# Les n-grams de caractères pour l'aide à l'extraction de connaissances dans des bases de données textuelles multilingues

Ismaïl Biskri & Sylvain Delisle

Université du Québec à Trois Rivières
Département de mathématiques et d'informatique
C.P. 500, Trois-Rivières, Québec, Canada, G9A 5H7
{ismail_biskri, sylvain_delisle}@uqtr.uquebec.ca

## Résumé – Abstract

Une véritable classification numérique multilingue est impossible si on considère seulement le mot comme unité d'information privilégiée. En traitant les mots comme jetons, la *tokenisation* s'avère relativement simple pour le français et l'anglais, mais très difficile pour des langues comme l'allemand ou l'arabe. D'autre part, la lemmatisation utilisée comme moyen de normalisation et de réduction du lexique constitue un écueil non moins négligeable. La notion de n-grams, qui depuis une décennie donne de bons résultats dans l'identification de la langue ou dans l'analyse de l'oral, est, par les recherches récentes, devenue un axe privilégié dans l'acquisition et l'extraction des connaissances dans les textes. Dans cet article, nous présenterons un outil de classification numérique basé sur le concept de n-grams de caractères. Nous évaluons aussi les résultats de cet outil que nous comparons à des résultats obtenus au moyen d'une classification fondée sur des mots.

Real multilingual numerical classification is impossible if only words are treated as the privileged unit of information. Although it makes tokenisation (in which words are considered as tokens) relatively easy in English or French, it makes it much more difficult for other languages such as German or Arabic. Moreover, lemmatisation, typically used to normalise and reduce the size of the lexicon, poses another challenge. The notion of n-grams which, for the last ten years, seems to have produced good results both in language identification and speech analysis, has recently become a privileged research axis in several areas of knowledge acquisition and extraction from text. In this paper, we present a text classification tool based on n-grams of characters and evaluate its results and compare them with those obtained from a different classification tool based solely on the processing of words.

**Mots clés :** classification numérique de textes, n-grams, multilinguisme.

# 1. Introduction

Normalement la première étape dans un processus de traitement d'un gros corpus au moyen d'un outil statistique est de subdiviser le texte à traiter en plusieurs unités d'information appelées *tokens* qui sont, traditionnellement, des mots simples. Ce processus de *tokenisation* pose une question primordiale : sur le plan informatique, comment repérer un mot ? En d'autres termes, quels sont les indicateurs formels de surface, non ambigus, qui peuvent délimiter un mot ? Si pour le français ou l'anglais littéraire, ou des langues apparentées, la réponse est presque triviale — à savoir que toute chaîne de caractères précédée et suivie d'un espace est considérée comme un mot simple — il en va tout autrement pour d'autres langues. Dans le cas de termes composés en langue allemande comme, par exemple, *lebensversicherungsgesellschaftsangestellter* ("employé d'une compagnie d'assurance vie"), ou pour la langue arabe dans laquelle les pronoms sujets et compléments sont dans certains cas attachés aux verbes et une seule chaîne de caractères représente ainsi une phrase comme, par exemple, kathabthouhou ("je l'ai écrit"), cette notion de *tokens* devient inadéquate (Manning & Schütze, 1999).

Si le mot simple ne convient pas à toutes les langues, quelle est donc l'unité d'information atomique la plus adéquate pour segmenter un texte ? Balpe *et al*. (1996) soulignent que dépendant de l'objectif de lecture et de compréhension que nous nous donnons, la définition de l'unité d'information dépend de l'usage qui en est attendu. Dans une perspective de classification numérique à des fins d'extraction de connaissances, la définition d'une unité d'information est tributaire des contraintes suivantes :

- L'unité d'information doit être une portion du texte soumis à l'analyse numérique.
- Il doit être facile sur le plan informatique de repérer les unités d'information.
- La définition d'une unité d'information doit être indépendante de la langue dans laquelle le texte est écrit. Une telle définition permet à l'analyse numérique, moyennant des modifications minimes, de couvrir un large éventail de langues.
- Les unités d'information doivent être statistiquement comparables. Il doit être aisé d'en calculer les fréquences d'apparition dans les différentes parties du texte et par conséquent d'estimer leur distribution et la régularité à laquelle plusieurs unités cooccurrent dans les mêmes parties du texte.

Que l'unité linguistique dans une analyse de classification numérique soit linguistiquement comprise dans une certaine mesure hors de son contexte n'est pas en soit une contrainte lors de la phase de classification. Toutefois à l'affichage des résultats, ce facteur devient important dès lors que l'interprétation faite par l'utilisateur en dépend. La plupart des analyseurs statistiques fondés sur le calcul de la fréquence des cooccurrences utilisent le mot comme unité d'information, même si celui-ci ne répond pas à toutes les contraintes énumérées ici. Cependant, l'importance de l'ergonomie interprétative des mots a prévalu sur tout autre aspect, particulièrement ceux liées aux aspects multilingues de l'analyse. Ce dernier facteur devient aujourd'hui incontournable : l'essor du Web confirme ce besoin de multilinguisme. Il semble donc impératif que les modèles pour l'analyse de corpus, qu'ils soient numériques ou linguistiques, tiennent compte du caractère multilingue des textes à analyser.

# 2. Les n-grams de caractères

Bien qu'ayant été proposée depuis longtemps et utilisée principalement en reconnaissance de la parole, la notion de **n-grams de caractères** prit davantage d'importance avec les travaux de

Greffenstette (1995) sur l'identification de la langue, et de Damashek (1995) sur le traitement de l'écrit. Autre autres, ils prouvèrent que ce découpage, bien que différent d'un découpage en mots, ne faisait pas perdre d'information. Parmi des applications récentes des n-grams on retrouve des travaux sur : l'indexation (Mayfield & McNamee, 1998) ; l'hypertextualisation automatique multilingue avec les travaux de Halleb et Lelu (1998) qui, à travers une méthode de classification thématique de collections de textes, indépendante du langage, construisent des interfaces de navigation hypertextuelle ; ou l'analyse exploratoire multidimensionnelle en vue d'une recherche d'information dans des corpus textuels (Lelu *et al*., 1998).

On définira un n-gram de caractères par une suite de n caractères : bi-grams pour n=2, tri-grams pour n=3, quadri-grams pour n=4, etc. Il n'est plus question de chercher un délimiteur comme c'était le cas pour le mot. Un découpage en n-grams de caractères, quelque soit n, reste valable pour toutes les langues utilisant un alphabet et la concaténation comme opérateur de construction de texte. Le choix des n-grams apporte un autre avantage très important : il permet de contrôler la taille du lexique et de la maintenir à un seuil raisonnable. La taille du lexique était jusqu'à présent l'aspect le plus controversé et considéré comme une limite des techniques fondées sur la comparaison des chaînes de caractères. En effet, un découpage en mots fait que la taille du lexique est d'autant plus grande que le corpus est grand. Cette limite subsiste malgré certains aménagements tels le "nettoyage" des mots fonctionnels, la lemmatisation et la suppression des hapax. Un lexique obtenu suite à un découpage en n-grams de caractères ne peut dépasser la taille de l'alphabet à la puissance n. Le choix d'un découpage en quadri-grams pour une langue de 26 caractères donnerait une taille maximale de $26^4$ entrées, soit un lexique de 456 976 quadri-grams possibles. Si on élimine les combinaisons qu'il est impossible de rencontrer (p.ex. AAAA, ABBB, BBBA, etc.), ce nombre diminue de façon considérable. D'ailleurs ce nombre est estimé par Lelu *et al*. (1998) à quelques 13 087 quadri-grams pour un texte de 173 000 caractères.

Dans une approche avec découpage en n-grams de caractères, contrairement aux approches avec découpage en mots, il n'est pas question d'utiliser la lemmatisation pour réduire le lexique. La lemmatisation (qui consiste à remplacer une forme fléchie par son lemme) est, d'une part, relativement lourde à mettre en œuvre sur le plan informatique mais en plus, impose un traitement spécifique à chaque langue. Qui plus est, plusieurs lemmatiseurs ne semblent pas être en mesure de ramener des termes comme informatisation, informatique, et informatiser à un même concept qu'est l'informatique. Or souvent dans les corpus, on utilise des expressions ayant quasiment le même contenu informationnel comme, par exemple, dans les segments suivants : "l'informatisation de l'école", "informatiser l'école" et "introduire l'informatique à l'école". Le découpage des trois segments en n-grams est suffisant pour classer les trois segments dans la même classe car, outre le mot école qui est redondant dans les trois expressions, les tri-grams inf, nfo, for, orm, rma, mat et ati, permettent par un calcul de similarité d'affirmer que c'est d'informatique dont il est question. Par ailleurs, les tri-grams susmentionnés apparaissent aussi dans le découpage des mots information, informationnel, etc., ce qui peut être considéré à juste titre comme du bruit, à moins bien sûr que l'on évoque une interprétation sémantique particulière de l'informatique comme étant une science de l'information.

# 3. GRAMEXCO : n-GRAMs pour l'EXtraction des COnnaissances

GRAMEXCO est un outil logiciel que nous avons développé pour la classification numérique des gros corpus et l'extraction de connaissances sur le contenu des textes. La classification

numérique s'effectue au moyen d'un réseau de neurones ART comme celui utilisé dans Biskri et Delisle (1999). L'unité d'information considérée est le n-gram de caractères, la valeur de n étant paramétrable. L'objectif visé est de fournir la même chaîne de traitement, peu importe la langue du corpus, avec toutefois des aménagements dans la présentation des résultats pour en permettre une relative facilité de lecture comme nous le verrons plus loin. Le fonctionnement de GRAMEXCO n'est pas totalement automatique. Le choix de certains paramètres est fait par l'utilisateur en fonction de ses propres objectifs. Du choix de ces paramètres dépend l'interprétation des résultats qui se fait par l'utilisateur en fonction de sa subjectivité. GRAMEXCO prend en entrée un texte brut (non indexé) sous format ASCII. Il s'en suit trois grandes étapes où l'utilisateur peut paramètrer certains traitements.

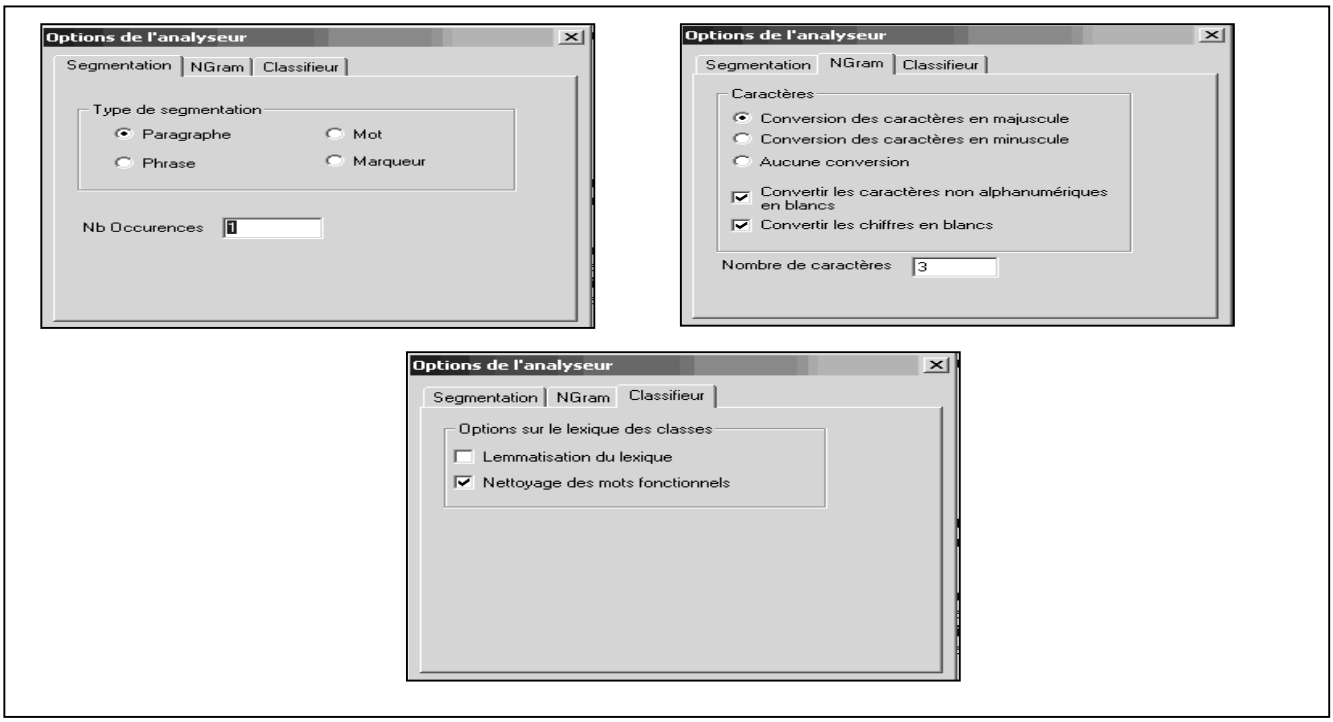


**Figure 1 : Paramétrage de l'outil GRAMEXCO**

1. La **première étape** consiste à construire la liste des n-grams de caractères contenus dans le texte ainsi qu'à partitionner le corpus en plusieurs segments. Les deux opérations se faisant simultanément, nous récupérons en sortie une matrice où seront répertoriés les fréquences d'apparition de chaque n-gram dans les différents segments. Le choix de la valeur du n (bi-gram, tri-gram, quadri-gram, etc.) dépend de l'utilisateur et de l'expertise qu'il veut mener. Outre la valeur du n, d'autres paramètres (voir Figure 1) sont la possibilité d'effectuer la conversion des caractères non alphanumériques en caractère espace, ou encore la conversion des chiffres en caractère espace. Ces deux paramètres répondent aux besoins d'une analyse pour laquelle les chiffres, la ponctuation ou encore d'autres caractères spécifiques seraient importants pour la qualité des résultats. Dans un texte technique par exemple, il serait peut être intéressant de savoir si version1 est différente de version2 et, par conséquent, les chiffres pourraient avoir autant d'impact informatif que les caractères alphabétiques. Le dernier paramètre pour les n-gram est en rapport avec la conversion des caractères majuscules en minuscules, ou vice versa. Si aucune de ces conversions n'est choisie, alors GRAMEXCO distinguera les lettres minuscules des majuscules. L'autre aspect important de cette première étape est le paramétrage de la segmentation. Ainsi, nous pouvons partager le texte soit en des

sections formées d'un nombre déterminé de phrases, de paragraphes ou de mots, ou tout simplement des sections séparées par un caractère spécial. Ce paramètre est toujours choisi par l'utilisateur. Le pseudo-lexique formé de n-grams subit au cours de cette première étape un nettoyage (voir Figure 2) soit, l'élimination des "n-grams hapax" dont la fréquence est inférieure à un certain seuil ou supérieure à un autre seuil, l'élimination de n-grams spécifiques sélectionnés dans la liste (par exemple des n-grams contenant des espaces) ou encore, si on veut pousser les choses plus loin, l'élimination de certains n-grams considérés comme fonctionnels, particulièrement les suffixes.

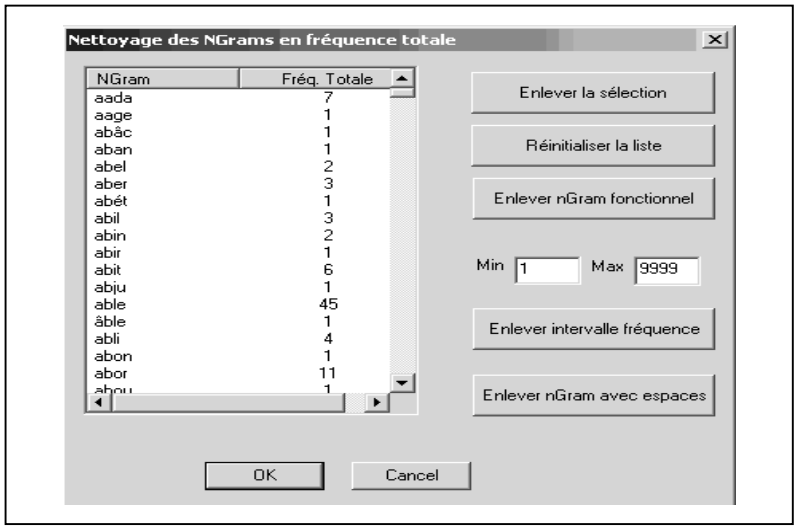


**Figure 2 : Nettoyage de la liste des n-grams produits par GRAMEXCO**

2. Dans la **deuxième étape**, les segments représentés dans la matrice obtenue à l'étape précédente sont comparés entre eux au moyen d'un réseau de neurones (ART dans notre cas). Les segments qui sont semblables, étant donnée une certaine fonction de similarité, sont classés dans les mêmes groupes. En simplifiant, on peut dire que deux segments sont semblables s'ils sont constitués des mêmes n-grams avec des fréquences presque identiques. Le choix du réseau ART pour la classification n'est pas dicté par des raisons de performances particulières car tel n'est pas notre objectif pour le moment. Nous aurions tout aussi bien pu choisir un autre réseau neuronal qui aurait certes donné des résultats différents — nous recommandons Turenne (2000) au lecteur intéressé aux méthodes et outils de classification pour le texte. De telles variations apparaissent dans les résultats d'une étude expérimentale et comparative des méthodes statistiques et des champs de Markov pour l'analyse de textes par ordinateur présentés dans Benhadid *et al*. (1998). Comme suite à ce travail, nous gardons d'ailleurs l'idée de paramètrer GRAMEXCO afin de permettre à l'utilisateur de choisir d'autres réseaux de neurones.

3. La configuration du résultat de la classification numérique se présente par l'affichage des classes de segments et, pour chaque classe, l'affichage des segments qui la constituent d'une part, et du lexique qui la forme d'autre part (voir Figure 3). À cette **troisième étape** la notion de n-gram n'est plus de mise. Il serait en effet impossible à un utilisateur d'interpréter des résultats et de donner des thèmes aux différentes classes à partir d'une seule liste de n-grams. Comme le souligne Turenne (2000), l'interprétation de telles classes est déjà un exercice non trivial en lui-même, dépendant *des* points de vue de l'utilisateur : il ne serait donc pas utile de lui rendre cette phase moins intuitive en utilisant une liste de n-grams. Le lexique de chaque classe est formé par les mots que

contiennent les différents segments de cette classe. L'utilisateur pourra considérer le lexique comme l'union des mots des segments pour déterminer le thème global des classes, leur intersection pour déterminer le thème commun partagé par les segments, leur différence pour identifier des gains informationnels, ou encore tous ceux dont la fréquence est au dessus d'un certain seuil, etc. L'utilisateur peut également lemmatiser le lexique des classes comme il peut en retirer les mots fonctionnels. L'utilisateur peut appliquer l'opération de lemmatisation à l'ensemble des lexiques de toutes les classes ou seulement au lexique d'une seule classe, ceci en fonction de contraintes de temps. Il est à retenir que la lemmatisation et la suppression des mots fonctionnels n'interviennent que pour améliorer l'aspect des résultats et n'interviennent nullement avant la classification à proprement parler. Toutes ces configurations du lexique sont à même d'aider l'utilisateur à proposer son interprétation des résultats. En effet, il demeure que GRAMEXCO, comme nous l'avons souligné plus haut, ne propose pas d'interprétation automatique. Il ne fait que faire ressortir les similarités et les régularités découvertes dans le corpus.

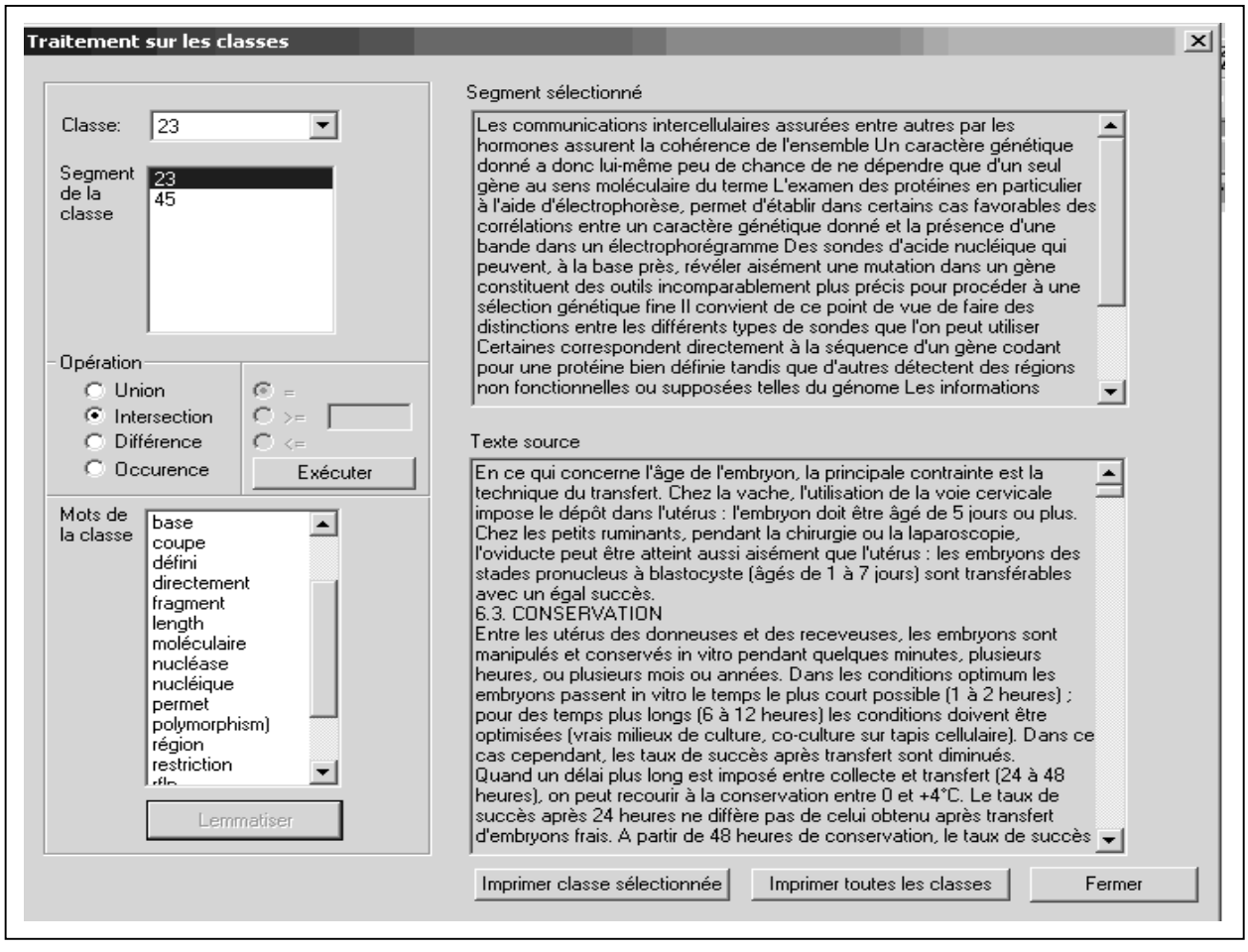


**Figure 3 : Configuration des résultats de GRAMEXCO**

Dépendant des paramètres choisis, les résultats de GRAMEXCO peuvent servir à plus d'une finalité. Comme nous le verrons à l'aide des exemples de la prochaine section, nous pouvons :

- déterminer le contenu lexical des segments similaires, et ainsi connaître le thème principal de ces segments ;
- déterminer l'acception et la signification d'un mot de par les mots qui lui sont associés dans une classe donnée ; et

- construire des classes de mots formés à partir d'un radical commun comme pour l'exemple avec informatiser, informatisation, et informatique.

# 4. Évaluation

Nous avons mené deux évaluations principales qualitatives (et deux complémentaires). La première voulait montrer le comportement d'une classification numérique fondée sur les n-grams de caractères. Elle a été réalisée sur un corpus formé d'une cinquantaine de pages (format ASCII) construit à partir d'extraits de documents trouvés sur le Web. Ces documents couvrent divers domaines et permettent une hétérogénéité du contenu du corpus et, par conséquent, une meilleure compréhension des résultats de la classification. La deuxième évaluation avait pour but d'expliquer pourquoi la classification avec les n-grams pouvait être aussi performante sinon plus performante qu'une "classification + lemmatisation". Cette évaluation a été réalisée sur un texte de deux pages. Sa finalité n'en exigeait pas plus pour construire des classes de mots ayant un même radical.

Pour les opérations préliminaires de la **première évaluation principale**, soit la segmentation et l'extraction des n-grams, nous avons opté pour les paramètres suivants : 10 (phrases) pour déterminer la taille d'un segment et 4 (caractères) pour déterminer la taille des n-grams[1]. De plus, à l'aide des paramètres de GRAMEXCO, nous avons considéré les lettres majuscules identiques aux lettres minuscules et nous avons remplacé les caractères non alphanumériques et les chiffres par des espaces. Nous avons ainsi récupéré 174 segments et 4 857 quadri-grams, après un "ménage" de la liste des n-grams qui a consisté à supprimer les n-grams contenant un ou plusieurs espaces et les n-grams ayant une fréquence égale à 1. La classification elle-même, au moyen du réseau de neurones ART avec un paramètre de vigilance de 0.1, donne lieu à la production de 100 classes de segments présentants des similarités. Examinons maintenant quelques résultats :

- La classe 100 regroupe les segments 137 et 157. Le lexique de cette classe formé de l'intersection des lexiques des deux segments est constitué par : {bourse, francs, marchés, millions, mobile, pdg, prix}. On constate, au regard de ce lexique, que le mot francs désigne la monnaie française et n'a aucun rapport avec la franchise ou avec les fameuses tribus "les francs". Ce même lexique nous renseigne également sur le thème commun que se partagent les segments 137 et 157, en l'occurrence le domaine financier.
- La classe 54 regroupe les segments 141 et 143. L'intersection des lexiques des segments de la classe 54 est formée de : {appel, cour, décidé, juge}. Ainsi pour le mot cour, une seule signification est possible au regard des mots qui l'accompagnent : cour de justice. On écarte aisément les sens suivants : la cour qu'on fait à une demoiselle, la cour de récréation, ou encore les toilettes des Belges. Le thème de la classe 54 est par ailleurs bien identifié en ce sens qu'il s'agit de segments dont le contenu traite d'affaires judiciaires.
- La classe 98 regroupe les segments 71 et 73. Le lexique issu de l'intersection des lexiques des deux segments est formé des mots : {culture, économiques, eurasistes, matérialiste, occident}. Dans ce contexte, le terme culture ne peut signifier que culture économique, et son utilisation n'est pas pour introduire une quelconque notion d'agriculture. Ce qui se confirme d'ailleurs avec le mot occident qui est utilisé ici dans le sens bloc géopolitique et non "là où se couche le soleil". Le thème de la classe 98 traite sans conteste d'options économiques ce que nous pouvons d'ailleurs vérifier au travers de la lecture des segments 71 et 73.

[1] Selon Damashek (1995), les quadri-grams donneraient les meilleurs résultats pour l'anglais. Lelu *et al*. (1998) semblent confirmer cela pour le français.

- La classe 64 regroupe les segments 166 et 167. Le lexique qu'on retiendra pour cette classe est formé de tous les mots dont la fréquence dans les segments 166 et 167 est supérieure ou égale à 2, en l'occurrence les mots : {chance, dernière, dire, match, stade, supporters, vélodrome}. Le mot stade, du fait particulièrement de la présence des mots match, supporters et vélodrome, est compris comme étant un stade de football. Par ailleurs, pour un public averti qui sait que le vélodrome est le stade de Marseille, on comprend aisément que les deux segments 166 et 167 traite des supporters de l'Olympique Marseillais.
- La classe 13 regroupe les segments 32, 35, 41 et 48. Le lexique de cette classe formé de l'intersection des lexiques des quatre segments est constitué du seul mot : russe. Celui-ci est suffisant pour nous permettre de conclure que le thème partagé par les quatre segments se rapporte à la Russie. L'union des lexiques, formée entre autres des mots : conservateur, socialisme, marxiste, conservateur, révolutionnaire, Dostoievski, doctrine, impérial, slavophile, etc., nous permet de préciser que le thème de la classe 13 est dédié aux slavophiles et à la culture politique russe du 19$^{ième}$ siècle. Une remarque s'impose : on imagine mal comment une classification fondée sur les mots aurait pu arriver à regrouper les segments 32, 35, 41 et 48 dans la même classe sans avoir recours à la lemmatisation étant donné que le seul mot commun est russe. Reste que la lemmatisation est relativement coûteuse en temps d'exécution et est une opération spécifique à chaque langue. Nous évitons ces inconvénients en utilisant les n-grams de caractères.

Les raisons d'aussi bonnes performances nous les retrouvons dans les résultats de la **deuxième évaluation principale**. En effet, celle-ci consistait à passer un texte de deux pages formé d'extraits d'un corpus sur les biotechnologies (utilisé dans Biskri et Delisle, 2000) par une classification basée sur les n-grams avec, comme paramètre, n=4 et la taille du segment ramenée à un mot seulement. Ainsi les segments regroupés dans une même classe seraient constitués des mots ayant des points communs, en particulier un radical commun et, donc, référant à une notion commune. Cette évaluation a permis de construire l'échantillon de classes suivantes :

Classe 101 : {survécu, survie}
Classe 102 : {utilisée, outil}
Classe 110 : {congelé, décongelé, congelés, congélateur}
Classe 112 : {simple, simplifier, simplifiée}
Classe 162 : {avenir, devenir}
Classe 4 : {principale, principalement}
Classe 48 : {optimisées, optimum}
Classe 60 : {cellules, cellulaire}
Classe 65 : {collecte, collectifs}
Classe 7 : {transfert, transférables, transférés, pénétrant, transferts, retransfert}
Classe 81 : {glycol, glycérol}
Classe 88 : {déshydratées, déshydratation}

Si nous prenons la classe 110 par exemple, nous nous apercevons que non seulement congelé et congelés sont regroupés, ce qu'aurait d'ailleurs fait une lemmatisation standard, mais en plus la classification leur associe décongelé et congélateur. En somme la classe 110 regroupe tout ce qui se rapporte à la notion de congélation. Il en est de même pour les autres classes qui chacune regroupe des mots partageant des notions communes. Ainsi la classe 101 porte sur la notion de survie, la classe 102 sur la notion d'outils utiles, la classe 112 sur la simplicité, la classe 48 sur l'optimalité, la classe 65 sur la notion de cellule, la classe 65 sur celle de la collection, la classe 7 sur la notion de transfert, la classe 81 sur un concept chimique particulier, et la classe 88 sur la notion de déshydratation.

Nous avons aussi effectué une **troisième évaluation**, complémentaire aux deux premières. Elle a consisté particulièrement à comparer les résultats de GRAMEXCO lors de la première

évaluation à des résultats obtenus avec NUMEXCO (Biskri et Delisle,1999 ; Meunier *et al*., 1997) — NUMEXCO est un autre outil logiciel que nous avons développé et qui, contrairement à GRAMEXCO, considère comme unité d'information le mot (et non les n-grams de caractères). Nous avons soumis à NUMEXCO le même texte ASCII que lors de notre seconde évaluation et ce, avec le même paramètre de segmentation. Nous avons obtenu un lexique de 4 884 mots, après lemmatisation et suppression des mots fonctionnels. La suppression des hapax (mots dont la fréquence est égale à 1) aurait diminué le lexique à 1 755 unités d'information. Notons cependant que la suppression des hapax renvoie en quelque sorte à une suppression de n-grams dont la fréquence pourrait dépasser 1, ce qui n'était pas le cas dans notre première évaluation où nous n'avons supprimé que les n-grams dont la fréquence était égale à 1 — ce facteur donne lieu à une comparaison biaisée. Ceci dit, il est important de souligner que pour un texte ne dépassant pas les 200 pages environ, la taille du lexique et le nombre de n-grams ne diffèrent pas de beaucoup. Pour certains textes, le nombre de n-grams peut dépasser la taille du lexique. Cependant, dès lors que la taille du texte dépasse les 200 pages, voire les 300 pages, la taille du lexique tend à augmenter alors que le nombre de n-grams se stabilise. Sur le plan de la classification à proprement parler, les classes que nous sommes arrivés à construire avec GRAMEXCO ont été impossibles à reproduire avec NUMEXCO et ce, en raison du peu de mots communs après lemmatisation se trouvant dans les différents segments des classes.

Finalement, dans le cadre d'une **dernière évaluation** complémentaire, nous sommes revenus à la question du multilinguisme et avons soumis à GRAMEXCO un corpus constitué de deux textes anglais portant respectivement sur la météo et la mécanique des petits moteurs — nous ferons prochainement d'autres évaluations sur un corpus constitué de plusieurs textes anglais provenant de différents sites Web. Nous avons obtenu des résultats tout à faire similaires à ceux des évaluation précédentes effectuées sur des corpus français et ce, *sans aucun traitement spécifique à la langue anglaise*. Il reste toutefois à l'utilisateur d'interpréter les résultats issus d'un corpus d'une ou de plusieurs autres langues. Cependant, avec la disponibilité de certains outils de traduction automatique modernes, on peut penser que l'utilisateur pourrait d'abord s'intéresser à la traduction du lexique construit à partir de l'occurrence des mots dans les segments d'une même classe pour décider s'il y a lieu de poursuivre son examen des textes correspondants.

Enfin, nous sommes conscients que les évaluations que nous présentons ici sont plus qualitatives que quantitatives. De plus, elles ne tiennent pas compte de l'influence des paramètres. Nous aurons certainement l'occasion dans nos prochaines publications de présenter d'autres évaluations complémentaires à celles présentées dans le présent article.

# 5. Conclusion

Notre travail semble très concluant quant à l'importance des n-grams de caractères dans l'acquisition des connaissances à partir de grands corpus, ce qui va dans le même sens que d'autres travaux récents. Comme nous l'avons montré, l'utilisation des n-grams dans une classification numérique permet de penser l'outil comme étant multilingue : *aucun module n'est spécifique à une langue particulière*. Les deux grandes contraintes que nous avions avec des outils de classification fondés sur les mots, à savoir la définition informatique du mot et la lemmatisation, se voient définitivement écartées, du moins aux étapes de classification. La configuration des résultats a ses propres contraintes et nous devons nécessairement choisir de présenter les résultats dans une forme admissible par l'utilisateur, ce qui n'est pas une

question simple au plan ergonomique. Nous avons opté pour les classes de segments et les classes de mots. Les deux résultats fondamentaux auxquels nous sommes arrivés, outre l'aspect multilingue, sont d'une part la qualité remarquable de la classification et d'autre part le fait que le choix des n-grams restreint la taille des vecteurs soumis à la classification pour des textes d'une taille supérieure à 200-300 pages. Comme suite à ce travail, nous envisageons de greffer GRAMEXCO à un système d'aide à la recherche d'information sur le Web et ce, dans le but de classifier les sites obtenus suite à une requête formulée à l'aide d'un moteur de recherche. L'idée serait de cerner les sites contenant les mots clés de notre requête avec la bonne acception des mots d'une part et, d'autre part, la possibilité de reformuler la requête en fonction des résultats de la classification. De telles perspectives ne sont envisageables de façon réaliste que grâce à des outils supportant le multilinguisme, comme GRAMEXCO.

# Références

Balpe, J.P., Lelu, A. Papy, F. (1996), *Techniques avancées pour l'hypertexte*. Paris, Hermes.

Benhadid, I., Meunier, J.G., Hamidi, S., Remaki, Z., Nyongwa, M. (1998), "Étude Expérimentale Comparative des Méthodes Statistiques pour la Classification des Données Textuelles", *Actes de JADT-98*, Nice, France.

Biskri, I., Delisle, S. (2000), "User-Relevant Access To Textual Information Through Flexible Identification Of Terms: A Semi-Automatic Method And Software Based On A Combination Of N-Grams And Surface Linguistic Filters", *Actes de RIAO-2000*, Paris, France, 1059-1068.

Biskri, I., Delisle, S., (1999), "Un Modèle Hybride pour le Textual Data Mining : Un Mariage de Raison entre le Numérique et le Linguistique", *Actes de TALN-99*, Cargèse, France, 55-64.

Damashek, M., (1995), "Gauging Similarity with n-Grams : Language-Independent Categorization Of Text", *Science*, 267, 843-848.

Greffenstette, (1995), "Comparing Two Language Identification Schemes", *Actes de JADT-95*, 85-96

Halleb M., Lelu A., (1998), "Hypertextualisation Automatique Multilingue à Partir des Fréquences de n-Grammes", *Actes de JADT-98*, Nice, France.

Lelu A., Halleb M. , Delprat B. (1998), "Recherche d'information et Cartographie dans des Corpus Textuels à Partir des Fréquences de n-Grammes", *Actes de JADT-98*, Nice, France.

Manning, C.D., Schütze, H., (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Mayfield, J., Mcnamee, P., (1998), "Indexing Using both n-Grams and Words", *NIST Special Publication 500-242 : TREC 7*, 419-424.

Meunier, J.G., Biskri, I., Nault, G., Nyongwa, M. (1997), "Aladin et le Traitement Connexionniste de l'Analyse Terminologique", *Actes de RIAO-97*, Montréal, Canada, 661-664.

Turenne, N. (2000), *Apprentissage statistique pour l'extraction de concepts à partir de textes (Application au filtrage d'informations textuelles),* thèse de doctorat en informatique, Université Louis-Pasteur, Strasbourg, France.