Indexation d'un corpus de littérature web : problématique, méthodologie et usage des mesures vectorielles appliquées au texte littéraire.

Christian Cote

Indexation d'un corpus de littérature web : problématique, méthodologie et usage des mesures vectorielles appliquées au texte littéraire.

COTE Christian

MARGE

Christian.cote@univ-lyon3.fr

Introduction.[1]

Notre objectif est l'élaboration d'une méthodologie de structuration d'un ensemble massif de données, expérimentée ici sur un corpus archive de la littérature web francophone. Cet ensemble de données est relativement hétérogène : le texte littéraire web est une production originale et particulièrement diverse. Du fait également d'un référencement limité et aléatoire, les méthodes d'indexation partant de la description bibliographique sont inapplicables.

Cette expérimentation vise à renouveler les méthodes d'indexation, en proposant un ensemble d'opérations automatisables pour identifier les traits signifiants de ces productions et à les organiser afin de structurer le corpus et donc la navigation des utilisateurs. On n'élabore pas prioritairement un **outil de recherche mais un outil de structuration et de guidage** par proximité ou opposition entre les textes et non sur une recherche documentaire par mots-clés. L'appareil descriptif n'a pas à être nécessairement visible ni à reposer sur des dénominations classificatoires.

L'analyse que l'on propose est inductive puisque l'on part d'observations à partir de calculs, comptages et opérations. Nous privilégions une approche par les données[1], où les données pour l'analyse sont produites dans des cadres théoriques (linguistiques, mathématiques) à partir des données textuelles. Néanmoins, ces données pour l'analyse permettent de construire des catégories d'index, et c'est ce que l'on présente maintenant.

Par ailleurs, concernant les textes littéraires, il serait vain de vouloir nécessairement construire un système de classification, un thésaurus ou une ontologie, d'autant plus que le corpus sur lequel cette structuration s'appuie est constitué de textes originaux et hétérogènes. Par conséquent, en se référant à un usage relevant de la navigation, voire du jeu vidéo, on proposera une représentation sous forme de treillis, en reprenant les principes des FCA. (En dehors de la RI, les FCA sont utilisées dans le cadre de la recherche de données dans des corpus issus des médias sociaux et plus généralement du web[2]). Les concepts des FCA intégrent une dimension extensionnelle (ce qui les distingue de leur définition dans les ontologies), de façon à relier les objets par des communautés et des distinctions de propriétés. La structure a la forme d'un treillis où les textes sont les objets, les phénomènes identifiés des propriétés, et tout objet a des propriétés distinctes.

---

[1] Cet article s'inscrit dans l'ANR LIFRANUM https://anr.fr/Projet-ANR-19-CE38-0007 , collaboration MARGE/Lyon3, ERIC/Lyon2 et BNF. Pour ce travail, nous remercions particulièrement Enzo Terreau et Julien Velcin du Laboratoire ERIC. L'ensemble des données, qu'il s'agisse du corpus ou des résultats d'analyse, est disponible sur l'espace partagé du projet.

L'analyse est manuelle et est contrôlée par la représentation de phénomènes pertinents pour les études littéraires. Les observations sont formalisables et visent à être automatisables et reproductibles.

Objet d'expérimentation et contraintes

Notre corpus de travail est une archive de la littérature web francophone, acquis par un crawl et constitué d'un ensemble de fichiers WARC[3] reproduisant le contenu HTML de chaque URL crawlée. Du fait des contraintes des fichiers WARC, l'unité traitée est le texte : les fichiers WARC enregistrent les données par URL et à chaque changement de caractère dans une URL correspond un fichier distinct. Les fichiers WARC captent les données HTML des pages sans prendre en charge les APIs particulières des sites et des blogs. Tous les fichiers sont de format identique sans structuration.

Nous distinguons le corpus et l'archive par des usages différents : le corpus est une « donnée primaire de la science » et l'archive un patrimoine accessible. Le corpus se définit par la capacité qu'a une grande quantité de données à permettre d'observer des phénomènes réguliers derrière des variations de surface[4]. Il s'ensuit la détection d'évolutions et l'identification de régularités qui n'apparaissent qu'avec des données massives et des outils quantitatifs. Le corpus est composé des œuvres de 1000 écrivains environ (avant crawl), la plupart inconnus[5].

Le texte littéraire manifeste un lien au réel différent de l'article scientifique ou journalistique : l'ambigüité et la pluralité des interprétations constituent un fondement du texte littéraire. Il ne se prête pas à une analyse à fondement référentiel et terminologique : les approches lexicales et thématiques utilisant les méthodes quantitatives pour déceler l'immanence des textes ne sont guère utiles pour l'indexation. Depuis la création des premiers corpus numériques massifs, le domaine de la littérature a été largement investi par les analyses quantitatives[6] mais aucune n'a de visée d'indexation : elles saisissent des particularités des textes qui ne peuvent être identifiées par une lecture humaine. On ne cherche pas à spécifier tel ou tel texte, mais à identifier des indices permettant d'assembler des textes et de les distinguer les uns des autres. L'indexation ne peut guère s'appuyer sur des organisations de connaissance spécifiques à la littérature : d'une part celles-ci sont lacunaires du fait de la difficile catégorisation des objets littéraires, et d'autre part cette littérature est profondément originale par rapport aux formes canoniques. Enfin, l'analyse littéraire se caractérise par le fait que les concepts utilisés, et qui peuvent provenir de différents champs disciplinaires, sont utilisés non pour catégoriser des textes mais comme outils d'analyse.

Nous suivons l'hypothèse selon laquelle la structure du texte permet de le catégoriser. Cette approche fondée sur le contexte[7] stipule que ce sont des marqueurs de structuration (énonciation distribution, etc.) qui construisent les propriétés du texte. En effet, ce qui rapproche et distingue des textes, ce n'est pas un imaginaire mais deux phénomènes concrets : des choix de d'organisation et de communication, d'autre part des choix d'usage de la langue.

Deux approches différentes permettent de saisir les dimensions discursives et textuelles relatives à ces choix à partir d'analyses automatique des données textuelles :

Les approches supervisées choisissent des unités discursives et les règles associées : on indique à la machine les entités discursives auxquelles elle doit s'intéresser et les règles qu'elle doit appliquer. Dans notre cas les propriétés pragmatiques des connecteurs et l'expression d'états mentaux représentés par les verbes identifient la structuration et la **visée communicationnelle d'un texte.**

Pour les **approches non-supervisées,** l'outil réalise des mesures sans contraintes et limitations liées à une théorie de la structure de la langue ; l'analyse permet de caractériser l'emploi de la langue dans le texte à l'aide de **l'ensemble des mesures effectuables sur un texte.** Là encore, on caractérise des choix d'expression.

Méthodologie d'analyse

Les mesures issues des différentes analyses ne sont considérées que comme des données. Elles deviennent signifiantes par une analyse qui spécifie leur portée. La méthode adoptée avec l'équipe ERIC-Enzo Terreau et Julien Velcin- est la suivante : ERIC réalise les calculs et nous analysons les données. Les mesures effectuées sont éditées sous forme de tableaux EXCEL et JSON à partir d'un sous-corpus de blogs : comme il est impossible de travailler directement avec des fichiers WARC, ERIC a élaboré un fac-similé d'une partie de notre corpus pour effectuer une capture utilisant les APIs de BLOGGER et WORDPRESS, soit 2722 textes).

Le choix d'analyse est fondé sur des problématiques littéraires :

- La question des genres littéraires traduite dans celle des structures de discours.
- Le lyrisme, l'éthos et les passions intégrés dans l'analyse de certains verbes.
- Les questions de poétique traitées par l'analyse non supervisée.

Ces choix constituent des entrées pour explorer les données : nous ne produisons pas des résultats d'analyse littéraire mais des accès pour une approche littéraire du corpus.

Approches des mesures et des traits.

Nous cherchons à établir des phénomènes spécifiques aux textes et dont la mise en évidence requiert la corrélation de données. Ces corrélations ne sont pas liées à une théorie quelconque mais sont produites par l'observation :

- Les mesures constituent des données sur lesquelles on observe des traits par poids, corrélation, opposition et convergence.
- Ces traits caractérisent des phénomènes textuels ou discursifs permettant d'attribuer des propriétés caractérisant partiellement le texte.
- Un type de texte classe un ensemble de textes ayant suffisamment de propriétés communes pour être considérés comme équivalents.

Les traits représentent des phénomènes[8]qui ne relevent pas systématiquement d'une discipline mais d'un niveau d'observation, lui-même dépendant d'une ontologie, à savoir une organisation élémentaire des discours et des textes en tant qu'ils possèdent une visée communicationnelle et

opèrent des choix dans l'usage de la langue. Les traits sont des indices de choix opérés par l'auteur dans l'élaboration de son texte.

Structuration des traits par corrélation.

Une mesure est considérée comme un indice, et c'est sa corrélation avec d'autres qui construit le trait. La convergence des indices donne un certain poids au trait et donc atteste le phénomène. Les **marques d'observation, compte tenu de la longueur des textes, sont les suivants :**

- Isolation : une entité particulièrement saillante par rapport aux entités comparables.
- Corrélations d'entités par double et triples.
- Absences de poids significatifs de marqueurs/ Faible présence de marqueurs.

Notre démarche est inductive, elle ne présuppose pas des catégories classificatoires préalables : on élabore des propriétés à partir des traits identifiés sur les textes. Les marques déterminent à partir de quelle quantité (ou poids, ou ratio) une donnée est signifiante ; les appartenances relatives sont fixées par des corrélations. Par exemple, une entité sera saillante tant qu'elle ne sera pas de même quantité qu'une autre de même unité de mesure, une corrélation d'entités sera pertinente tant qu'elle ne sera pas rejointe par une autre de même unité, etc.

Traits communicationnels

Les grammaires électroniques des connecteurs et des verbes sont fondées sur des théories du discours. Les connecteurs comme les verbes sont des ensembles d'entités relativement limitées par rapport aux entités nominales et sont structurants du discours parce qu'ils organisent les clauses et les lient dans la construction de la portée illocutoire des discours[9] : ils marquent à la fois leur structuration (leur cohérence et leur cohésion) et leur visée communicationnelle. Les verbes jouent ce rôle dans la limite de la structure de la clause[10]. Les analyses de Crible[11] montrent une cohérence entre l'usage des connecteurs et celui des verbes. Bien que les grammaires de Fillmore[12] et les SDRT, qui fondent respectivement VERBNET et LEXCONN ne décrivent pas les mêmes dimensions linguistiques (d'une part sémantique, de l'autre pragmatique), la corrélation des deux analyses reste néanmoins pertinente.

Les théories linguistiques qui fondent les analyses de connecteurs entraînent des segmentations et des distinctions importantes : les catégories classant les connecteurs changent entre les différentes grammaires[13]. Néanmoins, ces dénominations d'usage de connecteurs reposent sur des modèles différents de la prédication et ne permettent pas d'inférer des types de textes : un connecteur de narrativité ne permet pas d'inférer que le texte est un récit. Par ailleurs, notre méthodologie repose sur des corrélations, et nous n'inférons pas des types de textes à partir des dénominations de connecteurs ou de verbes.

Structuration par les connecteurs. Prototype et textes spécifiques.

Le principe des prototypes consiste à structurer ces oppositions à partir d'un comportement standard, lequel serait ensuite spécifié par des écarts relativement à un exemplaire-type. La théorie du prototype[14] classe par une différenciation progressive avec le prototype, marquée par l'affaiblissement de traits prototypiques et l'apparition d'autres.

Les productions littéraires web se caractérisent par des marques de distinction par rapport aux autres types de discours et ainsi valident l'hypothèse d'une écriture web spécifique. Ainsi, un certain nombre de connecteurs (« concession », « digression », « rephrasing », « summary », « detachment ») sont quasiment absents voire complètement pour le dernier. (Ce sont essentiellement les marqueurs de reformulation : le discours du web littéraire se distingue ainsi facilement des discours scientifique, journalistique, etc.).

Certains connecteurs sont répartis sur l'ensemble des textes et donc leur fréquence et leur présence sont significatives de la littérature web, à la différence des résultats obtenus dans la littérature générale sur le corpus FRANTEXT[2] : [explanation] (« après quoi, ensuite, attendu que, au bout du compte, finalement, au cas où, des fois que, au contraire … ») [goal] (« afin de, pour, afin que, pour que, ainsi, alors, donc, alors même, tandis que, après, … ») et [continuation] (« finalement, au bout du compte, au cas où, des fois que, au contraire, à propos, au fait, … »). CONTINUATION/EXPLANATION structure le texte prototypique, défini comme un modèle de la dynamique textuelle, un enchaînement de situations. Ensuite, on voit apparaître un nombre important de textes marqués par le prototype et GOAL : CONTINUATION/GOAL, CONTNUATION/EXPLANATION/GOAL et EXPLANATION/GOAL.

Lorsque l'on s'éloigne du prototype, d'autres paires et triples de connecteurs réguliers apparaissent : les marqueurs de **temporalité** d'une part, les marqueurs de **dialogue** d'autre part. Enfin, d'autres textes contiennent les connecteurs distinctifs précédents, mais surtout d'autres types de connecteurs : **indexicalité, reformulation méta discursive, structuration complexe**.

Si l'on envisage le texte prototypique comme une sorte de description dynamique, alors l'ajout d'une temporalité et de relations logiques (argumentatives) apparait comme un écart par rapport au prototype. Cette analyse est corroborée par le fait que les textes comportant des connecteurs appartenant aux deux groupes distinctifs du prototype et aux trois marqueurs de spécification sont particulièrement rares : les catégories connecteurs sont structurantes.

Le schéma suivant représente cette structuration du corpus par les connecteurs :

[2] Analysé par C.Rose, L. Danlos et P. Miller, op. cit.)

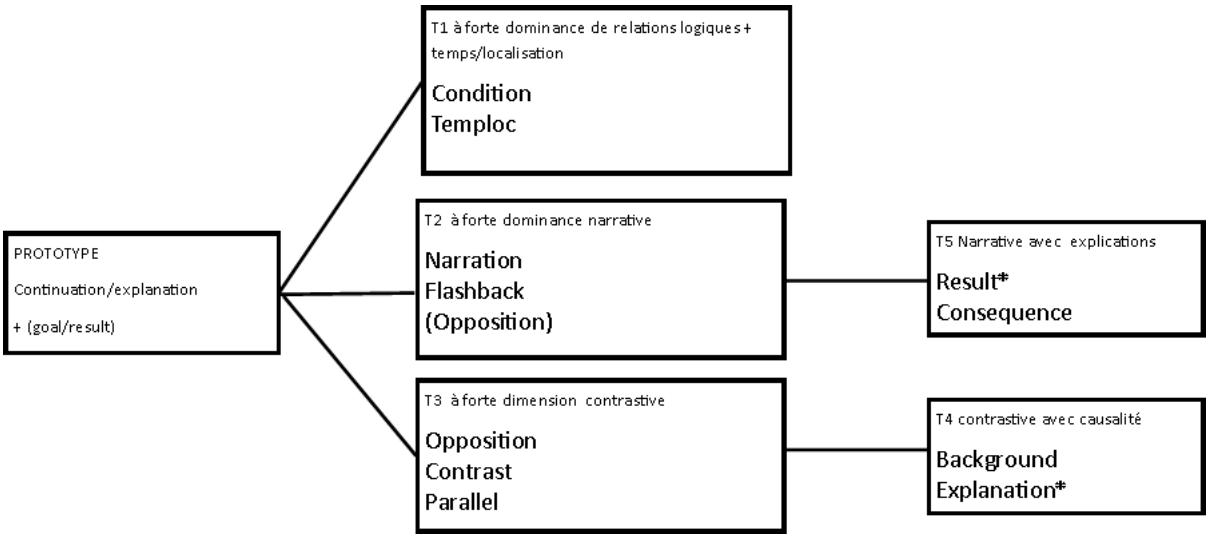

Schéma 1. Modèle des types de textes en fonction des connecteurs.

Ainsi, on obtient cinq types de textes distincts du prototype et marqués par un éloignement progressif par rapport à lui. Les pavés représentent les connecteurs formant un certain type (mais pouvant être de catégories différentes dans LEXCONN) et les traits des fréquences d'usage commun de connecteurs de différents types.

Stratégie de structuration par les verbes.

VERBNET[15] est structuré en 47 classes de bases, 193 au seconds et troisièmes niveaux. Ces classes sont fondées sur le comportement syntaxique et sémantique du verbe à partir de la langue anglaise et adaptées pour le français. Nous traitons uniquement des verbes qui indiquent un statut du texte et mettant en évidence une activité cognitive, marquée par des émotions, mais aussi l'engagement ou la représentation d'interactions, de l'activité créatrice, mentale ou communicationnelle.

Le but de l'étude de ces types de verbe consiste à approcher les notions de lyrisme, d'éthos et de passions[3] qui caractérisent certains types de textes littéraires. Ce domaine de travail est aussi celui des états mentaux représentés dans les textes.

Nous opérons avant l'analyse des données une sélection des verbes qui représentent les activités concernées dans le cadre des catégories de VERBNET. Cette sélection initiale a été testée sur le corpus et a mis en évidence dans les textes des groupements de verbes appartenant à des catégories distinctes. Certains verbes ont pu apparaitre dominants et d'autres beaucoup plus rares, constituant un appoint à ces catégories. Nous avons ainsi proposé quatre catégories d'activités cognitives à partir des corrélations de classes de verbes observables (ces dernières sont indiquées entre crochets et reprennent les dénominations anglaises) :

[3] Nous reprenons les définitions communes de ces concepts issus de la rhétorique classique.

- **Activité émotionnelle** : [amuse : engagement], [admire : affection + négatif], [appeal : apprécier, déprécier]
- **Activité mentale** (compléments prédicatifs, états mentaux) : [characterize : métadiscours à propos d'une œuvre] [conjecture : métadiscours sur la connaissance] [consider : évaluation d'un objet]
- **Activité communicationnelle** : [transfert-message : commentaire d'un texte], [send, hire, give : échanges]
- **Activité créatrice** : [SCRIBBLE, BUILD, ENGENDER : transformation], [indicate, discover : présentation, monstration]

Corrélation des types de textes définis par les connecteurs et par les verbes.

L'observation des corrélations entre les structurations de discours fondées sur les connecteurs et les expressions d'activités cognitives associées au comportement des verbes montre une concentration sur le T3, qui donc emploie le plus de verbes d'activité cognitive (quelle qu'elle soit). Le T5 se marquent également par une importante activité émotionnelle et mentale. L'expression émotionnelle apparait comme un attribut important à un certain type de textes (T3 surtout, T4 et T5), et très peu à d'autres (prototype, T1 et T2).

Sur l'ensemble du corpus, les différents types de discours n'accueillent donc pas de la même façon les marques d'engagement cognitif. Par extrapolation, les marques de lyrisme, les revendications d'activité créative ne sont pas nécessairement attestées dans tous les types de textes. L'inégale répartition des verbes sélectionnés sur les différents types de textes montre une certaine corrélation entre types de discours et engagement cognitif et émotionnel, ce qui servira à marquer des proximités dans le treillis entre les résultats des deux sortes d'analyse.

Traits de choix esthétiques

Le traitement non-supervisé repose sur un modèle inverse du précédent : aucune théorie linguistique ne vient structurer les données d'observation. Le cadre théorique de cette construction de données est mathématique.

Modèle de traitement automatique des textes

L'outil de calcul s'inscrit dans une perspective de traitement des unités textuelles reposant sur des traitements massifs de données. Ces méthodes quantitatives traitent d'abord les unités linguistiques comme parties du discours. La grammaire est ensuite construite par apprentissage progressif, de façon à saisir les régularités des textes. L'ensemble de ces modèles fonctionne par vectorialisation, à savoir la traduction des entités lexicales et grammaticales dans des données numériques. Parmi ces modèles, initiés par WORD2VEC puis DOC2VEC, les outils BERT[16] apportent les possibilités de traitement les plus complètes.

Segmentation du corpus par la longueur des textes.

On sait que la longueur des textes a un impact sur les mesures de fréquences, les répétitions et les hapax notamment. Il est donc nécessaire de segmenter le corpus, et nous choisissons une segmentation fondée sur des critères rythmiques (total des mots courts, ponctuation et longueur de phrases) parce qu'ils caractérisent la dimension textuelle sur des critères formels communs à la totalité des textes. Cette régularité s'observe également avec les espacements et les adverbes de positionnement mais pas avec d'autres marqueurs.

Les marques rythmiques isolent 11 segments :

Longueur des textes (mots)	**Corrélation ponctuation, longueur de phrases, total mots courts**	**Longueur des textes (mots)**	**Corrélation ponctuation, longueur de phrases, total mots courts**
0 - 540	Court	1197 - 1552	Long
541 - 764	Long	1553 - 1692	Court
765 - 954	Court	1693 - 1833	Long
955 - 1087	Long	1834 - 2351	Court
1088 - 1196	Court	2352 – 2410	Long
		2411 - 2742	Court

Les segmentations restent quelque peu sommaires vis-à-vis de la diversité des situations mais constituent des segments formellement cohérents pour gérer le problème posé par la longueur des textes.

Caractérisation des marqueurs généraux de choix expressifs.

Les modèles BERT traitent les données en calculant des fréquences, des constances, mais également des indices (Simpson, lisibilité) et d'autres types de calculs comme l'entropie. En dehors de ces opérations sur la totalité du texte, des comptages, comme les hapax (mots identifiés une seule fois) et la dislégoménie (mots identifiés deux fois), caractérisent des comportements spécifiques d'entités lexicales dans le texte. La liste des calculs est déterminée par les capacités à représenter mathématiquement un phénomène (l'information, la lisibilité, etc.) et à l'appliquer dans un espace vectoriel.

Méthodologiquement, ces résultats constituent des données d'observation analysées selon les mêmes modalités que précédemment.

Analyse plus précise sur les textes « médians ».

Sur les textes « relativement longs », [1088-1833], nous avons identifié les phénomènes suivants :

Les marques de diversité et étendue lexicales caractérisent la pluralité, la flexibilité et la variabilité du vocabulaire utilisé dans un texte. Une faible probabilité de répétition et des hapax importants caractérisent une forte diversité lexicale. Inversement, une forte probabilité de répétition et phénomène DIS marqués indiquent une faible diversité lexicale.

Les marqueurs de fréquence d'usage indiquent si le texte a recours aux mêmes mots ou pas dans son déroulement. La fréquence d'usage, notamment des mots fonctionnels, est distincte de la diversité des vocabulaires. Mais ce comptage ne prend pas en compte la répartition des

unités dans les textes. Les fréquences de mots sont corrélées aux hapax : plus il y a de fréquence, moins il y a d'hapax et plus de DIS et inversement.

Marqueurs de ruptures et de continuité. L'un caractérise la capacité du texte à renouveler son information (l'entropie), l'autre au contraire à marquer des constances (la mesure de Yule). Les constances calculent la façon dont les phénomènes apparus sur une partie du texte se reproduisent sur la totalité du texte : elles mesurent l'homogénéité des textes, quelle que soit par ailleurs leur richesse lexicale et grammaticale. C'est donc un marqueur de continuité de forme. Lorsque les constances et les probabilités d'apparition d'entités de la même espèce sont corrélées, elles montrent une homogénéité des textes, et se distinguent d'une faible entropie, qui marque des ruptures informationnelles. Ainsi, si l'entropie est faible (forte informativité) et les hapax importants, on a affaire à un texte qui porte une forte diversité dans son déroulement.

Nous présentons maintenant le tableau synthétisant les traits précédents :

Diversité lexicale	Ruptures dans le texte				Homogénéité lexicale
HAPAX	Entropie	Constances (Yule)	Probabilités d'apparition (Simpson)	Fréquence	DIS
+	-	-	-	-	-
-	+	+	+	+	+

Ainsi, un texte qui a une forte fréquence et constance sera marqué par une **continuité** ou **monotonie** importantes. (Cette observation converge relativement avec l'entropie). A contrario, un texte qui a une faible fréquence et constance sera marqué par un **dynamisme** important. Un texte à forte constance et faible fréquence est un texte homogène avec une forte diversité thématique.

A une **faible fréquence** (donc un vocabulaire peu répétitif) correspondent de **très fortes constances** de Yules et de probabilités de Simpson (ou **dispersion**). Néanmoins, l'inverse n'est pas vrai : une forte fréquence n'implique pas de faibles constances ou probabilités. Par ailleurs, de **faibles probabilités de dispersion** entraînent de **faibles fréquences** et de plus **fortes constances**. Enfin, une très **faible constance** entraîne une plus **forte dispersion** et une très **forte fréquence.** On obtient la représentation suivante :

	Fréquences	Constances	Probabilités de dispersion
Fréquences (-)		+	+
Probabilités de dispersion (-)	-	+/-	
Constances (-)	+		+
Fréquences (+)		-	+ /-

De fortes fréquences lexicales montrent peu de constantes textuelles. La cohérence du texte repose dans ce cas sur le lexique et non sur les constantes. On oppose donc des textes dont le vocabulaire est diversifié à ceux où il est concentré sur un nombre d'entités limitées.

Si l'on observe maintenant la probabilité de dispersion par rapport à l'entropie, une **faible entropie entraîne une très forte dispersion** et inversement et une très **faible dispersion une très faible entropie**. Par contre, pour les textes à forte entropie, ou faible renouvellement informationnel, la dispersion est relativement faible : la découverte **de contenus distincts y est très limitée.**

Enfin, la faible probabilité de dispersion caractérise une **diversité importante du texte** (une faible répétition des structures et constituants) et une **faible entropie** : le texte possède une pluralité de formes tout au long de son déroulement.

Les corrélations entre entropie et dispersion sont résumées ici :

	Entropie	Probabilité de dispersion
Entropie -		+
Probabilité de dispersion -	-	
Entropie +		+/-
Probabilité de dispersion +	-	

Indices et propriétés.

Le tableau suivant récapitule sur quels indices (ensemble des calculs et mesures représentées sur des échelles) sont fondées les propriétés. Ces propriétés se marquent par des dualités mais recouvrent des situations plus diverses liées à la précision des échelles de mesure. On ajoute les marques de rythme associées précédemment à la segmentation du corpus :

Propriétés	**Indices**
Linéarité VS dynamique	Hapax VS dis + ensemble des marqueurs
Diversité lexicale VS homogénéité lexicale	Fréquence, constances, dispersion de Simpson
Monotonie VS renouvellement	Dispersion de Simpson et entropie
Rythme	(avg-s-lengh, ponctuation, total des mots courts, fréquence des mots fonctionnels, comptage des syllabes, pronoms et autres anaphoriques)

Amorce d'analyse des catégories linguistiques.

Nous intégrons ici les résultats de l'analyse par catégorie linguistique (emplois des verbes, noms, adjectifs, etc.). Nous identifions des textes où des choix de catégories grammaticales sont marqués ou pas avec des combinaisons possibles :

N+ : dénomination	V+ : événements	ADJ+ : évaluation
ADV + : modalité	Mots fonctionnels + : structure discursive.	

On peut établir une cartographie des textes qui :

- Concentrent les mots fonctionnels avec peu de noms et d'adjectifs.
- De nombreux noms et peu d'adjectifs.
- Beaucoup d'adjectifs et très peu d'adverbes et de mots fonctionnels (fréquence moyenne de noms et verbes).

- Beaucoup d'adverbes et de verbes, beaucoup de noms et moins d'adjectifs et de mots fonctionnels.

Cette partie, que nous faisons figurer dans le modèle du treillis, sera développée ultérieurement.

Des traits à leur représentation.

Nous présentons maintenant certaines contraintes et pistes à l'élaboration d'un modèle d'indexation à partir des analyses. Celles-ci produisent des traits traduits sous formes de propriétés : tout d'abord des questions de dénominations se posent puis nous présenterons le schéma général du treillis et la forme de la recherche d'information qui peut y être associée.

Questions de dénomination.

Le premier problème qui se pose à la dénomination est celle de l'originalité des textes : utiliser des catégories linguistiques ou des modèles de calcul pour identifier et caractériser des phénomènes relatifs au texte littéraire constitue un problème majeur. Le deuxième problème est lié à l'originalité des productions littéraires, hétérogènes aux catégories instituées des genres littéraires. Pour cette raison aussi, le recours aux experts est une option difficile à mettre en œuvre. La grammaire des propriétés et donc la justification des dénominations reste encore largement à élaborer.

Limites aux comparaisons de résultats.

Les mesures non supervisées identifient une catégorie sur la base des part-of-speech répertoriées dans https://universaldependencies.org/u/pos/index.html. Ces catégories ne prennent pas en compte la structuration interne des catégories linguistiques proposée par les théories utilisées par LEXCONN et VERBNET. Par conséquent, les propriétés discursives et textuelles restent différenciées sur l'ensemble de l'indexation. De plus, les dénominations de ces propriétés restent limitées aux phénomènes discursifs et textuels qu'ils indiquent et donc ne peuvent se confondre avec une classification de texte.

Représentation générale des traits de la littérature numérique.

Les propriétés sont des assemblages de traits réguliers observables sur les 2722 textes du sous-corpus ERIC. Nous synthétisons dans le tableau suivant les quatre ensembles de propriétés que nous avons dégagés.

Structure discursive
PROTOTYPE Continuation/explanation + (goal/result)
T1 à forte dominance de relations logiques + temps/localisation (Condition, Temploc, Concession)
T2 à forte dominance narrative (Narration, Flashback, (Opposition))
T3 à forte dimension contrastive (Opposition, Contrast, Parallel)
T4 contrastive avec causalité (Background Explanation*)
T5 Narrative avec explications (Result*, Consequence)

Marques expressives
Activité émotionnelle : [amuse : engagement], [admire : affection + négatif], [appeal : apprécier, déprécier]
Activité mentale (compléments prédicatifs, états mentaux) : [characterize : métadiscours à propos d'une œuvre] [conjecture : métadiscours sur la connaissance (ou son contraire)] [consider : évaluation d'un objet]
Activité communicationnelle : [transfert-message : commentaire d'un texte], [send, hire, give : échanges]
Activité créatrice : [SCRIBBLE, BUILD, ENGENDER : transformation], [indicate, discover : présentation, monstration]

Choix textuels
Linéarité VS dynamique
Diversité lexicale VS homogénéité lexicale
Monotonie VS renouvellement
Rythme

Choix de catégories (+)	Choix de catégories (-)
Fonctionnels	Noms, adj
Noms	Adj
Adj	Adv, fonctionnels
Adv, verbes	Noms, adj

Dans le schéma suivant, nous synthétisons les propositions précédentes par un modèle prenant en compte la segmentation des textes par leur longueur.

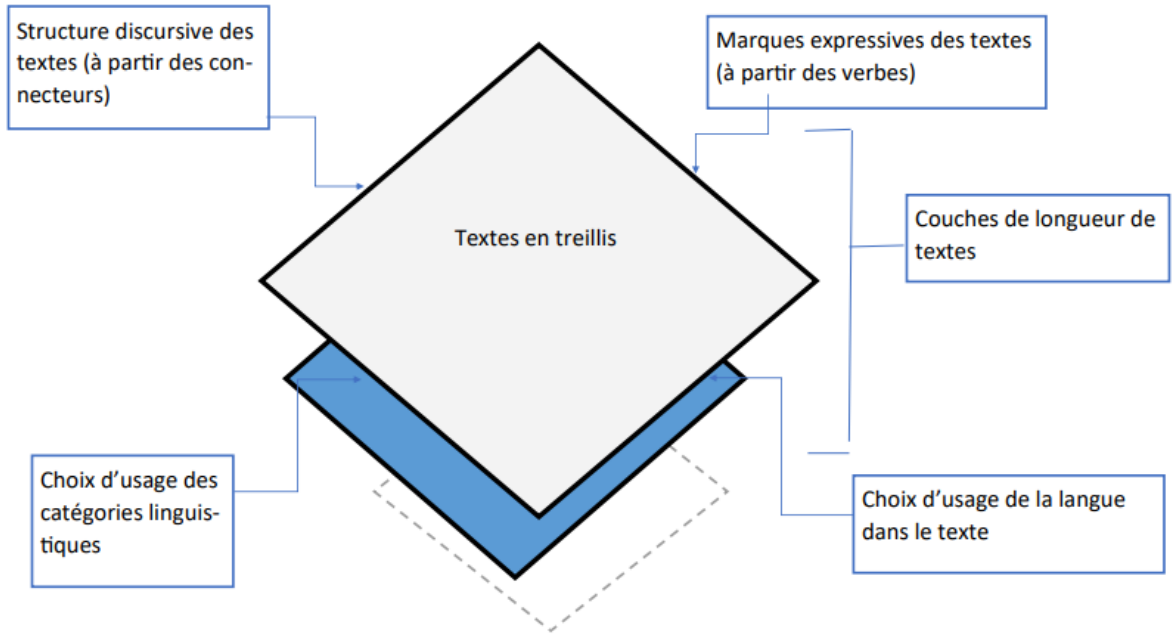

Schéma 2. Modèle général de l'indexation du corpus.

Dans les FCA, un concept formel est une paire associant des objets en extension et des propriétés en intension. Chaque texte est caractérisé par un volume et s'insère dans un segment de longueur, puis les propriétés lui sont affectées sous forme d'index inversé. Chacune des quatre sortes d'analyse constitue une propriété générique qui se décompose dans les différentes propriétés identifiées, auxquelles les textes sont indexés. Chaque texte est lié à d'autres par des propriétés communes, chacune étant relative à un type de choix différent. Les classes de textes sont caractérisées par des propriétés partagées qui permettent de visualiser des proximités.

Les mesures obtenues par les outils numériques sont reproductibles et fonctionnent pour tous les textes du corpus. Par conséquent, on peut envisager un workflow depuis les textes vers leur représentation en définissant exactement les règles par lesquelles des mesures deviennent des traits qui sont traduits dans des propriétés.

Dans le cadre de la modélisation, les règles d'obtention des traits seront formalisées et les propriétés obtenues intégrées dans les objets sous forme de métadonnées associées aux fichiers WARC.

Usages des treillis en IR.

La navigation dans un tel corpus ne peut être assimilée à une recherche d'information classique. Le mode de lecture de la littérature web est proche du jeu vidéo[17], ce qui accentue la recherche par navigation dans le corpus. La requête n'a pas à être formulée nécessairement à partir des propriétés et peut se faire directement à partir des extensions et par proximités. C'est cette approche que nous privilégions : l'usager littéraire s'intéresse d'abord aux textes…

Conclusion.

Nous avons présenté un workflow permettant d'indexer un ensemble massif d'informations hétérogènes et développé une démarche pour obtenir des propriétés de textes qui soient des entrées pour aborder le texte littéraire web. Il explore l'indexation de corpus massifs en articulant méthodes manuelles et automatiques alors que jusqu'à présent seuls des outils automatiques sont utilisés[18], ce qui, pour notre type de corpus et concernant des problématiques littéraires, ne pouvait être pertinent. Le modèle formel du processus est en cours de développement de même que la représentation précise du treillis. Enfin, nous nous sommes limités à des questions d'indexation et nous n'avons pas engagé d'analyses littéraires et linguistiques du corpus, même si de nombreuses pistes émergent tout au long du travail présenté ici.

---

[1] Glaser, B.G., & Strauss, A.L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago, IL : Aldine.

[2] Andrews, S., Gibson, H., Domdouzis, K., & Akhgar, B. (2016). Creating corroborated crisis reports from social media data through formal concept analysis. Journal of Intelligent Information Systems, 47, 287-312.
Andrews, S., Brewster, B., & Day, T. (2018). Organised crime and social media: a system for detecting, corroborating and visualising weak signals of organised crime online. Security Informatics, 7(1), 1-21.
Djouadi, Y. (2012). Généralisation des opérateurs de dérivation de Galois en recherche d'information basée sur l'analyse formelle de concepts. In CORIA (pp. 373-386).
Codocedo, V., & Napoli, A. (2015). Formal concept analysis and information retrieval–a survey. In Formal Concept Analysis: 13th International Conference, ICFCA 2015, Nerja, Spain, June 23-26, 2015, Proceedings 13 (pp. 61-77). Springer International Publishing.

[3] https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

[4] Teubert, W. (2005). My version of corpus linguistics. International journal of corpus linguistics, 10(1), 1-13.

[5] COTE Christian, Méthodologie pour l'élaboration d'un corpus et d'une archive du web littéraire francophone, in *Colloque international – Le web : source et archive* (3,4 et 5 avril 2023, Univ. de Lille), https://respadon.sciencesconf.org/resource/page/id/9, à paraitre dans « Les cahiers du numérique ».

[6] https://litlab.stanford.edu/

[7] Svenonius, E. (2004). The epistemological foundations of knowledge representations.

[8] Gnoli, C. (2008). Categories and facets in integrative levels. *Axiomathes*, *18*(2), 177-192.

[9] Jean-Jacques Franckel, « De l'énonciation à la méta-énonciation », Corela [En ligne], HS-31 | 2020, mis en ligne le 02 juillet 2020, consulté le 03 juillet 2020. URL : http://journals.openedition.org/corela/11607 ; DOI : https://doi.org/10.4000/corela.11607
Elizaveta Khachaturyan, « Marqueurs discursifs du dire (français, russe, norvégien) », Corela [En ligne], HS-31 | 2020, mis en ligne le 05 juin 2020, consulté le 03 juillet 2020. URL : http://journals.openedition.org/corela/11158 ; DOI : https://doi.org/10.4000/corela.11158
[10] Charlotte Roze, Laurence Danlos, Philippe Muller. LEXCONN: a French lexicon of discourse connectives (2012). *Discours - Revue de linguistique, psycholinguistique et informatique*, Laboratoire LATTICE, 2012, Multidisciplinary Perspectives on Signalling Text Organisation, pp.1-15. ⟨10.4000/discours.8645⟩. ⟨hal-00702542⟩
Vieu, L. (2018, March). A FrameNet lexicon and annotated corpus as DRD resource: Causality in the Asfalda French FrameNet. In *TextLink 2018 Final Action Conference* (pp. 172-178).
[11] Crible, L. (2022). The syntax and semantics of coherence relations: From relative configurations to predictive signals. *International Journal of Corpus Linguistics*, *27*(1), 59-92.
[12] https://framenet.icsi.berkeley.edu/about

[14] Melodie J. Fox. 2011. Prototype theory: An alternative concept theory for categorizing sex and gender? In Smiraglia, Richard P., ed. Proceedings from North American Symposium on Knowledge Organization, Vol. 3. Toronto, Canada, pp. 151-159
[15] Laurence Danlos, Quentin Pradet, Lucie Barque, Takuya Nakamura, Mathieu Constant. Un Verbenet du français. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), 2016, 57 (1), pp.25. hal-01392817
Laurence Danlos, Takuya Nakamura, Quentin Pradet. Vers la création d'un Verbnet du français. TALN 2014, 21ème conférence sur le Traitement Automatique des Langues Naturelles, Atelier FondamenTAL, Jul 2014, Marseille, France. hal-01084681
[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
Hou, Y. (2020). Fine-grained information status classification using discourse context-aware BERT. arXiv preprint arXiv:2010.14759.
[17] Rucar, Y., & Ganascia, J. G. (2019). An ontology and a memory island to give access to digital literature works. *Digital Scholarship in the Humanities*, *34*(Supplement_1), i150-i155.
[18] Duarte, J. M., & Berton, L. (2023). A review of semi-supervised learning for text classification. *Artificial Intelligence Review*, 1-69.