

5. Expérience 3. Annotation automatique du lexique en traits abstrait et concret

Après avoir mené les deux expériences précédentes, nous sommes arrivés à la conclusion que des variables telles que la fréquence, la longueur du mot, la position du mot dans la phrase, ainsi que la structure morphologique ne corrélaient pas avec le trait abstrait/concret (Goriachun, 2019a). Ces variables formelles n'étaient pas significatives, nous nous sommes alors tournés vers la variable « iconicité ». Ainsi, en fondant nos hypothèses sur des théories liées à la différence dans l'organisation des cadres sémantiques des mots abstraits et concrets (Crutch & Warrington, 2005), nous avons décidé de mener des recherches sur la possibilité d'annoter automatiquement des noms en fonction de leur niveau d'iconicité.

Cette notion n'était pas explicitement encodée dans les ressources existantes. Nous avons décidé d'annoter une liste restreinte initiale avec un petit nombre de mots annotés manuellement, dont on connaît le niveau d'iconicité (Ferrand, 2001; Ferrand et Alario, 1998) et de propager cette annotation à des mots liés par des relations syntaxiques (cooccurrences syntaxiques) et sémantiques (voisins distributionnels) Ainsi nous nous sommes fixés pour objectif de créer une

base de données contenant des noms français annotés automatiquement sur la base de deux méthodes distributionnelles en faisant la propagation du trait abstrait/concret d'un mot vers ses voisins distributionnels et cooccurrences syntaxiques.. Nous visons à proposer une méthode fiable validée par une évaluation humaine importante de plus de 1000 participants pour une annotation automatique qui permet de créer une base de mots annotés de taille importante.

Les cooccurrences syntaxiques sont les mots qui apparaissent fréquemment à côté du mot cible ; les voisins distributionnels sont les mots qui partagent les mêmes contextes (van der Plas, 2009). Nous avons décidé de fonder nos recherches sur la base de ces deux méthodes car elles montrent deux relations distinctes : des liens sémantiques dans le cas des voisins distributionnels et une proximité syntaxique dans le contexte dans le cas de cooccurrences syntaxiques. Ceci peut être crucial pour la détermination et la distinction des mots abstraits et concrets. Par exemple, les mots 'plante' et 'fleur' sont les voisins distributionnels du mot concret 'arbre', tandis que 'branche' et 'ombre' sont ses cooccurrents syntaxiques, c'est-à-dire ils peuvent apparaître dans les mêmes contextes. 'Inquiétude' et 'peur' sont les voisins distributionnels du mot abstrait 'crainte', tandis que 'dissipation' et 'reflet' sont ses cooccurrents syntaxiques.

Par évaluation des résultats de l'annotation automatique nous avons voulu établir quel est le pourcentage d'unités de l'échantillon obtenues à partir de la méthode des voisins distributionnels et des cooccurrences syntaxiques correspond mieux aux résultats de l'annotation humaine.

Nous nous sommes également intéressés aux différences dans la précision de la prédiction entre les mots abstraits et concrets et aux différences de taille des réseaux sémantiques des mots abstraits et concrets, s'il y en avait. Crutch & Warrington (2005) suggèrent que les mots concrets sont organisés selon un principe de similarité sémantique, tandis que les mots abstraits sont organisés par leur association avec d'autres mots. Si on trouve une prévalence quantitative et plus grande homogénéité dans les résultats pour les mots abstraits ou concrets lors de l'extension de la liste principale ou lors de l'évaluation humaine, il sera possible d'indiquer quel modèle est plus pertinent de recréer à l'aide des outils du TAL.

5.1. Données

Pour créer notre liste de mots annotés automatiquement en utilisant les voisins distributionnels et les cooccurrences syntaxiques, nous avons d'abord créé une liste initiale de mots contenant des noms abstraits et concrets. Pour cette tâche, nous avons utilisé les mots issus de deux

études (Ferrand, 2001; Ferrand et Alario, 1998) qui contiennent les données annotées pour la langue française selon des échelles d'iconicité et d'abstractivité. Les listes des mots ont été réduites aux noms de haute fréquence et monosémiques.

Notre liste initiale contient 19 noms abstraits (Ferrand, 2001) et 42 noms concrets (Ferrand et Alario, 1998) avec un score de fréquence ≥ 40 (selon la base de données lexicale Lexique 3⁴). Les noms abstraits sont souvent monosémiques selon la ressource lexicale ReSyf (Billami et al., 2018). Les noms annotés comme très concrets et avec un indicateur haute fréquence dans Ferrand et Alario (1998) étaient souvent polysémiques. Pour cette raison, nous avons décidé d'utiliser des mots concrets qui n'ont pas de signification abstraite.

5.2. Méthodologie

Par la suite, nous avons extrait manuellement de la base de données distributionnelle *Les Voisins de Le Monde*⁵ 50 unités lexicales pour chaque nom. L'ensemble de données expérimental initial a été réduit à seulement 50 voisins les plus proches car nous avons identifié qu'après ces 50 premiers voisins les relations sont devenues trop distancées (selon le score de distance fourni par *Les Voisins De Le Monde*). Le type de relation distributionnelle était limitée aux voisins distributionnels car, à cette étape, cette méthode montrait une meilleure cohérence pour notre tâche. La sortie de la méthode des cooccurrences syntaxiques comprenait beaucoup des non noms.

Après les deux premières étapes, nous avons obtenu une liste complète composée de 369 unités (180 noms concrets et 189 abstraits) (cf. Table 4). Les deux listes ont été nettoyées des répétitions et des non-noms. Nous avons aussi supprimé de la liste des mots abstraits les noms avec un niveau élevé d'iconicité (par exemple, le mot 'chat'). De même, si dans la liste des mots concrets figurait le mot 'bonheur', il a été supprimé en raison de son haut niveau d'abstractivité. Cette expérience nous a également permis d'identifier la première différence entre les listes de noms abstraits et concrets : un nombre réduit de voisins des mots concrets, ainsi qu'un grand nombre de non-noms et de noms abstraits dans la liste des mots concrets (c'est pour cette raison que nous avons dû étendre la liste initiale à 42 noms concrets afin que le nombre de voisins pour

⁴ Lexique 3. Consulté le 19 février 2020, à l'adresse <http://www.lexique.org/>

⁵ Les Voisins De Le Monde. Consulté le 4 février 2020, à l'adresse <http://redac.univ-tlse2.fr/voisinsdelemonde/>

la liste expérimentale de noms concrets (180) soit approximativement égale à la longueur de la liste abstraite (189 noms)).

Les mots abstraits ont montré plus de cohérence en termes de catégorie grammaticale et de catégorie abstrait/concret : les unités supprimées étaient principalement des répétitions.

Catégorie	Abstrait	Concret
Listes initiales (Ferrand and Alario, 1998; Ferrand, 2001)	19	42
Avant le traitement manuel	910	1638
Après traitement (données expérimentales)	189	180

Table 5. Listes initiales obtenues manuellement à partir de la ressource *Les Voisins de Le Monde*.

L'étape suivante a été d'extraire automatiquement, pour chacune des 369 unités, leurs voisins distributionnels et cooccurrences syntaxiques à partir de la ressource *Les Voisins de Le Monde*. Nous avons par la suite comparé les résultats obtenus avec chaque approche.

5.3. Résultats

À partir des listes initiales presque égales en nombre (180 mots abstraits et 189 mots concrets), nous avons obtenu une sortie quantitativement différente : 62.174 mots abstraits et 31.333 mots concrets, ce qui signifie que le nombre des voisins distributionnels et les cooccurrences syntaxiques des mots concrets comprend la moitié du nombre des voisins distributionnels et cooccurrences syntaxiques des mots abstraits. Après avoir éliminé toutes les répétitions, nous avons obtenu 4.222 mots concrets uniques et 3.676 mots abstraits uniques (cf. Table 6), ce qui nous permet de vérifier la propriété selon laquelle les mots abstraits ont tendance à apparaître dans des contextes plus similaires. En d'autres termes, un mot abstrait aléatoire X est plus susceptible d'avoir un autre mot abstrait aléatoire Y comme voisin distributionnel (ou comme cooccurrent syntaxique) que deux mots concrets aléatoires Z et W, qui auront la possibilité d'apparaître dans le même réseau sémantique.

Catégorie	Abstrait	Concret
Données brutes	62.174	31.333
Données traitées	3.676	4.222

Table 6. Nombre de mots annotés Abstrait / Concrets.

Une différence entre les voisins distributionnels et les cooccurrences syntaxiques a également été observée. Pour les noms concrets, la sortie des voisins et des cooccurrences est presque égale, mais pour les noms abstraits, ces nombres sont très différents (cf. Table 7). Ce résultat montre que les voisins distributionnels (VD) s'avère être la méthode qui fonctionne le mieux pour les mots abstraits, et les cooccurrences syntaxiques (CS) est la méthode qui fonctionne mieux pour les mots concrets. Ce résultat confirme une hypothèse de différences de représentations sémantiques entre les mots concrets et abstraits.

Catégorie	Abstrait	Concret
Données brutes VD	45.340	16.223
Données brutes CS	16.834	15.110
Données traitées VD	2.129	1.631
Données traitées CS	1.546	2.592

Table 7. Nombre de mots Abstrait / Concrets obtenus après propagation des annotations à partir de la liste expérimentale (VD – voisins distributionnels, CS – cooccurrences syntaxiques).

Il est important de signaler que la polysémie était très forte dans les données collectées : beaucoup de mots dans la liste des mots abstraits existaient dans la liste des mots concrets et vice versa. Le phénomène de la polysémie est inévitable dans le lexique et on pouvait s'attendre à ce que nous l'ayons aussi dans la distinction des mots abstraits et concrets. Les mots qui apparaissent uniquement dans une liste abstraite ou concrète après l'application de la méthode distributionnelle, sont indiqués ci-dessous (cf. Table 8).

Catégorie	Abstrait	Concret
VD	551	400
CS	654	1.352
Total	1.205	1.753

Table 8. Nombre des mots Abstrait et Concrets uniques.

Comme on peut le constater, le nombre des mots concrets uniques est plus important que le nombre des mots abstraits uniques. Le nombre des cooccurrences syntaxiques est beaucoup plus grand que le nombre des voisins distributionnels pour les mots concrets. Dans les mots abstraits cette différence vers la plus grande quantité des cooccurrences syntaxiques est moins importante, ce qui confirme les résultats précédents sur les différences de représentations sémantiques entre

mots concrets et abstraits selon les méthodes distributionnelles appliquées. Cependant, on suppose que les mots uniques peuvent avoir une fréquence plus basse que les mots dans les données traitées.

5.4. Évaluation

5.4.1. Stimuli et procédure

Nous avons voulu évaluer les annotations obtenues automatiquement à partir de *Les Voisins De Le Monde* par rapport à des annotations humaines. Nous avons donc créé un échantillon de 120 noms de la liste initiale : 60 noms concrets (30 voisins distributionnels et 30 cooccurrences syntaxiques des mots de la liste initiale de 180 noms concrets) et 60 abstraits (30 voisins distributionnels et 30 cooccurrences syntaxiques des mots de la liste initiale de 189 noms abstraits), que nous avons soumis à des juges (Annexe 14). L'échantillon a été choisi au hasard dans les données globales pour éviter les biais d'échantillonnage. Nous avons alors créé, à l'aide de C. Zelinski du Centre de Ressources Expérimentales de l'ILCB (CREX⁶) un questionnaire en ligne qui a été distribué parmi les étudiants et le personnel de l'université d'Aix-Marseille via le réseau universitaire (cf. Figure 2). Pour cette expérience nous avons utilisé une échelle de plus grande amplitude, modélisée sous la forme d'une glissière (slider), au lieu de l'échelle de Likert utilisée dans les deux expériences précédentes (cf. Figure 3). Ce choix s'explique par le fait que les attitudes des participants pour un stimulus particulier ont la forme d'un continuum multidimensionnel. L'échelle de Likert est unidimensionnelle et donne seulement quelques options de choix réduites. De même, souvent, les participants évitent de choisir les options « extrêmes » sur l'échelle, en essayant de rester au milieu ce qui est encore plus probable dans la tâche de décision sémantique ou les stimuli causent beaucoup d'hésitation. Dans cette expérience les participants ont dû annoter chaque mot selon une échelle graduée entre -100 (très abstrait) et 100 (très concret). Le test a été lancé le 26/03 à 17h. A 21h plus de mille participants y avaient répondu. Nous avons donc été capables de récolter 1083 réponses en quatre heures.

Les résultats de l'annotation ont été soumis à une analyse statistique avec le logiciel R. Tous les choix inférieurs à 0 ont été considérés comme des choix vers la notion d'abstrait et tous

⁶ ILCB | Institute of Language, Communication and the Brain. Consulté le 28 mars 2020, à l'adresse <https://www.ilcb.fr/about/crex/>

les choix supérieurs à 0 sont considérés comme des choix vers le concret. Les mots au milieu de l'échelle de -10 à 10 ont été considérés comme très ambigus. Cette échelle nous permet d'observer le degré d'abstractivité d'un mot selon le jugement humain et de le comparer avec le résultat de l'annotation automatique. Dans les consignes pour les participants au début de l'épreuve il était conseillé de ne pas utiliser trop souvent le choix au milieu de l'échelle. Nous avons fait l'hypothèse que l'obtention d'une valeur moyenne indiquera un fort degré de polysémie. Par ailleurs, les participants qui ont trop souvent validé une réponse avec le *slider* au milieu, ont été éliminés de l'étude (nous avons considéré que les réponses étaient invalides).

Figure 2. Page d'accueil d'expérience en ligne.

Figure 3. Présentation du stimulus.

5.4.2. Participants

1.083 participants ont passé l'expérience. Leur âge varie de 18 à 75 ans, âge moyenne = 26 ans. Nombre de femmes = 687, nombre d'hommes = 396. Les niveaux d'études des participants sont très divers de Bac à Bac+8 et BEP. 599 participants ont étudié d'autres langues étrangères et 74 participants n'étaient pas des francophones natifs ce qui n'a pas eu l'impact sur les résultats (voir partie C de la section suivante).

5.4.3. Analyse statistique et résultats

Le but de notre analyse statistique était de comparer la sortie obtenue par l'annotation automatique et le jugement humain. Premièrement, nous voulions savoir si les noms annotés par l'une des deux méthodes distributionnelles comme abstraits ou concrets appartiennent ou non à ces catégories selon l'annotation des participants. Deuxièmement, notre but était d'identifier quelle méthode (voisins distributionnels ou cooccurrences syntaxiques) donne les résultats plus précis dans la tâche de l'annotation du lexique abstrait et concret.

Nos stimuli ont été choisis pour représenter chaque catégorie et chaque méthode. Au début c'étaient 30 voisins distributionnels abstraits, 30 cooccurrences syntaxiques abstraites, 30 voisins distributionnels concrets, 30 cooccurrences syntaxiques concrètes.

Étant donné que certains mots de la catégorie des noms abstraits se trouvent également dans la catégorie concrète, il existe pour eux deux options. Par exemple, le nom 'mairie' est le voisin distributionnel du nom concret 'banque' mais en même temps c'est le voisin distributionnel du nom abstrait 'institution' ; 'milieu' est le cooccurrent syntaxique du nom concret 'salle' mais aussi le cooccurrent syntaxique du nom abstrait 'origine'. Certains stimuli ne sont pas ambigus, ils apparaissent dans une seule catégorie. Par exemple, le nom 'orge' a été annoté dans la catégorie des voisins distributionnels du nom concret 'betterave'. On a décidé de ne pas se débarrasser de ce type d'ambiguïté et de voir si les participants hésiteraient aussi pendant l'annotation des stimuli qui apparaissent aux deux catégories.

A. Analyse de la précision des méthodes distributionnelles

La première étape était de comparer les résultats obtenus par l'annotation par le jugement humain avec la sortie obtenue après l'annotation automatique. Comme c'était mentionné ci-dessus, les stimuli ont été également distribués dans chaque catégorie.

Pour convertir les données reçues de l'annotation nous avons calculé les moyennes pour chaque stimulus en se basant sur les résultats de toutes les 1.083 passations. Tous les stimuli avec le niveau moyen supérieur à zéro ont été considéré comme concrets et tous les stimuli avec le moyen inférieur à zéro comme abstraits (cf. Figure 4).

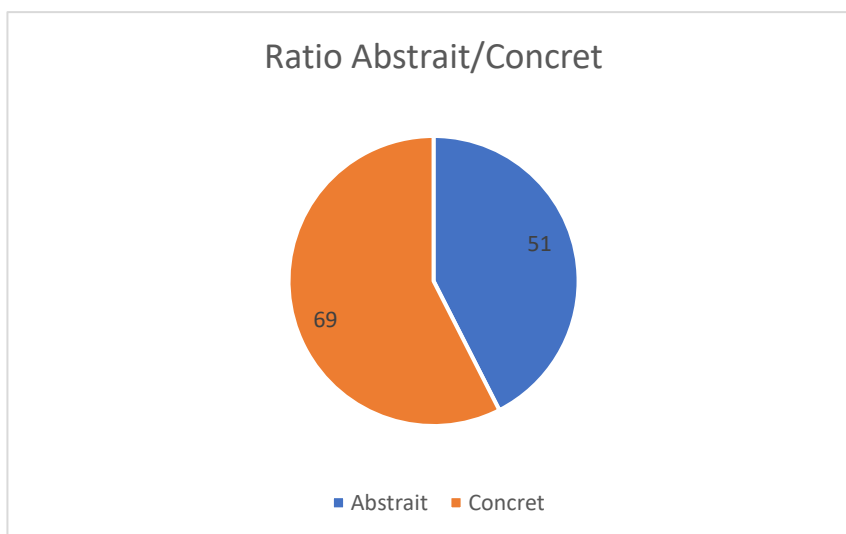


Figure 4. Ratio des noms abstrait et concrets dans les résultats d'annotation.

Catégorie	Abstrait	Concret
VD	21 (70%)	25 (83,3%)
CCS	15 (50%)	19 (63,3%)

Table 9. Nombre de réponses correspondant au sorti et le niveau de précision

En utilisant la formule Kappa de Fleiss (1981), nous avons calculé l'accord entre les annotateurs qui est égal à 0.256 (un accord faible mais attendu dans une expérience de type décision sémantique lexicale avec une grande échelle de -100 à 100). La moyenne des écarts types de stimuli était égale à 56 ce qui confirme une fois de plus la difficulté de la tâche. Les stimuli avec les écarts types moins grands (<40) étaient tous les noms concrets, parmi les stimuli avec les écarts types plus grands (>65) il y avait des noms abstraits et des noms concrets polysémiques. La dépendance linéaire entre les valeurs moyennes des stimuli (leurs niveaux de concrétude) et les écarts types correspond à une corrélation forte ($r = -0.6210$) ce qui signifie que plus grand est le degré de concrétude, plus bas est la valeur des écarts types. En général, plus un mot est concret, moins les participants hésitent lors de l'annotation.

Partant du fait que la tâche de décision sémantique est complexe et ambiguë même pour un humain, nous avons obtenu de meilleurs résultats en utilisant la méthode des voisins

distributionnels. Pour les deux méthodes, les mots concrets ont donné de meilleurs résultats que les mots abstraits, comme le montre la Table 10.

Catégorie	Abstrait	Concret	Total
Voisins Distrib.	21 de 30	25 de 30	46
Précision	70%	83,3%	77 %
Cooccs. Syntaxiques	15 de 30	19 de 30	34
Précision	50%	63,3%	57 %

Table 10. Nombre de correspondances jugement humain / annotation automatique et score de précision.

B. Analyse de l'impact de temps de réaction

La plateforme utilisée pour le test nous a également permis de mesurer le temps de réaction des participants. Cependant, notre méthode pour la mise en place du test de la décision lexicale était différente de la méthode habituelle en raison de la présence d'une échelle avec un curseur et un bouton pour confirmer la réponse. Les données obtenues sur le temps de réaction sont très différentes des données fournies par la base Open Lexicon⁷ de la ressource Lexique. Par exemple, le temps de réaction pour le mot 'abri' est égal à 570.13 ms dans l'Open Lexicon, dans notre expérience on a obtenu la moyenne du temps de réaction pour le mot 'abri' égal à 3570.064 ms. Nous avons calculé les valeurs de la corrélation entre les variables : moyennes et écarts types et la variable de temp de réaction. La corrélation entre les moyennes et les temps de réaction est faible ($r = -0.216$) ainsi que la corrélation entre les écarts types et les temps de réaction ($r = 0.251$), ceci nous permet de dire que dans notre cas les participants n'ont pas répondu plus ou moins vite selon la catégorie abstraite ou concrète d'un mot ou la valeur des écarts types.

C. Analyse de l'impact de la langue maternelle et des troubles du langage

Dans le questionnaire que les participants ont rempli après l'expérience, ils ont indiqué leur langue maternelle et signalé des éventuels troubles du langage oral ou écrit, le cas échéant. Comme un nombre suffisamment important de participants, 141 personnes, ont déclaré qu'ils avaient des problèmes liés au langage oral ou écrit, nous avons décidé d'analyser ces données séparément pour comprendre si les résultats obtenus pour ces participants pouvaient affecter l'image globale. Après nos analyses ces tests n'ont pas montré des différences significatives qui pourraient nous faire

⁷ Open Lexicon. Consulté le 30 mars 2020, à l'adresse <http://www.lexique.org/shiny/openlexicon/>.

comprendre que les données de personnes ayant des troubles du langage auraient une incidence sur la totalité des données (moyennes globales et moyennes troubles du langage : $t = 0,10436$, $p = 0,917$; écarts types globales et écarts types troubles du langage : $t = -1,0765$, $p = 0,2828$; temps de réaction globales et temps de réaction troubles du langage : $t = 0,23679$, $p = 0,8131$).

Le groupe de participantes non-francophones natifs était moins nombreux, 74 personnes. Les résultats ont été également comparés avec ceux des francophones natifs. Les différences significatives n'ont pas été observées entre les moyens ($t = -0.2010$, $p = 0.8409$), écarts types ($t = 0.971$, $p = 0.3321$) et temps de réaction ($t = 1.6954$, $p\text{-value} = 0.092$) de deux groupes.

D. Analyse d'effet de l'ambiguïté abstrait/concret

En général, l'ambiguïté abstrait/concret que nous avons pu observer à la suite de l'implémentation des méthodes distributionnelles est confirmée par les résultats des jugements humains. Les stimuli dans les catégories dont les personnes doutaient appartenaient souvent à deux groupes sémantiques, abstraits et concrets, selon les résultats de l'annotation automatique. Comme dans Kwong (2013) nos données confirment l'hypothèse qu'en cas de polysémie, une personne choisit une signification concrète, plutôt qu'abstraite. Par exemple, les mots 'cadre', 'échelle', 'cote', 'espèce', 'réserve', 'secours' ont été classés comme concrets.

E. Analyse de l'impact de la fréquence des stimuli

Puisque nos stimuli ont été pris au hasard à partir de données annotées automatiquement, leurs fréquences sont différentes. Parmi nos stimuli, il y en a qui sont très rares selon la ressource Lexique 3 et d'autres qui sont fréquents en français. Une analyse de covariance a été effectuée pour confirmer que les participants ont choisi la catégorie d'un mot avec ou sans beaucoup d'hésitation indépendamment de sa fréquence⁸. La fréquence d'un stimulus n'a pas eu l'impact sur les trois variables : moyens ($r = -0.0039$), temps de réaction ($r = -0.041$) et écarts types ($r = -0.0901$).

5.4.4. Discussion

Au cours de cette expérience, nous avons effectué deux étapes. La première étape est l'expansion automatique d'une liste de 20 noms abstraits et 42 noms concrets (à partir d'une première vérification manuelle sur un nombre réduit de mots). La deuxième étape consiste à vérifier les résultats de l'annotation automatique à l'aide d'une expérience de comparaison avec des jugements humains.

⁸ Lexique.org Consulté le 7 mars 2020, à l'adresse <http://www.lexique.org/>