Rendu_Nguyen_Bachey

Analyse syntaxique

Démarche globale :

Nous souhaitons faire une analyse contrastive entre ces sites pour confronter la narrative sur une presse "libre" au vietnam. Est-ce que les seules ressources disponibles prouvent qu'il existe une pluralité idéologique de la presse ? Est-ce que nous retrouvons des formes de discours, des récurrences ? Celles-ci appuient-elles notre travail initial de recherche ?

Notre recherche est au-delà d'une simple analyse syntaxique. En effet, elle tend plutôt à s'inscrire dans une analyse du discours. Celle-ci ne considère pas ce corpus comme un support d'information mais bien comme nous avons tenté de le démontrer dans notre premier devoir est plutôt de considérer ce corpus et chacun des textes qui le composent comme un *texte*.
Ainsi, notre étude cherche à être sémiologique. En effet, notre première partie consistait à définir les "lois qui régissent l'univers raconté" (Claude Bremond, La logique des possibles narratifs)c'est à dire le contexte politique et médiatique de la presse vietnamienne sur internet et dans un second temps, celui du présent devoir qui est "l'analyse techniques de narration" que nous essayons de faire à l'aide d'une analyse du discours.

Le corpus :

Contrairement au premier semestre, nous avons travaillé sur deux corpus différents :

Le premier est constitué de 3 sites vietnamiens (disponible à l'international) : Nhan Dan, Vietnam Plus et Bao Tin Tuc. Nous avons scrap 500 articles par site en utilisant le mot clé "Quoc phong", c'est-à-dire "Défense".

A l'avant dernière séance, Mr. Valette nous a parlé de faire plutôt une comparaison avec un corpus français.
Nous avons donc constitué un nouveau corpus pour avoir des articles à des dates similaires. Ce fût un semi-échec : Bao Tin Tuc n'a pas de version française et Nhan Dan bloque le scraping à 30 URLs sur le site français. Nous n'avons pas eu le temps de trouver une méthode pour contourner ça. Nous sommes donc ressorti avec un corpus déséquilibré : le corpus Nhan Dan est composé de 525 articles vietnamien mais seulement 30 français. Par contre, pour VietnamPlus, tout a marché : nous avons donc 500 articles vietnamiens et (500) articles français. (C'est ce que nous pensions)
En appliquant les outils d'analyses, nous nous sommes rendus compte que les articles étaient soit vides, soit en doublons. Il était trop tard pour tout modifier, nous ne pouvons donc pas proposer une étude vraiment comparative.

Rappel premier semestre :

Au premier semestre, nous avions fait une analyse LDA au corpus constitué des trois sites. Nous avons refait la même chose sur le nouveau corpus.
Pour le corpus VNP, en vietnamien, les 5 mots qui ressortent le plus sont les suivants :

- Điện Biên Phủ : le nom d'une ville
- phát triển : (se) développer

- hợp tác : coopérer, collaborer
- chiến thắng : victorieux
- quốc phòng : défense nationale

en français, ce sont les suivants :

- défense
- Laos
- Dien Bien
- industrie
- sécurité

Pour le corpus ND, les résultats étaient les suivants :

- quốc phòng : défense nationale
- Viet Nam
- hợp tác : coopérer, collaborer
- cơ_quan : établissement
- xây dựng : batir, construire*

et pour le français :

- Vietnam
- Coopération
- Défense
- paix
- Chine

Ces résultats montrent que même en France, les articles sont portés vers la "coopération internationale" pour la "paix" et la "sécurité" du pays.

Quels sont les patterns auxquels nous nous attendons à rencontrer ? Pourquoi cette démarche ?

L'analyse syntaxique consiste à mettre en évidence la structure d'un texte, généralement une phrase écrite dans une langue naturelle.
Dans la phrase syntaxique, que nous considérons comme une unité de sens, nous retrouvons 2 constituants obligatoires : le sujet et le prédicat. Nous pouvons également retrouver un troisième constituant : le complément de phrase.
Pour déterminer le nombre de phrase syntaxique il faut compter le nombre de verbe conjugué. (Premier traitement à faire ?)
Dans la grammaire vietnamienne la construction de la phrase est la même qu'en français.
Nous pouvons aussi regarder les phrases annotées comme passives ?

Quels sont les outils que nous utilisons pour mettre en évidence les patterns syntaxiques ?

Nous utilisons dans un premier temps un annotateur qui nous permet d'annoter syntaxiquement les phrases.
Pour ce faire, nous utilisons l'outil : VnCoreNLP

Une fois que le texte est annoté, il faut réflechir à comment nous pouvons mettre en avant les différentes structures.
Les annotations sont dans un fichier .pdf dans le même dossier que ce document.

Une fois cette annotation réalisée, nous obtenons un fichier txt que nous processons avec différents scripts python qui nous permettent d'extraire des patterns.

Sous les conseils de Monsieur Valette, nous avons tenté d'utiliser TXM. Pour ce faire nous avons convertis nos sorties de fichiers txt (après le traitement avec vncorenlp) en fichier xml. Mais nous n'avons pas réussi à faire en sorte qu'il soit pris en charge par le logiciel. Ainsi, une piste d'amélioration pour notre travail serait de générer un fichier XML qui serait pris en charge par TXM pour faciliter les analyses. En revanche, le compte-rendu de notre travail actuel nous permets dors et déjà de constater des résultats significatifs sur les procédés syntaxiques volontairement (ou non) utilisés dans la presse vietnamienne ont comme influence sur les narratives.

Dans un soucis de traitement égalitaire, nous avons, malgré les outils existant pour le français, choisi de le traiter avec des outils similaires et en recherchant les mêmes patterns.

Premiers résultats

Pour la première recherche, nous cherchons à identifier les sujets et le COD qui leur est associé pour mettre en evidence leur position.
J'essaye dans un premier temps sur quatre urls. Une fois les résultats dans un fichier csv, nous produisons et executons un script qui permet de mettre en valeur les combinaisons qui apparaissent le plus.
Ce test sur quelques urls nous pousse à nous interroger sur la généralisation de ces outils semi-automatiques qui rendent l'analyse pénible et coûteuse. De même, est-ce que nos machines seront capable de supporter un si gros corpus.
Nos craintes se sont confirmés puisque l'analyse de corpus aussi gros empêche de fiables vérifications humaines.

Résultats sur le corpus vietnamien (3 sites)

Nous avons pris les 10 mots les plus fréquents dans le corpus :

ND :
('Thủ_tướng', 252) : premier ministre
('ông', 157) et ('Ông', 99) : titre de respect équivalent à "Monsieur"
('bên', 152) : côté, parti
('nước', 144) : pays
('Đây', 116) : ici, voici
('người', 95) : personne
('Việt_Nam', 90) : Vietnam
('phụ_nữ', 87) : femme
('doanh_nghiệp', 61) : entreprise

VNP :
('Thủ_tướng', 228) : premier ministre
('Chủ_tịch', 217) : président
('Việt_Nam', 210) : Vietnam
('Ông', 190) et ('ông', 170) : titre de respect équivalent à "Monsieur"

('bên', 140) : côté, parti
('nước', 140) : pays
('người', 120) : personne
('Bộ', 107) : ensemble
('cơ_quan', 99) : établissement

BTT :
('ông', 286) et ('Ông', 141) : titre de respect équivalent à "Monsieur"
('người', 226) : personne
('họ', 176) : "ils"
('nước', 128) : pays
('Tổng_thống', 126) : président
('việc', 122) : travail
('Mỹ', 98) : Amérique
('cuộc', 91) : partie
('lực_lượng', 83) : forces

Résultats sur notre corpus (français et vietnamien)

Comme précédemment, nous avons pris les 10 mots les plus fréquents :

ND :
('cơ_quan', 419) : établissement
('hiệu_quả', 120) : effet, efficacité
('Thượng_tướng', 105) : général de corps d'armée
('Thủ_tướng', 105) : premier ministre
('bên', 105) : côté, parti
('ông', 105) : titre de respect équivalent à "Monsieur"
('đơn_vị', 90) : unité, division
('Bộ_trưởng', 90) : ministre
('đại_biểu', 75) : délégué, représentant
('Chủ_tịch', 75) : président

VNP :
('Thủ_tướng', 400) : général de corps d'armée
('lực_lượng', 300) : forces
('Việt_Nam', 275) : VN
('bên', 275) : parti, côté
('Brazil', 250) : brésil
('ông', 225) : titre de respect équivalent à "Monsieur"
('nước', 225) : pays
('Bộ_trưởng', 200) : ministre
('Đại_tướng', 200) : général d'armée
('Đây', 175) : ici, voici

Comme pour l'analyse LDA, des mots qui tournent autour du contexte "défense" ressortent mais plus pour parler de défense de l'armée. Ici, il n'est plus question de coopération ou de paix mais plutôt d'efficacité, de forces.
On remarque que souvent, les deux entités autour du verbe sont des

Pour le peu d'information que nous obtenons sur le corpus français, nous pouvons témoigner de prémice de "tendance", notamment que pour le thème défense, on retrouve souvent des grades. Ce qui peut montrer que souvent les protagonistes sont plutôt des individus ("ministre", "ambassadeur",...) et non des entités même si celles-ci ne sont pourtant pas absentes. Nous retrouvons donc une volonté? d'identifier les protagonistes par des figures plus "humaines". La défense étant une question plutôt de diplomatie, on peut s'interroger sur le sens que pourrait avoir cette utilisation.

Limites de notre étude

Notre travail d'analyse rencontre plusieurs problèmes méthodologiques majeurs qu'il est important de signaler.
Tout d'abord notre corpus n'est pas parfaitement équilibré puisque certains sites sont équipés de pare-feu ce qui nous empêche d'avoir un corpus véritablement représentatif et équilibré. Ainsi les prémices de nos recherches seraient à confirmer en les transférant sur un corpus plus fiable.
Et, étant limité dans l'utilisation de logiciels de textométris, nous avons exclusivement traité la syntage à l'aide de scripts python. Ce qui forcément oriente notre recherche et empêche en un sens de "laisser parler" le corpus.
Malgré cela, nous pensons que notre étude garde une pertinence et qu'elle est encourageante pour des travaux futurs. Ce sujet de recherche mérite plus de temps et de moyens afin de vraiment "creuser" la question. Il nous aurait fallu plus de temps pour reconstituer un corpus comme nous avions pu le faire au premier semestre.

Bibliographie

Communications, 8, 1966. Recherches sémiologiques : L'analyse structurale du récit (https://www.persee.fr/issue/comm_0588-8018_1966_num_8_1)
L'analyse française du discours (https://shs.hal.science/file/index/docid/396398/filename/index.html)