'erreurs associatives, comme *vol – punition*, que d'erreurs sémantiquement similaires, comme *vol - crime* en réponse à des mots cibles abstraits et plus d'erreurs sémantiquement similaires que des erreurs associatives en réponse à des mots cibles concrets.

les notions d'abstrait et de concret

La caractérisation des mots en concrets et abstraits reste une tâche difficile. Premièrement, par des mots concrets on comprend des mots qui ont un degré élevé d'iconicité. Selon Tellier et al., (2018), les mots concrets sont associés à une grande iconicité, notamment en termes de représentation mentale, tandis que les mots abstraits sont plutôt encodés verbalement (Paivio, 1986). Les mots concrets sont davantage associés aux informations contextuelles et aux expériences sensorimotrices que les mots abstraits, dans la mesure où les mots concrets sont liés à une haute iconicité et les mots abstraits à une faible iconicité.

La notion de nom ‘concret’ fait référence aux objets, matériaux, sources de sensations relativement directes (Gorman, 1961); la notion de nom ‘abstrait’ fait référence à des objets, des matériaux et des sources de sensations indirectes. Un mot peut être générique (nommer un groupe ou une catégorie) ou spécifique (nommer une idée ou un objet spécifique) et abstrait, ainsi que générique ou spécifique et concret. On classe comme ‘abstrait’ tous les noms de mesures, processus, types d'humains, avec un trait sensoriel. Les noms des créatures mystiques sont classés comme concrets. Les états, périodes et qualités, phénomènes et événements sont classés comme abstraits.

La notion de concrétude concerne aussi les mots qui peuvent être ressentis par l’un des cinq sens (Dove, 2016). Les mots concrets se réfèrent généralement à des concepts qui sont

spatialement et physiquement perceptibles, alors que les mots abstraits se réfèrent souvent à des concepts composés d'information sociale ou introspectif (Danguecan & Buchanan, 2016) (cf. Table 1).

Mots abstraits		**Mots concrets**	
Processus, états et periodes	confinement, espoir, mois	Perceptibles spacialement	table, arbre
Mesures et qualités	degré, gentillesse	Physiquement perceptible par l'un des cinq sens	musique, arc-en-ciel, amertume
Phénomènes et événements	conseil, soirée	Tous les êtres vivants	femme, chat
Traits d'humains	menteur, génie	Creatures mythologiques	troll, dragon

Table 1 : Typologie des noms abstraits et concrets.

Une classification binaire des mots en abstraits et concrets, cependant, reste assez subjective, premièrement, parce que chaque personne a une expérience linguistique différente, et deuxièmement, parce que dans le vocabulaire de chaque langue, il y a beaucoup de mots polysémiques qui souvent ont des significations liées à différentes catégories sur l'échelle de l'iconicité.

Même si la nature binaire d'une telle division peut sembler être un obstacle à la classification, dans cette étude, nous adhérons à une telle binarité. On suppose que si des études précédentes ont pu prouver la différence dans la perception des mots abstraits et concrets par le cerveau humain, la ligne entre l'abstrait et le concret existe dans le lexique et peut se refléter dans des caractéristiques spécifiques inhérentes au vocabulaire. En revanche, cette binarité n'est pas absolue : à la lumière des résultats de notre évaluation par des humains (cf. section 6.3) il y a une certaine gradation dans la perception de l'iconicité. Par exemple, 'gare' sera perçu comme très concret, 'signe' ou 'nation' au milieu de l'échelle, et 'manie' comme plutôt abstrait.

Il existe cependant quelques bases de données contenant des informations sur les mots abstraits. Elles reflètent généralement les résultats d'annotations humaines, contiennent moins d'un millier de mots, peu de traits sémantiques ou lexicales. (Brysbaert et al., 2014; Bonin et al., 2003). Par exemple, Ferrand et Alario ont utilisé une base de données contenant 260 mots abstraits

(Ferrand, 2001) et 366 mots concrets (Ferrand & Alario, 1998) afin de mener une expérience d'associations de mots. Ces listes de mots hors contexte avec le niveau d'iconicité ont été compilées sur la base des corpus américains et canadiens traduits et approuvés par des francophones. Une ressource comme JeuxDeMots (Lafourcade, 2007), réseau lexical de référence pour le français, ne contient pas, à ce jour, des informations de ce type.

2.3. Méthodes d'annotation de mots abstraits et concrets

Différentes tentatives de construction de listes annotées de mots abstraits et concrets sont décrites dans la littérature. Rabinovich et al. (2018) utilisent une approche faiblement supervisée pour prédire l'abstractivité des mots et des expressions en l'absence totale de données étiquetées. Ils exploitent uniquement les indices morphologiques en tant que suffixes et préfixes et l'environnement contextuel d'un mot tel qu'il apparaît dans le texte. Leurs résultats montrent que les indices proposés sont suffisamment puissants pour obtenir une forte corrélation avec les marqueurs humains. Les résultats démontrent également qu'un indice morphologique minimum et un corpus textuel sont suffisants pour fournir quelques prédictions. Les auteurs ont utilisé l'ensemble des « indicateurs d'abstractivité » en anglais, comme les suffixes *-ness, -ence, -ety, -ship* etc.

D'autres recherches en anglais montrent différents degrés de concrétude pour les formes de mots construits dans la représentation mentale. Les mots à structure opaque (‘*departement*’) peuvent être plus difficiles à catégoriser que les mots qui peuvent être facilement décomposés en une racine avec une forte signification sémantique et un morphème qui forme le dérivé (‘*happiness*’) (Marslen-Wilson et al., 2013).

Avec l'essor récent des techniques de plongement de mots (ou *word embeddings*), les méthodes de construction ont évolué permettant d'étendre automatiquement les réseaux de distribution en utilisant les informations de proximité sémantique comme vecteurs. Des études impliquant l'utilisation des algorithmes de *word embedding* pour prédire le caractère concret des mots dans une langue et entre les langues ont été proposées par Ljubešić et al. (2018). Pierrejean & Tanguy (2019) ont également étudié le problème de la stabilité du plongement des mots en fonction de l'affectation à la catégorie concrété ou abstraite. Les résultats de cette étude ont montré que la propagation de mots concrets est plus performante que la propagation de mots abstraits. Enfin, Abnar et collaborateurs (2018) ont mené des expériences en utilisant plusieurs algorithmes

pour comparer leurs performances aux résultats de l'activité cérébrale dans le but de trouver une meilleure solution pour arriver à la classification des noms en abstraits et concrets.

L'approche par plongement de mots est très puissante en TALN. Cependant, elle a des inconvénients, de même que de nombreux autres mécanismes d'apprentissage automatique, à savoir, le fait qu'il représente souvent une 'boîte noire' pour le chercheur : ce qui se passe à l'intérieur de l'opératoire de l'algorithme reste vague et limité à l'interprétation des résultats (Chen et al., 2018). Dans notre étude, nous nous intéressons non seulement à ce qui se passe après l'application d'un algorithme de TALN, mais aussi quelle est la différence entre les résultats de l'annotation automatique et du jugement humain, et pour quelle catégorie, abstrait ou concret, on peut obtenir moins de différence dans les résultats. Notre objectif est de rendre possible une propagation à partir d'une liste de mots abstraits et concrets annotée manuellement et de savoir si cette propagation fonctionne mieux pour les noms abstraits ou pour les noms concrets. Notre hypothèse est que les noms abstraits sont sémantiquement liés à d'autres noms abstraits et que les noms concrets sont sémantiquement liés à des noms concrets. On évite d'utiliser le terme 'synonymes' car les méthodes qu'on utilise dans cette étude en plus des synonymes incluent d'autres relations lexicales telles que les analogies, les antonymes et les associations de mots.

Le voisinage sémantique des mots peut être utilisé dans des algorithmes d'apprentissage automatique qui se concentrent sur la récupération de différents types d'informations sémantiques et lexicales afin d'améliorer la désambiguïsation des mots abstrait et concrets. Ces études sont souvent placées à la frontière de différents domaines scientifiques. Une étude de Hessel et al. (2018) a prouvé que les concepts concrets sont plus facilement reconnus par les algorithmes du TALN, et la méthode distributionnelle de k plus proches voisins (*k-nearest neighbors*) fonctionne mieux et est plus applicable pour les mots abstraits que pour les mots concrets. Cela peut être expliqué par l'existence de milliers d'images de mots concrets sur Internet, qui peuvent être facilement associés aux mots. Pour les mots abstraits, ces images auront une représentation moins homogène. Une autre étude (Reilly & Desai, 2017) soutient cependant que la densité de voisinage sémantique est plus élevée pour les mots concrets.

Comme on peut le voir, la variable sémantique de l'iconicité et son impact sur le comportement des mots associés restent non étudiés. Les recherches en sciences cognitives et en linguistique informatique tentent de faire un parallèle entre les schémas de traitement en cerveau et les algorithmes artificiels et statistiques. Hultén et al. (2018) ont réussi à montrer que le décodage neuronal du sens abstrait ou concret des mots est fondé sur le système cognitif verbal à

travers les régularités de l'usage et peut être recréé en utilisant des algorithmes de calcul mesurant ces régularités.

Dans la suite du mémoire, nous allons présenter différentes expériences visant à :

1) caractériser le lexique abstrait et concret ;
2) observer la dépendance de la structure morphologique des mots et leur niveau de l'abstractivité ;
3) créer une base de données des noms abstraits et concrets ou ce trait soit explicit.

Notre première expérience vise à observer l'impact de la fréquence des noms polysémiques sur la reconnaissance de ces noms comme abstraits ou concrets (Goriachun, 2019a). 36 stimuli abstraits de haute et basse fréquence et 36 stimuli concrets de haute et basse fréquence ont été placés dans les contextes cohérents et présentés aux participants sous la forme de questionnaire. La tâche des participants était de classer les noms en gras (stimuli) comme abstraits ou concrets. La deuxième expérience vise à tester l'hypothèse de la dépendance de la structure morphologique des noms sur leurs niveau d'abstractivité. Nous avons établi le questionnaire consistant en 10 mots abstraits construits avec 10 synonymes simples et 10 mots concrets construits avec 10 synonymes simples et 10 mots très concrets simples. Le questionnaire avec les stimuli sans contexte a été présenté à des juges humains sous la forme de Google Form et comprenait 4 choix pour le classement : abstrait, plutôt abstrait, plutôt concret et concret (Goriachun, 2019b). La troisième expérience vise à construire automatiquement la base des noms abstrait et concrets à partir de la liste initiale de 61 noms. L'expérience comprend deux étapes : la propagation des informations d'une liste initiale annotée manuellement à l'aide des méthodes distributionnels (voisins distributionnels et cooccurrences syntaxiques) et l'évaluation du résultat par comparaison avec des jugements humains. Utilisation de deux méthodes distributionnelles nous a permis de tester l'hypothèse de la différence dans l'organisation des réseaux sémantiques des noms abstraits et concrets. Nous avons pu identifier la pertinence des méthodes dans la tâche d'annotation automatique selon la catégorie abstraite/concrète. La deuxième étape consistait d'évaluer l'échantillon de 120 noms (60 noms concrets (30 voisins distributionnels et 30 cooccurrences syntaxiques) et 60 abstraits (30 voisins distributionnels et 30 cooccurrences syntaxiques). L'évaluation a été menée en ligne avec 1083 participants au total. Tous les participants ont eu pour a tache de classer les stimuli sans contexte selon l'échelle glissière de -100 (très abstrait) a 100 (très concret) (Goriachun et Gala, 2020.).