Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines

Journées scientifiques du GDR LIFT, 2020

Pablo Ruiz, Delphine Bernhard, Carole Werner

Université de Strasbourg

lilpa
linguistique, langues, parole

Plan

- **Objectifs**
- Corpus
- Encodage TEI
- Traitement automatique des variantes orthographiques
- FAIRisation

Objectifs

- Projet MeThAL
  « Vers une macroanalyse du théâtre en alsacien »

- Créer corpus TEI de théâtre dialectal
  - 50 pièces ; de 1870 à 1940

- Ressources publiques FAIR (Wilkinson et al., 2016)
  - Findable, Accessible, Interoperable, Reusable

Intérêt du corpus

- Accroître ressources numériques en alsacien
  - Cf. projet ANR RESTAURE

- Analyse dramatique quantitative
  - Cf. projets QuaDramA, « Théâtre classique », DraCor
  - Ici, focus sur une tradition « mineure »

- Sociolinguistique historique

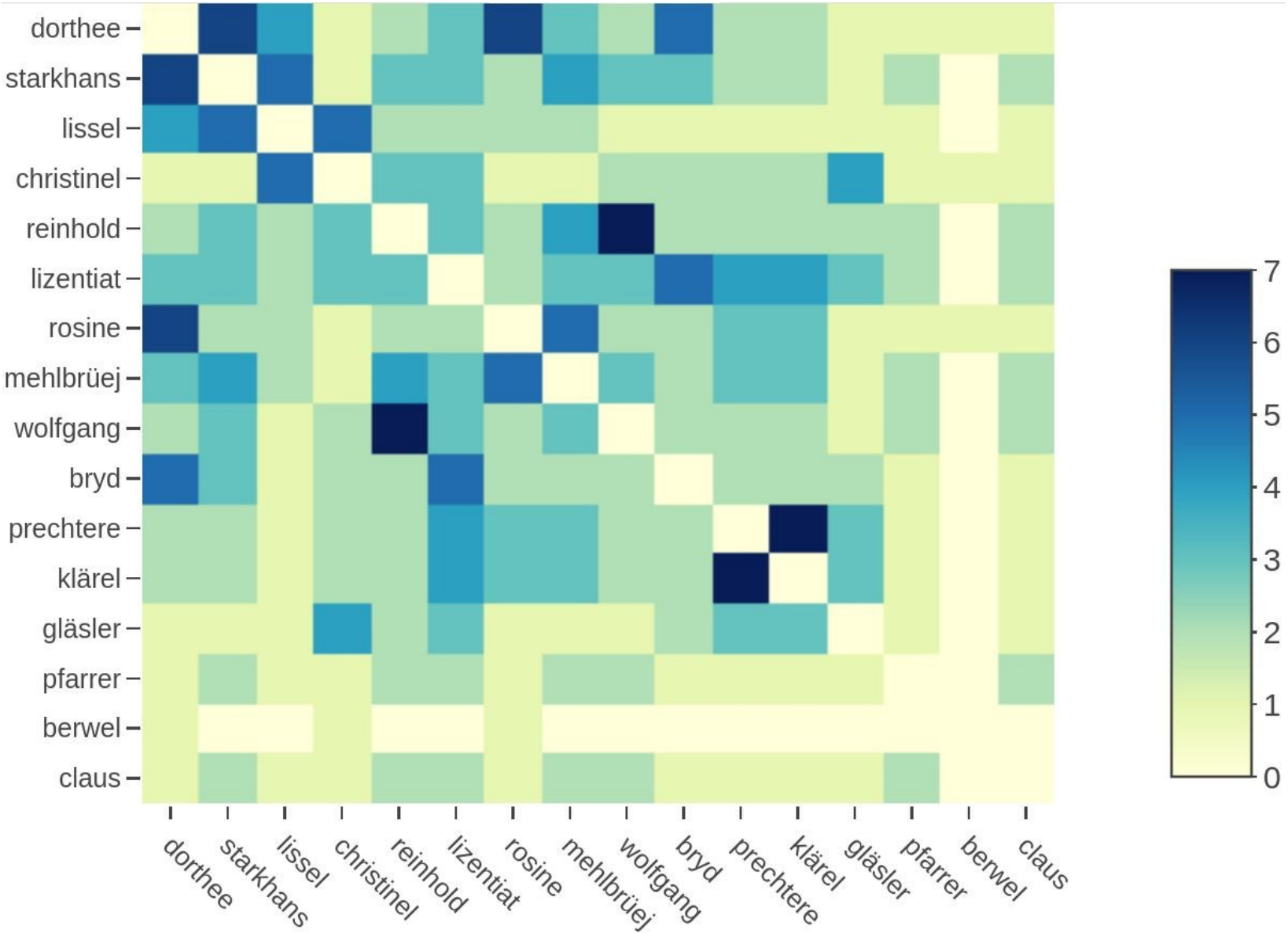
  (cf. Huck et al., 2007 ; Huck, 2015)

Analyse dramatique quantitative

- Patrons d'interaction entre
  - Personnages
  - Groupes de personnages

- Contenu des échanges entre les groupes

Interaction entre personnages



Pièce : Der Pfingstmontag · Source: Shiny DraCor

- Niveau d'abstraction suivant : Interaction entre groupes

Sociolinguistique historique

- Réserve : Corpus de *théâtre*
- *Dans la mesure où* il documente pratiques langagières de son époque, il permet :
  - Analyse de variation
    - Choix de scripturalisation
    - Lexique
    - Phonologie ? (Watson & Jensen, 2020)
  - Selon variables sociales des personnages
    - Statut, âge, sexe, origine

Plan

- Objectifs
- **Corpus**
- Encodage TEI
- Traitement automatique de variantes orthographiques
- FAIRisation

Corpus source

- Créateur : Bnu Strasbourg (Numistral)
- Période : 1870 – 1940
  - Décennies 1900 et 1920 prédominent
- Volume : 150 pièces par 32 auteurs

- Grande variation orthographique
  - Pratiques de scripturalisation obsolètes
- « Mélange de langues »

Variation orthographique

- Selon les auteurs ; selon la pièce
- Dans une même pièce, *selon le personnage*

	Forme (pour *Tag*, *jour*)	Variables sociales
(a)	Daö	variante « rurale » (Kochersberg)
(b)	Daa	variante strasbourgeoise
(c)	Tag	allemand standard (dans une lettre)

D'r Herr Maire (G. Stoskopf, 1898)

- Défi traitement automatique des langues (TAL)

Mélange de langues

- Alternance codique

(1) [...] d'Arweitsklass wurd üsgenutzt hytt am Daa. **„Parole d'honneur!"**
la classe ouvrière est exploitée de nos jours

In's Ropfer's Apothek (Stoskopf, 1907)

- Français « à l'alsacienne »

(2) Pong! Pong! Athiö ! Schmangwäh ! I kumm hyt owes widder.
Bon ! Bon ! Adieu ! Je m'en vais ! *Je reviens ce soir*

Der Pfingstmontag (Arnold, 1816)

- Coexistence de variétés alsaciennes, français, allemand et d'autres dialectes
  - Répliques *vs.* didascalies et liste de personnages

Plan

- Objectifs
- Corpus
- **Encodage TEI**
- Traitement automatique de variantes orthographiques
- FAIRisation

Encodage TEI : Personnages

- Variables sociales décrivant les personnages
- Formalisation TEI avec feature structures ?

Galleron (2017)

```
<fDecl name="age">
  <vRange>
    <vAlt>
      <symbol value="young" xml:id="AGY"/>
      <symbol value="old" xml:id="AGO"/>
      <symbol value="child" xml:id="AGC"/>
      <symbol value="adult" xml:id="AGA"/>
      <symbol value="unknown" xml:id="AGU"/>
    </vAlt>
  </vRange>
</fDecl>
```

Consortium Cahier (Idmhand & Galleron, 2020) [Exemple annexe aux guidelines]

```
<fDecl name="occupation" xml:id="OCC" optional="true">
   <vRange>
      <string/>
   </vRange>
</fDecl>
```

Encodage TEI : Personnages

- Métadonnées des personnages
  - Transcrites
  - Enrichies : normalisation de professions, classe sociale

persName	roleDesc	sex	profession	normProfession	class
Drohtspitz	Schuehmachermaischter	M	Schuehmachermaischter	master shoemaker	LMC
Eddes	Lehrbuewe biem Drohtspitz	M	Lehrbuewe (Schuehmacher)	shoemaker apprentice	LC
Schakob	e Büur	M	Büur	farmer	LC
Linser	Photograph	M	Photograph	photographer	LMC
Mlle Florentine Blind	rentière	F	rentière	rentier	UMC
Mlle Amélie	G'sellschaftre vu Mlle Florentine	F	G'sellschaftre	companion	UMC
Sophie	Kecha	F	Kecha	cook	LC
Elise	femme de chambre	F	femme de chambre	housemaid	LC
Fritz Grinsinger	Konservenfabrikant	M	Konservenfabrikant	manufacturer	UC
Olga Gauthier	stud. phil.	F	stud. phil.	philosophy student	UC

Encodage TEI : Personnages

- Métadonnées des personnages
  - Transcrites
  - Enrichies : normalisation de professions, classe sociale

persName	roleDesc	sex	profession	normProfession	class
Drohtspitz	Schuehmachermaischter	M	Schuehmachermaischter	master shoemaker	LMC
Eddes	Lehrbuewe biem Drohtspitz	M	Lehrbuewe (Schuehmacher)	shoemaker apprentice	LC
Schakob	e Büur	M	Büur	farmer	LC
Linser	Photograph	M	Photograph	photographer	LMC
Mlle Florentine Blind	rentière	F	rentière	rentier	UMC
Mlle Amélie	G'sellschaftre vu Mlle Florentine	F	G'sellschaftre	companion	UMC
Sophie	Kecha	F	Kecha	cook	LC
Elise	femme de chambre	F	femme de chambre	housemaid	LC
Fritz Grinsinger	Konservenfabrikant	M	Konservenfabrikant	manufacturer	UC
Olga Gauthier	stud. phil.	F	stud. phil.	philosophy student	UC

Encodage TEI : Personnages

- Métadonnées des personnages
  - Transcrites
  - Enrichies : normalisation de professions, classe sociale

persName	roleDesc	sex	profession	normProfession	class
Drohtspitz	Schuehmachermaischter	M	Schuehmachermaischter	master shoemaker	LMC
Eddes	Lehrbuewe biem Drohtspitz	M	Lehrbuewe (Schuehmacher)	shoemaker apprentice	LC
Schakob	e Büur	M	Büur	farmer	LC
Linser	Photograph	M	Photograph	photographer	LMC
Mlle Florentine Blind	rentière	F	rentière	rentier	UMC
Mlle Amélie	G'sellschaftre vu Mlle Florentine	F	G'sellschaftre	companion	UMC
Sophie	Kecha	F	Kecha	cook	LC
Elise	femme de chambre	F	femme de chambre	housemaid	LC
Fritz Grinsinger	Konservenfabrikant	M	Konservenfabrikant	manufacturer	UC
Olga Gauthier	stud. phil.	F	stud. phil.	philosophy student	UC

Encodage TEI : Variation orthographique

- Moyen de relier variantes entre elles ?

LEXIQUE STAND-OFF

```
<w xml:id="bn:00000086n">Tag</w>
<w xml:id="mt:00424242x">g'fröjt</w>
```

TEXTES

```
<w corresp="bn:00000086n">Daö</w>
   [...]
<w corresp="bn:00000086n">Daa</w>
   [...]
<w corresp="mt:00424242x">g'fröuit</w>
   [...]
<w corresp="mt:00424242x">g'frajt</w>
```

Encodage TEI : « Mélange de langues »

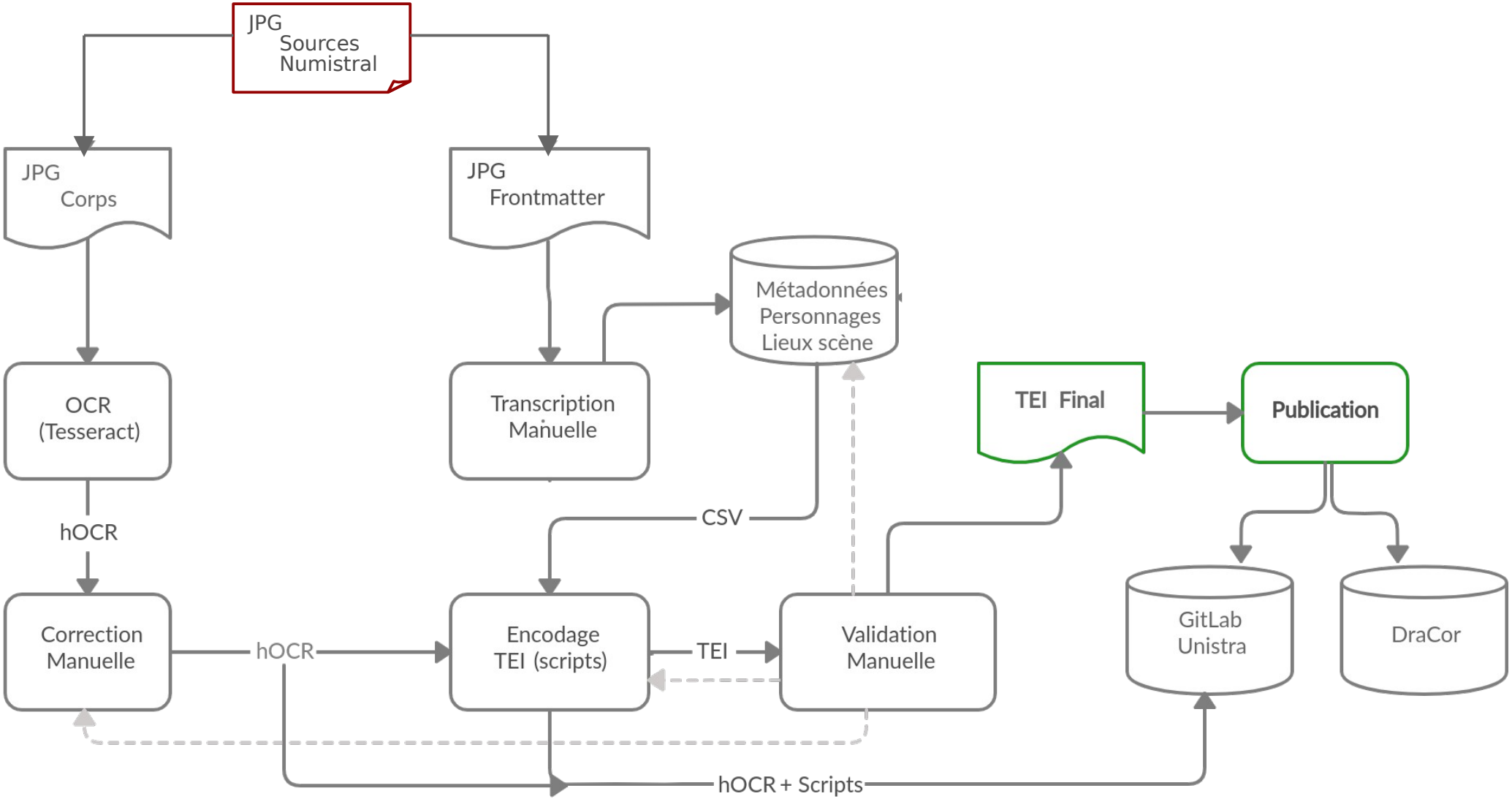
- Encodage de base prévu (`@xml:lang`)

```
<sp who="#madame_schweberle">
  <speaker>Mme SCHWEBERLE:</speaker>
  <p>
  <seg xml:lang="fre">Toutes mes félicitations</seg>; ich wuensch
  eich Glueck vun ganzem Herze.
  <stage>(Zue Schakob)</stage> Un Ihr, do hinte, kumme emol here.
  </p>
</sp>
```

```
<div type="front" xml:lang="ger">
  <head>Inhaltsangabe.</head>
  <p>Der Urtext stammt aus der Feder der
  bekannten Dichterin des « Stadtnarr »,
```

D'r poetisch Oscar (Marie Hart, éd. de 1937)

TEI : Chaîne de traitement



TEI : Chaîne de traitement

- Sources Numistral (images)
- Transcription manuelle du contenu peu prédictible
  - Couverture, liste de personnages
- OCR du reste
  - Tesseract + modèles RESTAURE → hOCR
  - Validation manuelle
- Création TEI automatique sur base des précédents
  - Validation manuelle

Plan

- Objectifs
- Corpus
- Encodage TEI
- **Traitement automatique de variantes orthographiques**
- FAIRisation

Traitement automatique de variantes orthographiques

- Requis pour analyses outillées du contenu ; garantir la comparabilité des textes
- Normalisation : ramener à une forme « standard »
- Identification de variantes
  - Expériences sur lexiques alsacien-français (Bernhard, 2014)
  - Approches supervisées (cf. Barteld et al., 2019)

Plan

- Objectifs
- Corpus
- Encodage TEI
- Traitement automatique de variantes orthographiques
- **FAIRisation**

FAIRisation

Findabilité, Accessibilité

- Disponible sur plateforme DraCor
- Non disponible sur une plateforme ouverte généraliste d'exposition de données (p. ex. Nakala)
- Pas d'identifiants persistants

Interopérabilité, Réutilisabilité

- TEI
- IDs Wikidata
- Scripts, lexiques et hOCR publiés
- Licence ouverte
- Métadonnées
- Wiki

Références

- Barteld, F., Biemann, C., & Zinsmeister, H. (2019). Token-based spelling variant detection in Middle Low German texts. Language Resources and Evaluation, 53(4), 677-706. https://doi.org/10.1007/s10579-018-09441-5
- Bernhard, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources : The Example of Alsatian. Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014), 23-29. https://hal.archives-ouvertes.fr/hal-00966820
- Galleron, I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm : Needs, Problems and Foreseen Solutions. Human and Social Studies, 6(1), 88-108. https://doi.org/10.1515/hssr-2017-0007
- Huck, D. (2015). Une histoire des langues de l'Alsace. La Nuée Bleue.
- Huck, D., Bothorel-Witz, A., & Geiger-Jallet, A. (2007). L'Alsace et ses langues. Eléments de description d'une situation sociolinguistique en zone frontalière. In A. Abel, M. Stuflesser, & L. Voltmer (Éds.), Aspects of Multilingualism in European Border Regions : Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol (p. 13-101). EURAC Research (Europäische Akademie / Accademia Europea / European Academy). http://ala.u-strasbg.fr/documents/Publication%20-%20L%27Alsace%20et%20ses%20langues.pdf
- Idmhand, F., & Galleron, I. (2020). Guide pour la FAIRisation des données des corpus d'auteurs préparé par Fatiha Idmhand et Ioana Galleron pour le [Groupe de travail Data_Cahier] [Research Report]. Huma-Num. https://halshs.archives-ouvertes.fr/halshs-02889777
- Watson, K., & Jensen, M. M. (2020). Automatic analysis of dialect literature : Advantages and challenges. In Dialect Writing and the North of England. Edinburgh University Press.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Projets cités

- DraCor : https://dracor.org
- QuaDramA : https://quadrama.github.io/index.en
- RESTAURE : https://restaure.unistra.fr
- Théâtre classique : http://www.theatre-classique.fr

Merci !

https://methal.pages.unistra.fr/

Ce travail a bénéficié d’un financement dans le cadre de l’IdEx Université de Strasbourg