

# Les récits nationaux autour de la figure de Staline dans les manuels d'histoire russes

## Analyse diachronique (1947–2023)

Anastasiia Belosevich

Université Sorbonne Nouvelle  
a.belosevich@gmail.com

### Résumé

Cette étude examine l'évolution des récits nationaux sur la figure de Staline et la période stalinienne dans 22 manuels d'histoire russes publiés entre 1947 et 2023, en combinant analyse diachronique et méthodes de TAL (embeddings SBERT, clustering HDBSCAN). À partir d'un corpus de 3 528 segments textuels filtrés thématiquement, la recherche identifie trois phases narratives distinctes : un discours de glorification axé sur l'industrialisation pendant la période stalinienne (1947-1953), une uniformisation du discours malgré la déstalinisation (1954-1999), et une évolution post-soviétique (2000-2023) où les répressions sont relativisées au profit d'une vision centrée sur l'État qui met en avant la puissance nationale et la souveraineté. Ces transformations montrent comment le passé est réécrit selon les priorités politiques du moment, soulignant le rôle actif de l'enseignement historique dans la formation de l'identité nationale russe.

**Mots-clés :** récits nationaux, analyse diachronique, extraction de narratifs

### 1 Introduction

"Nothing changes more constantly than the past", citation souvent attribuée à l'essayiste américain Gerald W. Johnson, résume l'enjeu central du rapport entre histoire, mémoire et récit national. Le passé n'est jamais figé : il est continuellement reconstruit à travers des récits reflétant les valeurs et priorités d'une société à un moment donné. Ces récits participent à la production du sens historique et à la stabilisation de représentations collectives. L'enseignement de l'histoire constitue le vecteur privilégié de cette transmission narrative.

Dans la Russie contemporaine, cette dimension narrative s'avère particulièrement marquée. Depuis le milieu des années 2010, l'État russe mène une politique de standardisation visant à réduire la pluralité des interprétations au profit d'une lecture uni-

fiée du passé national. Dès 2016–2017, des manuels conformes à un nouveau standard historico-culturel ont été mis en place, poursuivant une dynamique amorcée en 2013 avec le projet de manuels uniques, exempts de "doubles interprétations".

Cette normalisation narrative s'est intensifiée en août 2023, lorsque Vladimir Medinski (l'ancien ministre de la Culture) a présenté un manuel unique d'histoire de la Russie, illustrant la volonté de renforcer le contrôle sur les récits historiques transmis aux nouvelles générations[RBC, 2023].

Les manuels scolaires et universitaires apparaissent ainsi comme des dispositifs discursifs centraux dans la construction identitaire nationale. Par le choix des événements, des figures historiques, du lexique employé et des évaluations morales proposées, ils contribuent activement à la production de récits nationaux et à la configuration d'une mémoire collective.

Le présent travail propose d'analyser un corpus de manuels d'histoire russes publiés des années 1947 à 2023, afin d'examiner, diachroniquement, l'évolution des récits nationaux relatifs au passé soviétique. L'étude se concentre sur la figure de Joseph Staline, personnage central du récit soviétique, puis figure profondément problématisée à partir de la déstalinisation et de la perestroïka. En tant que figure à la fois fondatrice, controversée et continuellement réinterprétée, Staline constitue un observatoire privilégié des transformations narratives du passé.

La problématique est donc la suivante : comment les récits nationaux autour de la figure de Staline et de la période stalinienne évoluent-ils dans les manuels d'histoire russes en fonction des contextes historiques et politiques de leur production ?

Mon hypothèse postule une reconfiguration narrative progressive : d'une représentation idéalisée à l'époque soviétique, la figure de Staline évolue vers une incarnation des répressions durant la perestroïka et les années 1990–2000, pour aboutir ré-

cemment à un récit de stabilisation nationale où les violences sont relativisées au profit de la puissance étatique et de la souveraineté.

L’approche choisie s’inscrit dans le courant récent des travaux utilisant les embeddings textuels pour l’extraction et l’analyse de narratifs [Elfes, 2024].

L’ensemble du corpus utilisé, le code source permettant de reproduire les différentes étapes de l’analyse, ainsi que les fichiers de sortie (statistiques au format CSV et résultats intermédiaires au format JSON) sont disponibles sur mon GitHub : [github.com/abelosev/recits-nationaux](https://github.com/abelosev/recits-nationaux)

## 2 Méthodologie

### 2.1 Corpus et métadonnées

Le corpus analysé est composé de 22 manuels d’histoire russe et soviétique, publiés entre les années 1947 et 2023. Il comprend des manuels scolaires (collège et lycée) ainsi que des manuels universitaires. La période 1990–2000 est moins représentée dans le corpus, ce qui constitue une limite assumée de l’étude. Le corpus comprend un total de 1,805,203 mots.

Chaque document est associé à un ensemble de métadonnées (pays, année de publication, niveau d’enseignement, auteur, titre), stockées dans le fichier *metadata.csv* (créé manuellement). Les titres des ouvrages sont conservés dans leur langue originale (le russe) afin de garantir l’identification exacte des sources. Les noms d’auteurs sont indiqués selon une translittération uniforme afin d’assurer la cohérence entre les métadonnées et les fichiers du corpus.

L’utilisation des métadonnées permet d’optimiser la segmentation du corpus en sous-corpus. Il est ainsi possible de regrouper les manuels en fonction de leur année de publication, de leur niveau d’enseignement (primaire, secondaire, universitaire) ainsi que de l’espace narratif auquel ils se rattachent (URSS, Russie contemporaine ou pays anciennement intégrés à l’Union soviétique, tels que la Moldavie, l’Estonie ou l’Ukraine). Ce corpus n’est pas exhaustif et est destiné à être enrichi et élargi dans le cadre de travaux ultérieurs sur cette thématique.

### 2.2 Sous-corpus

Dans cette étude, je travaille avec des sous-corpus organisés par période de publication : la périodisation retenue est basée sur les principaux changements politiques. La période 1947–1953

Période	Nombre de manuels	Nombre de mots	Espace narratif
1947–1953	3	157651	URSS, RSS d’Estonie
1954–1984	7	348655	URSS, RSS d’Ukraine, RSS kazakhe, RSS moldave
1985–1999	2	213427	URSS, Russie
2000–2012	5	665015	Russie
2012–2023	5	420455	Russie

TABEAU 1 : Statistiques des sous-corpus

marque la fin du stalinisme, caractérisée par une gouvernance répressive. Entre 1954 et 1984, sous Khrouchtchev puis Brejnev, le système soviétique autoritaire se stabilise, fonctionnant essentiellement par inertie institutionnelle. La phase 1985–1999, initiée par Gorbatchev puis poursuivie sous Eltsine, constitue une transition marquée par une profonde crise des institutions soviétiques. De 2000 à 2012, sous Poutine et Medvedev, l’État se reconstruit autour d’un consensus autoritaire. Enfin, la période 2013–2023, dominée par Poutine, représente un tournant conservateur et impérial, avec une confrontation croissante avec l’Occident et une militarisation progressive du régime.

Le tableau 1 présente les sous-corpus et leurs principales caractéristiques.

### 2.3 Prétraitement

#### 2.3.1 Normalisation du corpus

La première étape du pipeline consiste en une normalisation légère des textes bruts extraits des manuels scolaires. Cette normalisation vise à nettoyer les artefacts techniques tout en préservant la structure discursive essentielle à l’analyse sémantique ultérieure.

#### Nettoyage effectué

Le script *step1\_normalize.py* procède à l’unification des caractères spéciaux (espaces insécables, tirets typographiques, sauts de page) et supprime les éléments paratextuels non pertinents pour l’analyse : numéros de pages isolés, références bibliographiques entre crochets et indications de pagination. Cruciale pour le traitement par SBERT, la fusion des retours à la ligne simples au sein des paragraphes transforme le texte fragmenté par l’extraction PDF en blocs cohérents, tout en conservant les doubles sauts de ligne comme marqueurs de frontières paragraphiques.

#### Traitement des fichiers défectueux

Certains fichiers du corpus (par exemple *pancratova\_1952.txt*) présentaient une structure altérée : le texte y apparaissait comme un flux continu sans

aucune segmentation en paragraphes (moins de 20 sauts de ligne pour plus de 900 000 caractères). Pour ces cas particuliers, une heuristique de reconstruction a été implémentée : le script détecte automatiquement les textes mal formatés et insère des frontières paragraphiques aux points suivis directement d'une majuscule, en excluant les abréviations courantes (par exemple, г. (année)) pour éviter les faux positifs.

### 2.3.2 Segmentation du corpus

Dans cette partie je procède au découpage des textes en unités d'analyse. Cette étape est essentielle pour l'application des méthodes de TAL.

#### Stratégie

L'approche choisie repose sur le paragraphe comme unité principale d'analyse. Ce choix se justifie par le fait que le paragraphe constitue généralement une unité thématique cohérente dans les manuels scolaires.

Cependant, les paragraphes présentent une grande variabilité de longueur. Or, le modèle SBERT que nous utilisons pour générer les représentations vectorielles impose une limite de 300-400 tokens[Snegirev et al., 2025]. Pour respecter cette contrainte technique tout en préservant la cohérence sémantique, les règles suivantes ont été appliquées :

- Les paragraphes courts (moins de 1500 caractères et 5 phrases maximum) sont conservés tels quels.
- Les paragraphes longs sont subdivisés en segments de 2 à 5 phrases, en veillant à ne pas dépasser la limite de caractères.
- Les segments d'une seule phrase sont acceptés lorsqu'ils correspondent à des paragraphes courts dans le texte original, car ils peuvent contenir des informations importantes (définitions, thèses centrales).

#### Traitement des spécificités du russe

La segmentation en phrases nécessite une attention particulière pour les textes en russe, en raison de la fréquence des abréviations utilisant des points. Des expressions comme т.е. (c'est-à-dire), т.д. (et cetera), г. (année) ou вв. (siècles) pourraient être interprétées à tort comme des fins de phrase. L'algorithme du script *step2\_segment.py* protège ces abréviations avant le découpage, puis les restaure dans le texte final.

### Filtrage et métadonnées

J'exclus de l'analyse les segments trop courts (moins de 15 mots), qui correspondent généralement à des titres, des légendes ou des fragments peu informatifs. Chaque segment conserve ses métadonnées d'origine : identifiant du document, année de publication, niveau scolaire et espace narratif (pays), issues du fichier de métadonnées. La période historique est déduite de l'organisation du corpus en répertoires thématiques. Ces informations permettront d'analyser les variations du discours selon différents axes.

#### Résultats de la segmentation

Le corpus segmenté est exporté aux formats JSON et CSV, accompagné de statistiques descriptives par période. Cette double exportation facilite à la fois l'inspection manuelle des données et leur traitement automatisé lors des étapes suivantes (génération des embeddings, clustering).

### 2.3.3 Filtrage thématique

Le corpus segmenté contient l'ensemble des passages des manuels, dont une grande partie ne concerne pas directement Staline ni la période stalinienne. Afin de constituer un sous-corpus pertinent pour l'analyse, j'ai mis en place un système de filtrage fondé sur des marqueurs lexicaux.

#### Catégories de marqueurs

Ainsi, 6 catégories de marqueurs thématiques ont été définies, chacune associée à un poids reflétant sa pertinence :

- **Mentions directes** (poids 3) : références explicites à Staline (Сталин (Staline), сталинск- (de Staline), Джугашвили (Djougachvili))
- **Mentions faibles** (poids 1) : le surnom révolutionnaire Коба (Koba), traité séparément en raison de sa possible ambiguïté contextuelle
- **Culte de la personnalité** (poids 2) : expressions liées à l'idéologie stalinienne (культ личности (culte de la personnalité), вождь народов (guide des peuples), отец народов (père des peuples))
- **Répressions** (poids 2) : vocabulaire de la terreur (репресс- (racine du mot "répressions"), НКВД (NKVD), ГУЛАГ (Goulag), Большой террор (la Grande Terreur))
- **Procès politiques** (poids 1) : noms des accusés et événements judiciaires (дело врачей (l'affaire des médecins), Ежов (Iejov))

- **Événements** (poids 1) : transformations socio-économiques de l'époque (коллективизация (collectivisation), индустриализация (industrialisation), пятилетка (plan quinquennal))

Un score est calculé pour chaque segment en additionnant les poids des marqueurs détectés. Un segment est retenu s'il mentionne explicitement Staline (score automatique) ou si son score atteint un seuil minimal de 2 points avec au moins un marqueur thématique fort.

### Traitement des dates

Les années de la période stalinienne (1924-1953) constituent un signal pertinent, mais insuffisant à lui seul. Par exemple, un segment évoquant uniquement "1941" ou "1945" peut traiter de la guerre sans aucune référence au régime stalinien. Pour éviter ce bruit, une règle de *gating* a été introduite : les dates ne contribuent au score (jusqu'à 3 points maximum) que si le segment contient déjà au moins un marqueur thématique fort. Cette approche réduit significativement les faux positifs.

Les années retenues sont : 1929 (lancement de la collectivisation), 1930 (répression contre les "koulaks"), 1932-1934 (famines), 1936-1939 (la Grande Terreur), 1946 (reprise de la répression après la guerre), 1948-1949 (montée de l'antisémitisme d'État et renforcement de l'isolement idéologique), 1952-1953 (fin du stalinisme).

### Difficultés techniques

Plusieurs adaptations ont été nécessaires pour le traitement du russe :

- Normalisation de la lettre ё : les textes utilisent parfois e à la place de ё. Une normalisation préalable (ё → e) garantit la détection de toutes les variantes
- Exclusion de Сталинград (Stalingrad) : le pattern Сталин\* (Staline) capturerait Сталинград et Сталинградская битва (bataille de Stalingrad), qui relèvent de la guerre et non du personnage. Une exclusion explicite a été implémentée
- Comptage par années uniques : pour éviter les répétitions, chaque année n'est comptée qu'une fois par segment.

### Validation et traçabilité

Pour faciliter la validation manuelle et l'analyse des erreurs, chaque segment retenu est enrichi de métadonnées :

- *filter\_score* : score total du segment
- *filter\_matches* : nombre de correspondances par catégorie
- *filter\_key\_years* : liste des années clés détectées
- *filter\_passed\_reason* : raison du passage (*stalin\_direct*, *strong\_repression*, *strong\_cult*, etc.)

Pour ce dernier champ, la catégorie dominante est déterminée par la contribution maximale au score (nombre de marqueurs × poids). En cas d'égalité, un départage est effectué par le nombre brut de marqueurs. Si l'égalité persiste, le segment est classé *strong\_mixed*.

Un rapport de qualité est généré contenant plusieurs échantillons stratifiés : les segments au score le plus élevé, les segments à la limite du seuil (*borderline*), les segments rejetés uniquement en raison du *gating* des dates, ainsi que des échantillons aléatoires. Cette approche permet une vérification manuelle et une estimation de la précision du filtre par catégorie de passage.

### 2.3.4 Déduplication

Cette étape vise à réduire la redondance des données par une déduplication des segments identiques, phénomène fréquent dans les manuels scolaires, qui reprennent souvent des formulations ou définitions similaires.

### Méthodologie

Le script *step4\_dedup.py* implémente une approche à deux niveaux pour identifier et supprimer les doublons :

- Déduplication exacte : basée sur le hachage MD5 du texte normalisé. Cette méthode détecte les segments identiques après suppression de la ponctuation, unification des espaces et conversion en minuscules.
- Déduplication floue : utilisant la similarité de Jaccard sur des trigrammes de caractères (n=3), avec un seuil de similarité de 0,85. Cette méthode permet de détecter les quasi-doublons présentant de légères variations (reformulations mineures, différences typographiques). Pour optimiser les performances, un filtrage préalable par longueur a été appliqué : les segments dont le ratio de longueurs est inférieur à 0,85 sont exclus de la comparaison, car mathématiquement, leur similarité de Jaccard ne peut atteindre le seuil requis.

## Stratégie de conservation

Lors de la détection de doublons, on conserve systématiquement le premier segment rencontré. Cette stratégie, bien que simple, présente une limite : elle peut privilégier une version de moindre qualité si l'ordre d'entrée est aléatoire. Une amélioration future pourrait consister à sélectionner le segment le plus long ou celui présentant le moins de caractères non alphabétiques.

## Déduplication intra-période

Afin de préserver les évolutions diachroniques du discours, la déduplication est effectuée séparément au sein de chaque période temporelle. Cette approche permet de conserver des formulations similaires apparaissant dans différentes périodes, ce qui est essentiel pour l'analyse de l'évolution des discours idéologiques.

## Paramètres et optimisations

Les segments de moins de 10 caractères après normalisation sont exclus de la déduplication floue, car trop courts pour produire des comparaisons fiables. Pour la déduplication floue, l'algorithme compare chaque nouveau segment avec tous les segments uniques déjà identifiés, garantissant ainsi la détection du meilleur match (similarité maximale) plutôt que du premier match acceptable. Cette recherche exhaustive, bien que plus coûteuse (complexité  $O(n^2)$ ), assure une déduplication optimale.

## Validation et traçabilité

Le script génère 3 fichiers de sortie : les segments dédupliques, un rapport statistique détaillé par période, et un journal des paires de doublons supprimés incluant les textes originaux et normalisés. Ce dernier permet une validation manuelle, particulièrement pour les cas limites (similarité  $< 0,90$ ) où la pertinence de la suppression peut être ambiguë.

## 2.4 Analyse

### 2.4.1 Embeddings

Pour générer les représentations vectorielles des segments, j'ai utilisé le modèle `sbert_large_nlu_ru`<sup>1</sup>, un modèle Sentence-BERT spécifiquement entraîné pour le russe [Snegirev et al., 2025]. Ce modèle a été choisi en raison de ses performances supérieures pour la capture de la similarité sémantique en russe.

1. [https://huggingface.co/ai-forever/sbert\\_large\\_nlu\\_ru](https://huggingface.co/ai-forever/sbert_large_nlu_ru)

Le modèle produit des vecteurs de 1024 dimensions pour chaque segment textuel. L'encodage a été réalisé par lots de 32 segments, avec une normalisation L2 des vecteurs. Cette normalisation permet d'utiliser directement le produit scalaire comme mesure de similarité cosinus lors de l'étape de clustering, simplifiant ainsi les calculs tout en préservant les relations sémantiques entre segments.

Les embeddings ont été sauvegardés à la fois globalement et séparément pour chaque période historique, facilitant ainsi les analyses diachroniques comparatives. Cette organisation permet d'examiner l'évolution des configurations narratives au sein de chaque période avant de procéder à une analyse transversale de l'ensemble du corpus. Chaque vecteur est associé à ses métadonnées d'origine (identifiant du document, période, année de publication, score de filtrage), permettant une traçabilité complète lors de l'interprétation des clusters thématiques.

### 2.4.2 Clusterisation

L'objectif de cette étape est d'identifier des thématiques narratives latentes à partir du corpus filtré, sans imposer de catégories prédéfinies. Pour ce faire, une approche non supervisée a été adoptée, combinant une réduction de dimensionnalité et un algorithme de clustering basé sur la densité.

## Réduction de dimension avec UMAP

Les segments textuels sont représentés par des embeddings sémantiques de dimension élevée (1024), ce qui rend leur traitement direct peu adapté à la clusterisation. Une réduction de dimension est donc appliquée à l'aide de l'algorithme UMAP (*Uniform Manifold Approximation and Projection*), choisi pour sa capacité à préserver les voisinages locaux dans l'espace sémantique et à fonctionner avec une métrique cosinus adaptée aux embeddings normalisés.

Les paramètres suivants sont fixés afin d'assurer la reproductibilité des résultats : `n_neighbors=15`, `min_dist=0.0`, `metric=cosine`, `random_state=42`. La réduction projette les données dans un espace de 50 dimensions, utilisé exclusivement pour la clusterisation. Une projection bidimensionnelle distincte (`min_dist=0.1`) est calculée à des fins de visualisation.

## Clusterisation avec HDBSCAN

La clusterisation est réalisée à l'aide de l'algorithme HDBSCAN (*Hierarchical Density-Based*



*Spatial Clustering of Applications with Noise*), appliqué à l'espace réduit par UMAP. Cet algorithme présente plusieurs avantages : il ne nécessite pas de spécifier le nombre de clusters à l'avance, détecte automatiquement les segments isolés (bruit), et s'adapte à des structures de densité variable.

Les paramètres retenus sont : `min_cluster_size=10`, `min_samples=5`, `metric=euclidean`, `cluster_selection_method=eom`. La métrique euclidienne est utilisée à ce stade, la clusterisation étant effectuée dans l'espace UMAP, qui est euclidien par construction.

### Probabilités d'appartenance et bruit

HDBSCAN fournit, pour chaque segment, une probabilité d'appartenance au cluster, indiquant le degré de confiance de l'algorithme dans l'affectation. Cette information permet de distinguer le noyau des clusters des segments plus périphériques. Les segments non intégrables dans un cluster cohérent sont automatiquement identifiés comme bruit, sans être supprimés du corpus.

### Reproductibilité et sorties

L'ensemble des paramètres, des statistiques de réduction et de clusterisation, ainsi que les résultats intermédiaires (labels, probabilités, projections, exports tabulaires) est systématiquement sauvegardé, garantissant la traçabilité et la reproductibilité de l'analyse.

## 3 Résultats

### 3.1 Prétraitement

Le pipeline de prétraitement a permis de constituer un corpus final de **3 528** segments à partir d'un ensemble initial de **31 968** segments extraits de 22 manuels d'histoire soviétiques et post-soviétiques.

La distribution par période révèle des caractéristiques stylistiques distinctes (tableau 2) :

Période	Segments	Mots/segment
1947–1953	2 976	52,4
1954–1984	6 442	52,3
1985–1999	3 679	54,6
2000–2012	10 398	61,6
2012–2023	8 473	39,8
<b>Total</b>	<b>31 968</b>	<b>52,3</b>

TABLEAU 2 : Distribution des segments par période après segmentation

Le **filtrage** par mots-clés a retenu 3 647 segments (11,41% du corpus initial) jugés pertinents

pour l'analyse du stalinisme. Cette sélectivité rigoureuse garantit un corpus hautement ciblé sur l'objet d'étude.

La répartition des segments retenus par période (tableau 3) révèle une concentration marquée dans les périodes contemporaines :

Période	Segments retenus	% du total filtré
1947–1953	880	24,1%
1954–1984	322	8,8%
1985–1999	341	9,4%
2000–2012	1 467	40,2%
2013–2023	637	17,5%
<b>Total</b>	<b>3 647</b>	<b>100%</b>

TABLEAU 3 : Distribution des segments après filtrage thématique

Les périodes 2000–2012 et 2013–2023 représentent ensemble 57,7% du corpus filtré, témoignant d'une intensification discursive de la référence au stalinisme dans l'historiographie post-soviétique.

L'analyse des années les plus fréquemment mentionnées dans les segments retenus confirme la pertinence du filtrage. Les 10 dates les plus citées correspondent exactement aux moments cardinaux du stalinisme : 1930 (133 mentions), 1929 (129), 1937 (124), 1939 (100), 1953 (99). Ces dates renvoient respectivement à la collectivisation forcée, à la Grande Terreur, aux purges et à la mort de Staline.

À l'inverse, les segments rejetés concentrent les mentions d'années périphériques à l'objet d'étude : 1917 (740 mentions), 1991 (495), 1941 (394), 1945 (336). Le filtrage a donc efficacement écarté les passages centrés sur la Révolution, l'effondrement de l'URSS et la Seconde Guerre mondiale, confirmant la précision thématique de l'approche choisie.

La **déduplication** a éliminé 119 segments (3,26% du corpus filtré), produisant un corpus final de 3 528 segments. Ce taux de déduplication relativement faible indique une diversité textuelle satisfaisante et l'absence de redondances systématiques.

La concentration massive des doublons dans la période 1954–1984 (36,3% des segments initiaux) est remarquable et témoigne d'une forte standardisation discursive caractéristique de l'historiographie soviétique tardive. Les manuels de cette époque reproduisent fréquemment des formulations canoniques identiques, reflétant le contrôle idéologique strict de la production historiogra-

prique. En revanche, l'absence totale de doublons exacts dans les périodes post-soviétiques (1985–2023) indique une diversification des approches narratives et une plus grande liberté d'expression.

Le **corpus final** comprend **3 528** segments représentant environ **184 000** mots, distribués sur 5 périodes historiques. La construction de ce corpus garantit :

- Une haute pertinence thématique : 58,7% des segments contiennent des références directes à Staline
- Une diversité textuelle : taux de déduplication minimal (3,26%)
- Une représentativité diachronique : couverture de la période 1947–2023
- Une granularité appropriée : segments de 40 à 62 mots en moyenne, suffisamment longs pour porter du sens sans mélanger les thèmes

Ce corpus constitue donc une base adaptée à l'analyse des évolutions du discours sur le stalinisme dans l'historiographie russe et soviétique.

### 3.2 Analyse

L'application du pipeline UMAP + HDBSCAN sur le corpus filtré de 3 528 segments a permis d'identifier 68 clusters thématiques. Parmi ces segments, 2 322 (65,8 %) ont été intégrés à un cluster, tandis que 1 206 (34,2 %) ont été classés comme bruit.

Ce taux de bruit reflète la nature du corpus : les manuels scolaires contiennent de nombreux segments transversaux ou faiblement spécifiques. La détection explicite du bruit par HDBSCAN constitue un résultat méthodologique en soi, évitant la sur-interprétation de fragments isolés.

La distribution des probabilités d'appartenance (figure 1) confirme la robustesse de la clusterisation : la grande majorité des segments clusterisés présente une probabilité supérieure à 0,8, avec un pic marqué autour de 1,0.

### 3.3 Trois phases narratives du discours sur le stalinisme

L'analyse croisée des tailles de clusters (figure 2) et de leur distribution temporelle (figure 3) permet d'identifier trois phases narratives distinctes, correspondant à l'évolution des cadres interprétatifs du stalinisme dans l'historiographie soviétique et post-soviétique.

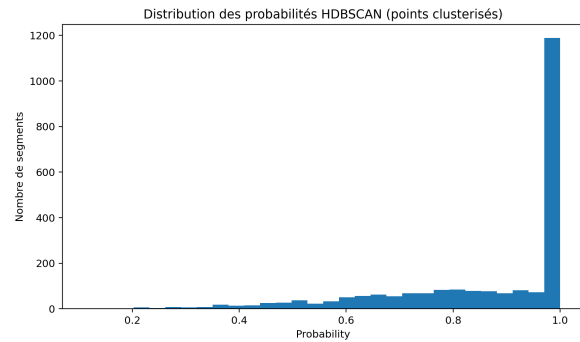


FIGURE 1 : Distribution des probabilités d'appartenance HDBSCAN pour les segments clusterisés

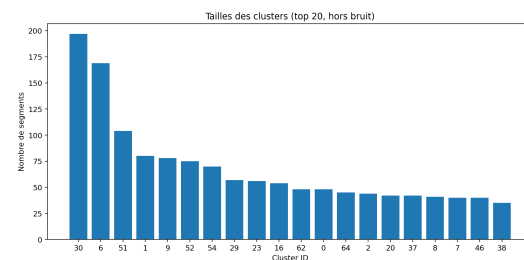


FIGURE 2 : Tailles des 20 clusters principaux (hors bruit)

### Phase 1 : Glorification et industrialisation (1947-1953)

La première phase correspond au discours de glorification caractéristique de la période stalinienne tardive. Elle est principalement incarnée par les clusters 6, 29 et 46, qui représentent ensemble plus de 30 % des segments de cette période.

Le cluster 6 (16,9 % de la période) concentre les segments mobilisant la parole de Staline comme argument d'autorité : citations directes, formules énonciatives canoniques ("товарищ Сталин говорил" (comme disait le camarade Staline)), et mise en récit téléologique des succès soviétiques. Les segments représentatifs font référence aux discours de Staline sur l'industrialisation, à ses interventions lors des conférences du parti, et à la nécessité de "rattraper et dépasser" les pays capitalistes.

Le cluster 29 (8,5 %) prolonge cette logique en structurant le récit autour des congrès du parti et des programmes politiques. Le cluster 46 (6,5 %) développe le thème du "génie militaire" de Staline, présentant la victoire de 1945 comme la démonstration de sa clairvoyance stratégique.

Cette configuration narrative correspond à un discours de *glorification* au sens strict : la figure de Staline y est présentée comme le guide infallible de l'industrialisation et de la victoire militaire, sans

aucune distance critique.

## Phase 2 : Uniformisation malgré la déstalinisation (1954–1999)

La deuxième phase, couvrant les périodes 1954-1984 et 1985-1999, se caractérise par une stabilité thématique malgré les bouleversements politiques (déstalinisation khrouchtchévienne, perestroïka).

L'analyse révèle que le cluster 30 (collectivisation et transformations socio-économiques) devient dominant dès 1954-1984 (11,7 %) et le reste durant la période 1985-1999 (10,1 %). Le cluster 6, associé au discours canonique stalinien, ne disparaît pas : il représente encore 9,3 % des segments en 1985-1999.

Cette persistance du lexique et des structures narratives constitue un résultat significatif. Malgré le rapport secret de Khrouchtchev (1956) et les réformes de Gorbatchev, le discours historiographique sur le stalinisme conserve ses cadres thématiques fondamentaux : industrialisation, collectivisation, transformation socio-économique. La critique du "culte de la personnalité" n'engendre pas une reconfiguration profonde des narratifs, mais plutôt une atténuation du ton apologetique.

Cette uniformisation peut s'expliquer par l'inertie institutionnelle des manuels scolaires, soumis à des processus de validation longs et à des contraintes éditoriales conservatrices. Elle témoigne également de la difficulté à construire un récit alternatif au sein du système soviétique.

## Phase 3 : Relativisation et vision étatiste (2000-2023)

La troisième phase, correspondant aux périodes 2000-2012 et 2013-2023, marque une reconfiguration significative du discours historiographique.

Le cluster 30 (collectivisation) reste important (10,9 % en 2000-2012), mais de nouveaux clusters émergent comme dominants : le cluster 51 (7,1 %) et le cluster 9 (7,4 % en 2013-2023).

Le cluster 51 regroupe les segments relatifs aux conférences interalliées (Téhéran, Yalta, Potsdam) et au rôle de l'URSS dans l'ordre international d'après-guerre. Les segments représentatifs mettent en avant la dimension diplomatique et géopolitique du stalinisme : réparations, zones d'influence. Cette thématique, quasi absente des périodes antérieures, traduit un recentrage du récit sur la puissance étatique de l'URSS stalinienne.

Le cluster 9 rassemble les segments consacrés à la culture soviétique et au réalisme socialiste. Les textes représentatifs évoquent les écrivains,

artistes et intellectuels de l'époque, ainsi que les mécanismes de contrôle idéologique. Ce cluster témoigne d'une diversification des angles d'approche, intégrant désormais la dimension culturelle du stalinisme.

Cette évolution correspond à ce que l'hypothèse initiale désignait comme une "relativisation des répressions au profit d'une vision centrée sur l'État". Les violences stalinienne ne sont pas niées, mais elles sont intégrées dans un récit plus large valorisant la construction de la puissance soviétique et la souveraineté nationale.

Cette observation appelle plusieurs interprétations complémentaires. D'une part, elle peut refléter une tendance inhérente aux manuels scolaires à atténuer les aspects les plus conflictuels de l'histoire nationale, privilégiant des récits consensuels et mobilisateurs. D'autre part, elle peut résulter d'un biais de corpus : la période 1985-1999, durant laquelle le discours sur les répressions était le plus intense dans l'espace public, est sous-représentée (seulement 2 manuels). Enfin, cette configuration peut traduire une évolution réelle du discours historiographique post-soviétique, où les répressions, sans être occultées, sont progressivement subordonnées à un récit de construction étatique.

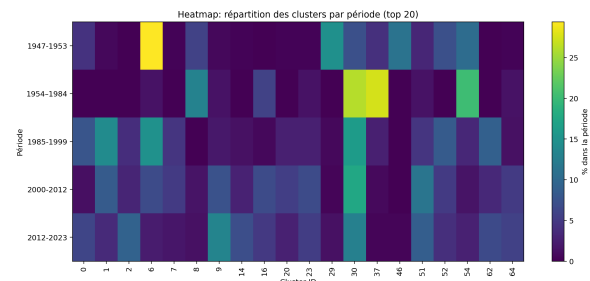


FIGURE 3 : Répartition des clusters par période (pourcentage intra-période, top 20 clusters). La heatmap visualise les 3 phases narratives : dominance du cluster 6 en 1947-1953, stabilité du cluster 30 de 1954 à 1999, émergence des clusters 51 et 9 après 2000.

## 3.4 Visualisation de l'espace sémantique

Les projections UMAP offrent une représentation complémentaire de la structure du corpus.

La figure 4 montre une structure partiellement différenciée : certains clusters forment des îlots compacts (périphérie du graphique), tandis que la zone centrale présente un chevauchement. Cette configuration est cohérente avec la nature du corpus : les manuels traitent des thèmes récurrents avec des variations de cadrage.

La figure 5 révèle une stratification diachronique



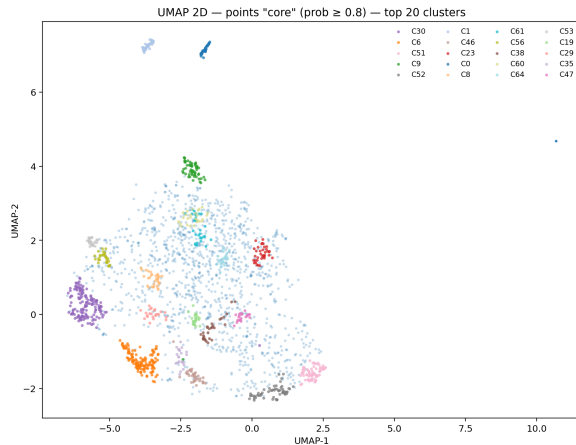


FIGURE 4 : Projection UMAP des segments "core" (probabilité  $\geq 0,8$ ), colorés par cluster

partielle. La période 1947-1953 présente des zones de concentration spécifiques, tandis que les périodes post-soviétiques occupent une surface plus dispersée. Cette distribution mixte correspond à ce que l'on attend d'un corpus historiographique : les mêmes objets sont traités différemment selon les périodes.

### 3.5 Bruit thématique et limites

L'analyse a identifié des clusters correspondant à du bruit thématique structurel. Notamment, le cluster 4, dominant en 1954-1984 (13,6 %), regroupe des segments relatifs à l'histoire militaire pré-soviétique (guerres cosaques, interventions). Sa présence s'explique par l'héritage des manuels soviétiques, qui inscrivaient le récit stalinien dans une continuité historique longue.

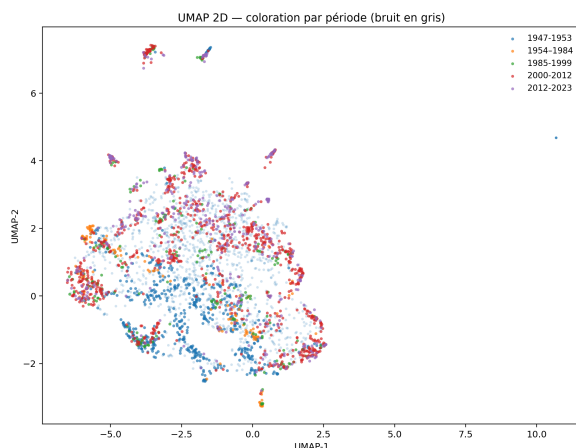


FIGURE 5 : Projection UMAP colorée par période historique

L'identification explicite de ce bruit renforce la robustesse méthodologique : ces segments ne sont pas intégrés de force dans l'interprétation théma-

tique.

## 4 Conclusion

Ainsi, l'analyse des clusters confirme l'hypothèse d'une reconfiguration narrative en 3 phases :

1. **Glorification (1947-1953)** : discours apologétique centré sur l'industrialisation et le génie de Staline (clusters 6, 29, 46).
2. **Uniformisation (1954-1999)** : persistance des cadres thématiques malgré la déstalinisation officielle (clusters 30, 6). L'inertie institutionnelle des manuels scolaires maintient une continuité discursive.
3. **Relativisation et vision étatiste (2000-2023)** : émergence de nouveaux narratifs centrés sur la puissance internationale (cluster 51) et la culture (cluster 9). Les répressions sont intégrées dans un récit valorisant la souveraineté nationale.

### Apports et limites

Le pipeline (filtrage pondéré, SBERT, HDBSCAN) offre un cadre reproductible. La détection du bruit et la déduplication (révélant 36,3% de doublons en 1954–1984) renforcent la robustesse. Les limites incluent le déséquilibre du corpus (1985–1999 sous-représentée), la tendance des manuels à atténuer les conflits, et la faible visibilité des répressions qui en résulte.

## Références

- Elfes. Mapping news narratives using llms and narrative-structured text embeddings. *arXiv preprint arXiv :2409.06540*, 2024.
- RBC. Medinski a présenté le manuel unique d'histoire pour les lycéens. <https://web.archive.org/web/20230811135722/https://www.rbc.ru/politics/07/08/2023/64d1016a9a794763ffc7ab63>, 2023.
- SberDevices. Sbert large nlu russian. [https://huggingface.co/ai-forever/sbert\\_large\\_nlu\\_ru](https://huggingface.co/ai-forever/sbert_large_nlu_ru).
- Snegirev et al. The russian-focused embedders' exploration : rumteb benchmark and russian embedding model design. *arXiv preprint arXiv :2408.12503*, 2025.