Expressions différenciées des besoins informationnels en Langue Naturelle : construction de profils utilisateurs en fonction des tâches de recherche d'informations

Marilyne Latour[1, 2]
(1) Université Grenoble-Alpes, GRESEC, F-38040 Grenoble
(2) ReportLinker, 4 Rue Montrochet, 69002 Lyon
marilyne.latour@ac-grenoble.fr, marilyne.latour@reportlinker.com

Résumé. Devant des collections massives et hétérogènes de données, les systèmes de RI doivent désormais pouvoir appréhender des comportements d'utilisateurs aussi variés qu'imprévisibles. L'objectif de notre approche est d'évaluer la façon dont un utilisateur verbalise un besoin informationnel à travers un énoncé de type « expression libre » ; appelé langage naturel (LN). Pour cela, nous nous situons dans un contexte applicatif, à savoir des demandes de remboursement des utilisateurs d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli via ce moteur, les demandes en LN sur 5 années consécutives totalisant un corpus de 1398 demandes. Nous avons alors comparé l'expression en tant que tel du besoin informationnel en fonction de la tâche de recherche d'informations (RI) de l'utilisateur.

Abstract. With the massive and heterogeneous web document collections, IR system must analyze the behaviors of users which are unpredictable and varied. The approach described in this paper provides a description of the verbalizations of the information need in natural language. For this, we used data collected (*i.e.* users' complaints in natural language) through a search engine dedicated to economic reports in French over 5 consecutive years totaling a corpus of 1398 natural language requests. Then, we compared the expression as such of the information need according to the IR task.

Mots-clés : Recherche informations ; Besoin informationnel, Expression et interprétation des besoins ; Formulation question ; Langage naturel ; comportement utilisateur ; tâches de recherche d'informations.

Keywords: Information Retrieval ; Information Need, Query formulation and Query Expression ; Query Formulation ; Natural Language ; User Behavior ; IR task.

1 Introduction

Le 26 septembre 2013, *Google* annonce lors d'une conférence de presse pour fêter ses 15 ans, son nouvel algorithme baptisé *« Hummingbird »*. Ce nouvel algorithme s'éloigne de la logique de la recherche d'informations (RI) des requêtes en mots-clés pour s'ouvrir aux requêtes en langage naturel (LN). C'est un changement majeur pour le géant de la recherche dont l'objectif affiché est d'être capable de traiter des requêtes plus complexes et plus longues tout en prenant en compte le sens des mots dans leur contexte. Si l'on reprend l'exemple de Danny Sullivan [1] à la requête « Quel est l'endroit le plus près de chez moi pour acheter un *iPhone* 5S ? », *Google* devrait pouvoir prendre en compte et lier les notions « endroit près de chez moi », « acheter », « iPhone 5S » sous-entendu : l'objet désiré, le type d'action exercée sur l'objet (celui d'acheter) et la couverture géographique (à proximité de là où est localisé l'internaute « chez moi »). L'objectif est donc double : *(a)* comprendre la demande dans sa globalité pour lui donner un « sens » plus exact et *(b)* capitaliser d'autres informations que celles apparaissant sur les pages de recherche. Plus largement et au travers de la demande exprimée, c'est l'interprétation du besoin informationnel de l'utilisateur qui est visé pour le premier point. Ainsi, dans l'exemple cité, cela reviendrait à comprendre que la demande concerne le domaine de la téléphonie mais aussi que le besoin s'étend également à une volonté d'acheter un appareil. Le second point, lui, s'intéresse au lieu d'habitation si l'internaute a déjà renseigné cette information. Également des requêtes précédemment effectuées sur le moteur peuvent aider à « contextualiser » la demande notamment sur les centres d'intérêts de l'utilisateur. Cet algorithme rend compte d'une prise de conscience de la part des développeurs des systèmes de recherche d'informations (SRI) de traiter plus efficacement les requêtes avec une

1. « FAQ : All About The New Google « Hummingbird » Algorithm », Danny Sullivan, 27/09/2013, disponible sur : http ://searchengineland.com/google-hummingbird-172816 .

meilleure contextualisation du besoin informationnel. Dans ce sens, plusieurs pistes d'améliorations sont en cours : les premières ont pour objectif de mieux définir le besoin informationnel ; les buts des utilisateurs ainsi que les tâches sous-jacentes à la réalisation de celui-ci ; les secondes portent sur une meilleure contextualisation du contexte de l'utilisateur et de son environnement.

2 Le besoin informationnel contextualisé par la tâche de RI

Pour [Cabanac11], un utilisateur est un individu qui, dans un contexte donné -professionnel ou personnel- a besoin des résultats de sa requête médiatisé par un système informatisé -un logiciel quelconque, ou un système de recherche d'informations- pour réaliser une tâche avec un objectif spécifique. Dans ce cadre, le besoin informationnel est la prise de conscience d'un utilisateur lorsqu'il est confronté à l'exigence d'une information qui lui est à la fois déficiente et nécessaire. Ce besoin apparaît comme étant ancré, déterminé par la position qu'occupe un individu dans son environnement social [Lecoadic98] ou de travail. Or, de nombreuses études dont notamment celles de [Ingwersen05], [Ramirez06] expriment un décalage entre le besoin informationnel et son expression à travers un SRI. Ce décalage s'exprime généralement à travers des requêtes uniquement. Or ces requêtes sont généralement courtes (entre 2 et 3 termes) exprimant un but plus ou moins explicite [Strohmaier08]. Celles qui ont un but explicite sont des requêtes qui décrivent avec précision leur intention de recherche, *i.e.* pouvant être reliées à un but spécifique, de manière reconnaissable et non ambiguë. Exemple : *« acheter une voiture »*, *« réparer une voiture »*, *« aller à Miami »*, alors que celles qui ont un but implicitement exprimé sont des requêtes où il est difficile d'obtenir le but spécifique des intentions de recherche. L'exploitation du contexte de la tâche de recherche permettrait de mieux prédire le type de besoin des requêtes traduisant la nature de la tâche de recherche, en exploitant les caractéristiques morphologiques des requêtes ainsi que le profil et le contexte de la session.

Un des premiers constat est que pour faire face à ce décalage, il est nécessaire de travailler sur la compréhension de la tâche de RI. Pour ce faire, une variété d'approches en RI ont vu le jour se basant sur une taxonomie des buts de l'utilisateur ou de la tâche à effectuer. Le type de besoin inhérent à la requête est alors défini comme étant **informationnel** -lié à la recherche du contenu informationnel de documents-, **navigationnel** -lié à la recherche des sites d'accueil des personnes, organisations ou autres- ou **transactionnel** - lié à la recherche des services en ligne [Broder02], [Rose04], [Jansen08]. L'exploitation du contexte de la tâche de recherche permettrait de mieux prédire le type de besoin des requêtes traduisant la nature de la tâche de recherche, en exploitant les caractéristiques morphologiques des requêtes ainsi que le profil et le contexte de la session. Alors que la plupart des approches exploitent seulement des caractéristiques morphologiques de la requête [Kang03], [Kang05] ou des indicateurs de comportement de l'utilisateur (clics, données de consultation de la page, nombre de documents consultés...), l'objectif sous-jacent des moteurs qui proposent la LN comme moyen d'interrogation voire d'interaction : proposer un mode d'interrogation non contraignant pour l'utilisateur, afin qu'il puisse exprimer librement son besoin informationnel.

3 Expérimentation

Nous nous intéresserons ici à la phase de formulation du besoin informationnel d'un utilisateur de documents spécifiques à savoir des études de marchés économiques et ceci à travers sa formulation en LN. Plus précisément, nous voulons évaluer quelles sont les informations comme la zone géographique, la date, le type de données ou encore le prix, etc présentées dans la demande en LN en fonction de sa tâche de RI. L'objectif est, d'un point de vue linguistique, de mieux connaître les stratégies de l'utilisateur pendant sa tâche de RI ainsi que la formulation de son besoin informationnel. Le but intrinsèque est de pouvoir établir à long terme et en fonction des profils identifiés des recommandations pour les systèmes de recherche d'informations (SRI) sur les fonctionnalités à développer (aide à la recherche ou à la navigation, pré-enregistrement de filtres de recherche, environnement adapté en fonction du profil utilisateur, etc.). Nous nous situons dans un contexte applicatif spécifique, à savoir un moteur de recherche dédié à des études économiques en français : www.plusdetudes.comDétenu par la société Ubiquick, Lyon (France) avec un autre moteur de recherche en anglais : www.reportlinker.com. . Nous avons recueilli, les demandes en LN en français effectuées sur 5 années consécutives (de 2002 à 2007) de ce moteur de recherche, ainsi que les données utilisateurs (identité, fonction, domaine d'activité) totalisant un corpus de 1398 demandes en LN. Le fonds documentaire de ce moteur est d'environ 10 000 études de marché, couvrant 450 secteurs d'activités économiques et organisé autour de six axes principaux dans le thésaurus sectoriel : Agroalimentaire, Technologies de l'information et Médias, Biens et services de consommation, Sciences de la vie, Industrie, Services.

3.1 Recueil du corpus

Deux types de données ont été recueillis : *(i)* les données utilisateurs *via* le champs « Contacts » du formulaire SAV : ces données concernent l'identité de la personne, le nom de l'entreprise (ou université) de laquelle il ou elle dépend, la fonction exercée ainsi que les coordonnées téléphoniques, mail et adresse ; *(ii)* les demandes en LN *via* le champs « Expression de la demande » du formulaire SAV : ce formulaire est utilisé quand les recherches s'avèrent infructueuses ou tout simplement lorsque l' utilisateur veut obtenir un remboursement car juge être non satisfait des résultats obtenus. En analysant ces deux types de données, une typologie en fonction de trois types de tâche de RI se distingue nettement de notre corpus :

– [TACHE-CREA] : **le but de la tâche de RI est la création d'entreprise ou le lancement d'un nouveau produit ou encore d'une nouvelle marque** ; les objectifs opérationnels sont de se procurer une étude de faisabilité, d'identifier la concurrence éventuelle. Elle concerne 379 demandes de notre corpus (soit 27,11%).
– [TACHE-SCO] : **le but de la tâche de RI est la réalisation d'une tâche scolaire** ; les objectifs opérationnels sont de préparer un examen, d'écrire un mémoire, de travailler sur une étude de cas.... Elle concerne 513 demandes de notre corpus (soit 36,69%).
– [TACHE-PRO] : **le but de la tâche de RI est l'obtention d'informations dans un cadre professionnel** ; les objectifs opérationnels sont de mieux connaitre le marché, ses éléments chiffrés, d'identifier les tendances d'un marché ou d'un produit, de faire de la veille stratégique...Elle concerne 506 demandes de notre corpus (soit 36,19%).

Nous pouvons supposer que si ces caractéristiques sont fortement liées à un aspect spécifique des demandes et besoins informationnels et de manière relativement stable en fonction du contexte de la recherche, alors il serait possible de trouver des corrélations entre ces caractéristiques et les buts des utilisateurs dans le contexte considéré.

3.2 Chaîne d'analyse d'une demande en LN

Nous avons développé un environnement d'analyse chaîne de traitement en deux étapes des demandes en LN ainsi obtenus. La première concerne la segmentation des demandes en LN en blocs d'informations ; la seconde permet d'extraire des concepts associés (zone géographique, scope temporel, etc).

Première phase : Segmentation des demandes en LN en blocs d'informations

Nous nous concentrons dans cette phase de segmentation sur un scénario de recherche particulier *i.e.* la recherche de données économiques via un moteur de recherche en français. Basé sur l'analyse des formulaires SAV, on constate que la plupart de ces demandes comportent une structure sous-jacente. Par conséquent, un ensemble de règles ont été écrites manuellement pour décrire les différents scénario de la structure de la demande en LN. Cette dernière s'articule autour de plusieurs éléments, désignés comme autant de *blocs d'informations* :
– [SALUTATION-DEBUT] : formules de salutations (début). Exemple : « Bonjour »
– [FONCTION-CLIENT] : présentation de la fonction. Exemple : « je suis étudiante »
– [CONTEXTE] : annonce du contexte de la recherche. Exemple : « en vue de la création de ma future entreprise »
– [INTENTION-RECHERCHE] : introduction d'une intention de recherche. Exemple : « je souhaiterais obtenir des données »
– [TYPES-DONNEES] : indications des types de données recherchées. Exemple : « les parts de marchés de »
– [REFERENT] : définition du ou des référent(s) : *i.e.* la verbalisation de l'objet même de la recherche, ce sur quoi porte le but de la démarche de RI. Exemple : « revêtement de sol »
– [PRECISIONS] : ajouts de précisions. Exemple : « type Haagen Dazs »
– [SALUTATION-FIN] : formules de salutations (fin) : « avec mes remerciements »
Ces blocs d'informations ne sont pas tous remplis par les utilisateurs ; certains n'utilisant qu'un schéma simple de demande d'informations.

Dans un premier temps, notre méthodologie a consisté à extraire de manière empirique les caractéristiques et les structurations des blocs d'information d'un corpus d'apprentissage constitué d'environ 15 % de notre corpus total soit 200 demandes en LN. Il s'agit d'utiliser les régularités que manifeste notre corpus pour effectuer des découpages et des structurations. Pour cela, nous avons relevé manuellement :
– le **vocabulaire** et les formulations d'une demande pour les différents blocs d'informations en s'appuyant sur des marqueurs sémantiques (« par exemple », « aussi », « concrètement », « en effet »...) ;

– des **marqueurs typographiques** comme la ponctuation (virgule, le point virgule) et les majuscules).
– des **lexiques** spécifiques selon le différents blocs comme des lexiques de verbes (« vouloir », « désirer »…) ou des lexiques de noms (informations, données…).

Par exemple, sont répertoriés dans le même bloc : « je voudrais obtenir des informations sur […] », « je désire toutes les données concernant […] ».

Les règles de segmentation s'effectuent sur les phrases se basent sur le triptyque suivant : *(a)* une syntaxe et un ordonnancement de règles (écrites manuellement) *(b)* un analyseur morpho-syntaxique[2] pour faire appel à des étiquettes (nom commun, nom propre, adjectif, déterminant, etc.) *(c)* du lexique (principalement issues de thésaurus interne à la société possédant le moteur de recherche).

Une première évaluation manuelle de cet échantillon a montré que dans 75 % des cas le découpage en motifs obtenu était correcte (*i.e.* correspondait à un découpage qui aurait été fait humainement) ; 20% des séquences restaient non étiquetées (*i.e.* pas d'appartenance aux différents blocs d'informations identifiés) et 5 % des séquences étaient étiquetées de manière erronée. Sur les 20% de séquences non étiquetées, nous avons pu remarquer que *(i)* des erreurs de segmentation pouvaient être rectifiées en retravaillant sur les règles (12%), *(ii)* des erreurs étaient tout simplement dues à un mauvais ou absence d'étiquetage de l'analyseur morpho-syntaxique (8%). Nous avons travaillé sur ce premier point notamment en les ajustant par profils utilisateurs ce qui a permis de supprimer les 12% de séquences non étiquetées mais non de remédier aux erreurs d'étiquetage de l'analyseur morpho-syntaxique.

Une fois les règles de segmentation ajustées, nous avons ensuite réalisé ce découpage de façon automatique sur l'ensemble des demandes en LN soit 1398 demandes. Ceci nous permet d'avoir un usage de ces blocs différencié : le bloc [REFERENT] nous livre un bon nombre d'informations sur l'objet même de la recherche alors que les blocs [FONCTION-CLIENT] et [CONTEXTE] contiennent des informations sur le profil utilisateurs. Le bloc [INTENTION-RECHERCHE] pour savoir comment les utilisateurs verbalisent leurs besoins. Également les blocs [TYPES-DONNEES], [PRECISIONS] et [REFERENT] nous donnent accès à des concepts associés comme la zone géographique recherchée, le scope temporel demandé, les critères de prix s'ils sont présents, ou d'autres éléments de contexte de la recherche comme le délai/l'urgence ou non de la demande. Enfin les [SALUTATION-DEBUT] et [SALUTATION-FIN] ne sont pas étudiés dans ce présent travail ; notre objectif est donc uniquement de les identifier. Une comparaison plus fine peut alors s'effectuer sur le bloc [REFERENT] qui représente le thème de la recherche et la requête.

Deuxième phase : Extraction de concepts associés

En parallèle de la segmentation de la demande en blocs d'informations, sont extraits plusieurs concepts importants : la zone géographique, les noms de marque, les expressions de temps, les valeurs monétaires ou encore le caractère urgent (le délai). Ces concepts peuvent apparaître de façon disparates dans les différents blocs d'informations. Pour la zone géographique et les noms de marque, nous les repérons grâce à l'étiquette nom propre de l'analyseur morpho-syntaxique. Nous confrontons alors ces noms propres aux thésaurus internes (thésaurus géographique et thésaurus entités nommées). L'extraction des autres concepts se fait à base de règles que nous décrivons ci-dessous. Les concepts sont relevés à partir des règles suivantes :

– la zone géographique : étiquette nom propre et appel à un lexique issu du thésaurus géographique interne à la société. Ce thésaurus regroupe à la fois les noms propres des noms de pays (« France ») ainsi que leurs formes adjectivales (« français » dans « marché français »),
– les noms de marques et/ou d'entreprises : étiquette nom propre et appel à la fois à un lexique comportant le nom et les marques des plus grandes entreprises et également appel à des règles linguistiques pour détecter celles qui ne sont pas renseignées. Ces règles linguistiques se basent sur plusieurs éléments : l'apparition d'une majuscule qui ne soit pas au début d'une phrase, des mentions de forme juridique des entreprises comme SA, SARL, SCI pour les formes françaises et enfin sur la fonction exercée dans l'entreprise comme « gérant de », « propriétaire de », « secrétaire chez »,
– les expressions de temps : combinaison de règles et de lexiques permettant de recueillir toutes les formes de dates *e.g.* simple mention de l'année, ou mois + année, etc ainsi que des notions de temporalité comme le semestre par exemple,
– les valeurs monétaires : combinaison de règles et de lexiques permettant la reconnaissance d'entités numériques suivies ou précédées d'un signe monétaire,

2. L'analyseur morpho-syntaxique utilisé est *xelda* (développé par *xerox*). Les grandes étapes de cet analyseur sont : il *(i)* identifie tout d'abord la langue (à partir des premiers caractères), *(ii)* segmente en phrases, *(iii)* tokénine (*i.e.* scinde un texte en unités lexicales élémentaires), *(iv)* analyse morphologiquement (renvoie les catégories grammaticales potentielles pour tous les mots identifiés durant la tokénisation) et enfin *(v)* désambiguïse morpho-syntaxiquement en déterminant la catégorie grammaticale d'un mot en fonction de son contexte. Cet analyseur a été utilisé pour déterminer tous les traits morphologiques et syntaxiques.

– le caractère urgent : combinaison de règles basées sur du lexique.

Sur ce présent corpus, nous avons évalué manuellement les résultats obtenus sur 200 demandes en LN. Il s'avère que la reconnaissance géographique est assez pertinente : moins de 2 % de bruit (sur le corpus de test de 200, 68 demandes en LN contenaient une géographie). Les données numériques et les format dates sont également bien reconnus : moins d' un 1% de bruit (sur le corpus de test de 200, 15 demandes en LN contenaient une donnée numérique ou une date). La reconnaissance des noms de marque et/ou de noms d'entreprise s'est avérée plus fastidieuse puisque celle-ci s'élève à 7 % de bruit (sur le corpus de test de 200, 20 demandes en LN contenaient un nom de marque et/ou de nom d'entreprise).

4 Quelle expression des besoins informationnels selon la tâche de RI ?

La [TACHE-PRO] formule avec moins de termes ses besoins informationnels : en moyenne 17,66 termes contre 21,51 pour la [TACHE-CREA] et 22,27 pour la [TACHE-SCO]. Cette distribution est représentée dans la Figure 1. Les histogrammes de la figure 1 différencient la distribution de la [TACHE-PRO] (histogramme de gauche) ; les deux premières catégories représentent plus de 150 demandes en LN qui contiennent moins de 20 termes. Peu de demandes sont longues dans ce groupe de personnes.

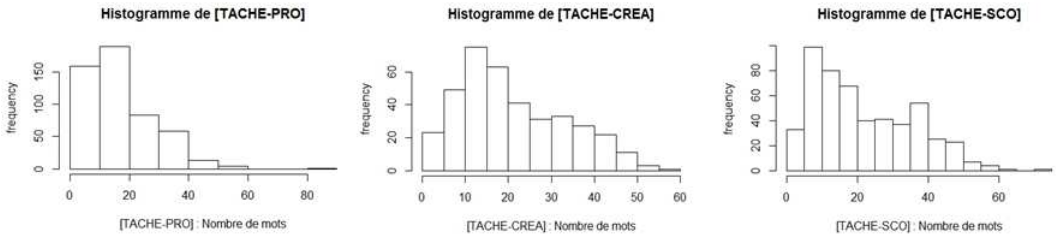

FIGURE 1 – Distribution du nombre de mots dans la demande en LN par type de tâche de RI, représentation en histogrammes

4.1 Traits morphologiques de la demande en LN différenciés par type de tâche en RI

En ce qui concerne les traits morphologiques, nous avons pu relever certaines différences entre la formulation des demandes en LN et la tâche de RI.

En ce qui concerne l'usage et la répartition des concepts à l'intérieur de la demande en LN, plusieurs caractéristiques se dégagent notamment à partir de la figure 2. Les deux concepts qui subissent le plus de variations entre les différents groupes d'utilisateurs sont la [FONCTION-CLIENT] et le [CONTEXTE]. En effet, le groupe [TACHE-SCO] utilise d'avantage le concept [FONCTION-CLIENT] pour indiquer qu'ils sont étudiants et éventuellement précise le degré d'étude (master, license) ou encore le nom de leur université. Le [CONTEXTE] est très peu utilisé par le groupe [TACHE-PRO] par rapport aux deux autres groupes d'utilisateurs : très peu d'informations supplémentaires sont données en plus du secteur recherché qui permettraient de mieux contextualiser leurs demandes avec des exemples ou en définissant plus précisément le cadre de leur recherche. Cette conclusion corrobore notamment le fait que le groupe tend à utiliser moins de termes pour formuler ses demandes en LN.

De façon moins prononcée, le concept [SALUTATION] est moins représenté par la [TACHE-PRO]. Cela suggère une démarche plus rapide et plus directe du formulaire SAV. Le groupe [TACHE-CREA] mentionne moins souvent que les deux autres groupes d'utilisateurs le type d'informations souhaité [TYPE-DONNÉES].

Nous relevons également d'autres écarts de comportements en fonction des types de tâche de RI sur les concepts de prix, de dates, de marques ou de géographies.

Ces concepts sont repris dans la Figure 3 à partir de laquelle nous observons que la géographie a une place plus importante pour le groupe d'utilisateurs ayant une [TACHE-CREA] ; ceci s'explique principalement par le fait que le développement d'une activité professionnelle est très liée à son emplacement géographique (*e.g.* « ouvrir un restaurant à Bordeaux »).

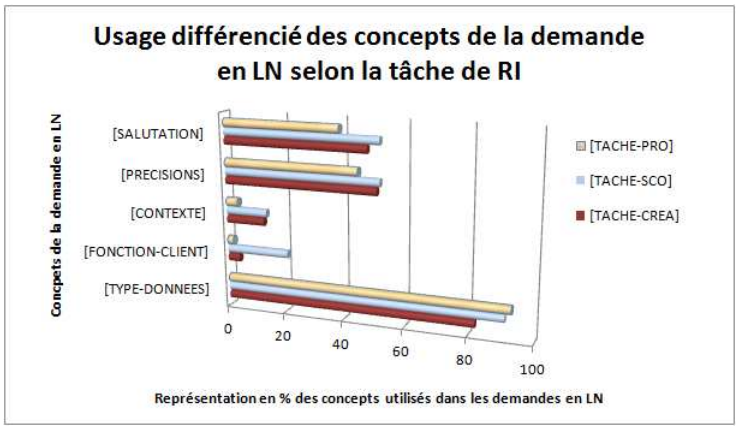

FIGURE 2 – Usage différencié des concepts de la demande en LN selon la tâche de RI

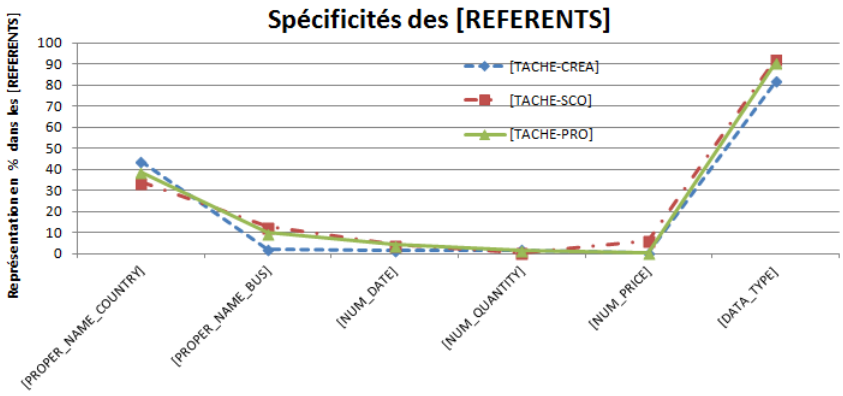

FIGURE 3 – Spécificités des (REFERENTS] dans les demandes en LN par type de tâche de RI

C'est le concept principalement développé dans les demandes en LN de ce groupe d'utilisateur : les concepts de dates, marques et prix apparaissent alors secondaires dans la façon de présenter leurs besoins. Le concept de prix est prédominant pour la [TACHE-SCO] qui demande une réduction voire la gratuité des études. Rappelons que la [TACHE-SCO] regroupe les utilisateurs qui ont une tâche scolaire à effectuer. Ce concept est moins présent pour la [TACHE-CREA] et inexistant pour la [TACHE-PRO]. Le concept de marque est également important pour la [TACHE-SCO] : les demandes comportent alors de nombreuses mentions de noms de marque ou de sociétés qui peuvent être l'objet même de leur travail (« réaliser une étude sur Coca-Cola ») ou sur un marché mais avec une demande particulière sur les principaux acteurs du secteur (*i.e* « les principaux leaders du marché des boissons énergétiques (CA de Isostar) »). Le concept de dates est utilisé de façon similaire par les groupes [TACHE-SCO] et [TACHE-PRO] ; ces utilisateurs ayant des contraintes de temps (délai) ou des demandes plus précises sur le scope temporel que doivent couvrir les études de marché. Les [TYPES-DONNEES] (non représentées dans le schéma) sont de 81,79% [TACHE-CREA], 92,39% pour la [TACHE-SCO] et 90,51% [TACHE-PRO].

4.2 Traits syntaxiques de la demande en LN différenciés par type de tâche en RI

Les principaux traits syntaxiques étudiés sont présentés dans la Figure 4. Notons tout d'abord que le groupe d'utilisateurs [TACHE-PRO] utilise davantage les tournures impersonnelles et moins de pronoms personnels (PP) que les deux autres groupes. Le groupe [TACHE-SCO] emploie plus souvent la première personne du pluriel, se situant dans une démarche de groupe (« nous effectuons un dossier sur »). Pour sa part, le groupe sur la [TACHE-CREA] utilise d'avantage la première personne du singulier, se situant dans une démarche plus personnelle (« je vais créer une entreprise »). Nous notons également que les utilisateurs ayant une [TACHE-SCO] s'expriment plus avec des phrases syntaxiquement correctes, notamment par rapport aux utilisateurs ayant une [TACHE-PRO] qui eux utilisent souvent une syntaxe incorrecte avec des

phrases partielles et/ou incomplètes (exemple : « étudier la croissance de Novartis »). Certaines de leurs demandes en LN peuvent s'apparenter d'ailleurs à des requêtes (exemple : « marché du parquet »).

Il semblerait que le groupe d'utilisateurs [TACHE-PRO] formule ses besoins en LN dans une structure moins formelle que les deux autres groupes ; leurs demandes en LN se rapprochent d'avantage à une formulation hybride entre une expression en LN et une requête.

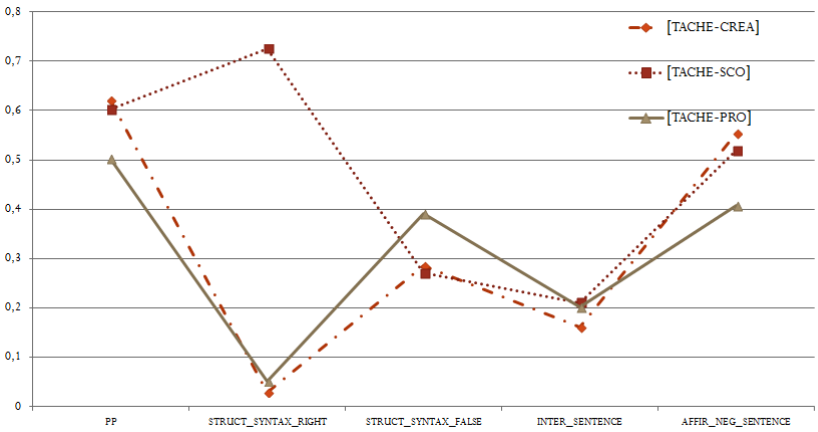

FIGURE 4 – Traits syntaxiques des demandes en LN différenciés par type de tâche de RI

4.3 Traits morpho-syntaxiques du [REFERENT] de la demande en LN selon les types d'utilisateurs

A partir du découpage en blocs d'informations, nous avons effectué un travail plus spécifique sur les référents. En effet, c'est dans ce bloc d'informations (qui peut contenir de 1 à 7 référents) que se retrouvent la plupart des éléments également formulées dans la requête du moteur de recherche. Le [REFERENT] est le concept porteur de l'information principale de la demande en LN ; il contient le thème de la recherche (secteur d'activité recherché). Afin de pouvoir être comparé plus finement avec la requête, ce bloc a fait l'objet d'une analyse qualitative avec une étude morpho-syntaxique de tous ses termes constitutifs.

Le nombre de [REFERENTS] dans les demandes en LN par type de tâche de RI : ce nombre est présenté dans la Figure 5. Cette figure représente la distribution du nombre de [REFERENTS] ; allant de $R = 0$ référent à $R = 7$ référents dans la demande en LN). Nous déduisons de cette figure que le groupe [TACHE-SCO], représenté en pointillé dans le schéma, formule davantage ses besoins informationnels avec un seul référent : 84,80% de leur demandes se verbalisent dans le [REFERENT-1] contre 67,28% pour le groupe [TACHE-CREA] et 65,61% pour le groupe [TACHE-PRO]. Les groupes d'utilisateurs [TACHE-CREA] et [TACHE-PRO] semblent avoir le même comportement par rapport au nombre de [REFERENTS] dans les demandes en LN.

La longueur des n-grammes des [REFERENTS] : cette longueur est présentée dans la Figure 6. La longueur des n-grammes dans les [REFERENTS] est de 1 terme ($n = 1$) dans 49,32% pour l'ensemble des [REFERENTS] du groupe [TACHE-SCO], 47,42% pour le groupe [TACHE-PRO] et de 38,92% pour la [TACHE-CREA]. Ce dernier groupe apparait en pointillé sur le graphique ; la distribution de la longueur des n-grammes est différente en donnant plus de poids aux formulations longues. Le $n = 0$ sur la Figure représente les 8 demandes en LN ne contenant pas de [REFERENT].

L'analyse morpho-syntaxique des [REFERENTS] : l'analyse morpho-syntaxique du bloc [REFERENT] de la demande en LN a été différenciée selon la tâche de RI des utilisateurs. Elle a relevé un usage différencié des catégories morpho-syntaxiques que nous présentons dans les Figures 7 et 8.

Il se dégage de ces graphiques certains traits caractéristiques récurrents par typologie d'utilisateurs :

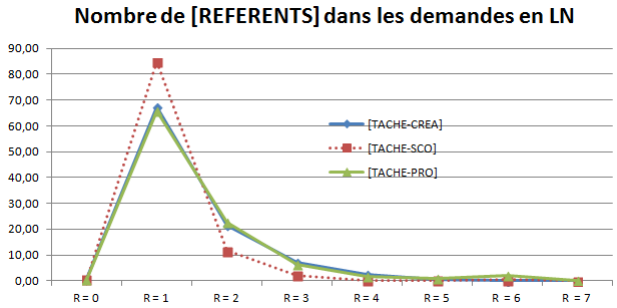

FIGURE 5 – Distribution du nombre de [REFERENTS] dans les demandes en LN selon le type de tâche de RI

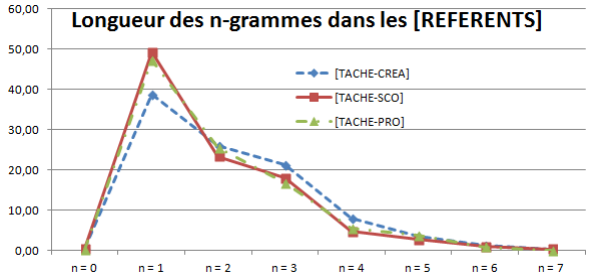

FIGURE 6 – Longueur des n-grammes dans les [REFERENTS] différenciée par type de tâche de RI

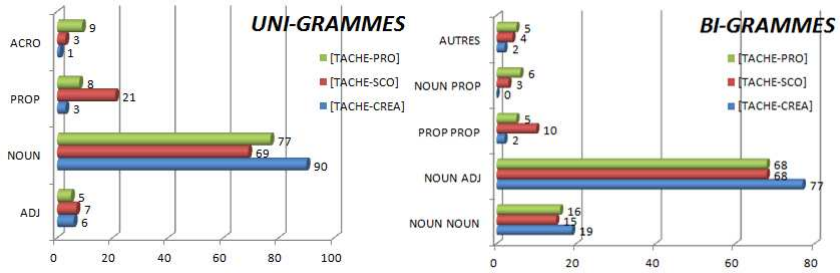

FIGURE 7 – Catégories morpho-syntaxiques pour les uni- et bi-grammes des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI (en pourcentages)

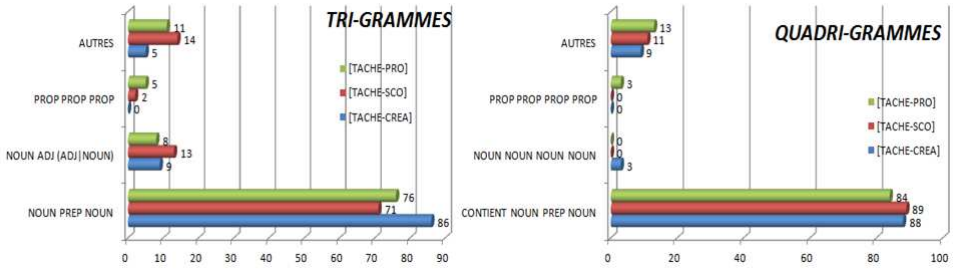

FIGURE 8 – Catégories morpho-syntaxiques pour les tri- et quadri-grammes des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI

– la [TACHE-CREA] utilise des noms et principalement des noms au singulier (NOUN-SG) dans le cas d'uni-grammes ainsi que des noms propres (PROP) et très peu d'autres formes, des NOUN NOUN dans le cas de bi-grammes ainsi que des

NOUN ADJ, des NOUN PREP NOUN pour les tri- et quadri-grammes ainsi que des expressions composées exclusivement de NOUN,
– la [TACHE-SCO] : utilise davantage les noms propres (PROP) à la fois dans les uni- et bi-grammes, puis introduisent fréquemment une forme adjectivale pour formuler ses référents,
– la [TACHE-PRO] a davantage recours à des acronymes, ainsi qu'à des noms pour les uni-grammes, des NOUN PROP pour les bi-grammes, des PROP PROP PROP pour les tri-grammes ou encore PROP PROP PROP PROP pour les quadri-grammes.

4.4 Traits sémantiques de la demande en LN différenciés par type de tâche en RI

Nous avons mesuré plusieurs aspects dans les traits sémantiques de la demande en LN : *(a)* la valeur polysémique [POLYSEMY-VALUE] : nombre de fois que le terme ou les termes du [REFERENT] apparaissant dans les thésaurus, *(b)* la complexité linguistique [LINGUISTIC-COMPLEXITY] : profondeur des nœuds dans le thésaurus du [REFERENT], *(c)* l'évaluation de l'ambigüité de la tâche : calcul en fonction des traits sémantiques évaluées comme favorisant l'ambiguïté de la tâche, *(d)* les secteurs d'activité mentionnés dans les [REFERENTS] : nombre de correspondances avec les thésaurus internes de l'entreprise.

(a) La valeur polysémique [POLYSEMY-VALUE] : cette valeur est calculée en fonction du nombre de fois que les termes apparaissent sous leurs formes lemmatisées dans les thésaurus internes de la société. Plus un terme apparait dans les thésaurus, plus sa valeur polysémique est élevée car non spécifique à un secteur donné particulier. Ainsi, la *polysemy value* à 0 indique que le terme n'est mentionné qu'une seule fois dans les différents thésaurus ; il est donc peu polysémique. Les *polysemy values* 1 à 3 sont présentées dans le Tableau 1. Les termes dont la *polysemy value* vaut $k(k \geq 0)$ apparaissent $k+1$ fois sauf si $k=3$ où le terme apparait au moins $k+1$ fois. La lemmatisation rend possible les correspondances quelle que soit la flexion ; la correspondance se fait sur le lemme et non sur sa forme. Un inconvénient est que cela peut entrainer des rapprochements non satisfaisants, particulièrement dans les cas des uni-termes.

Les *[unknown]* sont les termes qui ne sont pas présents dans les thésaurus. Un *token* inconnu désigne une entité (ou unité) lexicale qui n'a pas été reconnue lors de l'analyse et les comparaisons avec les termes des thésaurus. Cette information peut soit indiquer que le terme a été mal orthographié faussant l'analyse avec la correspondance des termes issus des thésaurus, soit que le terme n'est pas renseigné dans les thésaurus par manque de précision ou de recouvrement d'un secteur d'activité. Notons que sigles et noms des plus grandes marques ou entreprises peuvent être reconnus si ceux-ci correspondent à des entrées dans le thésaurus. La valeur polysémique, représentée dans le Tableau 1, permet d'avoir un aperçu de la polysémie relative à chaque groupe d'utilisateurs selon la tâche de RI.

[POLYSEMY-VALUE]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
0	62,77%	66,60%	78,10%
1	6,82%	7,91%	6,07%
2	2,34%	2,57%	2,11%
3	5,65%	7,11%	3,43%
UNKNOWN	22,42%	15,81%	10,29%

TABLE 1 – Valeur polysémique du [REFERENT] de la demande en LN différenciée par type de tâche de RI

Le groupe dont les termes utilisés dans les [REFERENTS] en LN est le moins porteur de polysémie est celui de la [TACHE-PRO] avec un taux de 0 égal à 78,10%. Il a aussi peu de valeurs fortement polysémiques au niveau 3 (3,43%). La [TACHE-SCO] a une valeur polysémique assez élevée en ce qui concerne le niveau 3 avec 7,11%. Les valeurs UNKNOWN sont importantes pour le groupe de la [TACHE-CREA] puisqu'ils sont à 22,42%.

(b) La complexité linguistique [LINGUISTIC-COMPLEXITY] : la complexité linguistique correspond à la profondeur des nœuds dans le thésaurus du terme ou de l'expression du [REFERENT]. Le [LEVEL 1] correspond aux catégories supérieures généralistes dans la hiérarchie du thésaurus sectoriel : Agro-alimentaire, Biens et Services de Consommation, Industrie lourde, Technologies de l'information et Médias, Sciences de la Vie et Services. Les niveaux suivants de [LEVEL 2] [...] [LEVEL 5] sont des niveaux du thésaurus hiérarchiquement descendants. Les termes sont plus spécifiques. Un même terme peut apparaître dans plusieurs branches de la hiérarchie. La distribution des [REFERENTS] par niveaux et par

type de tâche de RI est présentée dans le Tableau 2 ; l'usage du thésaurus donne un aperçu de l'utilisation de la profondeur des nœuds et donc de la spécificité de la recherche.

[LINGUISTIC-COMPLEXITY]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
1	2,53%	1,98%	1,06%
2	10,53%	16,40%	9,76%
3	40,74%	35,18%	41,69%
4	19,30%	22,92%	28,76%
5	4,48%	7,71%	8,44%
UNKNOWN	22,42%	15,81%	10,29%

TABLE 2 – Complexité linguistique du [REFERENT] de la demande en LN différenciée par type de tâche de RI

D'après le Tableau 2, nous pouvons relever que le groupe d'utilisateurs [TACHE-PRO] se distingue particulièrement aux niveaux 4 et 5 ; ils utilisent des termes qui sont plus spécifiques et plus fins que les deux autres groupes. *A contrario*, le groupe [TACHE-SCO] a une demande plus importante au niveau 2 du thésaurus, représentant des domaines plus généraux. Les nombres UNKOWN sont bien sur identiques à la dernière ligne du Tableau 1 de la [POLYSEMY-VALUE] ; il est difficile de dégager des tendances générales pour le groupe [TACHE-CREA]. Ce chiffre élevé de UNKOWN pour la [TACHE-CREA] peut toutefois révéler que les termes employés notamment les entités nommées de noms de marques, d'entreprises ou géographiques sont trop spécifiques (noms de petites entreprises ou noms géographiques d'une petite commune) pour figurer dans les ressources utilisées pour faire les comparaisons.

(c) L'ambiguïté de la tâche : nous avons défini un modèle reprenant certaines valeurs comme la polysémie ou encore la profondeur des nœuds dans le thésaurus afin de rendre compte de l'ambiguïté de la tâche par groupe d'utilisateurs.

Cette ambiguïté se calcule par la somme de fonctions de chacun des traits sémantiques présentés dans la Figure 3 : dans la moitié supérieure du tableau, les catégories font croître la complexité sémantique de la demande en LN. Dans la moitié inférieure du tableau, les catégories font au contraire décroître cette même complexité : plus elles sont renseignées moins la demande en LN est ambiguë pour les SRI, c'est-à-dire plus sa compréhension est aisée.

$$ambiguité(tâche) = \sum_{i=1}^{n} f_i(tâche), \tag{1}$$

où n est le nombre de traits, f_i est une valeur liée à la tâche et qui pondérera positivement ou négativement un de ses traits. La forme précise de chaque f_i est donnée dans le Tableau 4.

trait$_i$	f_i
[ACRONYM]	nombre d'[ACRONYM]
[UNKNOWN]	nombre de [UNKNOWN]
[POLYSEMY VALUE]	valeur de la [POLYSEMY VALUE]
[FOREIGN]	nombre de termes en langue étrangère
[STRUCT-SYNTAX-FALSE]	nombre de demandes avec une structure syntaxique fausse
[LINGUISTIC-COMPLEXITY]	valeur de [LINGUISTIC-COMPLEXITY]
[CONTEXT-PRECISIONS]	nombre de [CONTEXT-PRECISIONS]
[SYNT-DEPTH]	nombre de n-grammes [SYNT-DEPTH]
[TYPE-DONNEES]	nombre de [TYPE-DONNEES]
[PROPER-NAME-COUNTRY]	valeur de la [PROPER-NAME-COUNTRY]
[NUM-PRICE]	nombre de [NUM-PRICE]
[NUM-DATE]	nombre de [NUM-DATE]

TABLE 3 – Ambiguité de la tâche de RI en fonction des traits sémantiques

Il en ressort que les groupes [TACHE-SCO] et [TACHE-PRO] ont une valeur quasiment équivalente à 5.7 alors que le groupe [TACHE-CREA] a une valeur a 4.6. Il semblerait donc que l'ambiguïté de la tâche est moins importante pour le groupe

[TACHE-CREA]. Ce groupe doit en effet gagner en précision puisque la tâche est souvent assez claire : développer une activité, ouvrir un magasin, etc. La variance est un peu plus importante pour le groupe [TACHE-SCO] ; certains utilisateurs au sein de ce groupe ont une tâche plus ambiguë que d'autres.

[AMBIGUITY-TACHE]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
mean (moyenne)	4.604222	5.758285	5.693676
var (variance)	11.16041	12.41412	11.97133
SD (Ecart-Type)	3.340719	3.523367	3.45996

TABLE 4 – Ambiguité de la demande en LN en fonction de la tâche de RI des groupes utilisateurs

A partir de l'estimation de l'ambiguïté de la tâche, nous avons voulu tester si la longueur ([LENGTH] décrite à la page 5) de la demande en LN était révélatrice de cette ambiguïté. On peut effet émettre l'hypothèse que si une tâche est ambiguë (dans le sens difficile à expliciter) l'utilisateur aura tendance à utiliser plus de termes pour l'exprimer. Nous avons réalisé deux mesures pour tester cette hypothèse. Pour la première, nous avons utilisé le coefficient de corrélation de Pearson, indice statistique qui exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables quantitatives. Ce coefficient de corrélation a déjà été utilisé dans d'autres études équivalentes notamment par [Mothe2005]. Ces mesures ont été obtenues à l'aide du logiciel R. L'importance de la valeur de corrélation est exprimée par la valeur p associée . $P - valeur$ est une estimation de la probabilité que les résultats aussi extrême ou plus extrême se produisent par hasard. Un $p - valeur$ proche de 0 indique une grande confiance dans la corrélation , tandis qu'une $p - valeur$ proche de 1 indique une forte chance pour l'indépendance entre les variables . Nous retiendrons comme significatives les corrélations dont la $p - valeur$ est inférieure à $0,05$. Les résultats sont donnés dans le Tableau 5. Ils indiquent que les [TACHE-SCO] et [TACHE-PRO] ont une $p - valeur$ inférieure à 0,05 et ont donc des résultats significatifs : la longueur de la demande en LN est liée à l'ambiguïté de la tâche. Cette hypothèse est moindre pour la [TACHE-CREA] puisque sa $p - valeur$ est 0,0501.

	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
Corrélation de Pearson	-0,2067269	-0,2121073	-0,1681577
P-value	0,0501	0,01249	0,0001445

TABLE 5 – Ambiguité de la tâche de RI - Coéfficient de corrélation de Pearson

(d) Les secteurs d'activité mentionnées dans les [REFERENTS] : certains secteurs mentionnés dans les [REFERENTS] des demandes en LN sont prédominants selon le type de tâche de RI. Ainsi les utilisateurs du groupe [TACHE-SCO] recherchent davantage d'informations sur l'alimentation et le textile ; le groupe [TACHE-PRO] sur la construction et BTP ainsi que sur la chimie et les produits chimiques tandis que les utilisateurs de la [TACHE-CREA] recherchent davantage sur les secteurs liés aux loisirs et à la restauration. Ces résultats sont présentés dans la Figure 9.

5 Conclusions et Perspectives

Nous avons recueilli et analysé les besoins informationnels exprimés en LN, ceci dans un contexte bien particulier ; celui d'une demande de remboursement effectuée par des utilisateurs d'un moteur de recherche après des recherches a priori infructueuses. Nous nous avons ensuite établi des règles linguistiques et une analyse morpho-syntaxique qui nous a permis de schématiser l'énoncé en LN en fonction de la tâche de RI. Grâce à notre corpus très spécifique, nous avons observé des régularités sur la construction des demandes en LN en fonction de la tâche de RI et nous pouvons conclure qu'en fonction du profil et du type de tâche à réaliser, l'interface de navigation ainsi que les résultats à proposer à l'utilisateur doivent être différenciés par les SRI. Ainsi la [TACHE-SCO] mentionne davantage des critères de prix dans leurs demandes, la [TACHE-CREA] pour leur part indique davantage la zone géographique de leur recherche. La [TACHE-PRO] emploie plus facilement des tournures impersonnelles pour formaliser leurs demandes, etc. Ce sont autant d'indices (morphologiques, syntaxiques, sémantiques) qui peuvent aider à construire le profil utilisateur et à proposer des résultats plus spécifiques en fonction de la tâche de RI et des profils utilisateurs ainsi identifiés. Des améliorations sont envisagées notamment de reconnaissance d'entités nommées dans les demandes en LN ; un outil de reconnaissance et d'extraction d'entités nommées pourrait éventuellement améliorer les résultats ainsi obtenus.

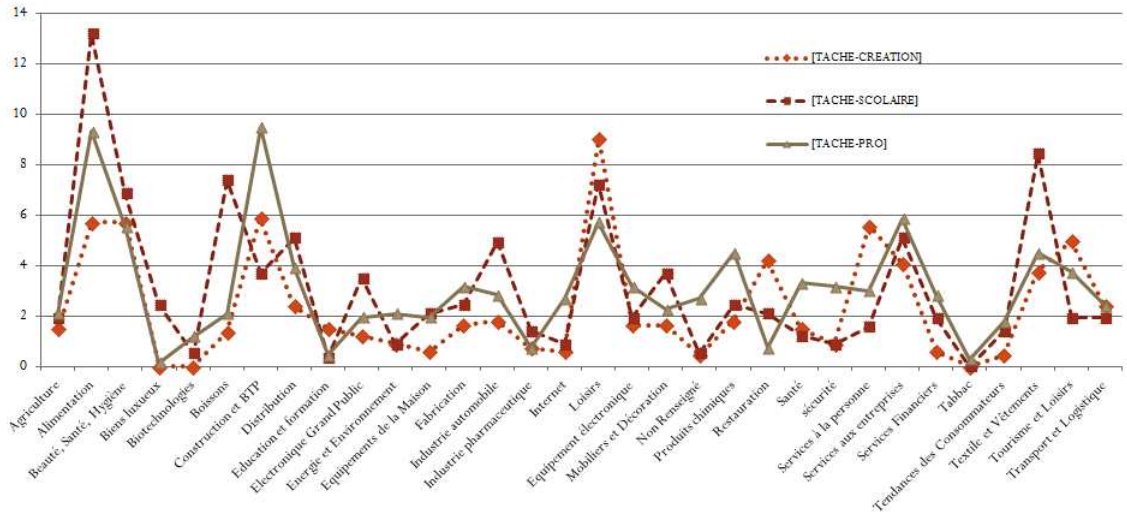

FIGURE 9 – Traits sémantiques des demandes en LN différenciés par type de tâche de RI

Références

[Broder02] BRODER Andrei. *A taxonomy of web search*. In : SIGIR FORUM, 2002, vol.36, n.2, pp.3-10.

[Cabanac11] CABANAC Guillaume, CHEVALIER Max, CIACCIA A. et al. *Recherche d'information et modélisation usagers*. In P. Bellot (Ed.) Recherche d'information contextuelle, assistée et personnalisée. Paris : Hermès, 2011.

[Ingwersen05] INGWERSEN Peter, JARVELIN Kalervo. *The Turn. Integration of information seeking and retrieval in context*. Dordrecht : Springer, 2005, 448 p.

[Jansen08] JANSEN Bernard J.,BOOTH L. Danielle, SPINK Amanda. *Determining the informational, navigational and transactional intent of Web queries*, Information Processing and Management, 2008, vol. 44, pp. 1251-1266.

[Kang05] KANG In-Ho. *Transactional query identification in web search*. In AIRS'05 : Proceedings Information Retrieval Technology, Second Asia a Information Retrieval Symposium, Jeju Island, Korea, 2005, pp. 221-232.

[Kang03] KANG In-Ho, KIM GilChang. *Query Type Classification for Web Document Retrieval*. In : Proceeding SIGIR '03. New York : ACM, 2003, pp.64-71.

[Lecoadic98] LE COADIC Yves-François. *Le besoin d'information : formulation, négociation, diagnostic*. Paris : ADBS Editions, 1998, 191 p.

[Mothe05] MOTHE Josiane, TANGUY Ludovic. *Linguistics features to predict query difficulty - A case study on previous TRES campaign*. In : ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications

[Ramirez06] RAMIREZ Goergina , DE VRIES Arjen P.. *Relevant contextual features in XML retrieval*. In : Proceedings of the 1st international conference on Information Interaction in Context. New York : ACM, 2006, pp. 95-110.

[Rose04] ROSE Daniel, LEVINSON Danny. *Understanding User Goals in Web Search*. In : Proceeding WWW '04 Proceedings of the 13th international conference on World Wide Web. New York : ACM, 2004, pp.13-19.

[Strohmaier08] STROHMAIER Markus, PRETTENHOFER Peter, LUX Mathias. *Different Degrees of Explicitness in Intentional Artifacts : Studying User Goals in a Large Search Query Log*. In CSKGOI'08, 2008, [10 p.]