Détection automatique de grandes thématiques de la propagande Nord Coréenne

Natalia Grabar*, Mason Richey**

* CNRS, UMR 8163, F-59000 Lille, France
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
natalia.grabar@univ-lille3.fr
http://natalia.grabar.free.fr/
**Graduate School of International and Area Studies
Hankuk University of Foreign Studies, South Korea
mrichey@hufs.ac.kr

Résumé. Les rhétoriques utilisées pour mener les politiques nationale et internationale sont liées entre elles, mais ne montrent pas souvent les mêmes priorités ni les mêmes modes de communication. Lorsque ces deux rhétoriques sont bien utilisées, elles permettent de maximiser la crédibilité et le potentiel de coercition d'un pays, tout en signalant son ouverture vers la coopération. La propagande Nord Coréenne est particulière parce qu'elle se base sur une rhétorique agressive dans sa politique internationale et parce qu'elle aborde des thématiques bien spécifiques. Nous proposons de distinguer automatiquement les grandes thématiques abordées dans la propagande Nord Coréenne en utilisant des scores de similarité lexicale entre deux articles et des méthodes de clustering non supervisé. L'évaluation des clusters obtenus indique que parmi les thématiques les plus présentes se trouvent par exemple: la propagande contre les États-Unis, le Japon et la Corée du Sud; les voeux et félicitations à destination de différents pays, présidents, institutions et personnalités; les honneurs reçus par les dirigeants de la Corée du Nord; les achèvements et avancées du pays et du peuple Nord Coréen; les relations officielles à différents niveaux; les nouvelles du quotidien Rodong Sinmun; les nouvelles inter-coréennes; la culture traditionnelle. Nous avons plusieurs perspectives à ce travail.

1 Introduction

Les politiques nationale et internationale, même si elle sont liées entre elles, ne montrent pas souvent les mêmes priorités ni les mêmes modes de communication. De plus, la rhétorique qui est utilisée avec succès pour l'une peut avoir plutôt un effet contraire sur l'autre. La plupart de dirigeants des pays arrivent à adapter leurs discours pour refléter leur orientation politique. L'exercice n'est pas évident. En voilà deux exemples :

– la politique intérieure est souvent appuyée par le protectionnisme de la nation dans différents domaines (*e.g.*, économie, emploi, sécurité), alors que la politique extérieure gagne

en montrant l'ouverture d'un pays dans différents domaines (*e.g.*, économie, emploi, sécurité) ;
– les politiciens emploient souvent un langage relativement dur en politique intérieure, tandis qu'ils utilisent un langage relativement diplomatique en politique extérieure.

Les théoriciens de relations internationales et les praticiens d'affaires étrangères comprennent très bien les enjeux et la difficulté de concilier les deux rhétoriques. Lorsque ces deux types de discours sont bien utilisés, ils permettent de maximiser la crédibilité et le potentiel de coercition d'un pays, tout en signalant son ouverture vers la coopération. Il existe cependant le danger de la perte de crédibilité d'un dirigeant dû au fait que, après une période d'escalade d'agressivité au niveau international, il revient en arrière et diminue cette agressivité dans son discours. En général, ceci est assez mal perçu au niveau national. C'est ce qu'on appelle *"audience cost"* (Fearon, 1994; Weeks, 2008; Weiss, 2013). La ligne politique doit donc rester bien cohérente.

Le bon dosage des deux modes de communication est important. Lorsqu'il n'est pas respecté, cela peut mener vers l'incompréhension des attentes et des intentions et même vers des situations de crise. La Corée du Nord est un exemple déconcertant de politique extérieure et intérieure, augmenté par les provocations militaires systématiques (Richey, 2015; Westby, 2014). D'une part, la Corée du Nord diffère d'autres pays par le fait que son audience et l'impact interne sont faibles à cause du régime dictatorial. D'autre part, la rhétorique politique est particulière à cause du niveau extraordinaire de grandiloquence et de belligérance, qui se manifestent dans la communication intérieure et extérieure. En effet, sa rhétorique belliqueuse au niveau international est fameuse pour ses locutions extrêmement créatives et hyperboliques :

– *Séoul va être transformé en "une mer de feu"*
– *l'armée de la Corée du Nord est "prête à mener la guerre sainte contre Séoul"*
– *la Corée du Nord va utiliser les armes nucléaires contre la Maison Blanche et le Pentagone, qui sont "les sources du mal"*
– *les États-Unis est le "colonisateur Yankee" et ses citoyens sont des "barbares"*
– *le président Obama est un "proxénète", Lee Myung Bak une "marionnette américaine", alors que le président Park se distingue par son "sifflement venimeux d'une mauviette"*

Ce qui est intéressant est que ces locutions sont systématiquement utilisées pour parler d'autres pays, typiquement des États-Unis, du Japon et de la Corée du Sud, et de leurs dirigeants, et font souvent partie de réponses que la Corée du Nord fait face aux actions politiques éventuelles, comme par exemple la condamnation de la Corée du Nord à cause du non-respect des droits de l'homme, les sanctions de l'ONU ou des entraînements militaires communs des États-Unis et de la Corée du Sud. Par ailleurs, même si la plupart de ces locutions sont créées en coréen, elles sont souvent traduites en anglais par la chaîne KCNA [1] pour être ensuite diffusées à l'international. Comme la grande majorité de Nord-Coréens ne parle pas anglais, on peut supposer que ces traductions sont vraiment destinées pour l'audience internationale.

Malgré la nature intéressante de la politique et du discours nord coréens, il existe très peu de travaux qui y sont consacrés. Nous proposons de contribuer à l'étude de la propagande politique, sur l'exemple de la propagande produite par la Corée du Nord. Dans la suite de cette contribution, nous présentons d'abord des travaux existants en relation avec le sujet traité (section 2), nous décrivons ensuite la méthode utilisée (section 3). Nous présentons et discutons les résultats (section 5) et concluons avec quelques perspectives à ce travail (section 6).

1. `http://www.kcna.us/`

2 Travaux existants

Les écrits journalistiques et politiques sont objet de plusieurs travaux en TAL, lexicométrie, étude discursive et étude de corpus de manière générale. Mentionnons par exemple : l'analyse générale du discours syndical (Habert, 1983; Salem, 1993) et politique (Salem, 1981), la propagation des informations sur les réseaux (Bourigault et al., 2014), la détection de buzz et de fausses rumeurs (Chou et al., 2015; Ma et al., 2015; Fuchs et Yu, 2015), la véracité des informations et les nouveaux modèles du journalisme sur les réseaux (Derczynski et Bontcheva, 2014; Sharma, 2015; Maigrot et al., 2016).

En relation avec la Corée du Nord, les chercheurs se concentrent sur différentes thématiques : le discours politique nord coréen et son impact sur la sécurité locale et mondiale (Myers, 2010, 2015; Richey, 2015; Ohn et Richey, 2015), le programme nucléaire nord coréen (Rich, 2012, 2014b), la rhétorique belliqueuse, essentiellement en relation avec la provocation militaire (Joo, 2015), l'émergence médiatique de leaders (Rich, 2014a), comparaison de discours de Kim Il Sung et Fidel Castro (Malici et Malici, 2005). Par ailleurs, les unités des analyses plus généralistes peuvent être les mois, les semaines (Zuell, 2010) ou les jours (Rich, 2012), avec des données qui s'étendent sur une à trois années. Dans la plupart de travaux existants, l'analyse du discours est effectuée manuellement par les chercheurs qui viennent essentiellement des domaines de relations internationales et d'études politiques. Dans de rares cas où des méthodes automatiques sont utilisées, elles exploitent : (1) des techniques de data mining et de lexicométrie, comme par exemple les pondérations (fréquences, tfidf ou jaccard) ou la régression binomiale d'un ensemble de termes prédéfinis (Haynes, 2001; Rich et Liu, 2012) ; (2) l'apprentissage supervisé pour détecter les articles provocatifs de la Corée du Nord. Ainsi, avec cinq mots-clés (*years, signed, assembly, June, Japanese*) une précision de 82 % est obtenue (Whang et al., 2016). Il existe également des travaux qui étudient la propagande et le discours extrémiste produit par d'autres pays ou groupes : détection automatique de commentaires commandités par le gouvernement de Chine (Blake et Miller, 2016) et détection de traces digitales d'extrémistes solitaires sur les réseaux (Chen, 2007; Brynielsson et al., 2012).

L'objectif de notre travail consiste à analyser les articles de l'agence de presse KCNA de la Corée du Nord pour détecter les grandes thématiques des articles de propagande. Nous abordons cette question de recherche comme un problème de clustering de textes.

3 Méthodes

Nous présentons d'abord les données étudiées et ensuite les différentes étapes de la méthode : pré-traitement, calcul de similarité entre les articles, le clustering et l'évaluation.

3.1 Données de la propagande Nord Coréenne

Nous avons collecté les articles du site KCNA qui fournit la propagande officielle de la Corée du Nord. Créée en 1946, l'agence KCNA effectue une publication journalière en ligne d'articles en anglais depuis 1997. Ce site ne propose que des données textuelles (pas d'images ni de vidéos). Il est admis que l'utilisation de la langue anglaise pour la propagande de KCNA cible une audience différente par rapport à la propagande en langue coréenne et conduit donc à

une différence de contenu aussi (Poneman et al., 2004). Néanmoins, l'exploitation de la propagande produite et traduite directement par une agence Nord Coréenne pour le public étranger permet d'éviter les biais de perception occidentale de ce type de littérature idéologique. De plus, les données sont accessibles en ligne pour plusieurs années maintenant.

Le corpus total contient 121 964 articles, soit 31 211 998 occurrence de mots. Le volume de la propagande va en augmentant au fil des années, autant en nombre d'articles que d'occurrences de mots. Le pic des années 2011 et 2012 est sans doute causé par les changements de leaders politiques dans le pays : l'émergence médiatique et politique de Kim Jong-Un, suite à la maladie de son père Kim Jong-Il et surtout par rapport à ses deux frères aînés, et la transition entre le père Kim Jong-Il et son fils cadet Kim Jong-Un. Nous proposons de nous concentrer sur deux années : 2003 (4 852 articles, 1 302 220 occ.) et 2013 (9 967 articles, 2 766 178 occ.). La méthode, ajustée sur ces données, pourra ensuite être testée sur le corpus entier.

3.2 Pré-traitement du corpus

Nous effectuons une série de pré-traitement du corpus :
- la conversion du format HTML au format texte ;
- la séparation des articles selon la langue dans laquelle ils sont écrits. Nous exploitons les listes de mots proposées dans un travail précédents pour distinguer l'anglais, le français et l'allemand (Grefenstette et Nioche, 2000) et y ajoutons des mots fréquents et typiques pour distinguer l'espagnol ;
- la suppression de motifs systématiquement insérés dans les articles, comme par exemple la date ou l'année de Juche, qui est l'idéologie autocratique développée par le 1er président de la Corée du Nord Kim Il-Sung et sur laquelle repose le régime de la Corée du Nord. À titre d'information, 2017 est l'année Juche 107 ;
- la création de trois ensembles de données : articles complets, titres des articles et corps des articles ;
- l'étiquetage morpho-syntaxique avec Treetagger (Schmid, 1994).

3.3 Calcul de similarité

La similarité entre chaque paire d'articles est calculée avec `Text::Similarity` [2], un module écrit en Perl et permettant de calculer l'intersection lexicale entre les textes traités. Ce module effectue des traitements supplémentaires : la suppression de la ponctuation, la minusculisation et la suppression de mots grammaticaux. La similarité est estimée être le nombre de mots communs dans les deux fichiers comparés, pondéré par la longueur de chaque fichier. Les valeurs de similarité sont entre 0 et 1.

3.4 Clustering

Le clustering est effectué avec l'algorithme `MCL` (Markov Cluster Algorithm), qui permet d'effectuer un clustering non supervisé sur des graphes (van Dongen, 2000). Cet algorthme a plusieurs avantages importants pour nous : il est simple d'utilisation, très rapide et il n'est pas nécessaire de lui indiquer le nombre de clusters attendus en sortie. En revanche, il est

2. `http://search.cpan.org/dist/Text-Similarity/lib/Text/Similarity.pm`

possible de modifier la valeur d'un des paramètre `-I`, qui permet de régler la granularité des clusters : plus cette valeur est élevée plus la granularité des clusters est fine. Cet algorithme a été exploité avec de différents types de données et nous proposons de le tester sur les données de la propagande.

3.5 Évaluation

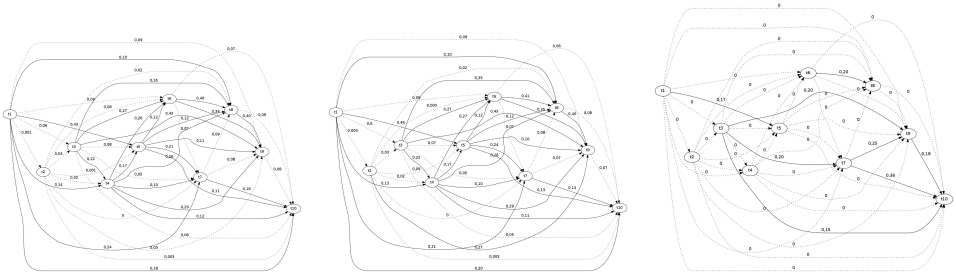
L'évaluation est effectuée manuellement et *a posteriori* des traitements automatiques. L'objectif est d'analyser le contenu des clusters obtenus. Pour ceci, nous analysons : (1) les mots les plus fréquents de chaque cluster et (2) les titres des articles de chaque cluster. Étant donné que les mots fréquents et les titres sont assez explicites, cela permet d'avoir un premier jugement sur la thématique des clusters et leur homogénéité.

4 Rationale de l'étude

Nous effectuons plusieurs tests, en variant plusieurs paramètres. Pour deux types d'unités linguistiques (formes et lemmes), nous étudions différents types d'unités textuelles : les articles entiers, les titres des articles, le corps des articles (sans leurs titres). Le recouvrement lexical entre deux unités textuelles est retenu comme valeur de similarité. Lors du clustering, nous testons plusieurs valeurs du paramètre `-I`, qui influence la granularité des clusters, dans l'intervalle [2,0, 9,5], en l'incrémentant par 0,5 point. Notons que 5,0 est la valeur par défaut de `I`. Les résultats présentés concernent les traitements effectués sur deux années, avec 10 ans de différence : 2003 et 2013. Le travail est effectué sur les articles écrits en anglais.

5 Résultats et discussion

Après la reconnaissance et le filtrage de la langue, nous retenons 3 926 articles en 2003 et 8 472 articles en 2013. D'autres articles de ces deux années sont soit en espagnol soit des articles trop courts pour lesquels la décision sur la langue ne peut pas être faite.



(a) Articles entiers (b) Corps des articles (c) Titres des articles

FIG. 1 – *Un exemple des valeurs de similarité au sein d'un ensemble de 10 articles.*

Nous supposons que les différents paramètres indiqués dans la section 4 vont avoir une influence sur les résultats. Par exemple, la séparation des articles en titre et corps de l'article et le calcul de similarité sur ces unités textuelles permet d'obtenir des valeurs de similarité différentes, comme présenté dans la figure 1 pour un ensemble de 10 articles. Sur cette figure, les noeuds représentent les articles alors que les arcs représentent les valeurs de similarité calculées entre chaque paire d'articles. Nous avons mis en pointillé les arcs dont les valeurs de similarité sont nulles ou inférieures à 0,10. Nous pouvons voir que les graphes obtenus avec les articles entiers et le corps des articles ont une forme similaire, avec toutefois des valeurs de similarité un peu différentes. Le graphe obtenu avec les titres est d'une part plus *"sélectif"* car il comporte beaucoup plus de valeurs de similarité nulles, et d'autre part il est différent car les similarités plus fortes ne relient pas toujours les mêmes articles. L'utilisation des formes et des lemmes produit les valeurs de similarité de mêmes grandeurs. Pour certaines paires d'articles, la similarité basée sur les lemmes est un peu plus élevée que celle obtenue avec les formes car les lemmes, en permettant de regrouper certaines formes ensemble, permettent également d'augmenter la similarité entre les textes.

`I`	*2003, Titres*	*2013, Titres*
4,0	99	70
4,5	118	83
5,0	143	107
5,5	176	121
6,0	196	145
6,5	220	165
7,0	240	181
7,5	266	199
8,0	292	222
8,5	316	246
9,0	344	263
9,5	363	276

TAB. 1 – *Nombre de clusters selon le paramètre `I` de `MCL`, avec les titres et les lemmes.*

Lorsque les graphes de similarités sont calculés, ils peuvent être exploités pour générer les clusters. Dans le tableau 1, nous présentons le nombre de clusters obtenus avec les valeurs de `I` allant de 4,0 à 9,5. Dans les tableaux 2 et 3, nous présentons les cinq clusters les plus gros, générés avec les titres des articles, les lemmes et `I`=5,0. Nous indiquons la taille de ces clusters et les mots-clés les plus fréquents. Par exemple, le 1er et 3eme clusters de 2003 et le 1er cluster de 2013 contiennent des articles dirigés contre les États-Unis, le Japon et la Corée du Sud. Les autres clusters de ces tableaux concernent la glorification de la Corée du Nord et de ses leaders à travers des commémorations, des cadeaux reçus et envoyé, des réunions, etc. À côté des clusters très et moyennement grands, nous avons 118 et 93 clusters de moins de 10 articles générés à partir des années 2003 et 2013, respectivement. Sur la base des informations comme celles présentées dans les tableaux 2 et 3, nous pouvons faire émerger quelques thématiques principales (dans l'ordre de leur importance) :

Taille	*Mots-clés*
1 156	*s./324 korea/307 dprk/217 korean/179 u.s./140 call/111 hold/61 people/58 national/50 struggle/44 japan/44 kcna/42 meeting/39 reunification/38 war/38 dispatch/36 troop/36 held/35 foreign/34 anti-u.s/30*
758	*greeting/193 president/169 minister/75 japan/64 dprk/62 prime/49 message/42 urged/38 foreign/32 koreans/32 anniversary/32 call/31 urge/30 korean/26 congratulation/25 envoy/24 special/24 national/23 sympathy/21 hold/20*
693	*u.s./408 dprk/220 fire/104 kcna/101 japan/66 urge/60 korean/47 move/47 s./46 war/42 policy/39 anti-dprk/39 talk/34 nuclear/33 military/30 blast/28 hostile/26 remark/25 urged/24 korea/24*
508	*kim/510 il/434 jong/358 sung/89 yong/60 nam/54 gift/54 kpa/43 floral/41 work/40 basket/39 president/36 unit/36 anecdote/29 inspect/29 anniversary/26 message/26 congratulatory/24 letter/23 publish/22*
84	*reception/60 ambassador/42 give/40 russian/16 host/15 chinese/8 hosts/7 military/7 dprk/7 iranian/6 performance/6 egyptian/4 embassy/4 palestinian/4 attache/4 general/3 cuban/3 participant/2 libyan/2 guest/2*

TAB. 2 – *Exemple de 5 clusters les plus gros en 2003, titres des articles, lemmes, I=5,0.*

1. *Propagande contre les États-Unis, le Japon et la Corée du Sud*. Si les armes, le feu ou la guerre nucléaire y sont très présents, d'autres sujets peuvent également provoquer l'écriture de ces articles, comme la politique étrangère, l'impérialisme, diverses sanctions pas forcément contre la Corée du Nord, des entraînements militaires, l'espionnage contre la Corée du Nord ou les droits de l'homme ;
2. *Voeux et félicitations à destination de différents pays, présidents, institutions et personnalités* (Fidel Castro, Yasser Arafat, President of Sudan, Slovenian President, President of Serbia and Montenegro, President of Trinidad and Tobago, Iranian President, Tunisian President, Tunisian Prime Minister, Syrian President, Uzbek Foreign Minister, Cyprian Foreign Minister, Russian President, Russian Prime Minister...) ;
3. *Glorification des dirigeants de la Corée du Nord* (Kim Il Sung, Kim Jong Il, Pak Pong Ju et ensuite Kim Jong Un). Il s'agit de messages, de félicitations, de lettres, de cadeaux, de fleurs, de cartes reçus mais aussi des oeuvres de ces dirigeants publiées à l'étranger. Notons qu'en 2003, il y a une série d'articles avec des anecdotes sur Kim Il Sung et Kim Jong Il, alors qu'en 2013 ce sujet est absent ;
4. *Achèvements et avancées du pays et du peuple Nord Coréen* dans différents domaines :
   - la nutrition (emballage sous vide du kimchi, nouveau ferment pour le kimchi, céréales riches en gras, thé nutritif, nouvelles variétés de pommiers, de concombres, etc.),
   - la santé (micromanipulateur biologique, peintures anticeptiques, purificateur de sang, produits contre différentes maladies, vaccins pour les animaux, etc.),
   - l'idéologie (timbres, posters, slogans, ouvrages, sites de propagande),
   - l'éducation (programmes et jeux informatiques, programmes éducatifs, dictionnaires),
   - l'industrie (tannage, station de marée motrice, broderie, brassage de bière...) ;
5. *Relations avec les ambassades et ambassadeurs* accompagnées d'événements sociaux ;

Taille	*Mots-clés*
7 147	*kim/1702 dprk/1624 korean/1369 s./1358 jong/1227 il/1051 korea/634 un/600 u.s./564 sinmun/512 rodong/512 war/437 sung/393 people/366 held/347 day/343 anniversary/342 foreign/296 party/278 organization/268*
610	*kim/302 jong/262 un/262 president/214 greeting/186 message/57 floral/51 congratulation/50 gift/43 receives/42 yong/40 basket/40 party/40 congratulatory/38 nam/37 sends/33 leader/32 political/24 letter/24 pm/21 russian/21*
104	*delegation/96 leaves/59 dprk/45 meets/19 government/13 returns/12 china/10 wpk/10 chinese/9 choe/6 mongolian/6 home/6 back/5 russia/5 visit/5 president/5 spa/5 hae/4 friendship/4 ryong/4*
104	*delegation/50 foreign/39 guests/13 chinese/12 arrives/12 leave/11 party/10 government/9 delegations/8 delegates/7 arrive/7 meet/6 indonesian/5 ministry/5 crewmen/4 president/4 meets/3 delegate/3 iiji/3 praise/3*
74	*exhibition/40 opens/40 national/29 held/20 art/14 contest/10 technological/8 scientific/8 photo/7 championship/5 fine/5 achievements/4 intl/4 sports/4 festival/4 song/3 martial/3 trade/3 pyongyang/3 presentation/3*

TAB. 3 – *Exemple de 5 clusters les plus gros en 2013, titres des articles, lemmes, I=5,0.*

6. *Réunions et rassemblements*, y compris Nord-Sud et y compris entre les scientifiques ;
7. *Réunions officielles*, y compris avec la signature de protocoles, accords et traités ;
8. *Nouvelles du quotidien Rodong Sinmun* (*Journal des Travailleurs*), qui est l'organe officiel du Parti du travail de Corée et le journal le plus lu dans le pays ;
9. *Aide aux fermiers* nord coréens venue de l'étranger et acceptée par les leaders du pays ;
10. *Visites des délégations de la Corée du Nord* en étranger ;
11. *Nouvelles inter-coréennes* (groupes de contact, familles séparées, chemins de fer...) ;
12. *Culture traditionnelle* (chants, légendes...).

Avec plusieurs clusters (par exemple, 2, 3, 5, 6, 7 et 10), il est possible de créer un réseau de pays vus comme amicaux par rapport à la Corée du Nord. En fonction de la symétrie des actions et des honneurs, il serait également possible d'établir une échelle de ces amitiés.

Notons que la plupart de ces thématiques sont bien différenciées au sein de clusters dédiés, même s'il est possible d'avoir des intrus dans ces clusters. D'autres thématiques (visites officielles, réunions, commémorations...) sont distribuées entre plusieurs clusters. En ce qui concerne le paramètre `I`, sa valeur par défaut 5,0 semble être optimale dans la génération non supervisée de clusters. Elle offre des clusters en un nombre raisonnable et assez homogènes quant à leur contenu. Nous pensons cependant qu'un `I` plus grand permet de générer des clusters plus fins et homogènes, mais aussi plus distribués. Ceci sera analysé dans un travail futur.

Cette analyse nous donne un aperçu de thématiques principales abordées par la propagande Nord Coréenne. Ces thématiques sont assez stables entre les deux années analysées. Il serait intéressant de comparer ces thématiques avec les sujets abordés dans la presse occidentale. Par exemple, l'économie semble être un des sujets manquants : elle n'est quasiment jamais mentionnée dans le corpus étudié, alors qu'elle occupe une place prépondérante dans la presse

mondiale. Une des raisons est sans doute que, selon la ligne officielle de la Corée du Nord, l'économie nord coréenne, telle que fondée au début de l'existence du pays, est parfaite. Il n'est donc pas nécessaire de discuter ce système économique ou d'essayer de l'améliorer.

Même s'il est difficile d'avoir un aperçu direct et précis de la politique et de la propagande interne de la Corée du Nord, la traduction de cette propagande par les organes de propagande officielle peut néanmoins donner une idée assez claire des lignes principales et privilégiées suivies. De plus, les articles sont écrits de manière très claire et explicite. Comme déjà indiqué dans d'autres travaux (Kim, 1998; Hachten, 1999), nous pensons également que la traduction locale est plus fiable.

6 Conclusion et Perspectives

Nous avons proposé un travail sur la distinction automatique de grandes thématiques de la propagande nord coréenne. Nous utilisons pour ceci un ensemble d'articles collectés sur un site de propagande locale. Nous abordons cette question de recherche comme le problème de clusterisation non supervisée. Le travail est effectué avec les articles en anglais. D'abord, nous calculons la similarité entre les articles par leur recouvrement lexical et ensuite nous effectuons un clustering. Nous travaillons essentiellement avec les titres des articles. L'évaluation est effectuée manuellement et *a posteriori*, en analysant le contenu des clusters obtenus (les mots-clés les plus fréquents et les titres).

Nos résultats permettent d'émerger les thématiques abordées dans les articles de l'agence de presse KCNA. Parmi les thématiques principales se trouvent par exemple : la propagande contre les États-Unis, le Japon et la Corée du Sud ; les voeux et félicitations à destination de différents pays, présidents, institutions et personnalités ; la glorification des dirigeants de la Corée du Nord ; les achèvements et avancées du pays et du peuple Nord Coréen ; les relations officielles à différents niveaux ; les nouvelles du quotidien Rodong Sinmun ; les nouvelles intercoréennes ; la culture traditionnelle.

Nous avons plusieurs perspectives à ce travail. La mesure de similarité utilisée actuellement est très simpliste : elle calcule seulement le recouvrement lexical. Nous allons exploiter des mesures plus sophistiquées et robustes (Dice, Jaccard, Word2Vec). Nous pensons que cela nous permettra également de raffiner les clusters obtenus actuellement et de travailler sur les articles complets.

Nous traitons actuellement deux années seulement, 2003 et 2013. La méthode proposée pourra être reglée sur cet sous-ensemble et pourra ensuite être appliquée à l'ensemble du corpus (1997 à 2015). Nous pensons obtenir ainsi des clusters plus complets, de faire une comparaison plus complète entre les années et de statuer sur la stabilité des thématiques de la propagande.

D'autres perspectives concernent la comparaison des thématiques abordées dans la propagande nord coréenne avec les sujets abordés dans la presse mondiale. Par ailleurs, il peut être très intéressant de comparer la propagande nord coréenne avec d'autres propagandes, comme par exemple la propagande russe, qui devient de plus en plus offensive et agressive. Comme le montre la figure 2, nous nous attendons à ce que la similarité se manifeste non seulement au niveau textuel, mais également au niveau des représentations.

(a) Kim Il Sung (b) Lénine

FIG. 2 – *Exemple de deux figures politiques en Corée du Nord et en Union Soviétique/Russie*

Remerciements

Ce projet est effectué dans le cadre de l'appel *Projet Partenarial* de la MESHS (Maison Européenne des Sciences de L'homme et de la Société) en Hauts-de-France. Nous remercions également Vincent Claveau et Thierry Hamon pour leurs conseils méthodologiques et le soutien logistique.

Références

Blake, A. et P. Miller (2016). Automated detection of chinese government astroturfers using network and social metadata. *SSRN's eLibrary*.

Bourigault, S., C. Lagnier, S. Lamprier, L. Denoyer, et P. Gallinari (2014). Learning social network embeddings for predicting information diffusion. In ACM (Ed.), *International Conference on Web Search and Data Mining*, New York, NY, USA, pp. 393–402.

Brynielsson, J., A. Horndahl, et F. Johansson (2012). Analysis of weak signals for detecting lone wolf terrorists. In *Intelligence and Security Informatics Conference (EISIC)*.

Chen, H. (2007). Exploring extremism and terrorism on the web : The Dark Web project. *Intelligence and Security Informatics 4430*, 1–20.

Chou, H.-I., G. Y. Tian, et X. Yin (2015). Takeover rumors : Returns and pricing of rumored targets. *International Review of Financial Analysis 41*, 13–27.

Derczynski, L. et K. Bontcheva (2014). Pheme : Veracity in digital social networks. In *Workshop on Interoperable Semantic Annotation (ISA)*.

Fearon, J. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review 88*(3), 577–592.

Fuchs, M. et P.-D. Yu (2015). Rumor source detection for rumor spreading on random increasing trees. *Electron. Commun. Probab 20*(2), 1–12.

Grefenstette, G. et J. Nioche (2000). Estimation of English and non-English language use on the WWW. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, Paris, pp. 237–246.

Habert, B. (1983). Études des formes spécifiques et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979). *MOTS, Presses de la Fondation Nationale des Sciences Politiques* (7), 97–124.

Hachten, W. (1999). *The World News Prism*. Ames, IA : Iowa State University Press.

Haynes, J. (2001). Red journalism as a keyhole : Evaluating discrepancies in news systems and inferring the political direction of North Korea based on a quantitative content analysis of the Korean central news agency website. Technical report, University of North Carolina. Master's Thesis.

Joo, H.-M. (2015). Predicting North Korean military provocations : Document classification analysis of KCNA news. In *ISA Conference 2015*, West Pasadena.

Kim, S. (1998). *North Korean Foreign Relations in the Post-Cold War Era*. New York : Oxford University Press.

Ma, J., W. Gao, Z. Wei, Y. Lu, et K.-F. Wong (2015). Detect rumors using time series of social context information on microblogging websites. In ACM (Ed.), *CIKM'15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Melbourne, Australia, pp. 1751–1754.

Maigrot, C., E. Kijak, et V. Claveau (2016). Médias traditionnels, médias sociaux : caractériser la réinformation. In *TALN 2016*, pp. 1–8.

Malici, A. et J. Malici (2005). The operational codes of Fidel Castro and Kim Il Sung : the last cold warriors ? *Polit. Psychol 26*(3), 387–412.

Myers, B. (2010). *The Cleanest Race : How North Koreans See Themselves and Why It Matters*. Hoboken, NJ : Melville House.

Myers, B. (2015). *North Korea's Juche Myth*. Busan : Sthele Press.

Ohn, D. et M. Richey (2015). China's evolving policy towards the Democratic People's Republic of Korea under the Xi Jinping leadership. *Asian Studies Review 39*(3).

Poneman, B., J. Wit, et R. Galluci (2004). *Going Critical : the First North Korean Nuclear Crisis*. Washington, DC. : Brookings Institution.

Rich, T. (2012). Deciphering North Korea's nuclear rhetoric : An automated content analysis of KCNA news. *Asian Affairs : An American Review 39*(2), 73–89.

Rich, T. (2014a). Introducing the great successor : North Korean english language news coverage of Kim Jong Un 2010-2011. *Communist and Post-Communist Studies 47*, 127–136.

Rich, T. (2014b). Propaganda with purpose : uncovering patterns in North Korean nuclear coverage, 1997-2012. *International Relations of the Asia-Pacific 14*(3), 427–453.

Rich, T. et T. Liu (2012). Reading between the lines : automated content analysis of North Korean nuclear rhetoric. *Rev. Glob. Polit 38*, 157–176.

Richey, M. (2015). Considering DPRK regime collapse : Its probability and possible geopolitical and security consequences. In *Egmont Security Policy Brief*.

Salem, A. (1981). Signalement et inventaire lexical : textes politiques français de 1793. In *Pratique de l'analyse des données*, pp. 183–197. Paris : Dunod.

Salem, A. (1993). De travailleurs à salariés. Repères pour une étude de l'évolution du vocabulaire syndical. *Mots 36*, 74–83.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, Manchester, UK, pp. 44–49.

Sharma, N. (2015). *Facebook journalism : An exploratory study into the news values and role of journalists on Facebook*. Thèse de doctorat, Indiana University, Indiana, USA.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. Thèse de doctorat, University of Utrecht, Utrecht, The Netherlands. http ://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm.

Weeks, J. (2008). Autocratic audience costs : Regime type and signaling resolve. *International Organization 62*, 35–64.

Weiss, J. (2013). Authoritarian signaling, mass audiences, and nationalist protest in china. *International Organization 67*, 1–35.

Westby, T. (2014). *North Korean Nuclear Deterrence : A Myth or a Reality ? An Analysis of North Korean Deterrence Credibility toward the United States and South Korea*. Master thesis, Institutt for Statsvitenskap, Oslo, Norway.

Whang, T., M. Lammbrau, et H. min Joo (2016). Detecting patterns in north korean military provocations : what machine-learning tells us. *Int Relat Asia Pac*.

Zuell, C. (2010). Using computer-assisted text analysis to identify media reported events. In *10th International Conference on Statistical Analysis of Textual Data*.

Summary

Rhetorics used for national and international politics are interlinked, although they do not have the same priorities nor they show the same modes of communication. When these two rhetorics are well used, they allow to maximize the credibility and the coertion potential of the countries, and to signal its opening to the cooperation. The Nord Korean propaganda is particular because it provides agressive rhetorics in its international politics and addresses very specific topics. We propose to distinguish automatically main thematics exploited by the North Korean propaganda using lexical similarity scores between each pair of articles and non-supervised clustering methods. The evaluation of the clusters obtained indicates that between the main thematics we can find for instance: propaganda against United States, Japan and South Korea; greetings and felicitations for different countries, presidents, institutions and persons; honnors received by North Korean leaders; realizations and achievements of the country and people; official relations at different levels; news on and from the Rodong Sinmun journal; inter-korean news; traditionnal culture. We have several directions for future work.