Théia

1 | 2024
Retours d'expérience en édition numérique de sources en histoire et histoire de l'art

La reconnaissance d'entités nommées dans les éditions numériques à l'exemple du récit de voyage du pharmacien Wagener

Recognition of named entities in digital editions: the example of pharmacist Wagener's travel account

Maximilian Görmar

Traduction de Angela Göbel

http://publications-prairial.fr/theia/index.php?id=144

DOI : 10.35562/theia.144

Référence électronique
Maximilian Görmar, « La reconnaissance d'entités nommées dans les éditions numériques à l'exemple du récit de voyage du pharmacien Wagener », *Théia* [En ligne], 1 | 2024, mis en ligne le 17 avril 2025, consulté le 15 septembre 2025. URL : http://publications-prairial.fr/theia/index.php?id=144

Prairial
Pôle éditorial
Lyon Saint-Étienne

La reconnaissance d'entités nommées dans les éditions numériques à l'exemple du récit de voyage du pharmacien Wagener

Recognition of named entities in digital editions: the example of pharmacist Wagener's travel account

Maximilian Görmar

Traduction de Angela Göbel

PLAN

NOTES DE L'AUTEUR

Le présent texte est issu d'une conférence donnée lors de l'atelier « Ouverture des récits de voyage historiques : mise en forme du texte, modélisation des données et visualisation » (20-21 juillet 2023), organisé par le projet « Éditions numériques des récits de voyage historiques » (Digitale Editionen Historischer Reiseberichte, DEHisRe) à l'Institut Leibniz de recherche sur l'Europe de l'Est et du Sud-Est (IOS) à Ratisbonne. Mes remerciements les plus chaleureux vont aux organisatrices Anna Ananieva et Sandra Balck, ainsi qu'à toute l'équipe du projet, pour leur invitation à intervenir, ainsi qu'à tous les participants pour les discussions stimulantes et les échanges. À l'origine rédigé en allemand, le texte a été publié en premier lieu sous la forme d'un billet de blog dans le blog associé au projet « Grand Tour digital. Numérisation, exploitation et visualisation des témoignages autobiographiques des voyages éducatifs de l'époque moderne grâce à l'application de procédés éditoriaux partiellement automatisés » à la Bibliothèque Herzog August de Wolfenbüttel, (https://grandtourdig.hypotheses.org/949). Pour cette publication, le texte a été légèrement étendu. Je tiens à exprimer ma gratitude à Angela Göbel pour avoir accueilli et traduit l'article en français.

TEXTE

Introduction

1 Les récits de voyage, en particulier ceux relatant des voyages éducatifs, constituent des témoignages importants qui peuvent fournir des informations sur les expériences et la vie de leurs auteurs. Ils documentent les rencontres avec de nouveaux lieux et personnes, les coutumes et les traditions, offrant ainsi un aperçu de l'appropriation mentale et matérielle ainsi que du traitement de l'étranger. Ils éclairent ainsi les processus de formation de soi et de la personnalité, souvent négligés jusqu'à présent par la recherche sur les témoignages personnels [2]. Particulièrement pour l'époque moderne, de vastes collections de cette catégorie de sources n'ont pas encore été suffisamment explorées ou analysées, telles que celles conservées à la Bibliothèque Herzog August de Wolfenbüttel [3]. Le projet d'édition « Grand Tour digital. Digitalisierung, Erschließung und Visualisierung frühneuzeitlicher Selbstzeugnisse von Bildungsreisen unter Anwendung teilautomatisierter Editionsverfahren » (Grand Tour digital. Numérisation, exploration et visualisation de témoignages personnels de voyages éducatifs de l'époque moderne en utilisant des procédés d'édition partiellement automatisés) part de ce problème fondamental et tente, comme le suggère déjà le titre complet, d'explorer également de nouvelles voies méthodologiques dans l'édition numérique en appliquant des procédures éditoriales partiellement automatisées [4]. Cela englobe d'une part l'utilisation de la reconnaissance automatique de l'écriture manuscrite (*Handwritten Text Recognition*, HTR) avec le programme Transkribus [5], et d'autre part, le traitement et la mise en évidence des textes identifiés avec des outils logiciels de reconnaissance des entités nommées (*Named Entity Recognition*, REN). L'étape suivante est ensuite de créer une liaison des entités nommées identifiées avec des données normalisées et d'autres sources de données. Pour ce deuxième aspect, la REN, les récits de voyage se prêtent particulièrement bien à l'étude de cas, car leur

densité en termes de noms de personnes et surtout de noms de lieux est particulièrement élevée pour alimenter une base de données.

Fig. 1 : Les voyages du pharmacien en formation Wagener, 1652–1659.

2 Ce type de travail sera montré à travers l'exemple d'un texte du projet « Grand Tour digital », rédigé par le jeune pharmacien Johannes Wagener entre 1652 et 1659 [6]. Wagener entreprit un voyage de compagnon (1652–1655), alors courant pour les futurs apothicaires à l'époque moderne [7], ainsi qu'un voyage de formation (1658/59) en compagnie du duc Ferdinand Albrecht I[er] de Brunswick-Wolfenbüttel-Bevern (Fig. 1) [8]. Le premier voyage le conduisit à travers le nord de l'Allemagne jusqu'à Königsberg en Prusse (aujourd'hui Kaliningrad), tandis que le second le mena à travers l'ouest et le sud de l'Allemagne, ainsi que la Suisse, jusqu'en France, où il décéda en 1659 à Angers des suites d'une fièvre ardente (« hitzigen Fieber ») [9]. Parallèlement, il entreprit également plusieurs voyages plus courts vers 1656/57, notamment au château de Gottorf, où il visita son célèbre cabinet de curiosités et put voir ainsi le globe géant qui s'y trouvait [10]. L'ampleur des informations géographiques, sociales et linguistiques du texte de Wagener ouvre de vastes possibilités, voire des défis, dans l'utilisation des méthodes de reconnaissance des entités nommées (REN) et dans le traitement des textes de voyage de

l'époque moderne. Celles-ci seront retracées ci-après en se basant sur les expériences et le flux de travail du projet « Grand Tour digital ».

Situation de départ

3 Tout d'abord, partons d'une définition générale : « La reconnaissance d'entités nommées (REN) désigne la détection des noms propres (entités nommées) dans les textes, ainsi que leur classification en différents types d'entités. […] par défaut, les personnes, les lieux et les organisations font partie de ces types d'entités » [11]. Il existe fondamentalement deux types de méthodes qui sont utilisées à cet effet : des approches basées sur des règles ou des connaissances, reposant par exemple sur des vocabulaires contrôlés ou des *gazetteers* (index géographique), et des approches basées sur l'apprentissage automatique. Ainsi, la REN peut être associée de manière similaire à la reconnaissance d'écriture manuscrite dans le vaste domaine de l'apprentissage automatique et de l'intelligence artificielle, auquel appartiennent également des outils tels que Chat-GPT, actuellement très populaires [12].

4 Contrairement aux grands modèles de langage (*Large Language Models*, LLM) avec des millions à des milliards de tokens et de paramètres, sur lesquels reposent Chat-GPT et d'autres outils de traitement du langage naturel pour les langues actuelles, les modèles correspondants et les ensembles de données disponibles pour les textes anciens sont beaucoup plus rares et plus petits.
Ainsi, Ehrmann *et al.*, dans un état des lieux récemment publié sur la REN dans de tels documents ont identifié 22 corpus annotés avec des données d'entraînement par rapport à 121 corpus pour les langues contemporaines [13]. Les données d'entraînement correspondent au corpus choisi pour étalonner la recherche et permettre ensuite au modèle de fonctionner. De plus, les modèles linguistiques pour les textes anciens, en raison de leur taille plus réduite et des défis spécifiques des stades linguistiques antérieurs sur lesquels nous reviendrons plus en détail, présentent généralement une fiabilité moindre (score F1 : 60-70 %) par rapport à ceux pour les langues actuelles (score F1 : généralement supérieur à

90 %)[14]. Bien entendu, même parmi eux, d'importantes différences subsistent en fonction des domaines spécifiques[15].

5 La situation en ce qui concerne la REN dans les textes anciens est donc certainement perfectible, mais il y a eu ces dernières années plusieurs initiatives et projets qui utilisent notamment des éditions numériques pour générer des ensembles de données d'entraînement correspondants. On peut citer, par exemple, le projet NERDPool en cours depuis 2020, porté par l'Académie autrichienne des sciences ainsi que les universités d'Innsbruck et de Graz, qui a rassemblé des ensembles de données d'entraînement pour la REN en allemand ancien à partir de six collections de textes ou éditions allant du XVI^e^ au début du XX^e^ siècle[16]. En outre, il existe des projets plus modestes, tels que ceux réalisés dans le cadre de mémoires de master ou de séminaires pratiques, où ont été élaborées des données d'entraînement grâce à de tels modèles et des tutoriels pour la REN dans les textes anciens[17]. Il s'agit donc d'un domaine de recherche extrêmement dynamique, particulièrement pour les sources du début de l'allemand contemporain et grâce aux récits de voyage[18].

Workflow et problèmes actuels

6 Dans ce contexte globalement positif, nous avons commencé à élaborer un travail pour la REN qui fait suite à l'établissement du texte intégral avec Transkribus. Tout a commencé par le choix d'un outil approprié pour la REN, avec l'évaluation de trois candidats au total : Stanford Named Entity Recognizer[19], WebLicht[20] et la bibliothèque Python appelé spaCy[21]. Les trois outils sont basés sur des systèmes techniques spécifiques, offrant chacun des avantages et des inconvénients ainsi que des approches différentes pour relever les défis des textes de cette époque et de leur édition numérique. Avec WebLicht, une sorte de boîte à outils en ligne pour le traitement du langage naturel (NLP), il est par exemple possible de compenser la variation linguistique des textes anciens par rapport à l'allemand moderne en normalisant les textes avant la REN à l'aide d'un outil spécialisé[22]. Une autre approche pour adapter le logiciel au matériel textuel ancien et à ses particularités consiste à former ses propres modèles de REN à l'aide de l'apprentissage automatique. Cette méthode est possible à la fois avec le Stanford Named Entity

Recognizer et avec spaCy. Cependant, avec le Stanford Named Entity Recognizer, ainsi qu'avec WebLicht, il est relativement complexe et fastidieux de traiter les données de transcription disponibles en TEI-XML [23], notamment pour que les annotations déjà effectuées restent telles quelles et que le résultat en sortie soit à nouveau en XML et conforme à la norme TEI de manière à inclure les entités reconnues et annotées par le programme. C'est pourquoi le choix s'est finalement porté sur spaCy, car grâce à cet outil il est non seulement relativement facile et rapide de mettre en œuvre un flux de travail pour la REN, mais aussi parce que la plupart des projets ayant travaillé sur l'application de la REN aux éditions numériques ont utilisé spaCy [24]. Les solutions aux problèmes déjà développées et documentées, sur lesquelles nous reviendrons plus tard, peuvent être réutilisées, ce qui non seulement représente un gain de temps considérable, mais est également conforme à une pratique de recherche durable dans le domaine de l'édition numérique.

7 SpaCy est une bibliothèque Python open source dédiée au traitement du langage naturel (NLP). En plus d'une gamme complète de scénarios d'utilisation en linguistique informatique (comme la tokenisation, la segmentation de phrases, le marquage grammatical), la bibliothèque comprend également des composants pour la REN et offre la possibilité de former des modèles personnalisés à cet effet. Parallèlement, ces modèles basés sur l'apprentissage automatique peuvent être combinés avec des listes de mots et des motifs de reconnaissance d'entités, intégrant ainsi une approche basée sur la connaissance ou sur des règles. Ses nombreuses fonctionnalités et sa facilité d'utilisation relative ont conduit, comme mentionné précédemment, à son utilisation dans plusieurs projets liés aux éditions numériques pour la NER. De là sont également nés quelques packages Python pour l'extraction de données à partir de textes annotés en TEI et pour l'intégration des résultats d'analyse, notamment l'acdh-spacytei [25] et le Standoff Converter [26].

8 Avant même de choisir un outil de REN, la décision avait été prise de sélectionner un texte pour les essais et l'entraînement. Par souci de simplicité, les premières pages du compte rendu de voyage de Wagener ont été choisies, environ 20 pages qui devaient de toute façon être transcrites manuellement pour l'entraînement de Transkribus [27]. Ainsi, sans trop d'effort supplémentaire, il a été

possible d'annoter les noms propres et les entités, formant ainsi la base de l'ensemble d'entraînement. Au total, 248 entités ont été étiquetées, dont 92 personnes, 148 lieux et 8 organismes. Il s'agit donc d'un ensemble d'entraînement relativement petit (en particulier pour les organismes), mais il semblait suffisant pour une première tentative et pour définir un modèle. Comme l'objectif initial était d'annoter le reste du compte rendu de voyage de Wagener avec cet ensemble, nous pouvions supposer que le matériel d'entraînement et le matériel cible présentaient une similarité suffisante pour obtenir des résultats acceptables.

9 Avant l'entraînement proprement dit, le texte devait être préparé d'une manière à pouvoir être traité. Ce fut le premier défi, car il s'est avéré initialement moins évident de convertir le texte TEI en un format JSON adapté à spaCy. En particulier, l'extraction de la position exacte des entités nommées s'est révélée initialement sujette à des erreurs en utilisant un script xslt, nécessitant ainsi un contrôle minutieux et chronophage [28]. Cependant, ce problème a pu être largement résolu en réutilisant des solutions et des packages de programmes existants provenant d'autres projets, comme mentionné précédemment. Après l'élaboration du flux de travail à partir d'un tutoriel d'Isabel Hansen issu de son mémoire de master rédigé à l'Université de Trèves, il a été assez simple de travailler avec le package Python acdh-spacytei [29], pour mettre le texte TEI dans la forme nécessaire pour l'ensemble d'entraînement. Ensuite, les données ont dû être divisées en ensembles d'entraînement, de validation et de test [30].

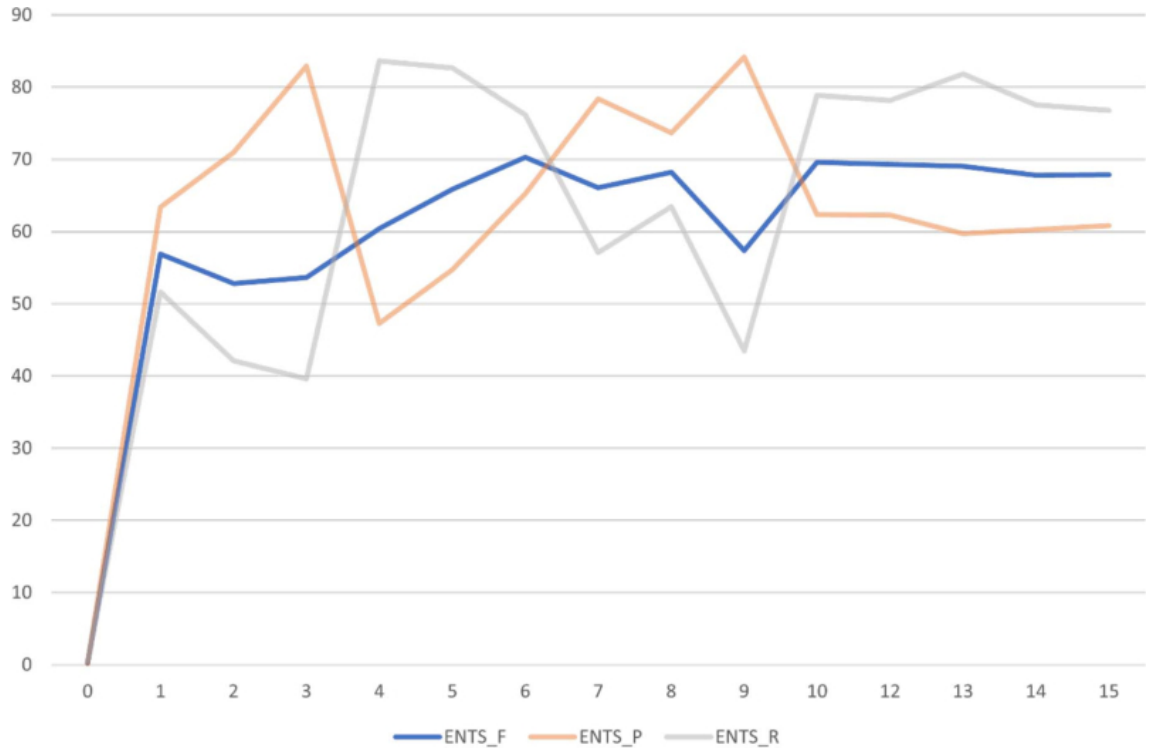
10 Ensuite, l'entraînement réel a eu lieu, ce qui peut parfois prendre beaucoup de temps. Au cours de plusieurs itérations, un réseau neuronal a été formé de manière à pouvoir prédire, en quelque sorte, quel mot ou quelle chaîne de caractères pouvait devenir une entité nommée. Cela se fait pratiquement par essais et erreurs, en attribuant d'abord aléatoirement des étiquettes aux entités nommées et en les comparant avec les annotations dans les données d'entraînement. Les erreurs qui surviennent sont mesurées et leur poids dans le réseau neuronal est ajusté en conséquence. Cela se poursuit jusqu'à ce qu'il n'y ait plus d'amélioration significative du modèle [31]. Les principales métriques utilisées pour évaluer ce processus sont la précision, le *recall* (rappel) et le score F1 calculé à

partir de ceux-ci [32]. Un aperçu de l'évolution de ces métriques au cours de l'entraînement est facilement compréhensible dans spaCy grâce à une présentation sur un tableau simple. Pour une meilleure visualisation, le déroulement de l'entraînement peut également être représenté sous forme de diagramme, montrant comment le score F1 se stabilise progressivement autour de 70 (Fig. 2 et 3). Bien que cette valeur semble relativement basse à première vue, on ne peut pas attendre trop du modèle formé. Cependant, il se situe dans la fourchette de valeurs couramment rencontrées dans la littérature pour des ensembles de textes anciens. En tenant compte de la petite taille de l'ensemble d'entraînement, la performance est en elle-même relativement bonne.

Fig. 2 : Vue d'ensemble de spaCy dans l'entraînement d'un modèle REN. Représentation simplifiée dans un tableau.

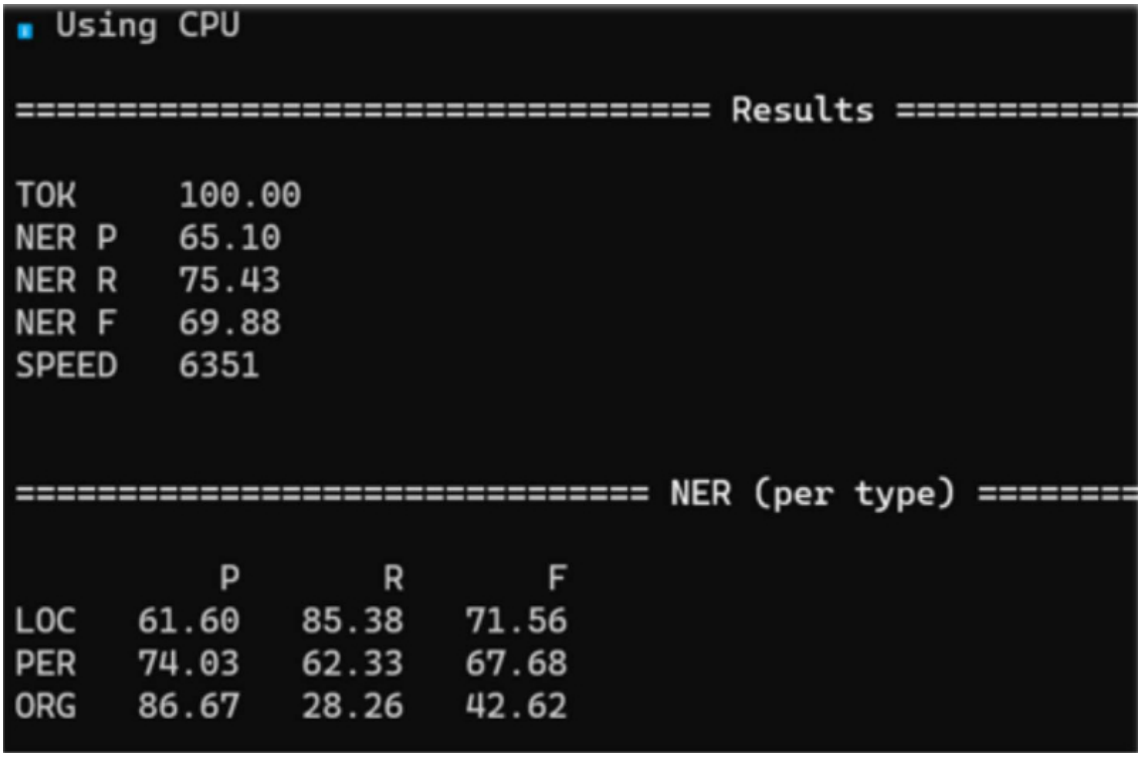
```
✓ Initialized pipeline

============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00    109.71    0.25    0.16    0.53    0.00
  1     200       1061.15   4155.34   56.90   63.41   51.61    0.57
  2     400         35.32   1680.59   52.84   70.96   42.09    0.53
  3     600         26.82   1498.22   53.63   82.91   39.64    0.54
  4     800        449.09   1410.32   60.40   47.28   83.60    0.60
  5    1000         34.84   1410.74   65.86   54.74   82.66    0.66
  6    1200         38.09   1273.08   70.27   65.22   76.18    0.70
  7    1400         44.37   1362.77   66.06   78.37   57.09    0.66
  8    1600         47.18   1360.55   68.19   73.69   63.46    0.68
  9    1800        200.29   1371.29   57.35   84.18   43.49    0.57
 10    2000         55.71   1265.83   69.62   62.32   78.87    0.70
 12    2200         63.78   1284.95   69.34   62.31   78.17    0.69
 13    2400         88.32   1200.16   69.06   59.74   81.84    0.69
 14    2600         88.65   1210.34   67.82   60.28   77.52    0.68
 15    2800         88.96   1217.98   67.89   60.85   76.77    0.68
✓ Saved pipeline to output directory
```

Fig. 3 : Visualisation des critères de qualité recall (ENTS_R), précision (ENTS_P) et f-score (ENTS_F) à partir du tableau ci-dessus au cours des itérations d'entraînement.



11 Une évaluation plus différenciée de la qualité du modèle peut se faire à l'aide des données du test. Il s'agit d'une partie des données annotées qui n'ont pas fait partie de l'ensemble d'entraînement. Avec elles, la précision, le *recall* et le score *F1* peuvent être calculés pour un texte inconnu, permettant ainsi d'évaluer le modèle. Là encore, spaCy fournit les valeurs correspondantes, cette fois-ci détaillées par type d'entité. Elles peuvent ainsi être facilement comparées entre elles (Fig. 4). Pour les personnes et les lieux, les valeurs sont d'environ 70, donc dans la fourchette du modèle global. On observe que la reconnaissance des lieux fonctionne légèrement mieux. En revanche, les organismes obtiennent des résultats nettement moins bons. Cela est lié à leur nombre nettement inférieur dans les données d'entraînement, et même pour les lieux et les personnes, les différences peuvent probablement être attribuées à leur fréquence respective dans l'ensemble d'entraînement.

Fig. 4 : Vue d'ensemble de spaCy des résultats de validation du modèle REN entraîné.



12 Après l'entraînement et l'évaluation, le modèle peut être appliqué au texte non annoté. Un extrait du résultat est visible dans la Fig. 5, qui montre la vue HTML générée avec displaCy, un module correspondant de spaCy [33]. Nous pouvons utiliser un simple fichier texte comme base pour cela, et le fichier HTML pourrait également être utilisé pour créer une version TEI rudimentaire du texte. Cependant, comme une partie de la balise d'origine serait perdue, le Standoff Converter de David Lassner a été utilisé [34]. Avec le convertisseur, le texte brut peut être extrait du XML, traité avec spaCy, et les nouvelles annotations (par exemple, les entités nommées) peuvent être insérées dans le document source sans que le balisage d'origine soit perdu. Le résultat est visible dans la Fig. 6.

Fig. 5 : Sortie HTML du résultat des REN (conçu avec displaCy).

Durch Durlach LOC eine feine Stadt vnd den March:

graffen von Baden LOC zugehörig, biß auff Stollhofen LOC

eine kleine hübsche Stadt auch denen von Baden LOC zuge:

horig. Alhier sahen wir den Hahnentantz zu;

welcher so langsahm, vnd mit fast zugesperten augen (maßen sie

keinen menschen ansehen mueßen) daß man kaum

siehet ob sie fortkommen oder nicht NB. Den 4. dieses

reiseten wir durch Hagenstege auff Saß

Strassburg LOC vnd kehrten zuem Bruder

hoff ein. NB. Peter Camehl schaffen. Den 4. durch hagenstege auff

ein Dorff Keul von Dannen vber der Rheinbru: LOC cken welche fast 1/4 Meil lang auf Straßburg LOC von

dieser Brückken ist hertzog Augusti Furst zu Br.

vnd Lb. H. Bruder PER mit Pferde her vnter gesturtzet vnd

ertruncken. . Straßburg eine schöne vnd alte

Reichsstadt LOC vnd weit berufffen nicht allein wegen

Fig. 6 : Sortie TEI du résultat des REN (conçu avec le Standoff-Converter).

```
</ab><pb facs="#facs_60" n="60" xml:id="img_0060"/><ab facs="#facs_60_r1">
   <lb facs="#facs_60_r111" n="N001"/>Die stadt <rs type="place">Franckfurt</rs> Jst eine sch&#246;ne handel stadt
   <lb facs="#facs_60_r112" n="N002"/>Ma&#223;en alle Jahr 2 gro&#223;e Me&#223;en allda gehalten werd[en]
   <lb facs="#facs_60_r113" n="N003"/>da dan viele 1000 Menschen zu handeln dahin komen
   <lb facs="#facs_60_r114" n="N004"/>der Mayn flie&#223;et mitten durch die stadt. Jst auch
   <lb facs="#facs_60_r115" n="N005"/>sonderlich beruemhbt wegen der Kayserlichen Wahl
   <lb facs="#facs_60_r116" n="N006"/>dadan sonderliche hauser vnd h&#246;ffe vor Jeden
   <lb facs="#facs_60_r117" n="N007"/>Churf&#252;rsten allezeit sein, darauff in Jhrer ankunfft
   <lb facs="#facs_60_r118" n="N008"/>zu wohnen. E&#223; hat 5 feine apothecken. Der elteste
   <lb facs="#facs_60_r119" n="N009"/>Jm hirsch H. Jacob Holtzapfel im Einhorn H. Sparr
   <lb facs="#facs_60_r1110" n="N010"/>im Kopff H. Banse im Schwan H. Saltzwedel
   <lb facs="#facs_60_r1111" n="N011"/>vnd im Engel H. Persebecher.
   <lb facs="#facs_60_r1112" n="N012"/>Oben an der Maynbrukke am thor siehet man
   <lb facs="#facs_60_r1113" n="N013"/>3 K&#246;pffe, derer, so vor diesen die stadt haben
   <lb facs="#facs_60_r1114" n="N014"/>verrathen wollen.
   <lb facs="#facs_60_r1115" n="N015"/>Den 29 Julij styl vet. haben Jhr F. G. <hi>He&#223;en</hi>
   <lb facs="#facs_60_r1116" n="N016"/>Ferdinand Albrecht hertzog zu Br. vnd <rs type="place">Luneburg</rs> mir
   <lb facs="#facs_60_r1117" n="N017"/>die gro&#223;e Gnade erwiesen, vnd von <rs type="place">Franckfurt</rs>
   <lb facs="#facs_60_r1118" n="N018"/>ab mit auf Jhre reise genommen, da wir dan
   <lb facs="#facs_60_r1119" n="N019"/>erwehnten dito auff H&#246;gst, ein klein stedlin
   <lb facs="#facs_60_r1120" n="N020"/>am Mayn lieget gereiset alhier ist ein altes zer
   <lb facs="#facs_60_r1121" n="N021"/>storetes schlo&#223;, den 30 dito haben wir vns da:
   <lb facs="#facs_60_r1122" n="N022"/>selbsten vber setzen la&#223;en vnd seind gefahren auf
   <lb facs="#facs_60_r1123" n="N023"/>Darmstadt worinnen ein schlo&#223; vnd F&#252;rstl.
   <lb facs="#facs_60_r1124" n="N024"/>Regierung. Alhier seind 3 apothecken, von hier
   <lb facs="#facs_60_r1125" n="N025"/>fuhren wir auf Everstadt ein Marckflecken
```

13 Ainsi, on distingue ici un procès rudimentaire pour la REN dans les documents TEI, et le résultat est réutilisable et peut être appliqué aux textes que nous traiterons encore dans le projet. Cependant, il reste quelques problèmes ouverts et du travail à faire. Le plus grand problème est certainement que le taux de reconnaissance avec le modèle développé jusqu'à présent peut encore être amélioré. Cela est dû à la fois à la petite taille de l'ensemble d'entraînement, mais aussi aux particularités linguistiques des sources. Ils posent un défi aux outils de traitement automatique du langage naturel (NLP) qui sont conçus pour des langues contemporaines relativement normalisées. En effet, la langue au XVII^{e} et XVIII^{e} siècle présente de grandes variations orthographiques, des différences de vocabulaire, de morphologie et de structure syntaxique par rapport à l'allemand actuel. On trouve également dans les textes anciens un multilinguisme récurrent [35]. Quelques mots en latin, français, italien ou d'autres langues étaient généralement intégrés sans distinction typographique. Souvent, des passages entiers en une ou plusieurs langues étrangères étaient également intégrés dans un texte principalement allemand. Il s'agissait fréquemment de citations étrangères utilisées par les voyageurs qui se formaient à l'apprentissage des langues étrangères [36]. Surtout, les noms propres étaient régulièrement traduits dans d'autres langues, sans parler du fait qu'il n'y avait pas de conventions uniformes pour leur écriture, y compris chez un même auteur. Par ailleurs, l'utilisation de majuscules n'était pas appliquée de manière cohérente à cette époque [37].

14 Cette incertitude sur les noms propres entraîne alors les deux autres problèmes : l'identification et la désambiguïsation des entités, ainsi que leur liaison avec des données normatives telles que Geonames [38], GND [39], VIAF [40] ou encore Wikidata [41]. Pour ces raisons, on recueille souvent, en plus d'une entrée principale, des formes de noms historiques alternatives ou des formes dans d'autres langues. Une source de ces formes de noms peut être des éditions (numériques), ce qui nous amène à notre troisième point, « Perspectives ».

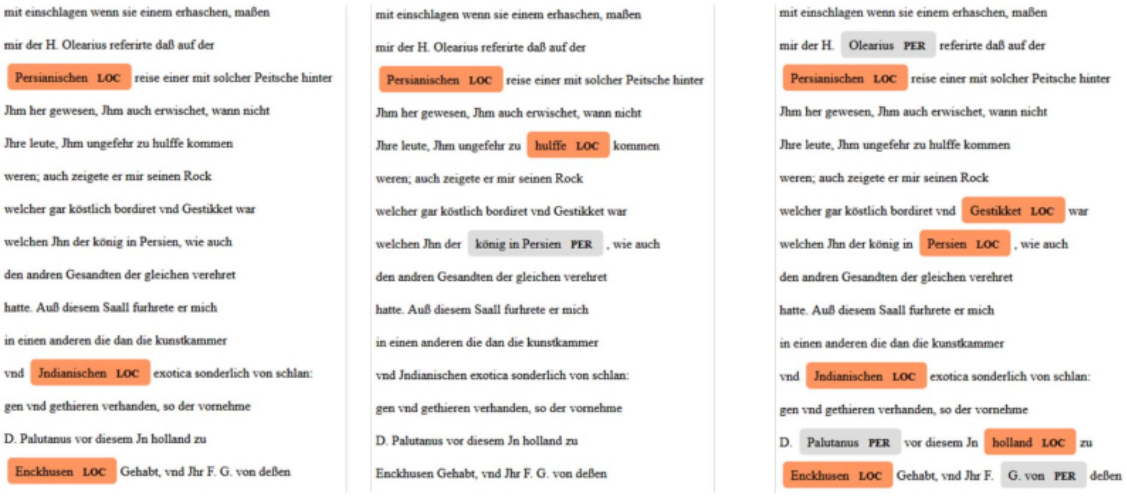
Perspectives

15 Quelles sont les solutions possibles pour les problèmes mentionnés ? Nous voyons surtout deux approches : l'élargissement de la base de

données pour l'entraînement de modèles linguistiques historiques et la combinaison de ces modèles basés sur des algorithmes d'apprentissage automatique avec des approches basées sur des règles ou des connaissances. Pour cette dernière approche, il est nécessaire de développer des bases de données ou des *gazetteers*, comme c'est déjà partiellement le cas. Ils peuvent et doivent certainement être étendus et complétés dans certains domaines, ce qui justifie des éditions numériques annotées. Dans les lignes qui suivent, nous allons brièvement présenter nos expériences faites jusqu'à présent dans ce domaine du projet et esquisser les perspectives qui en découlent.

16 L'entraînement de la REN peut être adapté à l'élargissement de la base de données d'autant plus qu'il existe déjà des collections de données similaires qui peuvent potentiellement être réutilisées. Un exemple concret serait le projet NERDPool mentionné précédemment [42]. Cependant, les tentatives menées jusqu'à présent pour entraîner des modèles plus importants avec des éditions de la Herzog August Bibliothek n'ont conduit qu'à des résultats légèrement meilleurs pour la REN. En plus des problèmes déjà formulés inhérents aux sources, des défis supplémentaires sont apparus en raison d'approches éditoriales différentes en fonction des sources écrites [43]. Ces défis concernent les différentes pratiques de balisage par rapport aux entités nommées, les différents niveaux de normalisation du langage source et les différences en ce qui concerne la segmentation des textes, ce qui peut être intégré en tant qu'information contextuelle dans le modèle entraîné [44]. Cela rend nécessaire une adaptation partiellement manuelle des textes et de leur codage, ce qui représente une charge de travail non négligeable. Par conséquent, il semble judicieux de ne pas simplement intégrer de nombreux textes dans un ensemble d'entraînement, mais plutôt de créer des échantillons représentatifs et spécifiques au domaine (par exemple, des récits de voyage) codés selon des normes uniformes. De plus, il peut être utile de soumettre les données textuelles à d'autres opérations de traitement automatique du langage naturel avant l'entraînement, telles que la division en phrases individuelles. Cela peut également réduire le temps de calcul nécessaire pour créer un modèle.

Fig. 7 : Sortie HTML pour différents modèles NER avec et sans découpage de phrases.

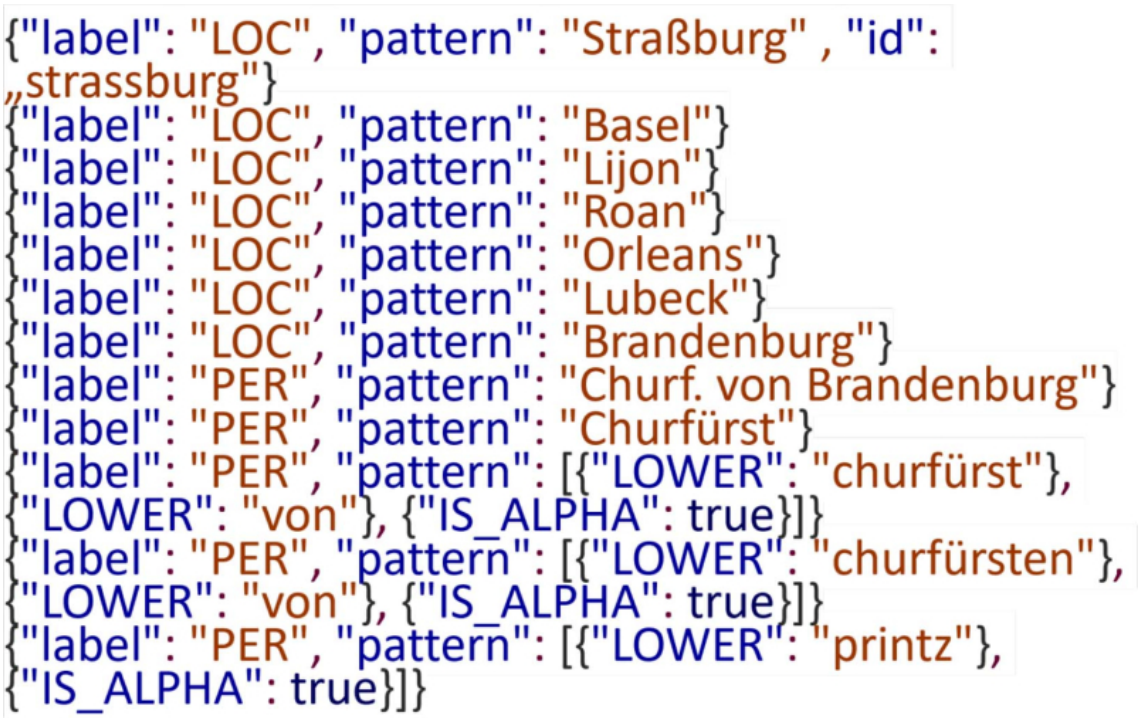


17 Un contrepoint comparatif des résultats obtenus jusqu'à présent, qui intègre déjà en partie ces considérations, est visualisé dans la Fig. 7 avec displaCy. À gauche on trouve le modèle de base. Au milieu, c'est le modèle qui entraîne le texte de Wagener combiné avec des journaux du duc Auguste le Jeune de Brunswick-Wolfenbüttel [45] et du prince Christian II d'Anhalt-Bernburg [46]. À droite, c'est un modèle basé sur Wagener et où le texte a de nouveau été divisé en phrases avec spaCy avant l'entraînement (ce qui repose bien sûr sur des modèles linguistiques pour des langues actuelles et est donc sujet aux erreurs correspondantes) [47].

18 L'intégration de *gazetteers* et de modèles de reconnaissance pour la REN est relativement facile à faire dans spaCy. Il suffit de mettre les données correspondantes, par exemple les formes de noms de lieux, dans un format JSON simplifié basé sur des lignes et de les intégrer dans le *pipeline* de traitement. Théoriquement, il est possible de construire un flux de travail REN uniquement à l'aide d'une telle liste [48]. Il serait alors possible de comprendre et de contrôler ce qui est reconnu comme entité et ce qui ne l'est pas. Ce n'est pas le cas avec les réseaux neuronaux auto-apprenants. Ils s'apparentent à une boîte noire à l'intérieur de laquelle même les développeurs ne peuvent pas regarder ou de manière très limitée. Il est aussi possible de réutiliser les ressources existantes telles que les éditions numériques. Et il est possible d'attribuer des identifiants, ce qui

permet la liaison avec des registres et des données normatives déjà existants (Fig. 8). Cela permettrait également d'automatiser dans une certaine mesure le lien avec les entités nommées, qui doit suivre la REN lors de la création de l'édition numérique [49].

Fig. 8 : extrait d'un fichier JSONL avec les modèles et notions pour la REN.



19 Les différentes formes de noms induisent l'essentiel des inconvénients. Dans la mesure où les personnes n'apparaissent souvent que dans des contextes spécifiques et où la variabilité des noms est encore plus élevée, il semble judicieux d'interroger des termes fixes surtout pour les lieux et les organismes, qui sont plus ou moins des entités de longue durée. Mises à part quelques personnalités importantes (Jésus-Christ, Luther, Calvin, Charlemagne etc.), il est plus judicieux pour les personnes de définir des modèles plus généraux, comme avec « Kurfürst von… » (prince-électeur de…). Cependant, pour les membres de la haute noblesse ou les souverains, le problème des entités dites « imbriquées » se pose souvent, où une partie du nom peut être simultanément un toponyme [50]. De plus, les indications de lieu peuvent être relatives à l'emplacement de l'écrivain et donc être fortement dépendantes du contexte, de sorte que des mots tels que « ici », « en cet endroit », « là-bas » peuvent déformer

la REN s'ils ne sont pas éliminés de la base de données. Il en va de même pour les pronoms personnels, qui peuvent également faire référence de manière contextuelle aux personnes. Malgré ces inconvénients, l'utilisation de *gazetteers* ou de listes de données peut améliorer de manière décisive la REN, en particulier lorsque peu de données d'entraînement sont disponibles.

Conclusion

20 Il est devenu évident que la REN est une composante exigeante mais judicieuse de la boîte à outils offerte par le traitement automatique du langage naturel pour l'élaboration d'éditions numériques. Cependant, il est important d'éviter la fausse allégation répandue selon laquelle « Big Data supposedly lets you get away with dirty data » [51]. Bien que nous ayons besoin de plus de données pour l'entraînement des modèles linguistiques anciens, notamment ceux de la REN, cela ne signifie pas nécessairement l'usage de *big data*. Des échantillons judicieusement choisis et soigneusement organisés devraient conduire à de bons résultats. Cela ne rend pas l'éditeur et le chercheur humains superflus, mais cela leur apportera un soutien significatif dans un avenir proche.

21 En outre, les problèmes liés à la REN ne doivent pas être considérés de manière isolée. Ils sont notamment liés à la modélisation des données et à l'enrichissement sémantique des éditions numériques, en particulier dans le domaine des récits de voyages anciens. Les défis mentionnés précédemment lors de la conversion des données textuelles et d'annotations vers et depuis le TEI-XML illustrent cette complexité. Ces problématiques prennent une place encore plus importante dans le cadre du développement d'ontologies visant à modéliser les relations entre les entités présentes dans les récits de voyages, ainsi que dans la création de jeux de données *Linked Open Data* réutilisables à partir d'éditions numériques. Ces efforts sont actuellement au cœur du projet de Ratisbonne, « Digitale Editionen Historischer Reiseberichte [52] » (DEHisRe, Édition numérique des récits de voyage historiques), et démontrent que des travaux de recherche sur la REN dans les récits de voyages sont essentiels pour la création d'éditions numériques fiables. Ces efforts soulignent la pertinence de la reconnaissance des entités nommées

(REN) dans le domaine de l'édition numérique. Il est évident que la question de la REN dans les récits de voyages historiques constitue actuellement un domaine de recherche très dynamique.

NOTES

1 Terme informatique désignant une unité de valeur numérique, un objet identifié numériquement.

2 Voir en introduction, et avec une bibliographie complémentaire Mathis Leibetseder, « Kavalierstour – Bildungsreise – Grand Tour : Reisen, Bildung und Wissenserwerb in der Frühen Neuzeit », dans Leibniz-Institut für Europäische Geschichte (IEG) (dir.), *Europäische Geschichte Online* (*EGO*), mis en ligne le 14 août 2013, consulté le 15 janvier 2024, (http://www.ieg-ego.eu/leibetsederm-2013-de).

3 Voir le projet d'édition en ligne : Inga Hanna Ralle (dir.), *Selbstzeugnisse der Frühen Neuzeit in der Herzog August Bibliothek. Digitales Selbstzeugnis-Repertorium*, Wolfenbüttel, Herzog August Bibliothek, mis en ligne en 2014-2017, consulté le 15 janvier 2024, (http://selbstzeugnisse.hab.de/repertorium/). Pour une présentation du projet, voir Inga Hanna Ralle (dir.) *Selbstzeugnisse der Frühen Neuzeit in der Herzog August Bibliothek : Digitale Edition des Diariums von Herzog August dem Jüngeren. Technische Konzeption und Begleitung durch David Maus, unter Mitarbeit von Jacqueline Krone*, Wolfenbüttel, Wolfenbütteler Digitale Editionen 1, 2015–2017, (https://hainhofer.hab.de/informationen-zur-edition/bibliographie/ralle_diarium_2015-2017).

4 Voir la présentation en cours de construction des résultats du projet, notamment des éditions en voie d'élaboration, sur le portail des témoignages autobiographiques de la Bibliothèque Herzog August à l'adresse suivante : http://selbstzeugnisse.hab.de/edition_grand_tour ; ainsi que le blog accompagnant le projet à l'adresse suivante : https://grandtourdig.hypotheses.org/.

5 Voir la contribution d'Angela Göbel dans ce dossier, de même que Tobias Hodel, « Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik », *Historische Zeitschrift*, 2023, n° 316, p. 151–180.

6 HAB Wolfenbüttel, Cod. Guelf. 267.1. Voir Wolf-Dieter Otte, *Die neueren Handschriften der Gruppe Extravagantes*, vol. 3, Francfort-sur-le-Main, Klostermann, 1993, p. 169 ; I. H. Ralle, *op. cit.*, voir ici l'entrée à propos du journal de voyage de Wagener, http://selbstzeugnisse.hab.de/repertorium/eintrag/267-1-extrav.

7 Voir Werner Gaude, *Die alte Apotheke. Eine tausendjährige Kulturgeschichte*, 3e éd., Leipzig, Koehler & Amelang, 1985, p. 25.

8 Voir Jill Bepler, *Ferdinand Albrecht, Duke of Braunschweig-Lüneburg (1636-1687). A Traveller and his Travelogue*, Wiesbaden, Harassowitz, 1988.

9 *Ibid.*, p. 293.

10 Voir Angela Göbel, Maximilian Görmar, « Zu Besuch in der Gottorfischen Kunstkammer. Berichte eines Apothekergesellen auf Reisen », *HABlog*, mis en ligne le 28 mars 2023, consulté le 15 janvier 2024, (https://www.hab.de/zu-besuch-in-der-gottorfischen-kunstkammer/).

11 Elisabeth Eder, « Named Entity Recognition (NER) », dans Helmut W. Klug (dir.), *KONDE Weißbuch*, mis en ligne en 2021, (https://gams.uni-graz.at/o :konde.141) : « Named Entity Recognition (NER) bezeichnet die Erkennung von Eigennamen (named entities) in Texten sowie auch deren Klassifizierung in verschiedene Entitätstypen. […] Standardmäßig zählen Personen, Orte und Organisationen zu diesen Entitätstypen ».

12 Voir de manière introductive Manuela Lenzen, *Künstliche Intelligenz. Fakten, Chancen, Risiken*, München, Beck, 2020.

13 Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, Antoine Doucet, « Named Entity Recognition and Classification on Historical Documents: A Survey », ACM *Computing Surveys*, n° 56/2, mis en ligne le 14 septembre 2023, p. 1–43, URL : https://doi.org/10.1145/3604931, ici p. 17.

14 *Ibid.*, p. 22.

15 Voir Mareike Schumacher, *Orte und Räume im Roman. Ein Beitrag zur digitalen Literaturwissenschaft*, Berlin, J.B. Metzler, 2023, p. 86–90. Le score F1 est calculé comme la moyenne harmonique des valeurs de la *precision* et du *recall*. La *precision* indique combien de *tokens* marqués comme noms propres ont été correctement reconnus. Le *recall* est la valeur représentant la proportion de *tokens* correctement identifiés par rapport à

l'ensemble des *tokens* appartenant à la catégorie d'entité respective, *ibid.*, p. 86.

16 *NERDPool : Data Pool for Named Entity Recognition*, consulté le 15 janvier 2024, (https://nerdpool-api.acdh-dev.oeaw.ac.at/). Voir également Peter Andorfer, Roman Bleier, Matthias Schlögl, « NERDPool. Datenpool für Named Entity Recognition », dans Michaela Geierhos (dir.), *DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts*, Potsdam, 2022, p. 333, URL: https://doi.org/10.5281/zenodo.630459.

17 Voir Isabel Hansen, *Named Entity Recognition. Einsatz und Verwendung der Technologie in den Digital Humanities am Beispiel historischer Texte*, mémoire de Master (M. Sc.), Université de Trèves, 2023 (un grand merci à Madame Hansen et Joëlle Weis pour la mise à disposition du travail !) ainsi que le tutoriel associé sous URL : https://easyh.github.io/NerDH/, et le modèle sous URL : https://huggingface.co/easyh/de_fnhd_nerdh ; Jacqueline More, *Theorie und Anwendung von Named Entity Recognition in den Digital Humanities mit Fokus auf historische Texte des 17. Jahrhunderts*, mémoire de Master, Université de Graz, 2021, URL : https://resolver.obvsg.at/urn :nbn :at :at-ubg:1-166482 ainsi que le dépôt GitHub correspondant sur URL : https://github.com/jackymore/NER_historical_texts ; voir de plus Tatiana Bessonova, Lisa Braune, Sarah Ondraszek *et al.*, *Reiseberichte karTRIERt*, mis en ligne le 24 avril 2023, (https://kartriert.github.io/index.html) ; voir encore d'autres tutoriels : Susan Grunewald, Andrew Janco, « Finding Places in Text with the World Historical Gazetteer », Programming Historian, 2022, n° 11, (https://doi.org/10.46430/phen0096 ; Mareike Schumacher, « Named Entity Recognition (NER) », *forTEXT. Literatur digital erforschen*, mis en ligne le 18 mai 2018, (https ://fortext.net/routinen/methoden/named-entity-recognition-ner).

18 Parallèlement, des journaux de l'époque moderne et des dossiers criminels ont également été examinés dans des études de cas à l'aide de la REN. Voir Thomas Kirchmair, Nina C. Rastinger, Claudia Resch, « Die historische ‚Wiener Zeitung' und ihre Sterbelisten als Fundament einer Vienna Time Machine. Digitale Ansätze zur automatischen Identifikation von Toponymen », *Wiener Digitale Revue*, 2022, n° 4, (https://doi.org/10.25365/wdr-04-03-04) ; Tobias Hodel, Ismail Prada Ziegler, Christa Schneider, « Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern », dans Anne Baillot, Toma Tasovac, Walter Scholger, Georg Vogeler (dir.), *Digital*

Humanities 2023: *Book of Abstracts*, Graz, 2023, p. 384.,URL : https://doi.org/10.5281/zenodo.7961822 ; voir également le modèle NER correspondant, consulté le 15 janvier 2024, (https://huggingface.co/dh-unibe/turmbuecher-lm-v1).

19 *Stanford Named Entity Recognizer*, Version 4.2.0, mis en ligne le 17 novembre 2020, consulté le 15 janvier 2024, (https://nlp.stanford.edu/software/CRF-NER.shtml). Voir aussi Mareike Schumacher, « Stanford Named Entity Recognizer », *forTEXT. Literatur digital erforschen*, mis en ligne le 20 septembre 2018, (https://fortext.net/tools/tools/stanford-named-entity-recognizer) ; *Id.*, « Named Entity Recognition mit dem Stanford Named Entity Recognizer », *forTEXT. Literatur digital erforschen*, mis en ligne le 26 août 2019, (https://fortext.net/routinen/lerneinheiten/named-entity-recognition-mit-dem-stanford-named-entity-recognizer).

20 CLARIN-D/Seminar für Sprachwissenschaft, Université de Tübingen : WebLicht : *Web-Based Linguistic Chaining Tool*, mis en ligne en 2012, (https://weblicht.sfs.uni-tuebingen.de). Pour une documentation et une introduction à l'utilisation, voir le wiki correspondant, consulté le 15 janvier 2024, (https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page).

21 ExplosionAI GmbH, *spaCy*, mis en ligne en 2016-2024 (https://spacy.io/). Voir aussi Peter Andorfer, Matthias Schlögl, , Helmut W. Klug (dir.), *KONDE Weißbuch*, « spaCy », mis en ligne en 2021, (https://www.digitale-edition.at/o :konde.170).

22 Il s'agit de l'outil de Bryan Jurish, *DTA::CAB – "Cascaded Analysis Broker" for error-tolerant linguistic analysis*, (https://kaskade.dwds.de/~moocow/software/dta-cab/). Voir aussi *Id.*, *Finite-state Canonicalization Techniques for Historical German*, thèse de doctorat, Université de Potsdam, 2012, consulté le 15 janvier 2024, URL : https://publishup.uni-potsdam.de/files/5562/jurish_diss.pdf. Le *Cascaded Analysis Broker* est utilisé, entre autres, dans le German Text Archive. Voir « Software im Deutschen Textarchiv », dans Berlin-Brandenburgische Akademie der Wissenschaften (dir.), *Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*, mis en ligne le 4 mars 2020, consulté le 15 janvier 2024, (https://www.deutschestextarchiv.de/doku/software).

23 TEI est le standard *de facto* pour la codification des textes dans le domaine des éditions numériques. Voir Text Encoding Initiative (dir.), *TEI: Guidelines for Electronic Text Encoding and Interchange. P5*, Version 4.7.0,

mis en ligne le 16 novembre 2023, consulté le 15 janvier 2024, (https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html).

24 Voir ci-dessus les notes 17 et 18.

25 Peter Andorfer, Matthias Schlögl, Saranya Balasubramanian, *acdh-spacytei*, Version 0.1.2, mis en ligne le 19 juillet 2019, (https://pypi.org/project/acdh-spacytei/). Voir aussi Peter Andorfer, Matthias Schlögl, Helmut W. Klug (dir.), *KONDE Weißbuch*, « acdh-spacytei », mis en ligne en 2021, consulté le 15 janvier 2024, (https://www.digitale-edition.at/o :konde.2).

26 David Lassner, *standoffconverter*, Version 0.8.11, mis en ligne le 28 avril 2022,(https://pypi.org/project/standoffconverter). Voir *Id.*, *Analysis of textual variants with robust machine learning methods: Towards novel insights for the digital humanities*, thèse de doctorat, TU Berlin, 2023. URL : https://doi.org/10.14279/depositonce-17717, p. 10–16 et 123–125.

27 Voir Angela Göbel, « Erste Schritte in Transkribus (1) – Modelltraining und Transkription der Handschrift Wagener », dans *Grand Tour digital. Ein Forschungsblog zum Editionsprojekt frühneuzeitlicher Selbstzeugnisse an der Herzog August Bibliothek Wolfenbüttel*, mis en ligne le 6 juillet 2023, (https://grandtourdig.hypotheses.org/290).

28 Ont été utilisés dans ce cadre Arunmozhi, *NER Text Annotator*, Version 1.3.0, mis en ligne le 12 mai 2023,(https://tecoholic.github.io/ner-annotator/).

29 P. Andorfer, M. Schlögl, S. Balasubramanian, *op. cit.*

30 Cela a été réalisé en préparant 70 % de l'ensemble de données comme données d'entraînement, 20 % comme données de validation et 10 % comme données d'essai.

31 Voir *Training Pipelines & Models*, (https://spacy.io/usage/training).

32 Pour expliquer ces indicateurs, voir la note 15 ci-dessus.

33 ExplosionAI GmbH, *displaCy Named Entity Visualizer*, (https://demos.explosion.ai/displacy-ent).

34 D. Lassner, *op. cit.*

35 Voir par exemple Joachim Peters, Sabrina Freund, « *Vormittags rien fait quj vaille.* Codeswitching im Fürstentagebuch Christians von Anhalt-Bernburg (1599–1656) », dans Elvira Glaser, Michael Prinz, Stefaniya Ptashnyk (dir.), *Historisches Codeswitching mit Deutsch.*

Multilinguale Sprachpraktiken in der Sprachgeschichte, Berlin/Boston, de Gruyter, 2021, p. 331–366.

36 Voir Arturo Tosi, *Language and the Grand Tour. Linguistic experiences of travelling in Early modern Europe*, Cambridge, Cambridge University Press, 2020.

37 Voir Johannes Volmert, « Geschichte der deutschen Sprache », dans *Id.* (dir.), *Grundkurs Sprachwissenschaft. Eine Einführung in die Sprachwissenschaft für Lehramtsstudiengänge*, 4e éd., Munich, W. Fink, 2000, p. 29–46, ici p. 42–48 ; de plus voir comme exemple Andreas Herz, « Zu den Schreibweisen Fürst Christians II. von Anhalt-Bernburg », dans Ronald G. Asch, Peter Burschel, Arndt Schreiber *et al.* (dir.), *Digitale Edition und Kommentierung der Tagebücher des Fürsten Christian II. von Anhalt-Bernburg (1599–1656). Christian II. von Anhalt-Bernburg*, Wolfenbüttel, Herzog August Bibliothek, 2013–2024, consulté le 15 janvier 2024, (http://diglib.hab.de/edoc/ed000228/id/edoc_ed000228_introduction_spelling_sm/start.htm).

38 https://www.geonames.org/.

39 https://gnd.network.

40 https://viaf.org/.

41 https://www.wikidata.org.

42 Voir la note 16 ci-dessus.

43 Outre le cadre fourni par la TEI, qui est déjà assez avancé et offre généralement plusieurs possibilités pour annoter le même phénomène textuel, de nombreuses éditions de la période de l'époque moderne s'orientent également d'après : Arbeitskreis Editionsprobleme der Frühen Neuzeit, « Empfehlungen zur Edition von frühneuzeitlichen Texten », *HEIMATFORSCHUNG-REGENSBURG.DE*, mis en ligne le 11 février 2016, (http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:355-rbh-280-9). Cependant, il existe généralement différentes déviations, de sorte que chaque édition établit fondamentalement ses propres directives de transcription et d'édition, qui ne sont pas nécessairement compatibles avec d'autres. Cette diversité, fondamentalement justifiée compte tenu de l'individualité des sources traitées et des axes de recherche possibles des projets d'édition respectifs, entraîne un niveau élevé de confusion et limite l'interopérabilité des données d'édition. Cependant, un degré élevé de différenciation et de spécialisation dans les directives d'édition et la confusion qui en résulte ne sont pas spécifiques aux éditions numériques,

mais concernent également depuis longtemps les éditions imprimées. Voir Bodo Plachta, *Editionswissenschaft. Eine Einführung in Methode und Praxis der Edition neuerer Texte*, édition complétée et actualisée, Stuttgart, Reclam, 2013 ; Hans-Gert Roloff, « Epochenprofilierung durch Editionen », dans *Id.* (dir.), *Editionsdesiderate der Frühen Neuzeit. Beiträge einer Tagung der Kommission für die Edition von Texten der Frühen Neuzeit*, vol. 1, Amsterdam/Atlanta, Rodopi, 1997, p. 1–13.

44 Pour la relation entre la segmentation du texte et l'analyse (numérique), voir Sabine Bartsch, Evelyn Gius, Marcus Müller, Andrea Rapp, Thomas Weitin, « Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft », *Zeitschrift für digitale Geisteswissenschaften*, 2023, n° 8, (https://doi.org/10.17175/2023_003).

45 Inga Hanna Ralle, Jacqueline Krone, David Maus (dir.), *Herzog August der Jüngere von Braunschweig-Wolfenbüttel (1579–1666) : Ephemerides. Sive Diarium (1594–1635)*, Wolfenbüttel, Herzog August Bibliothek, 2017, (https://doi.org/10.15499/edoc/ed000225).

46 Ronald G. Asch, Peter Burschel, Arndt Schreiber *et al.* (dir.), *Digitale Edition und Kommentierung der Tagebücher des Fürsten Christian II. von Anhalt-Bernburg (1599-1656). Christian II. von Anhalt-Bernburg*, Wolfenbüttel, Herzog August Bibliothek, 2013–2024, (http://diglib.hab.de/edoc/ed000228/start.htm).

47 Pour différentes possibilités techniques de segmentation de phrases dans spaCy voir *Linguistic Features*, consulté le 15 janvier 2024, (https://spacy.io/usage/linguistic-features).

48 Voir *Rule-based matching*, (https://spacy.io/usage/rule-based-matching) ; ainsi que plus loin S. Grunewald, A. Janco, *op. cit.*

49 Voir Sina Menzel, Hannes Schnaitter, Josefine Zinck *et al.* « Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten », dans Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, Hans Schürmann (dir.), *Qualität in der Inhaltserschließung*, Berlin/Boston, de Gruyter Saur, 2021, p. 229–257.

50 Voir Jenny Rose Finkel, Christopher D. Manning, « Nested Named Entity Recognition », dans Philipp Koehn, Rada Mihalcea (dir.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapour, Association for Computational Linguistics, 2009, URL: https://aclanthology.org/D09-1015, p. 141–150.

51 Tim Hitchcock, « Big Data, Small Data and Meaning », *Historyonics*, mis en ligne le 9 novembre 2014 (http://historyonics.blogspot.com/2014/11/big-data-small-data-and-meaning_9.html).

52 Voir Sandra Balck, Ingo Frank, Hermann Beyer-Thoma, Anna Ananieva, « Interlinking Text and Data with Semantic Annotation and Ontology Design Patterns to Analyse Historical Travelogues », *digital humanities quaterly*, n° 17/3, 2023 (http://www.digitalhumanities.org/dhq/vol/17/3/000726/000726.html).

RÉSUMÉS

Français

Dans le domaine des éditions numériques savantes, les applications de l'intelligence artificielle (IA) et de l'apprentissage automatique gagnent de plus en plus d'influence. De nos jours, la transcription ainsi que l'annotation de textes peuvent être facilitées par des outils puissants tels que Transkribus, avec lesquels des modèles d'IA peuvent être entraînés pour effectuer une reconnaissance de texte manuscrit (HTR). Dans une deuxième étape, les données textuelles peuvent être traitées par des logiciels conçus pour le Traitement du langage naturel (NLP) afin d'extraire et d'annoter des caractéristiques syntaxiques, morphologiques ainsi que des informations sémantiques. Par exemple, il est possible de baliser des personnes, des lieux et des organisations via la Reconnaissance d'entités nommées (NER), ce qui est particulièrement pertinent pour les éditions numériques de sources historiques.

Ces considérations sont l'une des principales préoccupations du projet de recherche et d'édition « Grand Tour digital » à la Bibliothèque Herzog August de Wolfenbüttel. Il vise à établir la faisabilité des méthodes d'IA pour l'édition savante de sources historiques et à adapter de telles méthodes dans un flux de travail complet et durable qui pourrait être applicable à d'autres projets également. Ce faisant, nous devons identifier les possibilités et, plus important encore, les défis potentiels de la technologie en ce qui concerne son adaptabilité aux sources historiques. Cet article présente un travail réalisable grâce à la NER en utilisant l'exemple d'un récit de voyage du milieu du xvii^{e} siècle écrit par un jeune apothicaire itinérant qui a voyagé dans la région baltique, à travers le nord de l'Allemagne, faisant fonction de laquais dans la suite du jeune duc Ferdinand Albrecht I^{er} de Brunswick-Wolfenbüttel-Bevern, en Suisse et en France.

En observant ce processus, apparaîtront certains problèmes et difficultés liés aux particularités des textes de l'époque moderne par rapport aux textes contemporains pour lesquels pratiquement tous les outils de NLP et de NER ont été initialement conçus. Tout d'abord, les textes actuels sont orthographiquement et grammaticalement beaucoup plus normalisés que

de nombreux textes anciens. À l'époque moderne, par exemple, le même scribe pouvait utiliser différentes orthographes du même mot sur la même page. De plus, de nombreux scribes et leurs textes étaient multilingues plutôt que monolingues, et même les noms de personnes ou de lieux pouvaient parfois être donnés dans différentes langues et versions, par exemple le prénom allemand Johannes ou Hans pouvait parfois apparaître dans sa forme française Jean même si la même personne était visée.
Il existe plusieurs approches pour atténuer ces difficultés et celles qui y sont liées, dont certaines seront évaluées dans cet article. La première consiste à utiliser des techniques d'apprentissage automatique pour former des modèles NER spécifiquement sur des textes de la période et de la langue qui nous intéressent. Le problème est qu'il existe, en général, très peu de jeux de données d'entraînement disponibles à partir de textes historiques qui peuvent être utilisés pour former des modèles spécialisés. Une autre approche est l'utilisation de gazetiers ou de dictionnaires de noms avec lesquels le programme peut reconnaître certains tokens [1] comme des noms. Encore une fois, il existe relativement peu de ressources pour les textes de l'époque moderne par rapport aux textes contemporains et surtout les noms de personnes montrent une grande variété et sont souvent très spécifiques à des textes individuels. Ainsi, l'approche basée sur le dictionnaire ou les règles ne peut être utilisée de manière significative que pour des entités qui existent sur une période relativement longue et peuvent apparaître, par conséquent, dans un plus grand nombre de textes, comme des lieux ou des organisations. Une troisième manière de faire face aux problèmes posés par les sources historiques pour la NER consiste en la combinaison des deux approches décrites ci-dessus. C'est possible avec certaines applications de NLP, par exemple spaCy, qui a été utilisé dans l'étude entre autre pour cette raison.
Dans l'ensemble, cet article présente une étude de cas pour l'application de méthodes NER aux éditions numériques savantes de textes de l'époque moderne. Il analyse les possibilités et les défis de cette entreprise et propose des solutions en cas de difficultés. Si ces réflexions peuvent être utiles à d'autres projets, elles sont encore à un stade préliminaire et nécessitent des tests et des améliorations supplémentaires.

English
In the field of scholarly digital editions applications of Artificial Intelligence (AI) and machine learning gain more and more influence. Nowadays, the transcription as well as the annotation of texts can be facilitated by powerful tools such as Transkribus with which AI-models can be trained to perform Handwritten Text Recognition (HTR). In a second step, the textual data can be processed by software designed for Natural Language Processing (NLP) to extract and annotate syntactical and morphological features as well as semantic information. For example, it is possible to markup persons, places and organisations via Named Entity Recognition (NER), which is especially relevant for digital editions of historical sources. These considerations are one main concern of the research and edition

project "Grand Tour digital" at the Herzog August Library Wolfenbüttel. It aims to establish the feasibility of AI-methods for the scholarly editing of historical sources and to adapt such methods into a comprehensive and sustainable workflow which might be applicable for other projects as well. In doing so, we need to identify the possibilities and, more importantly, potential challenges for the technology regarding its adaptability to historical sources. The present article outlines a possible workflow for NER using the example of a mid-17th century travelogue written by a young apothecary journeyman who travelled to the Baltic region, through northern Germany and, as some sort of lackey in the party of the young Duke Ferdinand Albrecht I. of Brunswick-Wolfenbüttel-Bevern, to Switzerland and France.

While discussing the workflow, some problems and difficulties will become apparent which are a result of the peculiarities of early modern in contrast to modern texts for which virtually all NLP and NER tools were originally designed. First and foremost, modern texts and languages are orthographically and grammatically far more normalized than many historical texts. In the early modern era, for example, the same scribe could employ different spellings of the same word on the same page. Additionally, many scribes and their texts were multilingual rather than monolingual, and even the names of persons or places could occasionally be given in different languages and different versions, e. g. the German given name Johannes or Hans would sometimes appear in its French form Jean even if the same person was meant by it.

There are several approaches to mitigate these and related difficulties, some of which will be evaluated in the paper. The first one is to use machine learning techniques to train NER models specifically on texts of the time period and language one is interested in. One problem regarding this solution is that there are, generally speaking, very few and rather small available training sets of suitable data from historical texts which can be used to train specialised models. Another approach is the use of gazetteers or dictionaries of names with which the program can recognize certain tokens as names. Again, there are comparatively few resources for pre-modern than modern texts to build and especially person names show a great variety and are often very specific for single texts. Thus, the dictionary- or rule-based approach can only be used in a meaningful way for entities that are existent over a relatively long time period and may appear, therefore, in a greater number of texts, such as places or organisations. A third way of addressing the problems posed by historical source for NER consists of the combination of the two approaches outlined above. This is possible with some NLP-applications, e. g. spaCy which was used in the present study because of that reason among others.

All in all, this article presents a case study for the application of NER methods to scholarly digital editions of early modern texts. It analyses the possibilities and challenges of this venture, and proposes some solutions for potential problems and difficulties. While these will be hopefully useful for

other projects with similar concerns, they are still in a preliminary state and need further testing and improvement.

INDEX

Mots-clés
éditions numériques savantes, reconnaissance d'entités nommées, récit de voyage, époque moderne, Grand Tour, apprentissage automatique, intelligence artificielle

Keywords
scholarly digital editions, named entity recognition, travelogue, early modern era, Grand Tour, machine learning, artificial intelligence

AUTEUR

Maximilian Görmar
IDREF : https://www.idref.fr/281724733

TRADUCTEUR

Angela Göbel
Herzog August Bibliothek Wolfenbüttel
IDREF : https://www.idref.fr/28172489X