ProtAnt

Un outil d'analyse de la prototypicalité des textes

Laurence Anthony et Paul Baker
Université de Waseda / Université de Lancaster

Les chercheurs basant leur travail sur l'étude de corpus ou sur les méthodes traditionnelles de recherche qualitative, en s'intéressant par exemple à l'analyse critique de discours, doivent souvent sélectionner, au sein d'un plus large corpus, des textes prototypiques présentant les traits de langage étudiés, en vue d'une lecture approfondie. Cette procédure de sélection s'est toujours appuyée sur des méthodes traditionnelles pour la plupart ad hoc. Dans cet article, nous proposerons une approche plus rigoureuse de sélection de textes destinés à une lecture approfondie qui repose sur un classement de ces textes selon le nombre de mots- clés qu'ils contiennent. Pour faciliter cette analyse, nous avons créé un logiciel gratuit et multiplateforme baptisé *ProtAnt*, qui analyse les textes, génère une liste ordonnée de mots-clés (en se basant sur la significativité statistique et la taille d'effet) et classe ensuite ces textes selon le nombre de mots-clés qu'ils contiennent. Nous décrirons également différentes expériences démontrant que l'analyse menée par *ProtAnt* est efficace non seulement pour identifier des textes prototypiques, mais également pour distinguer des textes non pertinents que l'on pourra éventuellement supprimer du corpus cible.

Mots-clés : *ProtAnt*, analyse critique de discours, prototypicalité, mots-clés, recherche qualitative

1. Introduction

Les chercheurs basant leur travail sur l'étude de corpus ou sur les méthodes traditionnelles de recherche qualitative doivent souvent sélectionner, au sein d'un plus large corpus, des textes présentant les traits de langage étudiés, en vue d'une lecture approfondie. Il faudrait idéalement que les textes choisis pour une lecture approfondie soient prototypiques, c'est-à-dire soient les meilleurs exemples, les plus purs, les plus typiques, les plus représentatifs (Labov 1973, Rosch 1975, Gries 2003). Toutefois, dans la réalité, une telle sélection est difficile à effectuer. Ainsi il arrive que les chercheurs soient taxés de partialité pour avoir glané, dans le corpus de travail étendu, des textes illustrant une idée ou un argument préconçu. Les analystes de discours critique sont particulièrement la cible de ce type de critiques (Widdowson 2004) et même les analystes de discours basés sur des corpus peuvent être accusés de manquer de partialité dans leurs choix, les hypothèses qu'ils développent à partir d'une analyse descendante d'un corpus étant généralement validées par une lecture approfondie ascendante de quelques textes soigneusement sélectionnés dans le corpus. Pour éviter ce type de critiques, les chercheurs pourraient choisir de sélectionner aléatoirement des textes au sein d'un corpus ou d'un ensemble de données (voir Sajjid 2013), mais une telle méthode de sélection pourrait ne pas révéler des caractéristiques inhabituelles, bien qu'importantes, identifiées lors de l'analyse descendante. Les chercheurs pourraient aussi tenter une lecture approfondie des textes du corpus, du premier jusqu'au dernier. Cependant, cette approche risquerait probablement de biaiser les observations et/ou de prendre un temps indument long.

Dans cet article, nous proposerons une approche plus rigoureuse pour la sélection de textes destinés à une lecture approfondie qui se base sur un classement des textes selon le nombre de mots-clés (mots inhabituellement plus fréquents dans le corpus cible que dans le corpus de référence) qu'ils contiennent. Pour faciliter cette analyse, nous avons créé un logiciel gratuit et multiplateforme baptisé *ProtAnt* (Anthony & Baker 2015), qui analyse les textes, génère une liste ordonnée de mots-clés (en se basant sur la significativité statistique et la taille d'effet) et les classe ensuite selon le nombre de mots-c$_i$lés qu'ils contiennent.

Dans la prochaine section, nous résumerons les travaux précédemment réalisés dans l'identification de textes

International Journal of Corpus Linguistics 20:3 (2015), 273–292. doi 10.1075/ijcl.20.3.01ant issn 1384–6655 / e-issn 1569–9811

prototypiques. La section 3 sera dédiée à la présentation de l'outil *ProtAnt* et de ses fonctionnalités. Dans la section 4, nous présenterons diverses expériences qui montrent la capacité de *ProtAnt* à classer correctement des textes, tant courts que longs, tout en offrant une visibilité sur leur degré de prototypicalité. Cette section nous apportera également l'opportunité de décrire une expérience montrant comment *ProtAnt* permet d'identifier des textes non pertinents ou mal classés qui pourraient être supprimés d'un échantillon cible avant l'étape de lecture approfondie. Enfin, la section 5 sera l'occasion pour nous de proposer des applications potentielles du logiciel dans les domaines de la recherche, de l'enseignement et de l'apprentissage ainsi que d'énoncer des pistes de réflexion pour de futurs développements de l'outil *ProtAnt*.

2. Méthodes de sélection de textes prototypes

Jusqu'à ce jour, les chercheurs étudiant les langues et les discours ont principalement opté pour deux approches dans la sélection de textes illustrant des caractéristiques cibles de la langue étudiée et donc candidats pour une étude approfondie. La première approche, qui est aussi la plus courante, pourrait être décrite comme la "sélection opportuniste". Elle consiste à effectuer un choix arbitraire, mais éclairé et prometteur, de textes à analyser. Dans le recueil d'études critiques de discours critique édité par Wodak (2013), cette approche peut être illustrée par les travaux de Caldas-Coulthard et al. (2003/2013 : 40), Chouliaraki (2000/2013 : 100), et Machin & Suilman (2006/2013: 229). Caldas-Coulthard et al. (2003/2013) décrivent leur sélection de textes cibles, regroupant des textes parlant d'ours en peluche, dans ces termes : « [...] nous avons acheté les 15 livres sur les ours dont disposait une librairie pour enfant à Londres. » Chouliaraki (2000/2013) s'est intéressée à la façon dont les pratiques des journaux télévisés produisent implicitement des positions politiques hégémoniques. Pour étudier cette théorie, elle a sélectionné un unique journal télévisé, diffusé le 16 août 2000, rapportant le décès d'un manifestant au cours d'une émeute. Machin & Suilman (2006/2013), pour leur part, se sont intéressés à l'effet que les jeux produisent sur le discours politique. Bien que leur discours cible n'est pas textuel, ils ont adopté la même méthode de sélection en choisissant uniquement deux jeux vidéo pour leur étude (le jeu américain Delta Force et le jeu Hezbollah - Special Force), et ce, en dépit du très vaste choix de jeux similaires existant sur le marché.

Quand le chercheur possède une connaissance et une expérience étendues du domaine étudié et de la langue, cette 'sélection opportuniste' est sans aucun doute efficace. Toutefois, cette méthode ne peut en aucun cas être décrite comme rigoureuse, ce qui soulève des questions quant à la partialité des chercheurs, les implications de leurs découvertes, et les possibilités de reproduire l'étude.

La deuxième approche de sélection de textes, beaucoup moins répandue parmi les chercheurs étudiant les langues et les discours, peut être décrite comme la "réduction sélective" et est illustrée par les travaux de Khosravinik (2010). Dans le cadre d'un projet commun financé par l'ESRC sur la représentation de l'immigration dans la presse britannique, Khosravinik (2010) utilise un corpus de 140 millions de mots issus de 170 000 articles, qui avaient antérieurement été analysés par Gabrielatos & Baker (2008) en utilisant des approches basées sur les corpus, et sélectionne des séries d'articles publiés sur cinq périodes d'une semaine durant lesquelles le nombre d'articles traitant de l'immigration atteignait des sommets. Pour chaque période, il a extrait d'un journal libéral de qualité, d'un journal conservateur de qualité et d'un journal à sensation l'ensemble des articles pertinents quant au thème de l'immigration. Il en a résulté un corpus de 439 articles que Khosravinik (2010) a ensuite lu en profondeur et examiné en termes de topoi (stratégies argumentatives) et selon quelques catégories socio- sémantiques proposées par van Leeuwen (1996) telles que l'agrégation, la collectivisation, la fonctionnalisation, l'humanisation et l'individualisation.

Un certain nombre d'études adoptant une approche de "réduction sélective" similaire apparaissent dans le recueil édité par Wodak (2013), telles que l'étude du discours argumentatif par Ehrlich & Blum-Kulka (2010/2013) et l'étude du discours de paix israélien par Gavriely-Nuri (2010/2013). Malheureusement, aucun détail n'est communiqué quant à la taille du « vaste corpus » réduit dans les travaux d'Ehrlich & Blum-Kulka (2010/2013 : 149), ni quant au nombre précis d'usages de métaphores dans les « centaines de mentions de métaphores » à partir desquelles s'opèrent les réductions dans les travaux de Gavriely- Nuri (2010/2013: 226).

De toute évidence, l'approche de "réduction sélective" est plus rigoureuse que l'approche de "sélection opportuniste" ou que l'approche encore plus simpliste fondée sur le choix de textes à partir de leur ordre dans un corpus précompilé ou sur le degré selon lequel un texte pourrait "sembler intéressant" pour l'analyste. Cependant, cette approche peut toujours résulter soit en un grand nombre de textes nécessitant une lecture approfondie, comme dans le cas de l'approche de Khosravinik (2010), soit en un petit nombre de textes qui peuvent paraitre avoir été, dans une certaine mesure, partialement sélectionnés comme c'est le cas dans les travaux de Ehrlich & Blum-Kulka (2010/2013) et de Gavriely-Nuri (2010/2013). La problématique de la sélection de textes appropriés est une question importante pour les entités possédant un vaste répertoire de données textuelles et qui ont besoin de trouver rapidement des textes dans un but de gestion, de prise de décision ou d'établissement de profil client (Durfee et al. 2007). Par conséquent, la sélection de textes prototypiques est devenue un domaine de recherche dynamique dans le

traitement automatique du langage naturel (TALN). Visa et al. (2001) et Kloptchenko et al. (2002, 2004) ont adopté une approche requérant d'abord le jugement d'un humain pour choisir des textes prototypiques représentatifs de différents types (ou classes) de textes au sein d'un corpus plus vaste. Ils analysent ensuite les textes prototypes en termes de structure de mots et de phrases, c'est-à-dire selon l'ordre des mots et les structures des paragraphes, pour créer un profil des textes prototypes et des autres textes du corpus. Puis ils ordonnent l'ensemble des textes du corpus en différentes classes établies selon leur distance par rapport au prototype initial. De nombreux autres algorithmes d'apprentissage automatique emploient également un concept similaire de prototypicalité. La méthode des plus proches voisins (par exemple Manning et al. 2008) nécessite une série d'échantillons (prototypes) connus et issus de classes particulières à partir desquels il construit une série d'attributs (par exemple des mots pour la classification de textes). Les nouveaux documents proposés sont alors assignés à une classe en fonction de leur "proximité" avec les documents d'entraînement d'une classe particulière. Chen et al. (2011) utilisent également une méthode basée sur les prototypes pour classer des messages de forum internet dans des catégories particulières. On retrouvera un compte-rendu utile des classifieurs basés sur les prototypes dans Fayed et al. (2007).

Fayed et al. (2007) proposent une méthode de classification qui requiert toujours une série d'échantillons d'entraînement pour une classe particulière mais qui s'appuie également sur un algorithme pour sélectionner automatiquement l'échantillon le plus prototypique de la série. Ces prototypes auto-générés sont ensuite utilisés dans la classification d'échantillons futurs inconnus. Bahrololoum et al. (2012) proposent un tel détecteur de prototypes basé sur un algorithme de recherche dit gravitationnel, qui modélise des attributs dans les données d'entraînement (par exemple les mots dans des ensembles de données textuelles) comme des masses de petite taille qui interagissent les unes avec les autres et se groupent ensemble.

Les méthodes de TALN décrites ci-dessus sont dépendantes du nombre de classes perçues comme existantes et dans le cas des travaux menés par Visa et al. (2001) et par Kloptchenko et al. (2002, 2004), elles reposent sur le(s) texte(s) prototype(s) sélectionné(s) en amont. En dépit de ces limites, les approches du TALN offrent au chercheur une méthodologie pour détecter de nombreux textes par nature similaires au(x) prototype(s) et qui surmontent donc partiellement le biais de la sélection initiale. Toutefois, lorsque l'objectif est de déterminer des textes prototypiques pour une analyse qualitative détaillée, ces approches sont peut-être moins intéressantes. Tout d'abord, le chercheur a besoin soit de sélectionner une série d'un ou plusieurs textes prototypiques (problématique source de nos recherches actuelles), soit de définir au moins les classes ou groupes possibles dans lesquels un texte pourra être classé. Les concepts mathématiques derrière ces approches de classification peuvent également se révéler extrêmement complexes, faisant de ces méthodes des "boîtes noires" pour les utilisateurs suivants. Un autre souci est que ces approches regroupent des textes sur la base de leur profil général plutôt que sur une série distincte de traits uniques qui pourraient se révéler intéressants en recherche qualitative (par exemple en analyse critique de discours).

Il est à noter qu'il existe également des techniques de TALN qui rassemblent les textes en groupes sans supervision (c'est-à-dire sans reposer sur aucune catégorie prédéfinie), mais encore une fois, ces techniques dépendent du profil général des textes. Elles peuvent aussi s'avérer extrêmement sensibles aux approches de regroupement adoptées (c'est-à-dire suggérant des groupes très différents si regroupés de façon descendante ou ascendante), poser des problèmes aux utilisateurs tentant d'étiqueter ces groupes, et encore une fois, elles peuvent paraître des "boîtes noires" pour les utilisateurs (Manning et al. 2008). Également, alors qu'il est possible d'identifier des textes prototypiques au sein de ces groupes, les méthodes d'identification reposent souvent sur des approches simples telles que le traitement de tous les échantillons d'un groupe comme prototypes, ou la sélection d'un échantillon représentant la moyenne ou le "centre" du groupe lorsqu'on le représente dans un espace géographique (pour une réflexion sur le sujet, voir Fayed et al. 2007).

Dans la section suivante, nous proposerons une nouvelle approche de sélection de textes prototypiques qui ne nécessite aucun groupe préalable ni aucune connaissance avancée en mathématiques pour en comprendre la méthodologie. Notre approche est également adaptée pour l'étude qualitative ciblant les traits uniques de textes et est exécutable grâce à un logiciel contenu dans un seul fichier et portable ne nécessitant aucun entraînement pour fonctionner.

3. Outil d'analyse ProtAnt

ProtAnt est un logiciel gratuit et portable conçu pour profiler les textes d'un corpus et les classer en fonction du degré selon lequel ils sont prototypiques de ce corpus dans son ensemble en se basant sur le concept de mots- clés. Nous partons de l'hypothèse qu'un texte contenant un grand nombre de mots-clés du corpus dans son ensemble joue aussi probablement un rôle de texte central ou typique de ce corpus.

On entend généralement par mots-clés des mots dont la fréquence est significativement plus élevée dans un corpus que dans un corpus de référence. Dans cet esprit, les mots-clés eux-mêmes peuvent être considérés comme distinctifs du corpus cible. Différents tests statistiques sont utilisés pour identifier les mots-clés et le log *likelihood* est

devenu de facto un standard dans de nombreux travaux sur corpus en raison de sa présence dans des concordanciers populaires, tels que *AntConc* (Anthony 2014) et *WordSmith Tools* (Scott 2014). Une fois identifiés, les mots-clés peuvent également être classés.

Traditionnellement, on classe les mots-clés en fonction de leur valeur de *keyness* directement issue d'une mesure statistique (en pratique, un classement selon leur valeur p). Cependant, il existe différentes possibilités pour classer les mots-clés. Scott (2014), par exemple, propose une approche mots-clés clés par laquelle des mots- clés sont générés pour chacun des textes du corpus pris individuellement, puis ces mots-clés sont classés en fonction de leur distribution à travers les textes du corpus dans son ensemble, les mots-clés apparaissant dans un grand nombre de textes figurent en haut de classement. Toutefois, l'inconvénient de cette approche est qu'elle peut uniquement fonctionner avec des textes individuels relativement longs. Sans quoi, les hypothèses de la statistique de *keyness* seront invalidées pour de nombreux termes candidats, c'est-à-dire ceux apparaissant moins de 5 fois dans chaque texte pris individuellement. Plus récemment, les mesures de taille d'effet telles que la fréquence relative (Demarau 1993) ou un log de la fréquence relative (par exemple Hardie 2014) ont été utilisés pour classer les mots-clés. Les mesures d'effet sont statistiquement rigoureuses, bien comprises, et produisent des valeurs de classement facilement comparables d'une étude à l'autre.

Les mots-clés sont habituellement utilisés dans la recherche sur corpus pour identifier des séries de lexies saillantes dans un ou plusieurs corpus, qui peuvent ensuite faire l'objet d'analyses plus qualitatives et interprétatives des collocatifs et des lignes de concordances. Par exemple, Leńko-Szymańska (2006) compare des corpus composés de dissertations écrites par des étudiants américains et polonais sur le même sujet, et trouve des mots-clés relatifs au style rédactionnel (par exemple, les étudiants polonais utilisent plus d'expressions de liaison telles que *moreover*, *however* et *thus* [*Ndt : de plus, cependant et ainsi*]), et à des éléments centraux (les étudiants américains utilisent moins de termes généraux et leurs mots-clés tendent à être liés à des situations spécifiques). Les mots-clés sont également utilisés pour indiquer les idéologies véhiculées par des textes. Par exemple, Baker et al. (2013) comparent les mots-clés de corpus composés d'articles de journaux au sujet de l'Islam. Ils découvrent que les journaux à sensation tendent à utiliser des mots-clés plus centrés sur le terrorisme et l'extrémisme (par exemple fanatics, terror, bomber) tandis que les journaux de qualité se focalisent plus sur le conflit en général par l'utilisation de mots-clés tels que *conflict*, *military*, *revolution*, et *occupation*.

Le logiciel *ProtAnt* que nous avons développé utilise les mots-clés d'une façon novatrice pour identifier des textes prototypiques. Notre approche de détection de prototypes basée sur les mots-clés fonctionne de la manière suivante. Les chercheurs souhaitant trouver des textes prototypiques au sein de leur corpus chargent tout d'abord leur corpus cible dans *ProtAnt* sous forme de fichiers individuels en texte brut (.txt) encodés en UTF-8. Un corpus de référence approprié encodé en UTF-8 doit ensuite être chargé. Il peut s'agir d'un fichier unique ou d'une série de fichiers séparés. Enfin, le chercheur doit spécifier la statistique de *keyness* (par exemple le log *likelihood*), une valeur statistique de seuil (par exemple $p<0{,}005$), une statistique de classement (par exemple le log de la fréquence relative) et plusieurs autres paramètres (voir ci-dessous). Une fois ces choix effectués, l'outil *ProtAnt* compare les fréquences des mots du corpus cible avec celles des mots du corpus de référence et calcule une série complète de mots-clés pour l'ensemble du corpus cible. À partir de cette liste, il calcule ensuite le nombre de mots-clés de l'ensemble du corpus cible présents dans chacun des textes dudit corpus et classe ensuite les textes selon le nombre de mots-clés qu'ils contiennent. Finalement, *ProtAnt* affiche sous forme de tableaux les mots-clés du corpus, les mots-clés du corpus apparaissant dans chacun des textes pris individuellement, et l'ensemble des classements des textes.

En suivant cette méthode, *ProtAnt* fonctionne à rebours de l'approche par mots-clés clés de Scott (2014), en trouvant d'abord les mots-clés qui sont distinctifs du corpus cible dans son ensemble, puis en comptant combien de ces mots-clés apparaissent au niveau du texte pris individuellement. Ce qui n'apparait pas immédiatement, cependant, est la raison pour laquelle notre approche pourrait sélectionner des textes prototypiques (c'est-à-dire centraux, typiques) d'un corpus cible au-delà de textes peut-être distincts, uniques dans le corpus. Pour comprendre notre raisonnement, il est important de se rappeler que les mots-clés sélectionnés sont distinctifs du corpus dans son ensemble selon de nombreuses dimensions possibles. Si un texte particulier du corpus cible contient plusieurs de ces mots-clés, il n'en ressort pas obligatoirement qu'il est spécialement distinctif. Un tel phénomène suggère plutôt qu'il représente plusieurs de ces différentes facettes. En d'autres termes, le texte peut être considéré comme prototypique du corpus dans son ensemble. En fait, un texte extrêmement distinctif (par exemple contenant quelques mots inhabituels répétés plusieurs fois) produirait seulement un petit nombre de mots-clés uniques et *ProtAnt* le classerait en bas de tableau. Nous éprouverons notre raisonnement par une série d'expériences dans la section suivante.

La figure 1 montre une capture d'écran de *ProtAnt* lors de l'analyse de 20 articles de journaux dont 10 abordent le sujet de l'Islam. Les résultats de cette analyse seront abordés en détail dans la prochaine section.

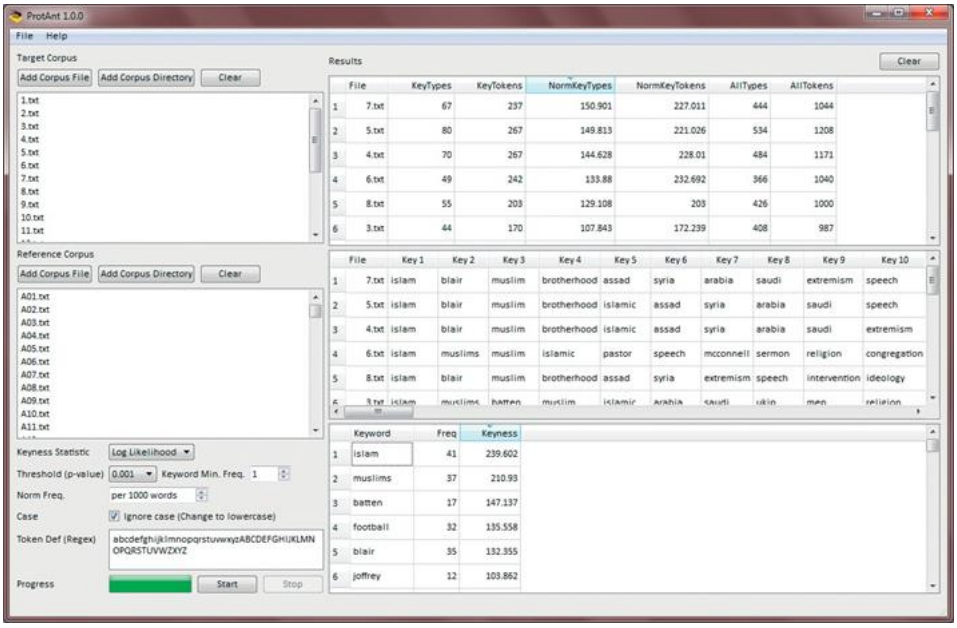

Figure 1. Écran principal de *ProtAnt*

Dans la figure 1, le volet supérieur gauche liste les fichiers cibles que le chercheur souhaite analyser. Le volet central gauche liste les fichiers du corpus de référence. Le volet inférieur gauche offre au chercheur diverses options pour changer la façon dont les mots-clés sont générés et classés, y compris la sélection de différentes statistiques et de différentes valeurs seuils pour les mots- clés. Il existe également des options affectant la façon dont les tokens des corpus cible et de référence sont comptés. Ici, la définition des tokens est indiquée comme une expression régulière où la casse est ignorée si besoin (réglage par défaut). La définition des tokens peut également comprendre des classes de caractères Unicode, telles que \{L}, qui représente n'importe quelle classe de caractères Unicode "lettre", ou \{N}, qui représente toute classe de caractères Unicode "nombre". Il en résulte que *ProtAnt* peut non seulement fonctionner avec des textes écrits en anglais mais aussi avec des textes rédigés en toute autre langue spécifiée dans le standard Unicode, y compris le japonais, le chinois, le coréen, l'arabe et l'hébreu. Enfin, il existe également une option pour changer la constante utilisée pour calculer la fréquence normalisée des mots (par exemple, fréquence par mille mots ou par un million de mots).

Le côté droit de la fenêtre principale de *ProtAnt*, montrée en Figure 1, affiche les résultats des analyses. Dans le volet supérieur droit, les fréquences et fréquences normalisées des types clés et des tokens clés peuvent être visualisées pour chacun des fichiers du corpus cible, accompagnées des fréquences de tous les types et tokens du fichier. Les valeurs affichées dans le tableau peuvent être triées selon n'importe quelle colonne en cliquant simplement sur le titre de la colonne. Par défaut, les résultats sont triés par types clés normalisés. Ainsi, ce tableau offre une visibilité immédiate du classement des textes selon le nombre de types clés qu'ils contiennent. Dans la Figure 1, le fichier 7.txt s'est avéré être le plus prototypique du corpus. Tout clic sur le nom d'un fichier provoque l'ouverture par le programme du fichier d'origine pour une lecture approfondie.

Le volet central droit de *ProtAnt* affiche l'ensemble des mots-clés apparaissant dans chacun des fichiers du corpus cible, classés selon leur valeur de *keyness*. Dans la figure 1, on peut observer que le mot-clé le plus saillant, *islam*, apparait dans les cinq premiers textes prototypiques (comme attendu), mais le second mot-clé le plus saillant, *muslims*, apparait uniquement dans l'un des textes (c'est-à-dire 6.txt). En observant la liste classée des mots-clés dans chacun des fichiers, un chercheur peut immédiatement saisir le sens du texte dans son ensemble. À nouveau, tout clic sur le nom d'un fichier provoque l'ouverture par le programme du fichier d'origine.

Le volet inférieur droit de *ProtAnt* affiche la liste complète des mots-clés du corpus cible, leur fréquence brute et les valeurs de *keyness* qui leur sont associées. À l'instar des autres tableaux de l'outil *ProtAnt*, tout clic sur le nom d'une colonne trie le tableau selon celle-ci.

4. Expériences utilisant ProtAnt

Dans le but de tester l'utilité de l'outil *ProtAnt* dans l'identification des textes prototypiques utiles, plusieurs expériences ont été menées. Celles-ci sont décrites dans les sections 4.1 à 4.5.

4.1 Expérience 1 : Identification d'articles de presse prototypiques

La première expérience a été élaborée pour étudier la capacité de *ProtAnt* à classer correctement une série de textes en fonction du cœur de leur sujet général. Pour cette première expérience, un corpus de 20 fichiers (Corpus 1) a été créé en utilisant la base de données journalistique consultable en ligne Nexis UK afin d'y récupérer des articles de journaux nationaux parus en mars 2014. Corpus 1 comprend 10 articles contenant le terme *Islam*, 5 articles contenant le terme *football*, et 5 autres articles sélectionnés au hasard en cherchant 5 mots fréquents issus du top 100 des mots les plus fréquents dans le British National Corpus (*time*, *people*, *other*, *know*, *see*). Cette sélection aléatoire a fait émerger des articles d'information sur le tennis, l'art, la science ainsi qu'une rubrique nécrologique et une critique télévisée. La taille du corpus est de 22 353 tokens et la taille moyenne d'un fichier est de 1 118 tokens. Pour le corpus de référence, nous avons utilisé le corpus BE06 (Baker 2009) qui contient 1 million de mots. Ce dernier comprend 200 échantillons d'écrits britanniques publiés en 2006 et appartenant à 15 genres.

Le Tableau 1 présente le classement des textes du Corpus 1 selon les valeurs normalisées des types clés et des tokens clés lorsque l'on mesure la statistique de *keyness* avec le log *likelihood*]. Dans le tableau, quatre différentes valeurs seuils de p furent utilisées, soit, 0.05, 0.01, 0.001, 0.0001. Les valeurs entre parenthèses dans les titres des colonnes indiquent le nombre de mots-clés obtenus pour chacune des valeurs seuils, et les nombres entre parenthèses dans les cellules indiquent le numéro du fichier. Un nom court de circonstance a été donné à chacun des fichiers, et les fichiers afférents à l'Islam ont été grisés pour être plus aisément identifiables. En partant du principe que la moitié des textes du corpus contiennent des articles à propos de l'Islam, on s'attendrait à ce que ces articles soient considérés comme typiques du corpus dans son ensemble et par conséquent soient classés en haut de tableau par rapport aux autres. On s'attendrait également à ce que les 5 articles sur le football apparaissent à la suite dans le classement et à ce que les 5 articles choisis aléatoirement apparaissent en bas de tableau.

Tableau 1 Ordre de classement des textes du Corpus 1 (articles de journaux) par types clés normalisés et tokens clés normalisés

	0.05(1069)		0.01(610)		0.001(234)		0.0001 (150)	
	Types	Tokens	Types	Tokens	Types	Tokens	Types	Tokens
1	Islam (5)	Islam (6)	Islam (5)	Islam (6)	Islam (7)	Islam (6)	Islam (5)	Islam (6)
2	Islam (2)	Islam (5)	Islam (7)	Islam (5)	Islam (5)	Islam (4)	Islam (4)	Islam (5)
3	Islam (7)	Islam (4)	Islam (6)	Islam (7)	Islam (4)	Islam (7)	Islam (7)	Islam (4)
4	Islam (4)	Islam (7)	Islam (4)	Islam (4)	Islam (6)	Islam (5)	Islam (8)	Islam (5)
5	Islam (6)	Islam (3)	Islam (2)	Islam (3)	Islam (8)	Islam (8)	Islam (6)	Islam (8)
6	Islam (3)	Islam (1)	Islam (3)	Islam (8)	Islam (3)	Islam (3)	Islam (2)	Football (11)
7	Review (19)	Islam (2)	Islam (1)	Islam (2)	Islam (1)	Football (11)	Islam (1)	Islam (3)
8	Islam (1)	Obituary (16)	Islam (8)	Football (11)	Islam (2)	Obituary (16)	Islam (3)	Islam (9)
9	Obituary (16)	Islam (8)	Review (19)	Obituary (16)	Islam (9)	Islam (9)	Islam (9)	Islam (2)
10	Islam (8)	Review (19)	Obituary (16)	Science (17)	Football (11)	Islam (2)	Obituary (16)	Football (14)
11	Science (17)	Football (11)	Football (14)	Islam (1)	Obituary (16)	Islam (1)	Football (11)	Islam (1)
12	Football (11)	Science (17)	Football (11)	Football (14)	Islam (10)	Football (14)	Islam (10)	Review (19)
13	Football (14)	Football (14)	Science (17)	Islam (9)	Review (19)	Science (17)	Review (19)	Obituary (16)
14	Islam (9)	Islam (9)	Islam (9)	Review (19)	Football (14)	Review (19)	Football (14)	Science (17)
15	Islam (10)	Tennis (18)	Islam (10)	Tennis (18)	Science (17)	Islam (10)	Tennis (18)	Football (12)

Tableau 1. (suite)

	0.05(1069)		0.01(610)		0.001(234)		0.0001 (150)	
16	Tennis (18)	Islam (18)	Tennis (18)	Islam (10)	Tennis (18)	Tennis (18)	Science (17)	Tennis (18)
17	Football (12)	Football (12)	Football (13)	Football (12)	Football (12)	Football (12)	Football (12)	Islam (10)
18	Football (13)	Football (13)	Football (12)	Art (20)	Football (13)	Art (20)	Football (12)	Art (20)
19	Art (20)	Art (20)	Art (20)	Football (13)	Football (15)	Football (15)	Art (20)	Football (13)
20	Football (15)	Football (15)	Football (15)	Football (15)	Art (20)	Football (13)	Football (15)	Football (15)

Les résultats présentés dans le Tableau 1 montrent que l'analyse menée par *ProtAnt* classe avec fiabilité les fichiers Islam comme étant les plus prototypiques du corpus dans son ensemble indépendamment des valeurs seuils utilisées. Par ailleurs, le classement des fichiers demeure relativement stable, qu'on utilise le classement par types ou par tokens. Il est rassurant d'observer qu'indépendamment des protocoles expérimentaux, les 5 fichiers les mieux classés sont ceux sur l'Islam. On peut observer dans le tableau que les fichiers 4,5,6 et 7 figurent systématiquement dans les cinq premières positions, classement qui suggère que ces quatre fichiers sont les plus prototypiques du corpus dans son ensemble.

Cependant, les résultats révèlent également que les fichiers 9 et 10 sont peut-être moins typiques que les autres fichiers sur l'Islam. Le fichier 9 reprend un discours de Tony Blair sur l'Islam, discours également mentionné dans les fichiers 4, 5, 7, 8. En revanche, le fichier 9 aborde à peine ce discours et la majorité de l'article traite d'une potentielle menace de la Chine et de la Russie contre l'occident plutôt que de l'Islam. Le fichier 10 diffère légèrement des 9 autres fichiers sur l'Islam en cela qu'il ne contient pas de points de vue d'hommes politiques sur l'Islam, mais narre plutôt l'histoire d'une école expliquant aux parents que les enfants devraient assister à des ateliers sur l'Islam ou être étiquetés racistes.

En observant les résultats du Tableau 1, on peut s'interroger sur les raisons justifiant la présence du fichier 19 (une critique) en milieu de classement dans la plupart des protocoles expérimentaux en dépit de ce qu'il est l'unique fichier de ce type. L'observation des 20 mots-clés les plus forts tout au long des 20 fichiers classés par ordre de force donne peut-être un début de réponse : *Islam, Muslims, Batten, Blair, football, Joffrey, Muslim, Mara, Brotherhood, Islamic, Assad, Kundnani, Syria, Arabia, Saudi, manager, Sansa, UKIP, pastor*. Tandis que 14 de ces mots- clés font référence à des éléments d'information sur l'Islam, 2 sont utilisés dans le fichier 19, la critique (*Joffrey* et *Sansa* avec respectivement 12 et 7 occurrences). Ces occurrences suffisent à propulser la critique en milieu de tableau.

Les cinq dernières positions dans le classement affiché par le Tableau 1 sont presque toujours occupées par les fichiers 12, 13, 15 et 20. Les trois premiers traitent du football et on aurait peut-être pu s'attendre à les voir apparaître plus proches du milieu de classement dans la mesure où le corpus contient cinq fichiers sur le football. Toutefois, bien que tous ces fichiers soient sur le football, chacun aborde un aspect différent de ce sport ; le fichier 12 est un article autobiographique sur les footballeurs changeant d'équipe, le fichier 13 fait référence à une condamnation pour mauvais comportement dans le monde du football et le fichier 15 évoque le début de la nouvelle saison de la ligue de football. À contrario des fichiers sur l'Islam qui contiennent principalement une série de mots-clés liés à Tony Blair, *the Muslim Brotherhood* [*Ndt : les Frères musulmans*], *Syria*, et Assad, les fichiers sur le football comptent peu de mots-clés et également peu de mots-clés partagés, indiquant qu'ils offrent une variété lexicale plus riche que celle à laquelle on aurait pu s'attendre.

4.2 Expérience 2 : Identification de romans prototypiques

Une deuxième expérience a été élaborée pour déterminer si *ProtAnt* était capable de classer correctement une série de textes plus longs d'un genre totalement différent (de la fiction plutôt que de l'information). Le Corpus 2 comprend 10 extraits du roman *Dracula* (1897), cinq de *Frankenstein* (1818) et cinq autres issus de romans individuels (deux datant du 19e s., trois écrits à la fin du 20e s. ou au début du 21e s.). La taille globale du corpus est de 40 759 tokens avec un nombre moyen de 2 038 tokens par fichier (environ la taille d'un fichier dans la famille Brown). À nouveau, le corpus BE06 (Baker 2009) a été utilisé comme corpus de référence pour obtenir les mots-clés.

Le Tableau 2 présente le classement des textes du Corpus 2 selon les valeurs normalisées des types clés et des tokens clés, à nouveau en utilisant le log *likelihood* pour la mesure statistique de *keyness* et quatre différentes valeurs seuils de p, c.-à-d. 0.05, 0.01, 0.001, et 0.0001. Comme dans le Tableau 1, les valeurs entre parenthèses dans les titres des colonnes indiquent le nombre de mots-clés obtenus pour chacune des valeurs seuils, et les nombres entre

parenthèses dans les cellules indiquent le numéro du fichier. Un nom court de circonstance a été donné à chacun des fichiers, et les fichiers en rapport avec *Dracula* ont été grisés pour être plus aisément identifiables. À nouveau, on s'attendrait à ce que les 10 fichiers Dracula apparaissent comme les plus typiques, suivis par les cinq Frankenstein, les 5 autres romans apparaissant en bas de tableau.

Tableau 2 Ordre de classement des textes du Corpus 2 (romans) par types clés normalisés et tokens clés normalisés*

	0.05 (1794)		0.01 (1175)		0.001 (442)		0.0001 (274)	
	Types	Tokens	Types	Tokens	Types	Tokens	Types	Tokens
1	D. (10)	D. (8)	D. (9)	D. (8)	D. (8)	D. (8)	D. (8)	D. (8)
2	D. (3)	D. (3)	D. (7)	D. (6)	D. (6)	D. (6)	D. (7)	D. (2)
3	D. (7)	D. (6)	D. (8)	D. (5)	D. (7)	D. (5)	D. (6)	D.(5)
4	D. (8)	D. (10)	D. (10)	D. (9)	D. (9)	D. (7)	D. (9)	F. (13)
5	D. (9)	F. (13)	D. (6)	F. (13)	D. (10)	D. (2)	D. (10)	D. (6)
6	F. (12)	D. (9)	D. (3)	D. (2)	D. (2)	D. (4)	D. (5)	D. (7)
7	D. (6)	D. (5)	D. (2)	D. (1)	D. (5)	D. (9)	D. (2)	D. (3)
8	F. (13)	D. (2)	D. (5)	D. (3)	F. (15)	F. (15)	D. (4)	D. (1)
9	D. (2)	D. (1)	F. (13)	D. (7)	D. (4)	F. (13)	D. (3)	F. (15)
10	F. (11)	D. (7)	F. (12)	D. (10)	D. (3)	D. (3)	F. (15)	D. (4)
11	D. (5)	F. (12)	D. (4)	D. (4)	F. (14)	D. (1)	F. (13)	D. (9)
12	D.(4)	F. (15)	F. (14)	F. (15)	F. (13)	D. (10)	D. (1)	F. (14)
13	F. (14)	D. (4)	F. (15)	F. (14)	D. (1)	F. (14)	F. (14)	D. (10)
14	D. (1)	F. (14)	D. (1)	F. (12)	F. (12)	F. (12)	M. (17)	F. (12)
15	F. (15)	F. (11)	F. (11)	F. (11)	M. (17)	F. (11)	F. (12)	F. (11)
16	J. (16)	J. (16)	J. (16)	J. (16)	F. (11)	T. (19)	F. (11)	T. (19)
17	M. (17)	M. (17)	M. (17)	M. (17)	J. (16)	M. (17)	J. (16)	J. (16)
18	H. (20)	H. (20)	H. (20)	T. (19)	H. (20)	J. (16)	H. (20)	M. (17)
19	T. (19)	T. (19)	I. (18)	H. (20)	T. (19)	H. (20)	T. (19)	H. (20)
20	I. (18)	I. (18)	T. (19)	I. (18)	I. (18)	I. (18)	I. (18)	I. (18)

D. = *Dracula*; F. = *Frankenstein*; T. = *The Intimate Adventures of a London Call Girl*; H. = *Harry Potter and the Deathly Hallows*; I. = *It*; M = *Moonstone*, J. = *Jane Eyre*

Comme observé lors de la première expérience, les résultats du Tableau 2 montrent qu'un changement de la valeur seuil p a peu d'incidences sur le classement des fichiers. De même, il existe peu de différences selon que l'on classe les fichiers par le nombre de types clés ou de tokens clés qu'ils contiennent. Dans six protocoles sur huit, le fichier 8 (*Dracula*) apparait comme le plus typique. Il est intéressant d'observer que le fichier 1 est classé comme le fichier *Dracula* le moins typique du corpus dans quatre des protocoles (son classement pouvant même descendre *ProtAnt* 9 jusqu'à la 14e position) indiquant qu'il ne s'agit pas d'un fichier très typique. Ce texte s'ouvre avec la description du séjour de Jonathan Harker au château de Dracula en Transylvanie, donc très tôt dans la narration et avant l'introduction de la majorité des personnages du roman. Par comparaison, les autres textes concernent des événements ayant lieu plus tard dans la narration, et se déroulant en grande partie au Royaume-Uni.

Dans cette expérience, l'analyse menée par *ProtAnt* se révèle à nouveau efficace dans l'identification de fichiers atypiques indépendamment de la valeur seuil puisque que les trois romans les plus récents (*The IntimatAdventures of a London Call Girl* (2007), *Harry Potter and the Deathly Hallows* (2007) and *It* (1986)) figurent presque systématiquement dans les trois positions les plus basses du classement. *It* est le seul roman américain du corpus, les autres ayant été écrits par des auteurs britanniques. Dans la mesure où certains mots-clés présentent une orthographe britannique, comme *sympathised* [Ndt : orthographié *sympatized* en anglais américain] ou des références à des concepts britanniques, tel que *solicitor*, on peut comprendre pourquoi ce fichier s'est généralement classé comme étant le moins typique.

4.3 Expérience 3 : Identification de textes prototypiques dans un corpus étendu

La troisième expérience que nous avons menée a été élaborée pour évaluer la capacité de *ProtAnt* à identifier des textes prototypiques dans un corpus plus étendu, dans lequel il ne serait pas possible de désigner avec fiabilité les fichiers les

plus typiques sans l'aide de l'analyse automatisée. Pour cette expérience, nous avons utilisé le corpus AmE06 (Potts & Baker 2012) d'un million de mots, en gardant le corpus BE06 (Baker 2009) comme corpus de référence. Le corpus AmE06 présente une structure identique à celle du BE06 (décrite précédemment) à l'exception du fait qu'il contient des textes américains plutôt que britanniques. En utilisant *ProtAnt*, l'analyse devrait identifier les textes les plus "américains" par nature (comparés au corpus de référence britannique). En classant les fichiers du corpus AmE06 selon le nombre de mots-clés qu'ils contiennent après comparaison avec le BE06, les fichiers les plus distinctifs lorsque cette comparaison est faite peuvent être considérés comme les plus typiques du AmE06.

En utilisant le même log *likelihood* pour la mesure statistique de *keyness* et une valeur seuil p de 0,0001, les trois fichiers qui se sont démarqués par leur nombre de mots-clés ont été H24, H17 et H13 (les fichiers H contiennent des textes du registre "Divers : documents gouvernementaux, rapports industriels, etc... "). Le fichier H24 provient du Département du Trésor américain et porte sur les États et les impôts généraux au niveau local. Un nombre élevé de références aux États américains et à leurs villes, qui apparaissent probablement beaucoup moins dans BE06, explique peut-être pourquoi ce fichier a été retenu comme le plus typique de AmE06 (comparé à BE06). Le fichier H17 est un document gouvernemental traitant de la nomination d'un du Bureau administratif des cours de justice américaines et contient également des références aux États américains ainsi qu'à des concepts propres aux États-Unis tels que *federal* et *congress* ou encore des graphies américaines de mots fréquemment cités tels que *program* [*Ndt : s'écrit programme en anglais britannique*], *toward* [*Ndt : l'anglais britannique utilise plus towards (vers)*] et *center* [*Ndt : s'écrit centre en anglais britannique*]. H13 est un document officiel émanant du Congrès et contient des mots-clés similaires à ceux de H17.

Les trois fichiers AmE06 apparaissant comme les moins typiques de l'anglais américain (comparé à BE06) sont N06, P27 et P19. Les fichiers N et P correspondent respectivement aux catégories "Aventure" et "Fiction romantique". Le texte N06 provient d'un roman d'aventure se déroulant au Vietnam en 1975, et l'extrait dont il est issu est un passage descriptif dans lequel le personnage principal saute d'un avion. Le texte P27 vient d'un roman d'amour historique dont l'action se passe en France en 1885, et par conséquent contient peu d'éléments de langage le désignant comme "américain". P19 est une description d'une rencontre charnelle entre deux personnages, et les mots-clés sont plus liés à cette rencontre (par exemple : *kissed*, *hear*, *cried*, *finger*, *willing* [*Ndt- embrassé, entendre, gémir, doigt, envie*]) qu'à un contexte spécifiquement américain.

Dans cette expérience, une lecture approfondie des textes identifiés par *ProtAnt* comme étant typiques et atypiques révèle clairement les raisons expliquant un tel classement des textes par l'outil. Ce qui confère à l'outil un haut degré de validité apparente (en plus de sa fiabilité intrinsèque).

4.4 Expérience 4 : Impact du corpus de référence sur l'identification de textes prototypiques

Dans l'analyse menée par *ProtAnt*, le choix du corpus de référence joue un rôle important dans la détermination des textes amenés à être sélectionnés comme typiques ou non. Par exemple, une comparaison entre AmE06 et BE06 se focalisera sur la typicalité en termes d'"américanité" des fichiers de AmE06. Pour étudier l'impact du corpus de référence sur la sélection des textes, nous avons conduit une quatrième expérience au cours de laquelle nous avons 10

utilisé le corpus Brown en guise de corpus de référence contre AmE06 et non plus BE06. Dans cette expérience l'utilisation de deux corpus de textes américains implique que c'est le temps qui devient le facteur important. En effet, le corpus Brown contient des échantillons de textes datant de 1961 tandis que AmE06 se compose de textes écrits 45 ans plus tard. Par conséquent, les textes typiques devraient être ceux qui étaient les plus typiques (représentatifs) de 2006.

Il est intéressant de noter que H24 apparait encore une fois comme le texte le plus typique, suivi par H21 et G40. Bien que H24 soit un texte gouvernemental relativement insipide au sujet des impôts, on s'y adresse directement au lecteur avec un usage important de mots-clés pronoms personnels de la deuxième personne, *you et your* (une caractéristique du discours personnalisé qui est finalement devenue plus populaire après 1961). H24 fait aussi référence à l'ouragan Katrina de 2005, thème abordé dans 15 fichiers différents au sein du corpus. AmE06.H21 contient des mots-clés spécifiques à la période qui s'avèrent révélateurs du climat politique de l'époque *terrorism, Bush, preparedness* [*Ndt : préparation*] *et Palestinian*, tandis que G40 est une autobiographie à la première personne écrite par une femme ayant grandi dans le Mississippi et qui, en plus de marques de discours personnalisé, contient des références aux problèmes touchant à la discrimination raciale et sexiste, ce qui suggère un point de vue plus proche de celui du 21e s. sur ces sujets.

Les trois fichiers les moins typiques de AmE06 après confrontation avec le corpus Brown sont N02, G21 et G71. N02 est une histoire d'aventure qui se déroule dans la jungle. Partant du principe que celle-ci se passe en

dehors du contexte américain et ne fasse pas référence aux technologies modernes ou à des événements récents, elle aurait tout aussi bien pu être écrite en 1961 qu'en 2006. G21 est extrait d'un texte sur la guerre civile américaine, donc à nouveau sans relation avec 1961 ou 2006, tandis que G71 est un texte sur un artiste du 20e s., Nozkowski, encore une fois sans lien avec les périodes étudiées. Au même titre que lors de l'expérience 3, les fichiers typiques et atypiques de AmE06 après confrontation avec le corpus Brown semblent fiables.

4.5 Expérience 5 : Identification de textes atypiques dans un corpus étendu

Une cinquième expérience a été élaborée pour évaluer la capacité de *ProtAnt* à identifier des textes non pertinents dans un corpus. Pour tester cette utilisation de l'outil, nous avons à nouveau utilisé AmE06 (avec BE06 comme corpus de référence), mais en sélectionnant cette fois-ci l'ensemble des fichiers de l'un des 15 registres plus un fichier d'un registre différent (en utilisant un générateur de nombre aléatoire). Nous avons répété ce protocole pour chaque registre. Idéalement, *ProtAnt* devrait classer les fichiers "non pertinents" vers le bas dans le classement des fichiers typiques. Le Tableau 3 montre les résultats basés sur un classement par types clés normalisés en utilisant le log *likelihood* pour la mesure statistique du *keyness* avec une valeur seuil p de 0,0001.

Les résultats présentés dans le Tableau 3 montrent que dans 10 cas sur 15, le fichier non pertinent est correctement identifié comme étant en bas de classement ou très proche de cette position, signe d'une excellente réussite dans deux cas sur trois. Toutefois, dans les cinq autres cas, c.-à-d. ligne F, G, J, M et R, l'outil est moins efficace dans la détection des fichiers non pertinents. La faible performance de F ("Popular lore") [*Ndt : "Connaissances populaires"*], G ("Belles lettres, biographies, essays") et R ("Humour") peut s'expliquer par l'emploi dans ces registres de termes quelque peu délicats à définir et par une plus grande variété de langage d'un fichier à l'autre que, par exemple, dans P ("Romance and love story ») ou A ("Press : Reportage").

Ces motifs peuvent rendre l'identification d'un texte non pertinent plus ardue car les similarités entre les fichiers d'un registre sont moindres. À l'opposé, on peut affirmer que le registre académique (J) est clairement défini. Ici, le texte non pertinent était issu de R ("Humour"). Le fichier R8 provient d'un roman dont le narrateur est un génie de neuf ans qui s'auto-proclame inventeur, entomologiste, archéologue amateur et origamiste. Le texte contient des mots-clés tels que *learning, dictionary, humans et age*. L'extrait de ce fichier aborde brièvement et tour à tour les sujets suivants : l'entomologie, le jujitsu, l'accouchement, la musique, le magazine *National Geographic*, les gratte-ciels et les limousines. Il est quelque peu inhabituel (même pour un texte de la catégorie R), ce qui peut expliquer pourquoi il n'est pas reconnu comme non pertinent par rapport au registre J. Enfin, la pauvre performance du registre M ("Science-fiction") peut être influencé par le fait que le texte non pertinent provient d'un registre similaire N ("Adventure and western"), c.-à-d. qu'ils appartiennent tous deux au domaine de la fiction.

Tableau 3 Classement des fichiers non pertinents par une analyse *ProtAnt* utilisant les types clés normalisés calculés grâce au log *likelihood* et une valeur seuil p de 0,0001.

Category	Register	Outlier file	Ranking of outlier file
A	Press: Reportage	K12	40/40
B	Press: Editorial	L9	28/28
C	Press: Reviews	P13	18/18
D	Religion	C8	18/18
E	Skills, trades and hobbies	N7	34/37
F	Popular lore	A3	28/49
G	Belles lettres, biographies, essays	M6	40/76
H	Miscellaneous: Government documents, industrial reports etc.	L13	30/31
J	Academic prose in various disciplines	R8	8/80
K	General fiction	E15	28/30

Cette expérience nous permet de conclure que *ProtAnt* est efficace dans l'identification de fichiers inconnus mal catégorisés dans le cas d'un corpus comprenant des fichiers issus d'un seul registre, aisément identifiable, et un fichier non pertinent issu d'un registre clairement différent.

5. Réflexions et conclusion

Les expériences que nous avons menées avec *ProtAnt* confirment l'utilité du recours au concept de mots-clés pour identifier des textes prototypiques, et présentent des résultats largement en adéquation avec les issues attendues.
En général, *ProtAnt* est efficace, il nous faut cependant noter que ses performances ne sont pas non plus parfaites. Toutefois, dans les cas où le classement des textes ne correspondait pas aux attentes, une lecture plus approfondie de ces derniers a permis d'apporter des explications raisonnables, telles que des idiosyncrasies inhérentes aux textes eux-mêmes.

Dans la mesure où, pour *ProtAnt*, la notion de typicalité des textes repose sur les mots-clés générés grâce à un corpus de référence, ce dernier peut lui-même être ajusté pour se prêter à des problématiques de recherche spécifiques. Lors des expériences menées avec AmE06, le choix du corpus de référence nous a permis d'étudier l'"américanité", en le comparant avec le corpus BE06, ou la "2006-ité", en le comparant avec le corpus Brown. Cette particularité de *ProtAnt*, c'est-à-dire cette capacité à évaluer la prototypicalité des textes individuels en fonction d'une sorte de catégorisation des caractères distinctifs au niveau du corpus offre des avantages, mais peut également être jugée contraignante si aucun corpus de référence pertinent n'est disponible. Dans ces derniers cas, une solution possible serait de traiter chaque fichier comme un pseudo corpus cible, et l'ensemble du corpus comme un corpus de référence. Cette méthode nécessiterait que chaque fichier (et le corpus dans son ensemble) soit suffisamment ample pour que la mesure statistique des mots-clés puisse être valable, mais l'interprétation des résultats serait simple (bien que peut-être contre-intuitive) : les fichiers comprenant le moins de mots-clés seraient les plus prototypiques du corpus dans son ensemble. En effet, les mots-clés générés par cette méthode désigneraient des usages plus atypiques du langage dans les textes individuels du corpus. Ainsi, les fichiers présentant le plus de mots-clés seraient les moins typiques et ceux contenant le moins de mots-clés se révéleraient les plus typiques. Indéniablement, l'un des inconvénients d'une telle approche est que la prototypicalité ne pourrait donc nullement être ajustée à une catégorie globale de caractères distinctifs, telle que l'"américanité" vu plus tôt. Nous espérons étudier cette méthode alternative de détection de prototypes et d'autres possibilités dans nos prochains travaux. Nous espérons également étudier plus en avant l'interaction entre les classements des textes et les valeurs seuils de significativité, le volume moyen des fichiers, le nombre des fichiers du corpus, et le choix du décompte des types clés ou des tokens clés. Dans cet article, nous avons abordé les possibilités offertes par *ProtAnt* dans la recherche de textes prototypiques destinés à une analyse approfondie plus qualitative à une étape ultérieure. Toutefois, d'autres applications de l'outil peuvent être envisagées. Par exemple, les linguistes de corpus pourraient utiliser l'outil pour identifier rapidement un texte typique à utiliser pour illustrer un exemple ou à examiner pour compléter leurs analyses sur corpus.

Dans l'enseignement des langues, l'outil serait utile aux formateurs et examinateurs pour identifier et examiner un petit nombre de dissertations d'étudiants ayant des niveaux de compétence différents afin d'avoir un aperçu de la disparité de ces niveaux. Les textes prototypiques identifiés par *ProtAnt* pourraient également faire office de modèles à proposer aux étudiants dans le cadre d'activités pédagogiques. *ProtAnt* pourrait également servir d'outil pour former les spécialistes de la police scientifique à décider si un texte devrait être catégorisé comme de la littérature extrémiste, ou encore pour identifier le style propre à un auteur. Autre exemple d'application dans la construction de corpus, les chercheurs du domaine pourraient utiliser l'outil pour identifier des textes qui sont atypiques d'un genre particulier et qui par conséquent devraient être catégorisés comme tels ou évincés. L'identification de textes atypiques pourrait également aider lors d'analyses de discours, en pointant des discours "résistants" ou "minoritaires" qui vont à contre-courant.

Nous avons souhaité laisser l'outil *ProtAnt* gratuit et disponible, dans l'espoir que des chercheurs et des formateurs puissent l'utiliser pour identifier rapidement des textes prototypiques et atypiques au sein d'un vaste ensemble de données et utiliser ces textes dans le cadre d'analyses et d'applications du langage plus sophistiquées.

Bibliographie

Anthony, L. (2014). *AntConc* (*Version 3.4.3*) [Logiciel]. Tokyo, Japon : Université de Waseda. Extrait de http://www.laurenceanthony.net/software/antconc/ (consulté pour la dernière fois en mai 2015).

Anthony, L., & Baker, P. (2015). *ProtAnt* (*Version 1.0*) [Logiciel]. Tokyo, Japon : Université de Waseda. Extrait de http://www.laurenceanthony.net/software/protant/ (consulté pour la dernière fois en mai 2015).

Bahrololoum, A., Nezamabadi-pour, H., Bahrololoum, H., & Saeed, M. (2012). A prototype classifier based on gravitational search algorithm. *Applied Soft Computing*, 12(2), 819-825. DOI : 10.1016/j.asoc.2011.10.008

Baker, P. (2009). The BE06 Corpus of British English and recent language change.*International Journal of Corpus Linguistics*, 14(3), 312-337. DOI : 10.1075/ijcl.14.3.02bak

Baker, P., Gabrielatos, C., & McEnery. T. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge, RU : Cambridge University Press. DOI : 10.1017/CBO9780511920103

Caldas-Coulthard, C. R., & van Leeuwen, T. (2013). Teddy bear stories. In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 35-60). Los Angeles, CA, É.U.A. : Sage. (Travail original publié en 2003).

Chen, L., Guo, G., & Wang, K. (2011). Class-dependent projection based method for text categorization. *Pattern Recognition Letters*, 32(10), 1493-1501. DOI : 10.1016/j.patrec.2011.01.018

Chouliaraki, L. (2013). Political discourse in the news: Democratizing responsibility or aestheticizing politics? In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 97-118). Los Angeles, CA, É.U.A. : Sage. (Travail original publié en 2000).

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, 29(4), 433-447. DOI : 10.1016/0306-4573(93)90039-G

Durfee, A., Visa, A., Vanharanta, H., Schneberger, S., & Back, B. (2007). Mining text with the Prototype- matching method. I*nformation Resources Management Journal*, 20(3), 19-31. DOI : 10.4018/irmj.2007070102

Ehrlich, S. Z., & Blum-Kulka, S. (2013). Peer talk as a 'double opportunity space': The case of argumentative discourse. In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 145-168). Los Angeles, CA, É.U.A.: Sage. (Travail original publié en 2010).

Fayed, H. A., Hashem, S. R., & Atiya, A. F. (2007). Self-generating prototypes for pattern classification. *Pattern Recognition*, 40(5), 1498-1509. DOI : 10.1016/j.patcog.2006.10.018

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press (1996-2005). *Journal of English Linguistics*, 36(1), 5-38. DOI : 10.1177/0075424207311247

Gavriely-Nuri, D. (2013). If both opponents "extend hands in peace", why don't they meet? Mythic metaphors and cultural codes in the Israeli peace discourse. In R. Wodak, (Ed.). *Critical Discourse Analysis Volume II: Methodologies* (pp. 169-186). Los Angeles, CA, É.U.A. : Sage. (Travail original publié en 2010).

Gries, S. Th. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, 1, 1-27. DOI : 10.1075/arcl.1.02gri

Hardie, A. (2014). *CQPWeb (Version 3.1.10)* [Logiciel]. Lancaster, RU : Université de Lancaster. Extrait de https://cqpweb.lancs.ac.uk/ (consulté pour la dernière fois en mai 2015).

Khosravinik, M. (2010). The representation of refugees, asylum seekers and immigrants in British newspapers: A critical discourse analysis. *Journal of Language and Politics*, 9(1), 1-28. DOI : 10.1075/jlp.9.1.01kho

Kloptchenko, A., Back, B., Visa, A., Toivonen, J., & Vanharanta, H. (2002). Toward content based retrieval from scientific text corpora. In *Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia*, 5-10 September 2002 (pp. 444-449). Washington, DC, É.U.A. : IEEE Computer Society. DOI : 10.1109/ICAIS.2002.1048170

Kloptchenko, A., Magnusson, C., Back, B., Visa, A., & Vanharanta, H. (2004). Mining textual contents of financial reports. *The International Journal of Digital Accounting Research*, 4(7), 1-29.

Labov, W. (1973). The boundaries of words and their meanings. In J. Fishman (Ed.), *New Ways of Analyzing Variation in English* (pp. 340-73). Washington, DC, É.U.A. : Georgetown University Press.

Leńko-Szymańska, A. (2006). The curse and blessing of mobile phones: A corpus-based study into American and Polish rhetorical conventions. In A. Wilson, D. Archer & P. Rayson (Eds.), *Corpus Linguistics around the World* (pp. 141-151). Londres, RU : Rodopi.

Machin, D., & Suleman, U. (2013). Arab and American computer war games: The influence of a global technology on discourse. In R. Wodak, (Ed.), *Critical Discourse Analysis Volume II: Methodologies* (pp. 229- 252). Los Angeles, CA, É.U.A. : Sage. (Travail original publié en 2006).

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *An Introduction to Information Retrieval*. Cambridge, RU : Cambridge University Press. DOI : 10.1017/CBO9780511809071

Potts, A., & Baker. P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics*, 17(3), 295-324. DOI : 10.1075/ijcl.17.3.01pot

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192-

233. DOI : 10.1037/0096-3445.104.3.192.
Sajid, F. (2013). Critical discourse analysis of news headline about Imran Khan's peace march towards Wazaristan. *Journal of Humanities and Social Science*, 7(3), 18-24. DOI : 10.9790/0837-0731824.
Scott, M. (2014). *WordSmith Tools (Version 6)* [Logiciel]. Liverpool, RU : Lexical Analysis Software. Extrait de http://www.lexically.net/wordsmith/index.html (consulté pour la dernière fois en mai 2015).
van Leeuwen, T. (1996). The representation of social actors. In C. R. Caldas Coulthard & M. Coulthard (Eds.), *Texts and Practices* (pp. 32-70). Londres, RU : Routledge.
Visa, A., Toivonen, J., Vanharanta, H., & Back, B. (2001). Prototype matching: Finding meaning in the books of the bible. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences* (HICSS-34), Hawaii, É.U.A., 3-6 Janvier 2001 (pp. 3002).Washington, DC, É.U.A. : IEEE Computer Society.
Widdowson, H. G. (2004). *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford, RU : Blackwell. DOI : 10.1002/9780470758427
Wodak,R(2013).*Critical Discourse Analysis*. LosAngeles, CA, É.U.A. : Sage. DOI : 10.4135/9781446286.

Coordonnées des auteurs

Laurence Anthony
Faculty of Science and Engineering
Waseda University
3-4-1 Ohkubo, Shinjuku-ku
Tokyo 169-8555
Japan
anthony@waseda.jp

Paul Baker
Department of Linguistics and English Language
Lancaster University
Bailrigg
Lancaster LA1 4YL
UK
j.p.baker@lancaster.ac.uk

Traduit de l'anglais par Jean-Yves Préault (jeanyvespreault@gmail.com)