



**HAL**  
open science

# D'un corpus à l'autre D'une étude reproductible et portable du discours direct nisvai à la comparaison linguistique

Jocelyn Aznar

► **To cite this version:**

Jocelyn Aznar. D'un corpus à l'autre D'une étude reproductible et portable du discours direct nisvai à la comparaison linguistique. 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), 2020, Montrouge, France. pp.44-53. hal-03047154

**HAL Id: hal-03047154**

**<https://hal.science/hal-03047154v1>**

Submitted on 3 Jan 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# D'un corpus à l'autre

## *D'une étude reproductible et portable du discours direct nisvai à la comparaison linguistique*

Jocelyn Aznar

(1) ZAS, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)

Schützenstr. 18, 10117 Berlin, Allemagne

(2) Aix-Marseille Université, CNRS, EHESS, CREDO UMR 7308, 3, Place Victor Hugo,

13331 Marseille, France

aznar@leibniz-zas.de

## RÉSUMÉ

---

À partir d'une étude comparative en cours portant sur le discours direct à travers des documentations linguistiques de langues orales peu documentées, nous proposons une réflexion sur la reproductibilité et la portabilité d'une recherche en linguistique. L'enjeu est de porter l'étude du discours direct réalisées sur le corpus de narrations nisvaies, une langue orale du Vanuatu, à d'autres corpus de langues orales. Les annotations de ces corpus ont été amendés et normalisées par les efforts combinés des projets DoReCo et QUEST. Nous verrons que si la reproductibilité d'une étude sur une langue facilite sa critique, la question de la portabilité d'une étude vers d'autres corpus requiert que ces derniers répondent à des normes et unités interopérables aussi bien d'un point de vue informatique que linguistique.

## ABSTRACT

---

### **From one corpus to another: From a reproducible and portable study of nisvai direct discourse to linguistic comparison**

Based on an ongoing comparative study of direct discourse through poorly documented linguistic documentation of oral languages, we propose a reflection on the reproducibility and portability of research in linguistics. The challenge is to bring the study of direct discourse carried out on the corpus of Nisvai narratives, an oral language of Vanuatu, to other corpuses of oral languages. The annotations of these other corpora have been corrected and standardised by the combined efforts of the DoReCo and QUEST projects. We will see that while the reproducibility of a study on a language facilitates its criticism, the question of the portability of a study to other corpora requires that the latter meet standards and units that are interoperable from both a computational and linguistic point of view.

---

**MOTS-CLÉS** : comparaison linguistique, langues orales, documentation linguistique, corpus de terrain, portabilité, reproductibilité.

**KEYWORDS:** cross-linguistics, oral languages, linguistic documentation, eldwork corpora, annotations, portability, reproductibility.

---

## 1 Introduction

La présente communication est une réflexion sur l'étude linguistique de corpus de langues orales à travers des techniques associées aux sciences des données. Le propos est de montrer comment les concepts de reproductibilité et de portabilité sont appliqués à l'étude de corpus de langues orales. Il s'agit de mettre en place des pratiques et des méthodes qui facilitent la critique, le partage et la réutilisation des données et des résultats d'une étude.

Cette réflexion, ainsi que l'étude sur laquelle elle s'appuie, sont issues de QUEST, un projet de recherche financé par le ministère des Sciences allemand afin de promouvoir et valider des normes et bonnes pratiques pour élaborer de corpus de documentation linguistique ré-utilisables. Plus particulièrement, cette étude s'insère dans le sous-projet RefCo, une initiative dont l'objectif est de mettre à disposition des corpus de référence adaptée à la comparaison linguistique. Les corpus de documentation linguistique utilisées pour cette étude sont ceux qui ont été amendés par les projets DoReCo (Paschen et al. 2020) et QUEST (<https://cutt.ly/quest-project>). L'étude porte sur la réalisation du discours direct au sein de ces corpus, en commençant par le corpus de narrations nisvaies (Aznar 2019) afin mettre au point une base de références techniques, pour s'étendre à un ensemble de cinq corpus documentant les langues beja, mojeno, sanzhi dargha et arapaho, corpus qui ont été également corrigés par DoReCo et QUEST. C'est dans ce cadre que s'inscrit la réflexion sur les concepts de reproductibilité et de portabilité : la comparaison linguistique de langues orales à travers des documentations réalisées avec le logiciel ELAN, dont les textes sont des narrations monologiques et dont les annotations ont été vérifiées et corrigées par DoReCo et QUEST.

Un premier point à aborder avant de rentrer dans le cœur du propos est la terminologie. Les termes portabilité et reproductibilité sont sujets à débat dans de nombreuses disciplines scientifiques<sup>1</sup>, l'enjeu n'est ici pas de proposer des définitions conceptuelles précises mais de réfléchir à leur implémentation pratique dans un cadre précis. L'acceptation de portabilité s'inspire du sens proposé par Bird et Simons (2003) que l'on peut résumer comme la mise en place d'un ensemble de moyens et de pratiques permettant de réutiliser une documentation linguistique sur le long terme. Le concept est appliqué ici à l'étude des données linguistique et non à la documentation. Il s'agit de réfléchir aux pré-requis qui sont à considérer afin qu'une étude linguistique assistée par des traitements automatisés sur un corpus puisse être réalisée sur d'autres corpus. La portabilité de l'étude du discours direct s'inscrit donc dans une démarche de reproductibilité et de répliquabilité de la recherche. Reproductibilité est entendu ici comme la possibilité de répéter une étude est fourni

---

<sup>1</sup> Les définitions des concepts de répliquabilité, reproductibilité, répétabilité varient en fonction de la langue où le concept est défini, de la discipline ou de la position théorique de l'auteur. Il ne s'agit pas ici de faire une proposition définitive quant à ses termes mais simplement de mieux expliciter notre proposition (voir par exemple (Berez-Kroeker et al. 2018; Drummond 2009; Rey-Coyrehourcq et al., s. d.; McArthur 2019).

aux lecteurs à travers l'accès aux données et aux algorithmes nécessaires pour arriver au résultat décrit. Quant à la replicabilité, elle est définie comme la possibilité de refaire une même étude dans un contexte différent afin d'obtenir des résultats permettant de discuter l'étude de référence.

Maintenant que nous avons vu le contexte dans lequel s'insère cette comparaison des discours direct, nous allons aborder différentes étapes de l'étude algorithmique à travers le logiciel Jupyter pour ensuite décrire différents problèmes auxquels nous sommes confrontés lorsque nous souhaitons porter cette étude d'un corpus à l'autre.

## **2 Étudier le discours direct au sein du corpus Nisvai reproductible avec Jupyter**

La présente description repose sur le corpus de pratiques narratives nisvaies, un corpus d'enregistrements audio annoté avec le logiciel ELAN (Aznar, 2019). Dans le cadre de mon travail au sein de QUEST, le corpus a été soumis à l'évaluation de DoReCo et RefCo. À cette fin, les enregistrements audio et leurs annotations ont été rendus accessibles sous licence Creative Commons BY-NC-ND, une licence qui permet de partager et d'étudier ces données.

### **2.1 QUESTCorpora: un module python pour créer des données linguistiques à partir des annotations du corpus**

Afin de fournir aux corpus de QUEST une interface informatisée unifiée et partageable, je développe un module python nommé QUESTCorpora qui permet d'accéder via des objets pythons à ses différents composants : Corpus, Textes, Annotations. L'enjeu du module est de produire des données linguistiques comparables à partir des corpus.

Pour ce faire, le module s'appuie sur un ensemble de corpus gérés par le projet RefCo qui ont été créés et retravaillés de manière suffisamment similaires pour en extraire des données linguistiques comparables. La production de ces données linguistiques comparables passent toutefois par une politique de gestion des annotations. Cette politique se traduit par un ensemble de paramètres dédiés à chacun des corpus qui configurent les algorithmes de gestion des annotations.

Concrètement, QUESTCorpora inspecte les fichiers ELAN des corpus gérés par RefCo, les transforme en instances Python pour fournir une interface permettant de manipuler les annotations et de produire des informations à travers des traitements automatisés (expressions régulières, calculs et statistiques) sur chacune des annotations. Les résultats provenant de ces traitements sont ensuite transformés en un tableau de données afin d'être manipulables via les modules Python dédiés aux traitements des données (voir notamment Pandas, Statsmodels). Ces modules permettent de produire des représentations statistiques et visuelles de ces annotations et des résultats des traitements automatiques.

## 2.2 Jupyter pour faciliter la reproductibilité d'une étude de corpus

Dans les domaines impliquant une chaîne de traitements informatisés des données, l'utilisation d'environnement de partage documenté des algorithmes employés se développent (Stodden, Leisch, et Peng 2014). L'étude comparée des discours directs repose en partie sur des traitements automatisés des corpus de données linguistiques. L'utilisation de Jupyter (Thomas et al. 2016), un « carnet » en ligne, permet d'annoter les différentes étapes d'une chaîne de traitements informatiques dans un format qui facilite leur partage sous la forme d'une narration. D'un côté, le logiciel aide le chercheur lorsqu'il explore, teste ses hypothèses sur ses données ou travaille en équipe, du côté des relecteurs, il facilite la critique et l'appropriation des étapes ayant conduit aux résultats. L'utilisation d'un module Python pour interagir avec les corpus facilite l'intégration de ces données dans une chaîne de traitements automatisés, reproductible et partageable.

```
Import des modules nécessaires à cette étude:

In [ ]: import pandas as pd # help handling the data
import matplotlib.pyplot as plt # for representing the data
import numpy as np # not used fro now, just in case
import statsmodels.formula.api as sm # not used yet, just in case
import seaborn as sns # to produce charts and visualizations,
import re # to parse the texts and segments of textual annotation
import os

In [ ]: import QuestCorpora

Analyse du corpus

In [ ]: #Creation du corpus de données linguistiques

Study = QuestCorpora.Study("pour_LIFT", "22.11.2020", "list_texts.csv", "corpus_LIFT_RefCo.csv")
Study.get_generic_data()
Study.re_analysis("(dire|parler)", "morphology", '_dire')
Study.re_analysis(":", "transcription", "_deuxpoints_transcription")
Study.re_analysis(";$", "transcription", "_deuxpointsFin_transcription")
Study.re_analysis(":.+<[^\>]+$", "transcription", "_deuxpointsFin_transcription")
Study.re_analysis("[^\>]+", "transcription", "_uniquFin_transcription")
Study.re_analysis("<.+>", "transcription", "_lesDeux_transcription")
Study.re_analysis("[^\>]+<.+>[^\>]+", "transcription", "_monoUniquement_transcription")
Study.re_analysis(":.+:", "transcription", "_Multi_transcription")
Study.re_analysis(":", "translation", "_deuxpoints_traduction")
Study.write_data("Quest_Nisvai_DirectSpeech.csv", "csv")

In [ ]: # Import des données dans Pandas
datafile = "Quest_Nisvai_DirectSpeech.csv"
data = pd.read_csv(datafile)
```

Figure 1: Exemples d'étapes pour la production de données linguistiques à partir du module *QUESTCorpora*

La figure 1 montre un exemple de chaîne de traitements automatisé réalisé sur un corpus du projet QUEST, le corpus nisvai. La première étape consiste en l'import des modules qui seront sollicités dans le cadre de l'étude. La deuxième étape consiste en l'instantiation d'une étude sur corpus à travers des expressions régulières. Dans le cadre de cette étude, des recherches à l'aide d'expressions régulières sont appliquées à chacune des unités d'annotation du corpus. Si la recherche est concluante, elle retourne un 1 dans la ligne du tableau<sup>2</sup> correspondant à l'annotation, dans le cas contraire, un zéro est retourné.

La figure 2 présente quelques colonnes contenant les données du corpus nisvai intégré dans le module *QUESTcorpora*. Les deux dernières colonnes correspondent aux résultats des requêtes réalisées avec des expressions régulières sur les annotations. Ce sont à partir de ces colonnes contenant soient les données des annotations du corpus, soit les métadonnée associées aux

<sup>2</sup> Il s'agit ici plus précisément d'un objet DataFrame proposé par le module Python Pandas (<https://pandas.pydata.org/>).

	index	time_1	time_2	reference	transcription	morphology	translation	coherency	_dire	_deuxpoints_transcription
0	nisvai_25_174	589390	590480	T41.2015.174	Kusvai. -- Kusvai.	--	Kousvé. -- Kousvé.	False	0	0
1	nisvai_25_173	587480	589060	T41.2015.173	Nabol nyn ga=qan nyn ni, ga=cub urun.	histoire DE.I.R 3SG=être_comme DE.I.R INT 3SG=te...	Cette histoire est comme cela, elle se termine...	True	0	0
2	nisvai_25_172	581170	586240	T41.2015.172	Ga=hub nabu-n naho-n wantaim, nahemac ili ga=m...	3SG=lapider ??-3SG face-3SG INT démon DET 3SG=...	Il la jette en pleine face d'un coup, le mécha...	True	0	0
3	nisvai_25_171	574770	580350	T41.2015.171	Naremac ili ga=kai : «Ui, asi na=han-!?» Ga=ka...	démon DET 3SG=dire INTER qui 1SG=manger-INTR 3...	Le méchant dit : «Eh, qui est-ce que j'ai mang...	False	1	1
4	nisvai_25_170	572560	573550	T41.2015.170	Haiq qa=han a=nanaq sa-q!»	2SG 2SG=manger ART.P=mère PA-1SG	Tu as mangé ma mère !»	True	0	0

Figure 2: Extrait d'un tableau Pandas de données d'un corpus QUESTCorpora dans l'interface de Jupyter.

locuteurs, telles que l'âge, le sexe, le genre du texte, ou au texte et à sa situation de production, genre discursif, date, corpus ou enfin les résultats des requêtes, que sont produites les analyses sur les corpus.

### 3 D'un corpus à la comparaison de corpus

Maintenant que nous avons vu les étapes d'une étude informatisée reproductible sur un corpus, nous devons nous intéresser au portage d'une étude vers un autre corpus.

La comparaison s'appuie sur l'utilisation de corpus ouverts, c'est-à-dire dont la licence d'utilisation permet au moins la copie et la consultation. Toutefois, si l'ouverture des corpus participe à la portabilité d'une étude, cette qualité n'est pas suffisante pour permettre la portabilité. Il est également nécessaire que les annotations de ces corpus soient documentées et que les unités linguistiques annotées soient comparables. Enfin, que ce soit pour la reproductibilité ou la portabilité de l'étude, il est nécessaire d'avoir accès à la méthode qui a permis d'arriver aux résultats. Lorsqu'un article linguistique décrit les résultats issus d'une étude et leur interprétation, les algorithmes employés pour réaliser l'étude ne sont pas nécessairement fournis, en particulier lorsque les algorithmes ne sont pas informatisés. C'est la combinaison de ces qualités, d'un côté des données accessibles et normalisées, et de l'autre des algorithmes ouverts, qui permet la mise en place d'études reproductibles et portables.

#### 3.1 Les étapes du portage d'une étude à d'autres corpus

Afin de pouvoir adapter le comportement du module QUESTCorpora aux spécificités de chacun des corpus, chaque corpus est associé à un fichier de configuration spécifique<sup>3</sup>. Chaque corpus doit être étudié en fonction de l'étude. Dans notre cas, il s'agit d'étudier le discours direct, et plus particulièrement les différents silences qui entourent sa mise en scène. Dans cette section sont rapportées les différentes sources de difficultés quant à l'observation du discours direct de manière automatisée à travers les corpus.

#### 3.2 L'indexation temporelle par rapport à l'enregistrement audio

<sup>3</sup> Les fichiers de configuration sont écrits en TOML (voir <https://toml.io/en/>), un format de fichier qui vise à être lisible tout en étant analysable d'un point de vue informatique.

La première source de difficultés sont les variations d'un corpus à l'autre au niveau de l'indexation des unités d'annotation par rapport aux enregistrements audio. Il s'agit de la stratégie qui a été mise en place lors de la transcription afin de segmenter le flot de la parole en unités d'annotation. Une première variation provient des possibilités offertes par ELAN, qui permet de produire des annotations soit continues, soit discontinues, c'est-à-dire soit collées les unes aux autres ou au contraire séparées par temps non annotés<sup>4</sup>. Plusieurs pratiques ont été relevées au sein des corpus DoReCo-QUEST :



Figure 3: Exemple d'indexation discontinue des unités d'annotation, avec une unité sur deux ne contenant qu'un silence.

- Une indexation par rapport aux groupes de souffles, voir figure 3 : les unités d'intonation correspondent aux moments où le locuteur parle, lorsqu'il marque une pause, que ce soit pour reprendre son souffle, structurer son texte ou simplement dans le cadre d'une dysfluence (réflexion ou hésitation), alors une segmentation est opérée.
- Une indexation syntactico-sémantique, voir figure 5: le linguiste segmente en fonction d'une unité linguistique qu'il identifie, potentiellement avec l'aide d'un locuteur de la langue. Cette segmentation correspond le plus souvent à une segmentation en phrases. Elle peut prendre en considération les pauses marquées par les locuteurs pour borner les unités d'annotation, mais ses unités d'annotation peuvent comprendre des pauses en leur sein.
- Une indexation continue, voir figure 4, produisant des unités d'annotation aussi bien pour les temps de paroles que pour les temps de silences.
- Une annotation continue, voir figure 6, intégrant les temps de silence dans unité d'annotation au sein de ses bornes.

<sup>4</sup> Nous verrons par la suite que la non-annotation peut être considéré comme une forme de silence.

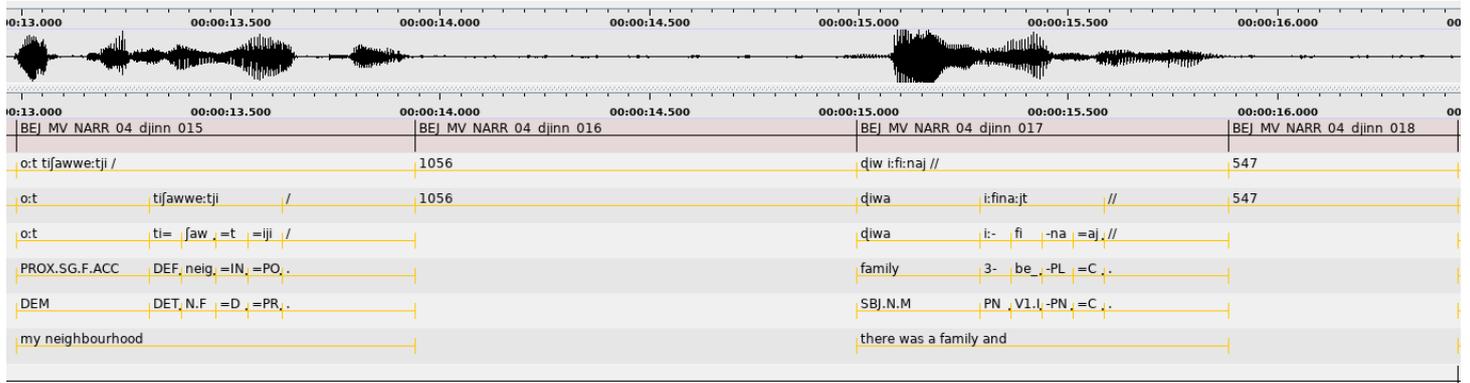


Figure 4: Exemple d'indexation continue des unités d'annotation, avec une unité sur deux ne contenant qu'un silence.

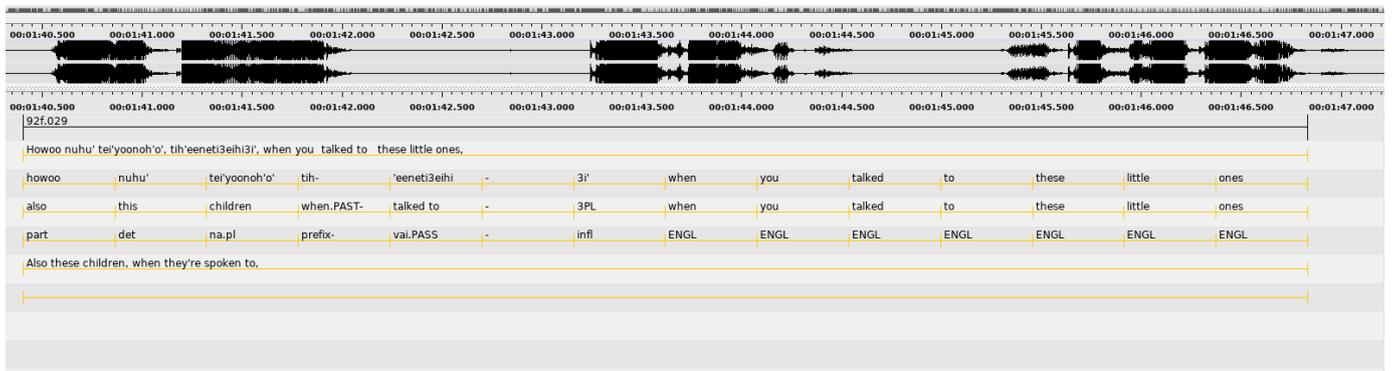


Figure 5: Exemple d'indexation discontinue avec des silences absorbés au sein de l'annotation et dont les limites ne sont que peu prises en compte.

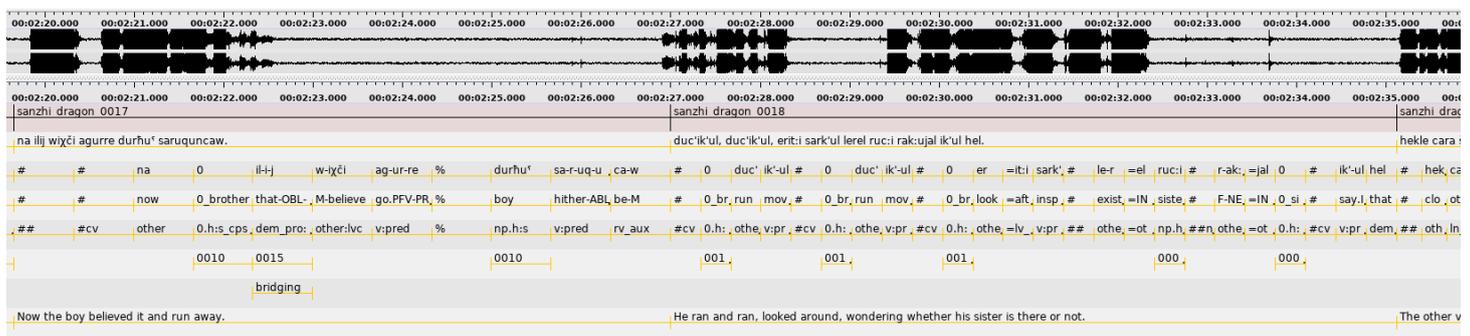


Figure 6: Exemple d'indexation continue des unités d'annotation avec intégration du temps de silence dans la suite de l'unité d'annotation

Dans le cadre de l'étude du silence de la réalisation du discours direct, deux paramètres sont identifiables afin de distinguer les pratiques des linguistes : la continuité ou non des unités d'annotation et la prise en compte du silence dans les unités d'annotation

## Discontinuité

## Continuité

<b>Prise en compte du silence</b>	Silence représenté par la non-Silence correspondant à des annotations	Silence correspondant à des annotations
<b>Non prise en compte</b>	Silence rogné par les unités d'annotation	Silence intégré avec le groupe de parole le précédent

À noter qu'une autre pratique a été réalisé par MAUS (Strunk, Schiel, et Seifart 2014), le système employé par DoReCo-QUEST afin d'indexer les graphèmes de la transcription à l'enregistrement. Dans ce cas, le logiciel indique par la balise <p:> les silences au sein d'un tier dont la segmentation est continue.

### 3.3 La transcription et l'interprétation des systèmes d'écriture

Parmi les corpus gérés par les projets DoReCo et QUEST, si à un texte est associé une couche de transcription, il est également possible dans certains corpus que plusieurs couches se rapportent à la transcription. Dans ce cas, il s'agit . L

Les systèmes d'écriture utilisés pour transcrire les différentes langues n'ont pas les mêmes valeurs. Ceci est résolu à l'aide d'un tableau associant les graphèmes employés par les linguistes à leurs valeurs phonétiques ou phonologiques.

En ce qui concerne la représentation du discours direct au sein de la transcription, un certain nombre de choix ont été opérés par les transcripateurs. En fonction de la langue maternelle de la personne ayant réalisée la transcription, les conventions typographiques pour représenter le discours direct peuvent variée : utilisation, ou non, des guillemets français ou anglais ; marquage du début du discours direct par deux points, transcription ou non de l'interjection employée par le locuteur. Ces variations ont pour conséquences que le discours direct ne puissent pas être identifié de manière systématique à travers la transcription.

### 3.4 Les traductions

La traduction de textes oraux est un aspect central de la documentation linguistique, mais elle ne fait l'objet finalement que de très peu de discussions de la part des linguistes travaillant sur des langues orales. La recommandation la plus courante est de traduire dans une des langues véhiculaires internationales, ou tout au moins, une langue plus connue que la langue documentée. Ce manque de discours sur la pratique explique que les textes ne soient pas traduits selon les mêmes principes en fonction des linguistes ou d'un projet à l'autre. Les destinataires des traductions ne sont pas les mêmes d'un projet à l'autre : produire une ressource bilingue pour l'école locale, annoter pour des linguistes. C'est en partie à travers l'identification de ces destinataires qu'un nombre de choix de traduction peuvent être réalisés : explicitation ou non du vocabulaire, la traduction systématique des termes, utilisation d'une typographie spécifique à un genre littéraire de destination, etc.

Du point de vue de DoReCo et de QUEST, la traduction est la couche la plus difficile à contrôler. S'il est demandé à la personne fournissant le corpus d'indiquer la langue de traduction et de fournir

un glossaire des termes de la langue source, vérifier l'adéquation entre ce qui est dit dans la langue source et ce qui est retranscrit dans la langue cible relève des compétences du linguiste sur le terrain et des personnes avec lesquelles il ou elle travaille.

En ce qui concerna l'étude du discours direct à travers la traduction, il apparaît au sein des corpus DoReCo-QUEST que si la couche de transcription ne représente pas systématiquement le discours direct, celui-ci est presque systématiquement représenté au sein de la couche associée à la traduction. La couche de traduction est un moyen plus fiable pour identifier les unités d'annotation contenant du discours direct. Certains corpus possèdent plusieurs couches de transcription qui se différencient alors au niveau de leur portée. Ainsi, pour le Beja, une première couche de traduction propose une correspondance 1:1 à la segmentation en unités d'annotation alors qu'une deuxième couche de traduction englobe plusieurs unités d'annotation afin de permettre une traduction moins littérale. Enfin, en fonction des projets dans lesquels les corpus ont été produits, les transcriptions ont pu être traduites en plusieurs langues véhiculaires.

### **3.5 L'annotation morphologique**

Une ou plusieurs couches d'annotation morphologique sont présentes dans chacun des corpus DoReCo-QUEST. Si les annotations morphologiques sont suffisamment formalisées d'un point de linguistique afin de pouvoir facilement associer visuellement un morphème à sa glosse, de nombreux corpus n'ont pas été annotés de manière systématique. Cela entraîne des difficultés dans les traitements automatisés des annotations. Ainsi, dans certains corpus, les linguistes peuvent indiquer des synonymes aux termes décrits en utilisant plusieurs conventions typographiques au sein d'un même texte. Il est alors nécessaire de les identifier, voire de réduire ses différentes conventions à un seul et unique procédé lorsque nous sommes certain qu'il s'agit d'un même et unique sens.

Un autre problème est l'hétérogénéité des couches d'annotation de la morphologie. Certains corpus possèdent plusieurs couches contenant des informations liées à la morphologie de la langue. La comparaison de ce contenu en devient alors plus difficile car les couches d'annotation ne sont pas équivalentes les unes aux autres. La solution retenue est de créer une couche d'abstraction contenant les différentes annotations afin de pouvoir comparer les informations morphologiques d'une langue à l'autre.

## **4 Conclusion**

Le portage d'une étude assistée par des traitements automatisés sur le discours direct à travers des corpus de langues orales requiert de prendre en compte les particularités de chacun de ces corpus. Il apparaît toutefois que ces particularités ne sont pas infinies et peuvent être identifiées afin de faire l'objet d'une gestion des annotations dédiées afin d'obtenir des données linguistiques comparables. C'est ce qui est mis en place à travers le module Python QUESTCorpora. Afin d'éviter de réinventer un outil qui existe déjà, la réutilisation de multitool, actuellement pris en charge par Delafontaine (<https://github.com/DoReCo/multitool>) dans le cadre du projet DoReCo, est en cours d'évaluation.

Nous concluons sur le fait que si la reproductibilité d'une étude est un critère de validation d'une étude, lorsque cette reproductibilité est combinée avec des corpus répondant aux mêmes normes, nous avons alors la possibilité de porter les études d'un corpus à l'autre afin de procéder à des comparaisons linguistiques. Finalement, en abordant la question de la portabilité d'une étude, lorsque cette problématique est combinée à la répétabilité d'une analyse algorithmique informatisée telle qu'offerte par le logiciel Jupyter, l'objectif est alors d'arriver à une répliquabilité typologique.

## Références

- Aznar, Jocelyn. 2019. « Narrer une nabol : La production des textes nisvais en fonction de l'âge et de la situation d'énonciation, Malekula, Vanuatu ». Marseille: EHESS.
- Berez-Kroeker, Andrea L, Lauren Gawne, Susan Smythe Kung, Barbara F Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. « Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field ». *De Gruyter, Linguistics*, 56 (1): 18. <https://doi.org/10.1515/ling-2017-0032>.
- Bird, Steven, et Gary F. Simons. 2003. « Seven Dimensions of Portability for Language Documentation and Description ». *Language, Language*, 79 (3): 557-82. <https://doi.org/10/d95m65>.
- Drummond, Chris. 2009. « Replicability Is Not Reproducibility: Nor Is It Good Science ». *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26 Th ICML*, 4.
- McArthur, Sally L. 2019. « Repeatability, Reproducibility, and Replicability: Tackling the 3R Challenge in Biointerface Science and Engineering ». *Biointerphases* 14 (2): 3. <https://doi.org/doi/10.1116/1.5093621>.
- Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, et Frank Seifart. 2020. « Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo) ». *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* 12: 2657–2666.
- Rey-Coyrehourcq, Sébastien, Robin Cura, Laure Nuninger, Julie Gravier, Lucie Nahassia, et Ryma Hachi. s. d. « Vers une recherche reproductible dans un cadre interdisciplinaire: enjeux et propositions pour le transfert du cadre conceptuel et la répliquabilité des modèles », 25.
- Stodden, Victoria, Friedrich Leisch, et Roger D. Peng, éd. 2014. *Implementing Reproducible Research*. The R Series. Boca Raton: CRC Press.

- Strunk, Jan, Florian Schiel, et Frank Seifart. 2014. « Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora Using WebMAUS ». *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3940--3947.
- Thomas, Kluyver, Ragan-Kelley Benjamin, Pérez Fernando, Granger Brian, Bussonnier Matthias, Frederic Jonathan, Kelley Kyle, et al. 2016. « Jupyter Notebooks &ndash; a Publishing Format for Reproducible Computational Work ows ». *Stand Alone*, 87–90. <https://doi.org/10/gf48c9>.