Les couleurs des gens

Mathieu Lafourcade, Nathalie Le Brun, Virginie Zampa

Les couleurs des gens

Mathieu Lafourcade1, Nathalie Le Brun2, Virginie Zampa3
(1) Lirmm, Université Montpellier 2, France
(2) Imagin@t, 34400 Lunel
(3) Lidilem, Grenoble 3 BP25, 38040 Grenoble cedex 9, France
mathieu.lafourcade@lirmm.fr, imaginat@imaginat.name, virginie.zampa@u-grenoble3.fr

Résumé. En TAL et plus particulièrement en analyse sémantique, les informations sur la couleur peuvent être importantes pour traiter correctement des informations textuelles (sens des mots, désambiguïsation et indexation). Plus généralement, connaître la ou les couleurs habituellement associée(s) à un terme est une information cruciale. Dans cet article, nous montrons comment le *crowdsourcing*, à travers un jeu, peut être une bonne stratégie pour collecter ces données lexico-sémantiques.

Abstract. In Natural Language Processing and semantic analysis in particular, color information may be important in order to properly process textual information (word sense disambiguation, and indexing). More specifically, knowing which colors are generally associated to terms is a crucial information. In this paper, we explore how crowdsourcing through a game with a purpose (GWAP) can be an adequate strategy to collect such lexico-semantic data.

Mots-clés: association couleur-mot, réseau lexical, crowdsourcing.
Keywords: Word Color Associations, Lexical Network, Crowdsourcing

1 Introduction

La couleur, en tant qu'élément de notre vie quotidienne, est une donnée intéressante en Traitement Automatique du Langage Naturel. En effet, fournir des informations relatives aux associations mots-couleurs, en plus des données classiques (hyperonymes, parties de, rôle sémantique, etc.) à un système dédié à l'analyse sémantique de textes, pourrait en améliorer grandement les performances. Bien que ce soit une observation marginale par rapport au sujet de cet article, il existe indiscutablement une très forte liaison entre couleurs et émotions. Ainsi, si on relève souvent des associations entre des couleurs et des termes abstraits relatifs à des émotions (comme la crainte, la colère, le danger, l'espoir, etc.), c'est encore plus vrai pour les termes désignant des choses concrètes (comme le ciel, un lion, la mer, la neige, etc.).

En psychologie, de nombreuses études concernent les associations entre mots et couleurs et leur impact sur la communication (verbale ou non-verbale). Saif (2011a) montre que la notion de couleur est d'une importance capitale pour la qualité du message délivré, que ce soit en marketing (Sable and Akcay, 2010), en conception de sites web (Meier, 1988; Pribadi et al., 1990), ou pour caractériser visuellement une information (Christ, 1975; Card et al.,1999). La couleur est indiscutablement une donnée très porteuse de sens. Mais pour de nombreux chercheurs, les significations attribuées aux couleurs dépendent de plusieurs facteurs. Luscher (1969), psychothérapeute auteur du test éponyme, le *Lüscher color test*, un outil qui permet d'évaluer l'état émotionnel d'une personne en fonction de ses préférences de couleurs, affirme que le choix de telle ou telle couleur par une personne dépend directement de (voire traduit) ses émotions et son état psychique. Child et al. (1968), et Ou et al. (2011) montrent que les préférences en matière de couleurs varient en fonction de l'âge et du sexe. De plus, pour certaines couleurs, on rencontre des variations lexicales en fonction des langues et des cultures (par exemple, en turc et en hongrois, il existe deux mots différents pour *rouge*).

Savoir si la relation établie entre couleur et sens est indépendante de l'âge, du sexe, ou de la nationalité, reste une question très controversée qui fait actuellement l'objet d'un vif débat. Cela pourrait être le cas, par exemple, pour la relation entre la couleur *rouge* et la notion de *danger*, étant donné que le *rouge* est associé aux idées de *sang* et de *feu* dans de nombreuses langues/cultures. Berlin and Kay (1969) disent que les différences pourraient être classées hiérarchiquement, et qu'on retrouve un nombre limité de termes de couleurs de base dans différentes cultures. Cette analyse est issue d'une comparaison des mots désignant des couleurs dans 20 langues, mais elle est particulièrement controversée, en particulier concernant la méthodologie employée pour la collecte des données. Dans un même ordre d'idée, on pourrait remarquer que de nombreuses expressions faisant référence à une couleur ont le même sens dans différentes langues, *a fortiori* si elles sont culturellement proches, ce qui n'est guère surprenant. Par exemple, *dark thoughts* en anglais et *idées noires* en français ont des sens pratiquement équivalents, de même que *to see red* et *voir rouge*. Cependant, il s'agit évidement d'exemples issus de langues tellement proches linguistiquement et culturellement, que cela ne peut guère valider une quelconque universalité des associations possibles entre lexiques et couleurs.

De nombreuses études, principalement en anglais, visent à établir des relations entre couleurs et mots, et en particulier couleurs et émotions. La plupart sont effectuées d'une manière classique en psycholinguistique, et leurs données brutes sont généralement de taille modeste et non librement accessibles. De plus, comme nous l'avons souligné précédemment, il est extrêmement délicat (et probablement imprudent) d'extrapoler directement le résultat de telles études d'une langue à une autre. Ainsi, il n'est actuellement pas possible d'aboutir à un consensus sur l'universalité des associations de mots et de couleurs quand il s'agit de sentiments, ou du moins de termes abstraits. Saif (2011b), au terme de diverses expériences sur un échantillon de 11 couleurs, a conclu que plus de 30% des termes issus d'un dictionnaire de vocabulaire courant étaient fortement liés à une couleur. Environ 33% des catégories de thésaurus (comme Roget) révèlent des associations de couleurs, et les termes abstraits sont associés à des couleurs (principalement de façon métaphorique) presque aussi souvent que ceux désignant des objets, des entités physiques. De même, Nijdam (2010) compare différents modèles issus de la recherche sur les liens entre couleurs et émotions, et propose une correspondance. Il conclut que si certains modèles de signification des couleurs ont des points communs et se recoupent partiellement, ils témoignent aussi d'une forte variabilité, sans doute parce que la couleur est appréhendée différemment suivant la situation affective/culturelle du sujet. En résumé, l'acquisition de données lexicales sur les associations mots-couleurs se heurte à une forte variabilité, même au sein d'une même langue: plus le terme est abstrait, plus on doit s'attendre à une variabilité des couleurs qui lui sont associées.

Ozbal et al. (2011) sont à l' origine d'une ressource qui répertorie des associations mots-couleurs en anglais. Sur une sélection de 200 mots (un sous-ensemble des mots utilisés dans Grefenstette (2005)), ils ont utilisé trois méthodes automatisées pour comparer leur ressource avec les associations faites par le service Amazon Mechanical Turk (10 annotateurs-11 couleurs): analyse d'images, modèles de langages sur les données du web, et similarité entre mots et couleurs en utilisant LSA. Actuellement, pour autant que nous le sachions, Ozbal et al. (2011) sont les seuls à avoir tenté de réaliser une ressource lexicale d'associations mots-couleurs en anglais. Ce type de ressource est quasiment inexistant dans d'autres langues, et notamment, il n'existe rien de similaire en français. Cela dit, comme l'a fait remarquer Grefenstette (2005), *s'il y a une caractéristique qu'en général les gens connaissent à propos des choses (concrètes ou même abstraites), c'est bien leur couleur typique*. C'est pourquoi cette information est généralement absente des dictionnaires et autres ressources lexicales, alors qu'elle serait d'un intérêt majeur pour de nombreuses applications informatiques, et en particulier en TAL.

L'information de couleur peut être très utile dans le cadre de la désambiguïsation lexicale automatique (*Word Sense Disambiguation – WSD*). Par exemple, le mot *tissu* est polysémique et peut signifier –entre autres- *étoffe* ou *tissu vivant*. Si, dans un texte, le mot *tissu* est associé à *bleu*, cette information va pouvoir aider à choisir la signification appropriée, ou du moins permettre d'éliminer celle(s) qui est (sont) non pertinente(s). L'information relative à la couleur est également précieuse dans la situation où on cherche à retrouver un mot momentanément oublié mais qu'on a *sur le bout de la langue*, à l'aide d'une application informatique à laquelle on fournit des indices (Lafourcade 2012 et Joubert 2012).

L'objectif de cet article est de montrer comment on peut générer efficacement et relativement vite une ressource lexicale d'associations entre couleurs et mots : c'est pour acquérir ce type de données que nous avons mis au point un GWAP (Game With A Purpose selon le concept défini par L. von Ahn. (2006)), baptisé ColorIt (accessible à http://www.jeuxdemots.org/colorit.php). Dans un premier temps nous évoquons l'intérêt que peut avoir l'information relative à la couleur dans le cadre de la *WSD*, puis nous exposons rapidement le fonctionnement du jeu ColorIt et poursuivons par une analyse détaillée des résultats obtenus.

2 ColorIt, un jeu pour associer des couleurs et des mots

Le but de ColorIt est de collecter des associations spontanées entre couleurs et termes, qu'il s'agisse de couleurs attribuées à des objets concrets, ou associées symboliquement et subjectivement à des verbes ou noms désignant des entités abstraites.

2.1 Intérêt des associations mots-couleurs pour la désambiguïsation lexicale

L'un des intérêts majeurs d'une ressource lexicale rassemblant des associations entre mots et couleurs est son exploitation dans le contexte de la désambiguïsation lexicale automatique (*WSD*). En effet, dans un texte donné, la précision apportée par l'indication de couleur est souvent le détail qui va permettre de sélectionner le sens approprié d'un terme polysémique. Dans de nombreux cas, l'association à une couleur spécifique est importante. Bien sûr, la technique a ses limites : il existe des mots pour lesquels l'information de couleur est sans objet (*sans couleur*), d'autres auxquels on peut associer de nombreuses couleurs (comme *voiture*), ou qui peuvent être simultanément de plusieurs couleurs (noir et blanc pour un *pingouin*, multicolore pour un *arc-en-ciel*). Sans prétendre à l'exhaustivité, nous donnons dans le tableau ci-après quelques mots pour lesquels l'information de couleur peut aider à lever l'ambiguïté résultant de la polysémie (les sens que l'on peut éliminer sont marqués par *, les ambiguïtés qui subsistent par ?). Les résultats présentés ci-dessous sont ceux obtenus par désambiguïsation endogène d'un terme associé à un contexte (la couleur) sur les données du réseau lexical JeuxDeMots.

lit blanc	lit > couche: blanc, rouge, * lit>rivière	**regard bleu/azur** **regard noir**	regard> expression (symbolique) * regard>trou
trompe grise **trompe dorée**	trompe>éléphant *? trompe>instrument de musique * trompe>anatomie humaine	**robe rose / bleu** **robe blanche /jaune** **robe pie/alezane**	robe>vêtement * robe>pelage (animal) ? ambigu pour certaines couleurs
bas noir	bas> sous-vêtement * bas>partie inférieure	**tableau noir**	tableau>école ? tableau>œuvre d'art (improbable)
manteau rouge	manteau>vêtement * manteau>géologie	**feuille blanche** **feuille rouge/jaune/vert**	feuille>papier (le plus probable) *? feuille>arbre (improbable) ? ambigu pour certaines couleurs
ours blanc/brun **ours rose**	ours>animal / peluche ours>peluche * ours>imprimerie	**canapé bleu**	canapé>meuble * canapé>petit four ? ambigu pour certaines couleurs
piste blanche/ verte / rouge	piste>ski (symbolique) *piste>chemin	**brioche dorée**	brioche>viennoiserie *? brioche>ventre
langue blanche / rose	langue>organe * langue>langage	**savon noir / rose / blanc**	savon>savonnette *savon>réprimande
sucre blanc/roux **sucre gris**	sucre>substance *? sucre>électricité (ambigu mais improbable)	**culotte bleu / rouge/ blanche**	culotte>sous vêtement * culotte>boucherie * culotte>défaite
costume anthracite	costume>tenue (le plus probable) *? costume>déguisement	**motif fleuri**	motif>dessin *motif> raison

Table 1: Dans environ 70 % des cas, l'information de couleur permet de déduire le sens convenable et d'éliminer les autres sans ambiguïté. Parfois, une incertitude demeure, mais un niveau de confiance peut être calculé.

2.2 Principe du jeu

Un terme est proposé au joueur avec un choix réduit de couleurs *via* une palette. Il doit cliquer sur la couleur qu'il associe au mot. Il peut également cliquer sur *sans couleur* ou "passer" (s'il ne connait pas le mot par exemple). Passer n'est pas pénalisant. Enfin, si aucune des couleurs de la palette ne convient, ou s'il souhaite en mettre plus d'une, il peut la ou les saisir dans le champ de texte. Le terme présenté peut être ambigu (*feuille)* ou contextualisé (*feuille>papier* / *feuille>plante*), le réseau JDM contenant pour chaque terme polysémique ses raffinements.

Une fois que le joueur a validé son choix, une notification apparait avec son score ainsi que les réponses données par les autres joueurs pour ce mot. Le score dépend de l'adéquation entre la réponse du joueur et la distribution des couleurs déjà affectées au mot *via* les réponses des autres. Il existe deux types de réponses : **une ou plusieurs couleurs** ou bien **sans couleur**. Si la réponse du joueur correspond à une des plus fréquentes, il gagne des points, sinon il en perd. Le nombre de points gagnés est covariant avec le poids des couleurs liées au mot dans le réseau lexical selon une fonction logistique sigmoïde *S*. Si une couleur est faiblement associée (par rapport à d'autres), le joueur va gagner peu de points. Par contre si elle est très fortement associée, le gain sera plus grand mais plafonné. Le poids associé à la relation mot-couleur est incrémenté de 1 à chaque réponse. Une relation de couleur nouvellement associée rentre dans le réseau lexical avec un poids de 1. Supposons par exemple, que dans le réseau le terme *éléphant* soit associé à *gris* avec un poids de p1 et à *blanc* avec un poids de p2. Si le joueur sélectionne *gris* il gagnera *S(p1)* points et le poids de la relation *gris* passera à p1+1. Si le joueur clique sur *pas de couleur* il perd *S(p1+p2)* points mais la relation *pas de couleur* est introduite et prend la valeur 1. Il n'y a pas d'avantage particulier à proposer des couleurs plutôt que *pas de couleur*.

Ce système incite les joueurs à être honnêtes en sélectionnant des couleurs appropriées, et à essayer d'être originaux quand c'est possible. L'honnêteté présupposée de la grande majorité des joueurs a été vérifiée empiriquement et se révèle être une constante dans les GWAP pour lesquels nous avons ce type d'information (Chamberlain, 2013). D'une part une erreur peut être coûteuse en termes de points, d'autre part un joueur qui chercherait à saboter les données en répondant systématiquement de manière incorrecte se lasserait très vite. Aucune tentative de sabotage n'a été détectée à ce jour. Tous les 100 points, le joueur passe au niveau supérieur et bénéficie d'un plus grand nombre de caractères autorisés dans le champ de texte libre. Les facteurs qui rendent le jeu motivant sont la possibilité de comparer sa (ses) réponse(s) avec celles des autres joueurs et la progression en termes de niveau, puisque plus le niveau est élevé plus on peut saisir de mots dans le champ de texte libre. L'expérience montre que la plupart des joueurs préfèrent entrer plusieurs réponses de couleur dans le champ de texte que se contenter de cliquer sur une des couleurs proposées. ColorIt s'avère être un jeu stimulant et très original si l'on en croit les nombreux retours positifs des joueurs. La variété du vocabulaire proposé et la présence de termes ambigus ou contextualisés (*ours>animal, frégate>oiseau, magistrat>singe, avocat>fruit, avocat>justice)* évite la monotonie. A côté de certains mots dont la couleur est évidente (comme *neige*, *nuit*, *charbon*, etc.) et ne se prête pas ou peu à une interprétation, d'autres peuvent avoir une large gamme de couleurs possibles, qu'elles soient objectives pour des noms concrets (comme *fleur*, *voiture*, etc.) ou subjectives pour des concepts abstraits (comme *colère*, *tristesse*, etc.), ou pas de couleur du tout (comme *augmentation*, *audace*, etc.). Choisir la couleur que l'on pense être la plus pertinente n'est pas une tâche facile et la confrontation avec les réponses des autres génère excitation, suspense, et… bonne ou mauvaise surprise ! Bien que le jeu ait pour but la

collecte de données lexico-sémantiques, aucune connaissance linguistique n'est requise pour jouer, et les données recueillies sont de bonne qualité (cf section 3) et parfois d'un niveau de précision et de nuance remarquable (voir liste des couleurs/apparences sur le site à jeuxdemots.org/colorit.php?action=colorlist). Ce niveau de précision va sans doute bien au-delà de ce qui est nécessaire pour de la désambiguïsation lexicale, mais reste parfaitement justifié en représentation des connaissances.

2.3 Dans les coulisses

Le projet JDM (Lafourcade, 2007 and Chamberlain, 2013) a pour but de construire un réseau lexical et sémantique le plus large possible par le biais de jeux (GWAP) et de *crowdsourcing*. Un tel réseau est composé de termes (les nœuds ou sommets) et de relations entre ces nœuds. Les relations sont typées, orientés et pondérées. La ressource contient actuellement plus de 320 000 termes et 9 millions de relations. Le réseau lexical JDM contient des termes, des significations de termes (raffinements), et différentes informations sémantiques (comme *personne*, *être vivant*, *abstrait*, *concret*, *artefact* etc.). Les relations sont pondérées (les plus évidentes ont les poids les plus élevés). Les relations affectées d'un poids négatif soulignent une impossibilité intéressante, comme une exception (par exemple: *autruche* agent: -100 *voler*). On notera qu'une telle approche peut être utilisée pour désigner des impossibilités en termes de couleurs. Les différentes significations d'un terme sont liées au terme principal par des *raffinements*. Par exemple, *frégate* est liée à *frégate>navire* et *frégate>oiseau*. Le raffinement *frégate>navire* est lui-même raffiné en *frégate>navire>moderne* et *frégate>navire>ancien*. Chaque raffinement ou sous-raffinement est lui-même connecté à d'autres termes dans le réseau.

Dans ce réseau, l'information relative à la couleur n'était jusqu'à maintenant présente qu'à travers la relation *caractéristique*, donc mélangée à des attributs de nature variée. Etant donnée l'importance de cette relation en représentation des connaissances, elle fait désormais l'objet d'une relation spécifique que le jeu ColorIt permet d'enrichir de manière conséquente. Un algorithme heuristique a été conçu de manière à avoir une probabilité raisonnable de sélectionner un terme pour lequel l'information de couleur a du sens. Ainsi, nous évitons de lasser les joueurs en leur proposant trop de termes pour lesquels c'est la réponse *pas de couleur* qui s'impose. Cet algorithme qui choisit le terme à proposer au joueur est le suivant : nous sélectionnons aléatoirement, dans le réseau JeuxDeMots, un mot qui a au moins une relation de couleur positive. Nous proposons ensuite au joueur, *via* un *pile ou face* virtuel, soit ce terme soit un de ses voisins dans le réseau. Ceci entraine une propagation rapide de l'acquisition des relations de couleur à travers le réseau. Un terme qui aura été caractérisé comme *sans couleur* plusieurs fois sera retiré de la liste des voisins sélectionnables. De la même manière, les raffinements des termes "colorés" deviennent rapidement "colorés" eux-mêmes. Le principe qui sous-tend cet algorithme de propagation est qu'il y a plus de chances de sélectionner un terme éligible pour l'information de couleur s'il est lié à un mot qui a déjà lui-même une caractéristique de couleur. Une sélection entièrement aléatoire serait contre-productive parce que la plupart du temps, aucune information de couleur ne pourrait être proposée par le joueur; en effet nous avons estimé par échantillonnage qu'environ 10% seulement des quelques 320 000 termes du réseau sont éligibles pour l'information de couleur.

Si une réponse fournie par le joueur existe dans le réseau mais n'est pas caractérisée comme une couleur ou une apparence, elle est alors proposée et sera validée (ou invalidée) en tant que telle par un administrateur. Si par contre le terme proposé n'existe pas dans le réseau, il doit d'abord y rentrer (toujours par validation), avant d'être caractérisé comme une couleur ou une apparence. L'interface Diko du réseau lexical JDM est un outil contributif permettant la validation des termes proposés (http://www.jeuxdemots.org/diko.php). L'algorithme de propagation a été amorcé avec les termes de couleurs proposés dans le jeu : *blanc, gris, noir, rouge, rouge foncé, orange, jaune, vert clair, vert, bleu, bleu clair, violet, rose, beige, marron.* Avant la création de ColorIt, ces couleurs étaient associées à de nombreux termes *via* d'autres relations que la couleur.

3 Collecte et évaluation des données

En neuf mois (août 2013 à avril 2014), plus de 28 000 associations mots-couleurs ont été créées et plus de 16 000 termes ont été dotés d'une information de couleur. Ainsi, en moyenne, un terme est associé à 2,2 couleurs (hors *sans couleur*). Le nombre total de votes de couleurs a excédé 192 000, le nombre de joueurs n'est pas connu avec précision (il n'est pas nécessaire de s'enregistrer pour jouer). D'après notre estimation sur le nombre de termes éligibles pour une information de couleur, nous pensons qu'actuellement la moitié des termes éligibles ont été dotés d'une information de couleur (soit environ 16 000 termes sur 10 % de 320 000).

Dès qu'une couleur a été associée à un terme, le réseau est mis à jour avec cette couleur ou apparence. Par exemple, le mot *eau* est associé avec différents bleus (*bleu clair* (26), *bleu* (9), *bleu lagon* (1), *turquoise* (1), etc.) plusieurs autres couleurs (*verdâtre* (4), *vert* (2), etc.) et différentes apparences (*trouble* (20), *transparent* (9+9), *limpide* (4), *incolore* (4), etc.). La possibilité d'ajouter des couleurs ou des apparences *via* le champ de texte libre permet de compléter le réseau avec des noms de couleurs spécifiques comme *jaune chartreuse, mandarine*, etc., des apparences comme *rayé*, *translucide*, etc., ou encore de nouvelles combinaisons de couleurs (la liste des couleurs est disponible à l'url http://www.jeuxdemots.org/colorit.php?action=colorlist). Actuellement, plus de 700 termes correspondent à des couleurs, des associations de couleurs ou des apparences.

Les données collectées ont été évaluées manuellement, sur deux échantillons d'environ 500 termes, représentant à peu près 2 500 associations mots-couleurs. Le premier échantillon S_C a été prélevé parmi les termes ayant au moins une association de couleur, le second S_{NC} parmi ceux ayant au moins un vote de type *sans couleur*. Les termes de l'échantillon S_C peuvent avoir des associations de type *sans couleur*, et vice-versa. Nous avons demandé à des volontaires de contrôler au hasard certaines associations mots-couleurs, et de dire si elles étaient correctes ou non, que ce soit dans le cas d'association de type *couleur* ou *sans couleur*. Cette méthode d'évaluation, où l'on demande à des volontaires de juger des associations (tâche fermée) est orthogonale avec la méthode d'acquisition où l'on demande aux gens de produire les associations (tâche ouverte). Notons que la méthode d'évaluation ne permet pas de se prononcer sur les associations de couleurs éventuellement omises (les silences).

	S_C Associations de couleur correctes	**S_C Associations de couleur incorrectes**	**S_{NC} Associations sans couleur correctes**	**S_{NC} Associations sans couleur incorrectes**	**Total**
# votes	13 308	366	708	81	14 463
%	92 %	2,5 %	5 %	0,5 %	100%
	sous-total 13308/13674 = 97%		sous-total 708/789 = 90%		

Table 2. Evaluation des associations mots-couleurs et mots-*sans couleur*.

L'évaluation par quelques dizaines de volontaires (*via* le jeu Askit : http://www.jeuxdemots.org/askit.php) a couvert environ 10 % des associations mots-couleurs générées par le jeu (acquisition). Les données collectées semblent contenir un pourcentage assez bas d'associations incorrectes (cf. Table 2). En outre, un examen plus approfondi du réseau lexical montre que les associations incorrectes sont affectées d'un poids faible par rapport aux associations correctes. Par exemple, pour le terme *soleil*, on trouve une fois l'association avec *bleu* pour 100 fois l'association avec *jaune*. De plus, l'examen des termes polysémiques montre que les couleurs associées aux raffinements sont le plus souvent adéquates et pourront éventuellement être discriminantes dans le cadre de la désambiguïsation lexicale (*WSD*).

Termes liés à	**politique (2702 termes)**	**zoologie (3191 termes)**	**arts (2815 termes)**	**émotions (304 termes)**	**botanique (4481 termes)**
Accord entre joueurs	0.55	0.65	0.71	0.28	0.76

Table 3. Kappa moyen entre les joueurs en fonction de différents champs sémantiques d'intérêt concernant les couleurs (réelles ou symboliques). Le chiffre entre parenthèses indique le nombre de termes reliés à ce champ sémantique dans le réseau lexical. Manifestement, les couleurs sont difficiles à accorder aux émotions.

La Table 3 présente l'évaluation de l'accord entre joueurs en fonction de l'appartenance des termes à tel ou tel domaine (champ sémantique). La valeur de l'accord pour un terme donné est la moyenne du Kappa de Cohen entre les associations des joueurs prises 2 à 2 pour ce terme. La valeur de l'accord pour un champ sémantique est la moyenne des accords des termes de ce champ qui ont été joués. Certains termes peuvent appartenir à plus d'un champ. C'est pour le champ sémantique relatif aux plantes que l'accord entre joueurs est le plus élevé. Il semble que certains joueurs se réfèrent à des sources extérieures (comme par exemple Wikipédia) pour savoir quelle(s) couleur(s) affecter à une plante ou un animal inconnu, jouant ainsi le rôle d'identificateurs d'informations qu'il serait très difficile d'extraire automatiquement (communication personnelle avec certains des joueurs). L'accord le plus faible est celui relatif au champ sémantique des émotions, ce qui s'explique aisément par le haut degré de subjectivité des associations correspondantes (et ceci bien que notre expérience n'ait été réalisée qu'en français, et sur une population relativement homogène constituée de 70 % de femmes entre 30 et 50 ans). Concernant le champ sémantique des arts, on trouve beaucoup de noms de couleurs exprimés au moyen d'un vocabulaire recherché, mais qui une fois décryptés présentent finalement assez peu de variations.

4 Conclusion

Nous avons conçu un jeu simple mais stimulant et efficace pour collecter des données relatives aux associations termes-couleurs. En quelques mois plus de 15 000 termes ont été associés à une ou des couleurs et environ 4 000 ont été caractérisés comme *sans couleur*. De nombreuses couleurs et apparences ont été introduites dans le réseau lexical JeuxDeMots *via* le jeu ColorIt. A notre connaissance, les données collectées constituent la première ressource (dynamique et libre) en français pour ce type d'association. L'une des difficultés dans la conception du jeu a été l'algorithme de propagation permettant de sélectionner des termes éligibles à une information de couleur, de façon à ne pas lasser les joueurs en leur proposant trop de termes auxquels il n'est pas pertinent d'attribuer une couleur.

Nous avons eu plusieurs surprises avec les données collectées grâce à ColorIt. Premièrement, le choix *sans couleur* a été initialement conçu pour servir d'échappatoire aux joueurs lorsque rien ne convenait. Mais il s'est rapidement avéré que la caractéristique *sans couleur* était précieuse pour éliminer certaines significations en cas de polysémie. En effet, dans le cadre de la désambiguïsation lexicale, la sélection de sens par les indications de couleur est plus souvent réalisée par élimination que par sélection ; cela s'explique, du moins partiellement, par la polysémie par métaphore: des sens abstraits (donc *sans couleur*) dérivent souvent de termes concrets. Deuxièmement, un effet collatéral et néanmoins positif du jeu a été d'enrichir le réseau JDM d'un grand nombre de couleurs et apparences qui n'y figuraient pas encore. En plus de vraies couleurs nouvellement introduites, le réseau contient maintenant de nombreux termes relatifs à l'aspect visuel, comme *translucide*, *trouble*, *rayé*, *tacheté*, etc.

Pour aller plus loin, une des perspectives est d'évaluer la corrélation entre couleur et sentiment associés à certains termes, en particulier les noms abstraits. Dès que nous disposerons d'un nombre suffisant d'associations mots-couleurs, nous pourrons envisager la comparaison de ces données avec d'autres ressources déjà disponibles sur la polarité, ou les sentiments/émotions associés, en plus de leur exploitation dans le cadre de la désambiguïsation lexicale. De plus, l'intérêt des informations de couleur recueillies par *crowdsourcing* dans les tâches de *WSD* doit faire l'objet d'une étude complète.

Références

VON AHN, L. (2006). *Games with a purpose.* Computer, 39(6):92–94, 2006.

BERLIN, B. & KAY, P. (1969). *Basic Color Terms: Their Universality and Evolution.* Berkeley: University of California Press. ISBN 1-57586-162-3, pp. 178.

CARD, S., MACKINLAY, J. AND SHNEIDERMAN, B., (1999). *Readings in information visualization: using vision to think.* Interactive Technologies, Morgan Kaufmann, ISBN-10: 1558605339, 712 p.

CHAMBERLAIN, J., FORT, K., KRUSCHWITZ, U., LAFOURCADE, M. AND POESIO, M. (2013) *Using Games to Create Language Resources: Successes and Limitations of the Approach.* Theory and Applications of Natural Language Processing. Gurevych, Iryna; Kim, Jungi (Eds.), Springer, ISBN 978-3-642-35084-9, 2013, 42 p.

CHILD, I, L., HANSEN, J,A., AND HORNBECK, F,W. (1968). *Age and sex differences in children's color preferences.* Child Development, 39 (1):237–247.

CHRIST, R. (1975). *Review and analysis of color coding research for visual displays.* Human Factors: The Journal of the Human Factors and Ergonomics Society, 17:542–570.

GREFENSTETTE, G. (2005). *The color of things: Towards the automatic acquisition of information for a descriptive dictionary.* Revue Française de Linguistique Appliquée, 2, pp. 83–94.

JOUBERT, A;, M. LAFOURCADE, M. (2012) *A new dynamic approach for lexical networks evaluation.* In proc of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 23-25h May 2012.

KAYA, N. & EPPS, H. (2004). *Relationship between color and emotion : a study of college students.* Academic journal article from College Student Journal, Vol. 38, No. 3.

LAFOURCADE, M. (2007). *Making people play for Lexical Acquisition.* In Proc. SNLP 2007, 7th Symposium on Natural Language Processing. Pattaya, Thailande, 13-15 December 2007, 8 p.

LAFOURCADE, M. JOUBERT, A. (2012) *Increasing long tail in weighted lexical networks.* In proc of Cognitive Aspects of the Lexicon (CogAlex-III), COLING, Mumbai, India, December 2012.

LUSCHER, M. (1969). *The Luscher Color Test.* Random House, New York, New York.

MEIER, B. (1988). *Ace: a color expert system for user interface design.* In Proceedings of the 1st annual ACM SIGGRAPH symposium on User Interface Software, UIST '88, pages 117–128, New York, NY, USA. ACM.

NIJDAM, N A. (2010). *Mapping Emotion to Color, Human Media Interaction Human Media Interaction.* University of Twente, the Netherlands. Available at http://hmi.ewi.utwente.nl/verslagen/capita selecta/CS Nijdam Niels.pdf

OU, L-C., LUO, M.R, SUN, P-L., HU, N-C., AND CHEN, H-S. (2011). *Age effects on color emotion, preference, and harmony.* Color Research and Application.

OZBAL, G., STRAPPARAVA, C., MIHALCEA, R. AND PIGHIN, D. (2011). *A Comparison of Unsupervised Methods to Associate Colors with Words.* Affective Computing and Intelligent Interaction. Lecture Notes in Computer Science. Volume 6975, 2011, pp 42-51.

PRIBADI, N, S., WADLOW, M,G., AND BOYARSKI, D. (1990). *The use of color in computer interfaces: Preliminary research.* Information Technology Center, Carnegie Mellon University, 49 p.

SABLE, P. AND AKCAY, O. (2010). *Color: Cross cultural marketing perspectives as to what governs our response to it.* Proceedings of ASBBS, vol 17:1, -Las vegas, CA. February 2010, pages 950–954.

SAIF, M. (2011a). *Colourful language: measuring word-colour associations.* Proceeding CMCL '11. Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. 97-106. http://www.aclweb.org/anthology-new/W/W11/W11-0611.pdf

SAIF, M. (2011b). Even *the Abstract have Colour : Consensus in Word–Colour. Associations.* In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technology. http://www.aclweb.org/anthology-new/P/P11/P11-2064.pdf

STRAPPARAVA, C., ÖZBAL, G (2010) *The color of emotions in texts.* 23rd International Conference on Computational Linguistics (COLING 2010), 28 p.