Utilisation de Numexco pour le repérage de termes-clés dans le domaine des télécommunications sans fil entre ordinateurs

Simon Lemieux

Université du Québec à Montréal

lemieux.simon@courrier.uqam.ca

ABSTRACT. This article deals with the first two series of tests that have been done on the corpus used for the GDST Project (Gestion et Diffusion du Savoir en Télécommunication) with the classification system Numexco created by the Lanci Laboratory (UQAM - Université du Québec à Montréal). The aim of the tests was to identify the key terms in the domain of wireless communication between computers and the potential relations between these terms. Based on these two tasks, another aim is to give computer assistance to the ontology building of the domain, the ontology itself being the final goal of the GDST Project. In addition to the fact that they have permitted the identification of some key terms and relations, these two tests have allowed us to discover the factors that can disrupt Numexco classification, factors that will as a result be handled more easily during forthcoming tests on the same corpus.

KEYWORDS : document management, document classification, key-term identification, wireless communication between computers.

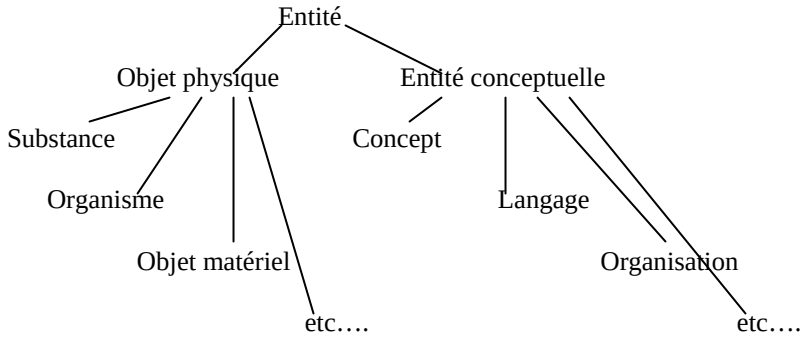
RÉSUMÉ. *Cet article traite des deux premières séries de tests qui ont été effectués sur le corpus utilisé dans le cadre du projet GDST (Gestion et Diffusion du Savoir en Télécommunications) par le biais de la chaîne de traitement Numexco du laboratoire Lanci de l'Université du Québec à Montréal. Le but de ces tests était d'en arriver à repérer les termes-clés propres à ce domaine ainsi que les potentielles relations présentes entre eux, et ce, afin de fournir une assistance informatique à la construction de l'ontologie de ce même domaine, cette ontologie étant en fait la visée finale du projet GDST. En plus d'avoir permis le repérage de bon nombre de termes et de relations, ces deux séries nous ont permis de prendre conscience des facteurs qui perturbent la classification de Numexco, ce qui ne pourra être que bénéfique lors des autres tests qui seront éventuellement faits sur ce corpus.*

MOTS-CLÉS : *gestion documentaire, classification documentaire, identification de termes-clés, ommunications sans fil entre ordinateurs.*

1. Introduction

Le domaine des ontologies a fait naître, au fil des années, plusieurs modèles de conception ou de design possibles, notamment ceux de Dahlgren (1988), de Lenat et Guha (1990), de Miller (1990) et de Sowa (1997), afin de représenter les éléments du

monde ou d'un domaine particulier de même que les relations qui relient ceux-ci[1]. Mais peu importe le modèle choisi, ils ont tous un point en commun, à savoir celui de présenter les éléments et les relations de façon *hiérarchique*, comme dans le schéma suivant:



Évidemment, d'une ontologie à l'autre ou d'un domaine à l'autre, les éléments mis en relation diffèreront; par exemple, une ontologie des mammifères terrestres contiendra des animaux, alors qu'une ontologie de l'automobile contiendra pour sa part toutes les pièces qui composent un tel véhicule. La conception et l'élaboration d'une ontologie peut se faire soit "manuellement", c'est-à-dire en recensant au fur à mesure tous les éléments qui feront partie de l'ontologie finale, soit par le biais de lectures, d'entrevues, etc., soit par le biais d'une assistance informatique qui permet de fouiller les textes retenus beaucoup plus rapidement et efficacement que par la simple intervention humaine, soit en combinant ces deux facettes. Dans le cadre du projet GDST (Gestion et Diffusion du Savoir en Télécommunications), mis en branle depuis mai 2001 et dans lequel sont impliqués l'UQAM, les HEC et le LUB, la combinaison humain-ordinateur a été privilégiée afin de construire un prototype d'ontologie sur les télécommunications sans fil entre ordinateurs, plus spécialement celles basées sur le protocole 802.11b. Globalement, les principaux utilisateurs de cette ontologie seront les CSE (Communications Systems Engineer) de Bell Canada; celle-ci leur donnera un soutien documentaire par rapport à deux de leurs tâches, à savoir les facettes "présentation-démonstration" et "conditions d'installation" de leur travail, la première consistant à présenter aux potentiels clients les avantages et les inconvénients de cette nouvelle technologie, le mode de fonctionnement de celui-ci, etc., et la deuxième à évaluer les emplacements, le matériel requis et les obstacles physiques à contourner en fonction des immeubles où ils seront éventuellement installés. Chaque CSE ayant bien sûr son propre bagage de connaissances de même que des tâches distinctes à réaliser, le système découlant de l'ontologie pourra être consulté en fonction et de leurs propres besoins et du travail qu'ils auront à accomplir; autrement dit, les CSE pourront repérer facilement la section de l'ontologie qui les intéresse (concepts visés + documents pertinents).

Pour construire ce prototype d'ontologie, quelques 141 documents traitant de ce domaine ont été recensés sur Internet afin d'identifier les concepts et les mots-clés qui feront partie de ce dernier. Extraire à la main des concepts et des termes-clés d'une base de documents plus ou moins volumineuse peut s'avérer une tâche à la fois ardue et de longue haleine; il est donc nécessaire voire essentiel de tenter d'automatiser le plus possible ce travail de

[1] v. FRIDMAN NOY et HAFNER (1997).

repérage et de classification des documents. Aussi, ce volumineux corpus fut soumis au sous-système de traitement Numexco de la plate-forme Satim du laboratoire Lanci (UQAM) dans le but, justement, de traiter plus rapidement et efficacement ce corpus et de vérifier dans quelle mesure un tel système permet de repérer les termes-clés de ce domaine tout en établissant des relations ou des liens entre ces mêmes termes, relations qui permettent quant à elles de classifier les documents ou les multiples sections de ceux-ci en fonction des associations de mots qu'ils contiennent.

1.1 Le corpus et son traitement préalable

Des 141 documents de format HTML ou PDF trouvés sur Internet pour le repérage de termes-clés portant sur les télécommunications sans fil entre ordinateurs, seuls 111 ont pu être convertis en format *txt*, ce qui fait que 30 d'entre eux, dû essentiellement au fait que leur affichage sur le Web est fortement sécurisé, n'ont pu être traités et convertis pour pouvoir être soumis par la suite à la plate-forme Satim (Numexco). Une fois la conversion des 111 documents effectuée, tous les caractères qui n'étaient pas des lettres ou des chiffres ont été supprimés des textes, pour ne pas perturber ou empêcher le traitement de ces textes par Numexco, étant donné que cette chaîne de traitement est sensible aux caractères qui ne sont pas alphanumériques. Ainsi, les 34 caractères suivants ont été éliminés des textes: @, ©, $, ¢, (,), ", ', «, », /, \, &, *, #, %, [,], _, +, =, <, >, ^, {, }, ¯, ¤, ¬, ¦, |, ~, ° et ±.

2. Tests et résultats

2.1 Première série de tests

Afin de soumettre les documents à la plate-forme Satim et de rendre le traitement de ceux-ci le plus efficace et clair possible, les documents, selon leur longueur, ont été divisés en sections ou parties variant entre 1 et 12 pages, chacune de ces sections correspondant à un paragraphe global. Par exemple, si un document de 30 pages contenait au départ cinq chapitres, les paragraphes appartenant à ces chapitres ont été compressés pour ne former qu'une seule et même partie, ce qui fait qu'un tel document était divisé en cinq sections dont le nombre de pages de celles-ci correspondait au nombre de pages des chapitres de départ. Au résultat, nous obtenons pour les 111 documents de la base de données un grand total de 257 segments différents. De plus, pour être capable de les repérer facilement en tout temps, chacune de ces parties fut identifiée par le biais d'un indice placé au tout début de celle-ci; par exemple, la section 3 du document 56 et la section 4 du document 107 furent respectivement marquées par les indices 56-3T et 107-4T, la présence de la lettre "T" étant pour éviter toute confusion avec des données potentielles dans les textes dont la valeur serait identique à celle de l'un des indices.

La plate-forme Satim ayant ses propres limites (mémoire, capacités de traitement, etc.), il fut impossible de traiter tous les documents en un bloc unique. Il a donc fallu divisé en

deux groupes le corpus global pour rendre le traitement possible; l'un, appelé *Fusion1-3-4*, comptant 565 pages et 174 sections et l'autre, *Fusion2*, comptant quant à lui 388 pages et 83 sections. Évidemment, puisqu'à chaque section correspond un paragraphe, le choix de segmentation lors du traitement des groupes dans Satim fut "paragraphe-occurrence = 1", pour faire en sorte que la segmentation de la plate-forme corresponde exactement à celle faite préalablement.

Le nettoyage et la lemmatisation des documents étant inutiles sur les textes anglais puisque les dictionnaires de Satim sont français, il a fallu éliminer les termes trop fréquents ou trop rares des documents par l'entremise de l'option "suppression par intervalles" de Numexco; les mots se situant dans les intervalles 1-8 et 1200-20000 furent donc éliminés. De plus, la suppression de mots non pertinents fut faite manuellement par la suite, car, bien sûr, bon nombre de mots se retrouvant dans l'intervalle 9-1199 n'étaient aucunement reliés ou pertinents au domaine visé. Une fois ces suppressions automatique et manuelle effectuées, le groupe *Fusion1-3-4*, qui contenait au départ 12 854 mots, n'en contenait plus que 705, alors que ne restaient que 844 des 11 541 mots de départ dans le groupe *Fusion2*. Finalement, l'étape de la classification donna au résultat un total de 38 classes (pour 83 segments) dans le cas du groupe *Fusion2* et de 116 classes (pour 174 segments) dans celui du groupe *Fusion1-3-4*, et ce, en utilisant la valeur 0.2 comme coefficient de vigilance dans ce même système de classification[2].

Toutefois, cette première série de tests nous a fait prendre conscience de deux problèmes: premièrement, qu'une segmentation de documents irrégulière basée sur la longueur des parties ou des chapitres des documents fait en sorte que la matrice de la chaîne de traitement doit constamment ajuster et compenser la longueur de tous les segments afin de les rendre uniformes, et deuxièmement, qu'il fallait enlever du corpus tous les documents dits "métas", les glossaires (au nombre de deux) dans notre cas, puisque dans ce type bien précis de documents, étant donné qu'on y retrouve une liste de définitions de termes qui ne sont pas nécessairement reliées entre elles, la proximité de deux ou plusieurs définitions de termes ne peut parfois être constatée que dans ce type de documents. Par exemple, les unités complexes *directed search* et *directory enabled networking*, qui font référence à deux sous-domaines différents du monde des télécommunications sans fil entre ordinateurs, ne peuvent à toute fin pratique apparaître l'un près de l'autre que dans un glossaire ou une liste de définitions présentées par ordre alphabétique:

- ***Directed search***: «Search request sent to a specific node known to contain a resource. A directed search is used to determine the continued existence of the resource and to obtain routing information specific to the node. See also broadcast search.» [3]

- ***Directory enabled networking***: «An LDAP-based information model for networked devices.»[4]

---

[2] Cette valeur étant celle qui avait offert les meilleurs résultats lors des tests qui avait été fait auparavant sur d'autres corpus avec cette chaîne de traitement, elle fut donc retenue pour les deux séries qui nous concernent.

[3] v. Cisco Systems (2001), lettre "D".

Aussi, ce type de proximité dans les glossaires risquant fort de biaiser les résultats ou d'avoir un impact sur le traitement de Numexco, il fut donc préférable de les retirer pour la deuxième série de tests.

2.2 Deuxième série de tests

2.2.1 Les multiples tentatives de classification avec Numexco

Pour les besoins de cette deuxième série de tests, les 997 pages du corpus ont été divisées en 11 grandes parties dont voici les données générales:

Tableau 1 – Caractéristiques des groupes de textes

	Nbre de textes	**Nbre de pages**	**Nbre de paragraphes**	**Nbre de mots**
Groupe 1	17	88	89	57 215
Groupe 2	11	92	89	63 279
Groupe 3	2	80	76	53 541
Groupe 4	22	93	93	59 902
Groupe 5	18	92	95	58 710
Groupe 6	7	94	89	58 006
Groupe 7	1	92	86	61 830
Groupe 8	1	82	76	45 969
Groupe 9	11	91	87	63 242
Groupe 10	20	90	93	61 071
Groupe 11	7	103	100	66 308

Comme nous pouvons le constater par rapport au nombre de paragraphes, les documents ont été cette fois-ci divisé en segments de plus ou moins une page, afin justement d'éviter que la matrice n'ait à faire d'énormes ajustements pour équilibrer ceux-ci, contrairement à ce qui avait été fait pour la première série. Mais malgré cet effort d'uniformisation, les résultats découlant de la segmentation "paragraphe-occurrence = 1" sont moins probants que ceux obtenus par le biais, par exemple, de la segmentation "mots-occurrences = 250". En plus de ces deux tentatives de segmentation, deux autres ont été faites sur chacun des groupes, soit "mots-occurrences = 500" et "mots-occurrences = 750". Les résultats obtenus en termes de classes (C) et de segments (S), toujours avec comme coefficient de vigilance 0.2, sont les suivants (le différentiel classes-segments étant affiché entre parenthèses):

[4] v. ibidem.

Tableau 2 – Résultats de multiples segmentations sur les 11 groupes de départ

	Paragraphe = 1	**Mots = 250**	**Mots = 500**	**Mots = 750**
Groupe 1	82C-90S (8)	168C-238S (70)	104C-119S (15)	70C-80S (10)
Groupe 2	85C-90S (5)	172C-239S (67)	109C-121S (12)	75C-80S (5)
Groupe 3	73C-76S (3)	162C-221S (59)	99C-111S (12)	72C-74S (2)
Groupe 4	92C-93S (1)	169C-249S (80)	110C-124S (14)	80C-83S (3)
Groupe 5	95C-95S (0)	170C-245S (75)	111C-123S (12)	81C-82S (1)
Groupe 6	84C-89S (5)	187C-244S (57)	116C-122S (6)	76C-82S (6)
Groupe 7	84C-89S (5)	183C-272S (89)	117C-136S (19)	85C-91S (6)
Groupe 8	58C-76S (18)	110C-193S (83)	72C-97C (25)	48C-65S (17)
Groupe 9	78C-87S (9)	187C-281S (94)	118C-141S (23)	81C-94S (13)
Groupe 10	92C-93S (1)	180C-253S (73)	114C-127S (13)	84C-85S (1)
Groupe 11	99C-100S (1)	184C-277S (93)	122C-139S (17)	92C-94S (2)
Moyenne du différentiel	5.09	76.36	15.27	6.0

Afin de vérifier les potentiels liens qui pourraient reliés des textes ou des parties de textes faisant partie de groupes différents, les 11 groupes précédents ont été combinés de multiples façons pour former des groupes plus volumineux, groupes sur lesquels les mêmes tests de segmentation que sur les groupes d'origine ont été faits. Le texte numéro 55 étant le plus long du corpus (242 pages), il fut de mise de l'analyser individuellement également, et ce, dans le but de vérifier quelles sections de ce document pouvaient être reliées entre elles:

Tableau 3 – Multiples segmentations sur les groupes fusionnés et le texte 55

	Paragraphe = 1	**Mots = 250**	**Mots = 500**	**Mots = 750**
Groupe 1-2-3-4-5	423C-442S (19)	Échec	505C-595S (90)	371C-397S (26)
Groupe 6-7-8-9-10-11	515C-531S (16)	Échec	664C-759S (95)	466C-506S (40)
Groupe 1-4-9-11	352C-369S (17)	Échec	448C-522S (74)	312C-348S (36)
Groupe 2-5-8-10	338C-353S (15)	634C-929S (295)	401C-465S (64)	289C-310S (21)
Groupe 3-6-7	241C-252S (11)	545C-736S (191)	329C-368S (12)	235C-246S (11)
Groupe 1-5-6-10	346C-364S (5)	680C-978S (298)	420C-489S (69)	303C-326S (23)
Groupe 2-4-9-11	354C-367S (13)	Échec	459C-522S (93)	331C-348S (17)
Groupe 3-7-8	224C-238S (14)	478C-686S (208)	309C-343C (34)	215C-229S (14)
Texte 55	188C-226S (38)	318C-641S (323)	226C-321S (95)	191C-214S (23)
Moyenne du différentiel	16.44	263.0	56.90	23.44

Comme le montre le tableau 3, le plus grand différentiel classes-segments parmi tous les tests effectués se situe au niveau de la segmentation "mots-occurrences = 250". Ce constat nous a donc poussé à utiliser les résultats de cette segmentation pour les fins de repérage des termes-clés et des relations entre ceux-ci, puisque plus le nombre de classes sera réduit par rapport au nombre de segments, plus il sera possible de retrouver un nombre plus ou moins grand de textes sous une classe donnée, ce qui est essentiel si l'on veut en arriver à relier des textes ou des parties de textes dont le lexique présente des similitudes. Dans le tableau 4 qui suit, les données concernant les classes (en termes de nombre de segments) des quatre grands groupes qui purent être traités avec la segmentation par 250 mots de même que celles du texte 55 sont présentées:

Tableau 4 – Nombre de segments par classes des groupes fusionnés et du texte 55

	Classes de 1 segment	**Classes de 2 segments**	**Classes de 3 à 5 segments**	**Classes de 5 segments et plus**	**Total**
Groupe 2-5-8-10	356 (56.1%)	256 (40.3%)	20 (3.2%)	2 (0.4%)	634
Groupe 3-6-7	370 (67.9%)	169 (31%)	5 (0.9%)	1 (0.2%)	545
Groupe 1-5-6-10	393 (57.8%)	280 (41.2%)	7 (1%)	0	680
Groupe 3-7-8	323 (67.6%)	131 (27.4%)	19 (4%)	5 (1%)	478
Texte 55	183 (57.7%)	90 (28.3%)	32 (10%)	13 (4%)	318
Moyenne	61.42%	33.64%	3.82%	1.12%	-------

Ce tableau rend compte de deux phénomènes importants; d'une part, que le texte 55 contient beaucoup plus de classes comptant plus de 5 segments que tous les quatre groupes de combinaisons de textes, ce qui est tout à fait normal puisqu'il s'agit d'un seul et même texte et que certaines parties de celui-ci sont susceptibles de traiter du même sujet, et que, d'autre part, même si la segmentation par 250 mots présentait les plus grandes différences entre le nombre de classes et de segments, plus de la moitié sont unaires (un segment) et le tiers des classes comptant plus d'un segment ne sont que binaires (deux segments).

Cette forte présence de classes unaires et binaires s'explique probablement par le fait que soit le corpus de textes est fortement hétérogène, ce qui fait en sorte que le système est incapable de regrouper plusieurs segments ensemble, soit le corpus est au contraire très homogène, de telle sorte que le système, devant tant d'homogénéité, ne sait trop comment en arriver à classer tous ces documents de façon conséquente, d'où la rareté des classes comptant plus de 3 segments dans les quatre grands groupes de textes[5]. Puisque tous ces textes portent sur un domaine bien particulier, la seconde hypothèse semble être la meilleure, comme nous le verrons dans la section suivante portant sur le repérage de termes-clés.

2.2.2 Premiers repérages de termes-clés et de relations inter-termes

Les premiers essais de repérage de termes-clés et de relations portent, entre autres, sur un mot très commun dans les textes, *channel*, et sur un acronyme propre au domaine étudié, *PLCP* (*Physical Layer Convergence Procedure*), dont voici les définitions respectives:

PLCP
«Physical layer convergence procedure. Specification that maps ATM cells into physical media, such as T3 or E3, and defines certain management information.»[6]

Channel
«1.Communication path wide enough to permit a single RF transmission. Multiple channels can be multiplexed over a single cable in certain environments.
2. The specific path between large computers. In IBM (such as mainframes) and attached peripheral devices.
3. Specific frequency allocation and bandwidth. Downstream channels used for television in the United States are 6 MHz wide.»[7]

Pour rendre compte de la grande homogénéité des documents, voici les données sur les termes qu'ont en commun trois segments du groupe 1-5-6-10 où l'on retrouve l'acronyme

[5] Évidemment, tel que déjà mentionné, l'isolement du texte 55 prouve le contraire, étant donné que 14% des classes de celui-ci comptent plus de 3 segments, ce qui est fort appréciable comparativement aux quatre groupes susmentionnés. Il faut également ajouter que, même si le coefficient de vigilance de 0.2 est le plus performant lors d'un traitement avec Numexco, il est possible qu'avec ce corpus, il soit ultérieurement nécessaire d'en utiliser un ou plusieurs afin de comparer les résultats qui découleraient de tels changements.
[6] v. idem, lettre "P".
[7] v. idem, lettre "C".

PLCP, à savoir le segment 566, formant à lui seul la classe 586, et les segments 568 et 569, formant quant à eux la classe 278:

Tableau 5 – Mots en commun entre trois segments du groupe 1-5-6-10 contenant l'acronyme PLCP

Mots pertinents en commun	**Segment 566**	**Segment 568**	**Segment 569**
Segment 566	----------	20	11
Segment 568	20	----------	18

Dans ce tableau, nous pouvons remarquer que le segment 566 a plus de mots pertinents en commun avec le segment 568 que n'en a le segment 569, ce dernier appartenant pourtant à la même classe que le segment 568 (20 termes contre 18). Conséquemment, au lieu d'être séparés en deux classes distinctes, ces trois segments auraient pu, voire auraient dû, être mis tout simplement dans une seule et même classe, ce qui ne fut pas le cas évidemment. D'ailleurs, le contenu de ces trois segments est très semblable, étant donné qu'ils correspondent tous les trois à des parties du sommaire du texte 55, le nombre de mots qu'ils ont en commun en étant la preuve.

L'impact de cette grande homogénéité entre les textes se reflètent aussi dans l'examen du terme *channel*, notamment dans les 13 segments du texte 7 où ce terme apparaît, comme le font foi ces données (le nombre de mots communs aux segments d'une même classe étant affiché en caractères gras de plus grande dimension):

Tableau 6 – Mots en commun entre les segments du texte 7 contenant le terme channel

Segment	**30**	**32**	**33**	**35**	**36**	**37**	**38**	**41**	**42**	**43**	**44**	**45**
30	----	5	8	9	6	2	3	9	7	7	3	3
32	5	----	8	8	8	8	11	11	7	7	7	7
33	8	8	----	14	12	7	11	13	9	8	8	7
35	9	8	14	----	14	7	10	16	12	10	8	10
36	6	8	12	14	----	8	12	13	9	9	10	5
37	2	8	7	7	8	----	7	6	6	3	6	4
38	3	11	11	10	12	7	----	11	5	6	8	4
41	9	11	13	16	13	6	11	----	**10**	9	8	6
42	7	7	9	12	9	6	5	**10**	----	10	10	6
43	7	7	8	10	9	3	6	9	10	----	9	7
44	3	8	8	8	10	6	8	8	10	9	----	**6**
45	3	7	7	10	5	4	4	6	6	7	**6**	----
Moy.	*5.6*	*8.0*	*9.5*	*10.7*	*9.6*	*5.8*	*8.0*	*10.2*	*8.3*	*7.7*	*7.5*	*5.9*

Si, dans certains cas, le nombre de mots pertinents que deux segments ont en commun est faible (segments 30-37, 30-45, 37-43), il arrive fort souvent que ce nombre soit égal sinon supérieur à celui que les segments 41-42, d'une part, et 44-45, d'autre part, ont en commun, et ce, en dépit du fait que chacune de ces deux paires de segments font partie de deux classes binaires. Donc, dû au fait que le nombre de termes communs aux multiples segments soit passablement homogène, il aurait été normal de retrouver davantage de segments de ce texte regroupés sous une même classe. La courte longueur de ce texte (6

pages, 4 120 mots) de même que le thème unique abordé dans celui-ci (comparaison entre le saut de fréquence et la modulation directe de spectre de diffusion[8]) viennent renforcer cette hypothèse, puisque si dans un texte les mêmes termes sont employés pour discuter d'un seul thème et que ce texte est, par surcroît, court, il n'y a aucune raison pour que les segments de celui-ci soient isolés dans des classes distinctes, comme ce fut le cas pour le texte 7 (12 segments pour 10 classes).

Le texte 55[9], dont les données de base figurent dans le tableau 4, montre toutefois qu'il est possible avec Numexco de regrouper sous une même classe les segments d'un texte portant sur un thème bien précis, puisque plusieurs classes de ce texte comptent un nombre plus ou moins considérable de segments traitant d'un même sujet, comme c'est le cas des classes 30, 172, 195 et 267 suivantes:

Classe	Nombre de segments	Thème abordé
30	46	Spécifications Contrôle d'accès
172	12	Spécifications MAC (MSDU)
195	22	Réglementations ANSIIEEE sur MLME et MPDU
267	12	Spécifications ATIM, PLCP et GFSK

De telles données montrent donc qu'il est sans aucun doute possible d'en arriver à un classement à la fois détaillé et cohérent d'un ensemble segments par le biais la chaîne de traitement Numexco. Ne reste qu'à identifier et corriger dans la mesure du possible les facteurs qui peuvent venir perturber les résultats de celles-ci dans certains cas, soit la valeur du coefficient de vigilance du classifieur, le contenu des documents, le style de ceux-ci de même que leur pertinence informationnelle, car il peut arriver qu'un document, en dépit du fait que plusieurs termes-clés peuvent être présents dans ce dernier, ne puisse apporter quoi que ce soit informationnellement si ces dits termes n'apparaissent qu'avec des mots qui ne sont pas vraiment reliés au domaine ciblé, ce qui est le cas dans des textes d'opinion ou des réflexions personnelles par exemple[10]. Dans les prochains tests qui seront effectués sur ce corpus, il sera nécessaire voire essentiel de vérifier l'impact de tous ces paramètres dans le processus de classification de Numexco.

3. Conclusion

Le premier examen du comportement de la chaîne de traitement Numexco sur le corpus du projet GDST s'est fait par le biais de deux séries de tests; la première, sommaire, nous a montré l'impact d'une segmentation irrégulière sur les résultats finaux, les capacités de traitement de Numexco, de même que l'influence que certains types de documents, notamment ceux dits "métas", peuvent eux aussi avoir sur la classification finale. La seconde, plus détaillée et plus orientée sur le repérage en soi de termes-clés et de relations inter-termes, nous a fait premièrement prendre conscience qu'une différence quant à la longueur des segments, aussi minime soit-elle, perturbe les résultats finaux et qu'il est

[8] v. ANDREN (1997).
[9] v. IEEE (1999).
[10] v. FLICKENGER (2001), TRUDEAU (2001) et ARBAUGH et al. (2001).

donc préférable d'y aller d'une segmentation extrêmement rigide (250, 500, 750 mots), et, deuxièmement, que l'homogénéité en ce qui a trait au contenu des documents influence tantôt positivement, tantôt négativement les résultats de la classification, comme l'ont montré les résultats portant sur *PLCP* et *channel*, puisque dans le premier cas, malgré le fait que les segments où apparaissaient *PLCP* avaient beaucoup de termes en commun, ils étaient tout de même rangés sous des classes distinctes, alors que le phénomène inverse s'est produit dans le cas des segments contenant *channel*.

Pour les tests qui seront ultérieurement faits sur ce corpus, d'autres facteurs devront être pris en compte afin de vérifier leur impact sur la classification finale, soit la valeur du coefficient de vigilance, le contenu des documents (textes très techniques, textes d'opinion, etc.) et la pertinence informationnelle caractérisant chacun de ces textes, car plus il y aura de facteurs qui seront contrôlés lors d'un tel processus de classification, plus les chances d'obtenir des résultats représentatifs du contenu des documents et des potentielles relations qui les caractérisent seront grandes.

Références

[ANDREN 97] ANDREN, Carl, , *A Comparison of Frequency Hopping and Direct Sequence Spread Spectrum Modulation for IEEE 802.11 Applications at 2.4 GHz*, 1997
URL: http://www.intersil.com/design/prism/papers/ds-v-fh.pdf
[ARBAUGH 01] ARBAUGH, William A., Narendar Shankar et Y.C. Justin Wan, 2001, *Your 802.11 Wireless Network has No Clothes*,
URL: http://www.cs.umd.edu/~waa/wireless.pdf
Cisco Systems, 2001, *Internetworking Terms and Acronyms*,
URL: http://www.cisco.com/univercd/cc/td/doc/cisintwk/ita/index.htm
[FLICKENGER 01] FLICKENGER, Rob, 2001, *A Wireless Long Shot*,
URL: http://www.oreillynet.com/lpt/a//wireless/2001/05/03/longshot.html
[FRIDMAN NOY 97] FRIDMAN NOY, Natalya et Carole D. HAFNER, 1997, "The State of the Art in ontology Design", in *AI Magazine (automne 1997)*, American Association for Artificial Intelligence, pp. 53-74.
[GEIER 01] GEIER, Jim, *Wireless LANs, Second Edition*, Indianapolis, Sams, 345 pages, 2001
IEEE (Institute of Electrical and Electronics Engineers), 1999, *ANSI/IEEE Std 802.11 Information technology, Telecommunications and information exchange between systems — Local and metropolitan area networks — Specific requirements*,
URL: [TRUDEAU 01] TRUDEAU, Pierre, *Building Secure Wireless Local Area Networks*, 2001
URL: http://www.colubris.com/en/support/whitepapers/whitepapers/WP-010712-EN-01-00.pdf