Sur les épaules d'un géant : utilisation des corpus et outils numériques pour l'histoire des discours antimodernes sur l'Europe dans la presse suisse 1900-1945

On a giant's shoulders: using corpus and digital tools for the history of anti-modern discourses on Europe in the Swiss press 1900-1945

Estelle Bunout[1]

[1]*Luxembourg Centre for Contemporary and Digital History (C2DH), Université du Luxembourg*

Abstract

Ce cas d'étude propose de montrer comment l'environnement de recherche créé par la numérisation des sources, la popularisation d'outils de traitement automatisé des textes et surtout le dialogue interdisciplinaire, font évoluer la pratique de la recherche en histoire. Au-delà de l'automatisation de certaines tâches typiques de la recherche historique, comme l'aide à la constitution d'un corpus de recherche ou l'annotation de ce corpus, la réflexion même sur les sources et leur contenu est stimulée par cet environnement. En l'occurrence : une analyse des discours se traduit par la mesurer une présence, ou l'estimation de la représentativité d'éléments distinctifs d'un discours. Ces tâches se prêtent particulièrement à une automatisation partielle, par le biais d'analyse automatisée des textes, rendue possible par leur numérisation. Ce papier présente ées étapes d'opérationalisation d'une analyse de discours en histoire, utilisant un corpus de presse numérisée, des outils de traitement des langues naturelles pour collecter un corpus de recherche, le classer et l'organiser par degré de similarité. L'itération entre conceptualisation, opérationalisation et analyse des résultats ouvre un nouvel angle d'observation des sources historiques.

Keywords

history, digital press, naive Bayesian classifier

1. Introduction

Avant de présenter nos étapes d'automatisation partielle de l'analyse de discours, il nous faudra dans un premier temps définir comment le discours antimoderne se matérialise dans la presse et trouver des textes « similaires ». L'analyse de discours implique un changement d'échelle entre la production d'idées et une certaine représentativité ou circulation de ces idées. Dans le même temps, ce changement d'échelle et la médiation qu'il implique risque également de décupler nos biais et de ne valider que les hypothèses émises au départ, en d'autres termes, ne va-t-on trouver que ce qu'on cherche ? La presse numérisée constitue pour cela, une source potentiellement formidable pour procéder à ce changement d'échelle, permettant de chercher à la fois de manière ciblée et de collecter une grande quantité de texte, simplement par un mot-clé.

Workshop on Digital Humanities and Semantic Web
estellebunout@gmail.com (E. Bunout)
https://www.c2dh.uni.lu/people/estelle-bunout (E. Bunout)

Il faut toutefois trouver des stratégies pour analyser cet accès facilité : comment identifier un discours dans la masse d'articles, publicités et annonces publiés au quotidien ?

C'est la richesse de cette source, que de présenter un kaléidoscope de l'usage du terme « Europe » et c'est un avantage indéniable, de sa forme numérisée, que de pouvoir mesurer relativement simplement, la proportion de ses différents usages et leur évolution au fil du temps. Le passage à une analyse de discours nécessite un traitement particulier de la masse collectée. Nous allons présenter comment la médiation par les outils de fouille de texte automatisée, implique une certaine transparence dans la démarche heuristique, dimension mise en avant dans la recherche en humanités numériques. L'aspect que nous voudrions souligner ici, est l'itération entre la formulation d'hypothèses de recherche, leur transposition en consignes aux algorithmes utilisés et la redéfinition des concepts utilisés [1]. En d'autres termes, comment la confrontation médiée ou distante à la source historique numérisée permet de poser des questions à des sources massives en offrant une interaction adaptée à leur format et envergure.

2. Annoter le corpus « Europe » dans la presse suisse : diversité thématique et discursive

Les articles qui ont servi dans cette analyse, ont été collectés via l'application impresso[1], en utilisant comme mot-clé « europ* », de façon à collecter les articles contenant tant les mots « Europe », qu' « européen », « européenne » etc., pour une diversité de titres de la presse francophone suisse (*l'Express, Gazette de Lausanne, l'Impartial, Journal de Genève, le Confédéré, l'Essor, la Liberté, la Lutte Syndicale, la Sentinelle, Solidarité*). Parmi ces titres, on retrouve des feuilles d'avis, dont la fonction était originellement la diffusion d'informations commerciales, des organes de partis ou syndicats et enfin, des titres de presse d'opinion. Au total, 227 351 articles ont été ainsi collectés.

Pour avoir une première impression du contenu des articles collectés, nous avons eu recours au topic modelling, qui est une méthode de calcul de probabilité de cooccurrence des mots dans une collection de document, sur la base d'observation partielle. Les informations produites par cet algorithme sont d'une part, les mots qui cooccurrent (probablement) au sein d'un document, qui forment un *topic* et la distribution de la présence des *topics* au sein de la même collection de documents. Cet outil permet d'avoir une description du contenu d'une vaste collection de documents, sans prédéfinir les thèmes mais en partant du contenu des documents. Par ces propriété et sa relative simplicité d'usage, il est particulièrement populaire pour la recherche dans la presse numérisée[2]

La détection de cooccurrence de mots au sein d'un document est moins simple à interpréter que les collocations, où on mesure une cooccurrence au sein d'une même phrase, mais jette un filet plus large de l'association récurrente de termes. On peut ainsi utiliser cet outil non

[1] Application développée dans le cadre du projet *impresso*: Media Monitoring of the Past, une colloboration entre les Universités de Zurich et du Luxembourg, et l'Ecole Polytechnique Fédérale de Lausanne, regroupant des collections de presse numérisée suisses et luxembourgeoise, enrichies par du *topic modelling*, la reconnaissance d'entitées nommées, la détection de textes dupliqués (*text reuse*), le plongement de mots (*word embedding*) et présentés dans un interface de recherche et d'exploration commun, accessible sous : https://impresso-project.ch/app/

[2] Comme l'indiquait déjà Blei [2] et comme est discuté dans Bunout et al. [3] à paraître.

seulement pour détecter des « thèmes » mais aussi des styles ou ton d'un texte.

Pour illustrer ce principe, prenons un exemple : le *topic* numéro 40, calculé pour le journal la Liberté se présente ainsi : « nos vos patrie sommes chers applaudissements devons avons avez voulons devoir noire dieu peuple-suisse confiance salut êtes messieurs bravos nous-mêmes ».

Ce sont les mots les plus récurrents pour ce *topic*. Ils forment une association qui reste à interpréter par le chercheur. Ici, on peut supposer qu'il s'agisse de discours prononcés lors de célébrations patriotiques en Suisse. On peut maintenant regarder les articles qui ont reçu une probabilité de contenir ce *topic* et observer parmi eux, un article du 15.09.1934 sur le Jeûne fédéral, des discours reproduits dans la presse, prononcés à l'occasion du Nouvel an de 1942 ou encore des articles de la rubrique d'annonces de contenus de la revue Semaine catholique. Ce qui n'était pas visible dans les mots décrivant le *topic* était donc la dimension religieuse des textes regroupés par celui-ci. Il est important de vérifier « manuellement/lecture humaine » le contenu des sous-collections d'articles créées par cet outil [4].

Les *topics* ont été utilisés pour cibler et sélectionner des exemples archétypaux, pour alimenter l'outil suivant : le classificateur naïf Bayésien. Le principe de ce classificateur est assez simple : sur la base d'une série de texte exemples et de textes contre-exemples, l'algorithme détermine une liste de mots qui est la plus discriminante pour un identifier l'une des deux catégories. Sur la base de sa présence dans l'une ou l'autre ou dans les deux catégories, chaque mot se voit attribuer une valeur de prédiction qui est utilisée pour mesurer la probabilité de textes inconnus (non utilisés pour définir ces catégories) à appartenir à l'une ou l'autre des deux catégories. Pour préparer le corpus d'entrainement, il est conseillé de choisir une diversité relative, pour couvrir les différents aspects d'une même catégorie, dont nous cherchons à identifier d'autres matérialisations. De manière symétrique, le corpus « neutre » ou de contraste, doit lui couvrir une grande diversité d'éléments à exclure ou minimiser dans le classement. Spontanément, on pourrait penser que choisir à partir d'un groupe d'article partageant le même *topic* pourrait donc s'avérer problématique, si la sélection reste trop homogène et ne permet pas de découvrir d'autres articles. Pour mieux comprendre comment cette question de l'« homogénéité » ou « diversité » a été traitée, nous allons détailler les critères retenus pour chaque catégorie et redonner quelques exemples pour illustrer les choix. L'établissement et la redéfinition des catégories de recherche résulte, sans trop d'originalité, d'un retour entre étude des sources et de la littérature scientifique. L'espoir ici est de tenter de matérialiser une pratique analogue : trouver des textes similaires au texte initialement repéré, estimer la proportion de la présence de textes similaires pour juger de l'importance du phénomène que le texte initial retenu reflète. Face à la masse de texte et l'impossibilité de redonner l'ensemble des sources pertinentes au lecteur, les analyses de discours se trouvent souvent amoindries au moment de la restitution des résultats.

Nous voudrions ici faire un chemin inverse à cette restitution traditionnelle en partant des idéaux-types que nous avons sélectionnés et discuter de la définition et mesure de la « similarité » via un classificateur naïf Bayésien (NBC)[3]. Ce faisant, nous rendons explicite les éléments utilisés pour définir l'idéal-type de chaque catégorie. Pour donner ici un exemple de cette démarche, nous nous concentrons sur la catégorie « diplomatique », la moins ambivalente. Pour cette catégorie, nous avions décidé de retenir les articles rapportant des faits avec des

[3]Pour une présentation plus détaillée des étapes, voir [5]

commentaires minimaux sur des rencontres diplomatiques, politiques à l'échelle européenne, ou encore faisant la chronique sans valorisation particulière, des initiatives diplomatiques de coopération européenne, notamment celle d'A. Briand en 1929. Le matériel collecté pour définir cette catégorie se compse d'éditoriaux, notamment de Maurice Muret, parus dans sous le titre « Bulletin politique », ou d'articles paraissant dans des rubriques type « vie internationale ». Ces rubriques sont souvent constituées de brèves et dans la reconnaissance automatique des contours des articles, produite lors de la numérisation de la source, ce type de format est souvent identifié comme un article entier. On pourrait considérer que ce type d'article contient beaucoup de « bruit » et pourrait fausser nos mesures de similarité. Nous avons cependant choisi de garder la rubrique entière pour « entrainer » la classification, car la collection dans laquelle nous allons chercher des articles similaires, en sera également composée. Il y aura donc des articles qui ne traitent pas d'« Europe » mais restent dans la tonalité que nous cherchons à identifier. Nous avons également sélectionné des articles qui rendent compte de conférences d'organisations telles que la Fédération des Unions intellectuelles portant des titres du type « La civilisation européenne en danger », qui aurait pu indiquer une thématique antimoderne, mais sont contenues dans des articles sans commentaires, plutôt informatifs. Enfin, nous avons ajouté des articles contenant des commentaires de type « géopolitique », sur les ambitions hégémoniques d'un pays particulier par exemple. Ainsi, les premiers mots de la liste des mots qui caractérisent cette catégorie semblent refléter le contenu ciblé : « convention », « union-douanière », « accords », « souscommission », « locarno », « délégué » etc. Le mot « convention » est accompagné d'une probabilité 95%, en d'autres termes, sur la base des articles donnés pour l'entrainement, un article qui contient ce mot a une très forte probabilité d'appartenir à la catégorie « diplomatie », et ainsi de suite pour tous les mots identifiés dans les articles du corpus d'entrainement.

La préparation de ces quatre catégories a donc été l'occasion de matérialiser des définitions et de rendre cette matérialisation communicable. Ces collections d'articles servent de base pour mesurer une similarité textuelle avec l'ensemble des articles contenant simplement le mot *europ**. Nous allons maintenant nous pencher sur les résultats de cette mesure de similarité.

3. La similarité lexicale comme brique de l'analyse discursive

Qu'en est-il des résultats de cette mesure de similarité ? La première information à retenir est la proportion d'articles annotés par cette mesure. Pour une première vérification manuelle, nous avons retenu les articles ayant reçu une mesure de 100%, c'est-à-dire que la probabilité d'appartenir à la catégorie mesurée est de 100%. Chaque article étant soumis séparément à la mesure de chaque catégorie (diplomatique, fédéraliste, utopique, antimoderne), il peut recevoir potentiellement une probabilité d'appartenir à chaque catégorie simultanément. Nous reviendrons sur cette dimension plus tard, mais dans un premier temps, il nous faut souligner qu'une partie réduite des articles ont été retenus pour être manuellement inspectés.

Dans la figure 1 ci-dessus, nous voyons que la proportion d'articles annotés est la plus forte en 1922, et la moins pour la période précédente. La quantité absolue reste en revanche la plus importante pour les années 1930 et 1940. Ceci signifie que les textes choisis pour incarner les idéaux-types ressemblent moins aux textes des années 1900-1920. Cette défection de l'utilisation

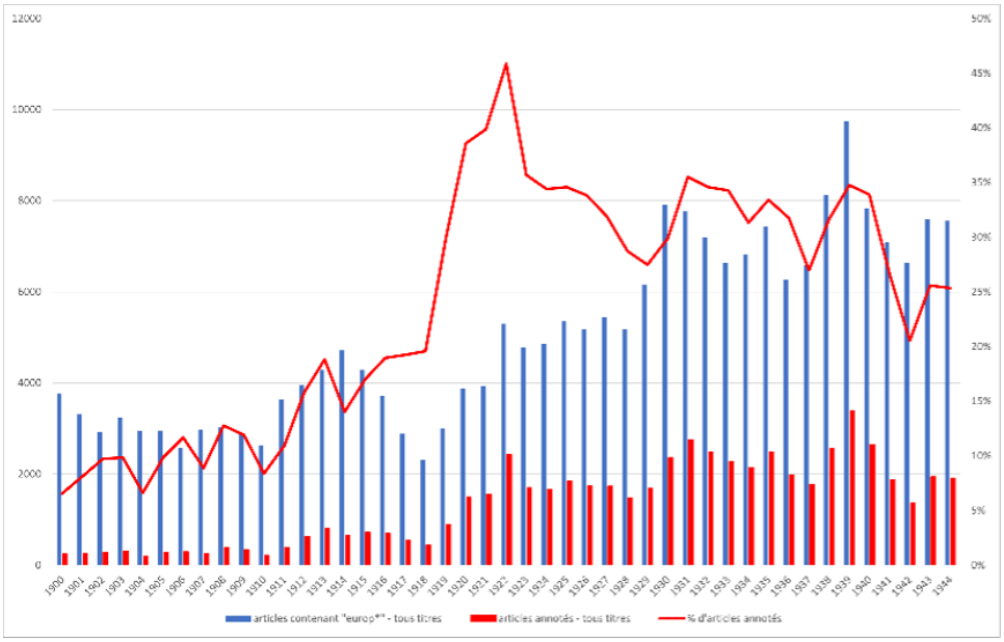

Figure 1: Distribution chronologique en valeurs absolue et relative des articles annotés, pour tous les titres de presse sélectionnés. La colonne bleue représente la totalité des articles collectés, contenant le mot-clé « europ* », la colone rouge, la totalité des articles annotés par au moins un classificateur, tandis que la courbe rouge indique la proportion des articles annotés dans les articles collectés, le tout par année. Par exemple, pour l'année 1944, un peu moins de 8000 articles ont été collectés, un peu moins de 2000 articles ont été annotés, correspondant à environ 25% des articles collectés.

de l'outil peut tout de même nous indiquer une évolution de vocabulaire à ce moment et appeler à une itération supplémentaire reflétant ces spécificités chronologiques, en utilisant des articles sélectionnés dans la même décennie où la similarité sera recherchée. Nous avons opté pour une recherche de similarité tout au long de la période, pour justement forger un point de repère commun pour ces cinq décennies. Toujours dans cette optique de vue d'ensemble, regardons à présent la distribution par titre des annotations.

Dans la figure 2, nous pouvons voir les degrés de similarité des articles retenus pour définir chaque classificateur avec les articles collectés pour chaque titre, contenant le terme « europ* », ce que nous voudrions utiliser pour déterminer la présence de discours respectivement utopique, antimoderne, diplomatique ou fédéraliste sur l'Europe. On remarque tout de suite que le point de référence commun produit une image très différenciée de la similarité pour chaque titre. On ne peut pas en conclure une domination d'un discours antimoderne pour l'Essor ou diplomatique pour le Journal de Genève, mais on peut considérer ce résultat comme indicatif d'une différence de ton, de vocabulaire utilisé dans ces différents titres (tout en gardant à l'esprit que ces articles ne sont qu'une partie des articles parus à ce moment dans ces titres). On voit aussi que les titres qui sont plus représentés dans la collection d'articles idéaux-types, ont une plus grande proportion d'articles annotés à 100%. Notamment, la Sentinelle score remarquablement faiblement, ce qui appelle à une itération supplémentaire nécessaire, reflétant plus nettement ce titre dans les

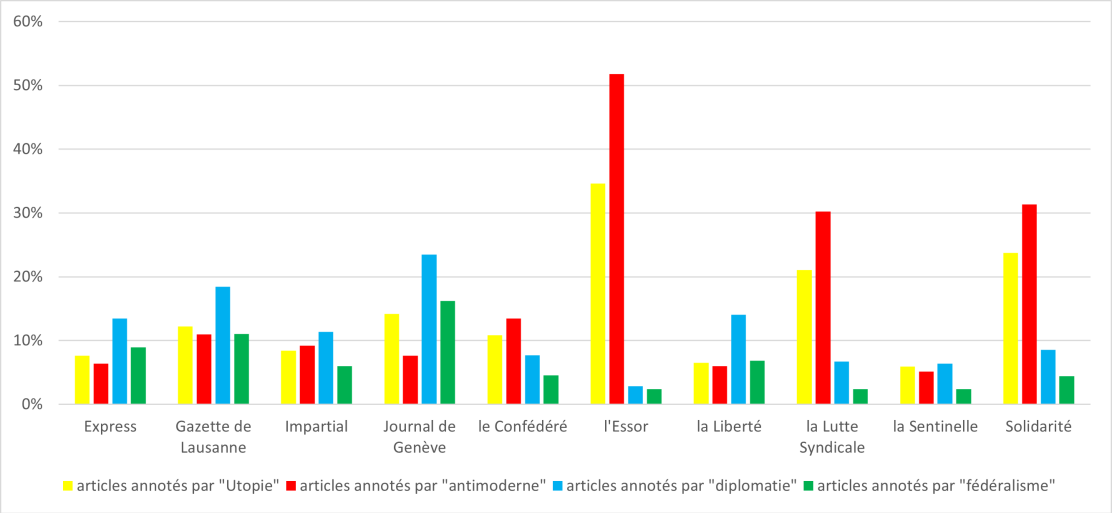

Figure 2: Proportion des annotations d'articles par les quatre classificateurs (utopique, antimoderne, diplomatique, fédéralisme) dans la totalité des articles collectés, par titre.

idéaux-types, et en particulier ses contenus d'utopie socialiste. D'autres titres, tels Solidarité, qui sont également faiblement présents dans la collection initiale, sont réceptifs à la mesure de similarité (ou du moins dans des proportions comparables ou supérieures aux titres mieux représentés, tels la Gazette de Lausanne). On remarque aussi nettement une présence forte de l'annotation « utopique » lorsque l'annotation « antimoderne » est forte, conformément à nos attentes concernant le ton, un peu plus surprenante du point de vue du contenu. Il semble que le classificateur soit plus sensible au style ou ton qu'au contenu. Il nous faut maintenant revenir à la distribution dans le temps de ces annotations, d'abord tous titres confondus, ensuite, nous avons sélectionnés quelques titres présentant des distributions distinctives.

L'espoir de ce travail de collecte et de mesure de similarité était que l'utilisation de ces exemples divers, dont serait extrait automatiquement le vocabulaire saillant, pourrait aider à identifier des articles contenant des discours similaires. Pour nous aider à traiter la masse qui, malgré sa réduction drastique, reste importante (de 227 351 à 60 348 articles), nous avons de nouveau recours aux annotations du topic modelling, présenté plus tôt. Le recroisement des quatre annotations du classificateur et du topic modelling permet de constituer des petits groupes d'articles, au sein de chaque titre, dont le contenu partage a priori certaines caractéristiques de thème et de style. L'accumulation de ces annotations sont utilisées comme faisceaux d'indice pour guider la lecture et l'analyse, avec toutes les précautions mentionnées au cours de cette présentation.

Ainsi, pour le Confédéré, le *topic* 32 est décrit par des mots « européenne souveraineté droit etats briand l'union fédération droit-international l'union-européenne… » et regroupe 14 articles contenant au minimum 30% de ce *topic*. Il aurait déjà retenu notre attention en tant que tel, mais les annotations invitent à interroger la diversité interne de ce groupe d'article : avec 5 articles annotés de « fédéralisme », « antimoderne » et « utopique », tandis que 8 sont annotés comme « diplomatique », « fédéralisme » et « utopique ».

Si les groupes d'articles du Confédéré restent de taille raisonnable, ceux de l'Express sont plus volumineux et les indices apportés par les classificateurs, plus utiles dans ce contexte. On peut ici utiliser cette accumulation d'annotations pour sélectionner par exemple, les *topics* qui contiennent la plus forte proportion d'annotation « antimoderne » par exemple, ou sélectionner parmi les autres qui paraissent pertinent, ceux qui ont cette même annotation. Par exemple, le *topic* décrit par « prix domicile suisse abonnements » n'a reçu aucune annotation d'aucun classificateur. Sa description avait déjà permis de potentiellement exclure de la vérification les 1706 articles qu'il regroupe, mais cette absence d'annotation permet de confirmer la probabilité que ces articles réfèrent à l'Europe comme simple espace de tarification pour des abonnements. Il en va de même pour les *topics* « bourse banque neuchât crédit suisse » et « concert musique disques », regroupant respectivement 1357 et 1280 articles. Par contraste, le *topic* « gouvernement une londres aux France », rassemblant 1271 articles, qui pouvait également sembler couvrir des thématiques diverses, ne contient aucun article annoté à 100% comme « antimoderne » mais 582 articles annotés comme « diplomatiques » et « fédéralistes », qui pourraient être vérifiés. Plus utile encore, le *topic* décrit par ces mots « faire politique lui contre pays guerre » et comptant 2300 articles, compte 389 articles annotés comme « antimodernes ».

Petit à petit, par cette sélection guidée par les annotations et les hypothèses que leur préparation a soulevé, s'ajoutant aux questions initiales, nous pouvons accumuler ces groupes d'articles, qui peuvent rassembler des articles venant d'une rubrique récurrente ou d'une chronique qui est ainsi identifiée, et dont le groupe initial pourra être élargi, ou autour d'événements, comme une discussion parlementaire, ou des rituels politiques, tel le tir fédéral. Certains groupes d'articles ne s'avèrent pas être caractérisés une homogénéité reconnaissable à l'inspection et sont écartés.

Cette démarche aide cependant à dépasser la citation anecdotique et mesurer plus clairement, de manière plus transparente, comment un discours se propage dans un corpus de textes d'archives, comment il cohabite, se superpose à d'autres discours et au contraire, comment il se distingue d'autres. En l'espèce, de la fabrication des catégories et des premières analyses de ces résultats naissent la nécessité de construire une autre mesure de similarité avec des sous-catégories jusqu'à présent intégrées aux quatre catégories étudiées jusqu'alors. Il semble aussi plus facile de distinguer les articles rapportant les événements diplomatiques ou liés aux efforts de coopération institutionnelle européenne des discours à visée utopique ou contenu antimoderne, que de distinguer ces deux dernières catégories. Et finalement, il apparaît que des textes au contenu antimodernes, mentionnant l'Europe au détour d'une phrase, sont plus fréquents que des discours visant à promouvoir une conception antimoderne de l'Europe.

Remerciements

Ce travail a pu être mené grâce aux efforts de numérisation des bibliothèques nationales au Luxembourg et en Suisse, au projet *impresso*[4], aux plateformes de popularisation des outils de traitement de textes, et finalement un script né de longs échanges avec Milan van Lange, chercheur au NIOD (NL)[5].

[4] https://impresso-project.ch/app/
[5] https://www.niod.nl/en/staff/milan-van-lange

Références

[1] D. Nguyen, M. Liakata, S. DeDeo, J. Eisenstein, D. Mimno, R. Tromble, J. Winters, How we do things with words: Analyzing text as social and cultural data, Frontiers in Artificial Intelligence 3 (2020). URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00062/full#h3. doi:10.3389/frai.2020.00062.

[2] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84. URL: https://dl.acm.org/doi/10.1145/2133806.2133826. doi:10.1145/2133806.2133826.

[3] E. Bunout, M. Ehrmann, F. Clavert (Eds.), Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization, De Gruyter Oldenbourg, 2022. URL: http://www.degruyter.com/document/isbn/9783110729214/html?pds=422022164415299843667093853399 6, publication Title: Digitised Newspapers – A New Eldorado for Historians?

[4] J. Grimmer, B. M. Stewart, Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, Political Analysis 21 (2013) 267–297. URL: https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20. doi:10.1093/pan/mps028.

[5] E. Bunout, Grasping the anti-modern discourse on europe in the swiss digitised press, or can text mining generate a research corpus from an article collection?, Journal of Open Humanities Data 7 (2021) 21. URL: http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.37/. doi:10.5334/johd.37, number: 0 Publisher: Ubiquity Press.