

Projet CONCORDANCE – « Concordances lexicales en langue tahitienne entre le dictionnaire en ligne de l'Académie tahitienne et la bibliothèque scientifique numérique Ana'ite de l'Université de la Polynésie française »

Bilan scientifique août 2021

Rappel du contexte et des objectifs du projet

Le projet Concordance s'inscrit dans le cadre d'une collaboration entre plusieurs partenaires : l'Université de la Polynésie française (UPF), le Service du patrimoine archivistique et audiovisuel de la Polynésie française (SPAA), l'Académie tahitienne et la société d'informatique Tahiti ingénierie.

Depuis 2017, l'UPF et le SPAA se sont entendus sur la nécessité de créer une bibliothèque numérique qui permette d'accéder librement en ligne à des ressources archivistiques rares tombées dans le domaine public concernant la Polynésie française. Ainsi est née *Ana'ite*¹, la bibliothèque scientifique numérique polynésienne (anaite.upf.pf), administrée conjointement par l'UPF et le SPAA, dans le cadre d'une convention de coopération numérique². *Ana'ite* a pour ambition d'accroître la capacité de recherche et d'analyse des contenus grâce à la conversion numérique des documents au format texte et à l'intégration de métadonnées de qualité. Parmi les collections déposées sur *Ana'ite*, on trouve de nombreux documents en tahitien. Choisi comme langue d'évangélisation par les missionnaires anglicans, le tahitien est l'une des rares langues océaniques pour lesquelles on dispose d'une production écrite significative, y compris laïque, depuis le XIX^e siècle. À l'époque du Protectorat français sur le royaume de Pōmare, à partir de 1842, puis des Établissements français de l'Océanie créés en 1880, un corps d'interprètes assistait les gouverneurs successifs dans l'administration de la colonie naissante, consolidant ainsi la place du tahitien comme langue véhiculaire de cette région du Pacifique centro-oriental. Une collection déposée sur *Ana'ite* illustre particulièrement cette situation sociolinguistique, celle du périodique *Te Vea no Tahiti*. Il s'agit du premier journal hebdomadaire en langue tahitienne, créé en 1850 par une ordonnance de Louis Alphonse Bonard, commandant de la division navale de l'Océanie et commissaire aux îles de la Société. Cet hebdomadaire paraîtra jusqu'en 1859. Les premiers numéros sont hélas manquant, mais le SPAA dispose de copies numériques de 332 numéros, de janvier 1851 à janvier 1859, désormais déposés sur *Ana'ite*.

L'UPF est aussi partenaire de l'Académie tahitienne depuis 2013 pour le développement de la version numérique de son dictionnaire tahitien-français. Ce dernier est hébergé sur une base, nommée *Anareo*³, en accès réservé sur le Cloud et conçue avec la société d'informatique Tahiti Ingénierie. Importés à partir d'un fichier MS-Word mis à disposition par l'Académie tahitienne, les 12 864 articles qui composent le dictionnaire ont été systématiquement structurés avec

¹ *Ana'ite* signifie littéralement « la grotte des savoirs ». Les anciens Polynésiens conservaient les reliques de leurs ancêtres dans des grottes.

² Convention n°03091 MCE/SPAA du 12 mai 2017 entre l'UPF et le ministère de la Culture de la Polynésie française, dont dépend le SPAA.

³ Littéralement, « la grotte des langues ».

des balises conformes aux standards communs des logiciels lexicographiques (types Toolbox ou Flex). La base ainsi constituée nourrit directement le dictionnaire tahitien-français en ligne de l'Académie tahitienne (www.farevanaa.pf/dictionnaire.php).

Le projet Concordance avait pour objectif principal de développer l'interopérabilité entre la bibliothèque numérique *Ana'ite* et le dictionnaire tahitien-français en ligne de l'Académie tahitienne. Plus précisément, il s'agissait de constituer un corpus de ressources textuelles en tahitien, déposé sur *Ana'ite* et rendu exploitable pour une recherche plein texte, et d'implémenter une fonction de concordance qui permette à un utilisateur qui consulte un mot tahitien dans le dictionnaire en ligne d'accéder à un répertoire d'occurrences référencées de ce mot dans le corpus.

Dans la conception initiale du projet, trois périodiques étaient plus particulièrement envisagés comme cibles prioritaires :

- 62 numéros du *Te Vea no Tahiti*, hebdomadaire, 2 p., 33 cm, de 1852 à 1853 ;
- 11 numéros du *Te Vea Maohi*, mensuel créé par le gouverneur Bouge, 8 à 16 p., 32 cm, de 1937 à 1941 ;
- 30 numéros du *Torea*, mensuel privé d'informations générales et locales, 12 à 16 p., 25 cm, de 1937 à 1939

Les trois principales actions nécessaires à la réalisation de la fonction de concordance étaient les suivantes :

1. Établir par transcription une première version anastatique au format texte du corpus sélectionné, ce corpus étant fourni initialement au format image JPEG par le SPAA. Cette version devait correspondre exactement, pour son orthographe, à la forme originelle du document ;
2. Produire une seconde version de ces textes, cette fois dans la graphie officielle contemporaine. En effet, la graphie du tahitien a évolué depuis le XIXe siècle, pour la notation de l'occlusive glottale et de la longueur vocalique, mais aussi pour le découpage morphémique, plusieurs suffixes n'étant pas accolés graphiquement à leur base lexicale dans les premières écritures. L'appariement avec les entrées du dictionnaire contemporain imposait donc un travail préalable de conversion orthographique ;
3. Développer et implémenter une fonction informatique de concordance entre la base de données *Anareo* qui héberge les articles du dictionnaire en ligne de l'Académie tahitienne et le corpus de textes déposé sur *Ana'ite* et ses métadonnées. Plus précisément, un utilisateur du dictionnaire en ligne, en partant d'une entrée en tahitien, doit pouvoir visualiser un répertoire d'occurrences de cette entrée, issues du corpus et donnant pour chaque occurrence un contexte restreint (50 signes avant et après) et un contexte élargi (400 signes avant et après), le mot cible figurant au milieu de ces contextes, chaque contexte étant référencé.

Pour rappel, la subvention attribuée par COLLEX-Persée à l'UPF, soit 24 912 €, devait contribuer aux deux premières étapes de production d'une version anastatique, puis d'une version orthographique contemporaine des textes sélectionnés, à partir des fichiers images, grâce au financement de contrats étudiants dédiés aux tâches de transcription et d'adaptation orthographique.

Bilan des actions réalisées

Production de la version anastatique au format texte

Suite au lancement du projet en mars 2019, nous avons obtenu auprès du SPAA, outre les documents cités plus haut, l'intégralité des copies des numéros du *Te Ve'a no Tahiti* (TVNT), de janvier 1851 à janvier 1859, soit 332 numéros, au lieu des 62 initialement prévus.

Plusieurs considérations nous ont conduit à prioriser le traitement de ce périodique. Les copies numériques du TVNT ont été réalisées par le SPAA dans les années 1980. La plupart des documents originaux ont depuis disparu des réserves du SPAA. Il n'a donc pas été possible de réaliser de nouvelles numérisations. Même après manipulation pour renforcer le contraste des caractères, la qualité des copies disponibles, striées et comportant parfois des inscriptions manuscrites ou des motifs du verso en transparence (cf. Figure 1), ne permettait pas de recourir à un logiciel de reconnaissance de caractère (OCR) de manière satisfaisante.

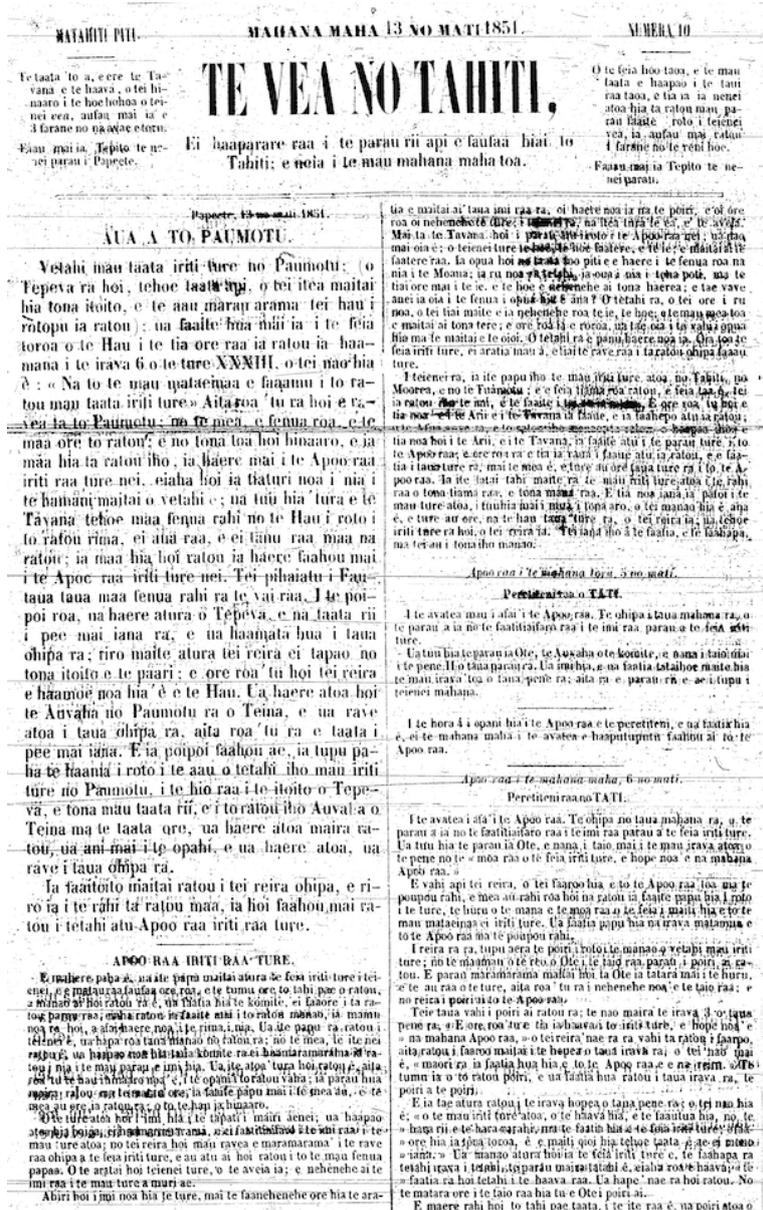


Figure 1 – Exemple d'une copie numérique du *Te Ve'a no Tahiti*. Première page du n°10 du 13/03/1851

Les tests réalisés avec différents OCR ont produit des résultats médiocres. Lors de phases de prétests, nous avons comparé le temps moyen nécessaire à la correction d'une version texte obtenue par océrisation et celui d'une retranscription à l'œil humain produite directement à partir du fichier image par un locuteur expert du tahitien. La seconde option s'est avérée pratiquement deux fois plus rapide et génératrice de moins d'erreurs dans la version finale, que la première.

Les copies des deux autres périodiques, en revanche, réalisées plus récemment, offrent un OCR de bien meilleure qualité et ne nécessitent pas une retranscription complète dans la production du fichier texte anastatique. Nous avons donc reporté à une étape ultérieure leur intégration au corpus de la fonction de concordance.

Cette première étape de production d'une version texte anastatique du corpus a nécessité l'identification d'un outil de transcription. Notre choix s'est porté sur la plateforme TACT (<https://tact.demarre-shs.fr/>) développée par l'Université Grenoble Alpes grâce à un financement IDEX. Après prise de contact et accord des opérateurs de cet outil, les copies des documents du corpus y ont été déposées au format JPEG (cf. Figure 2).

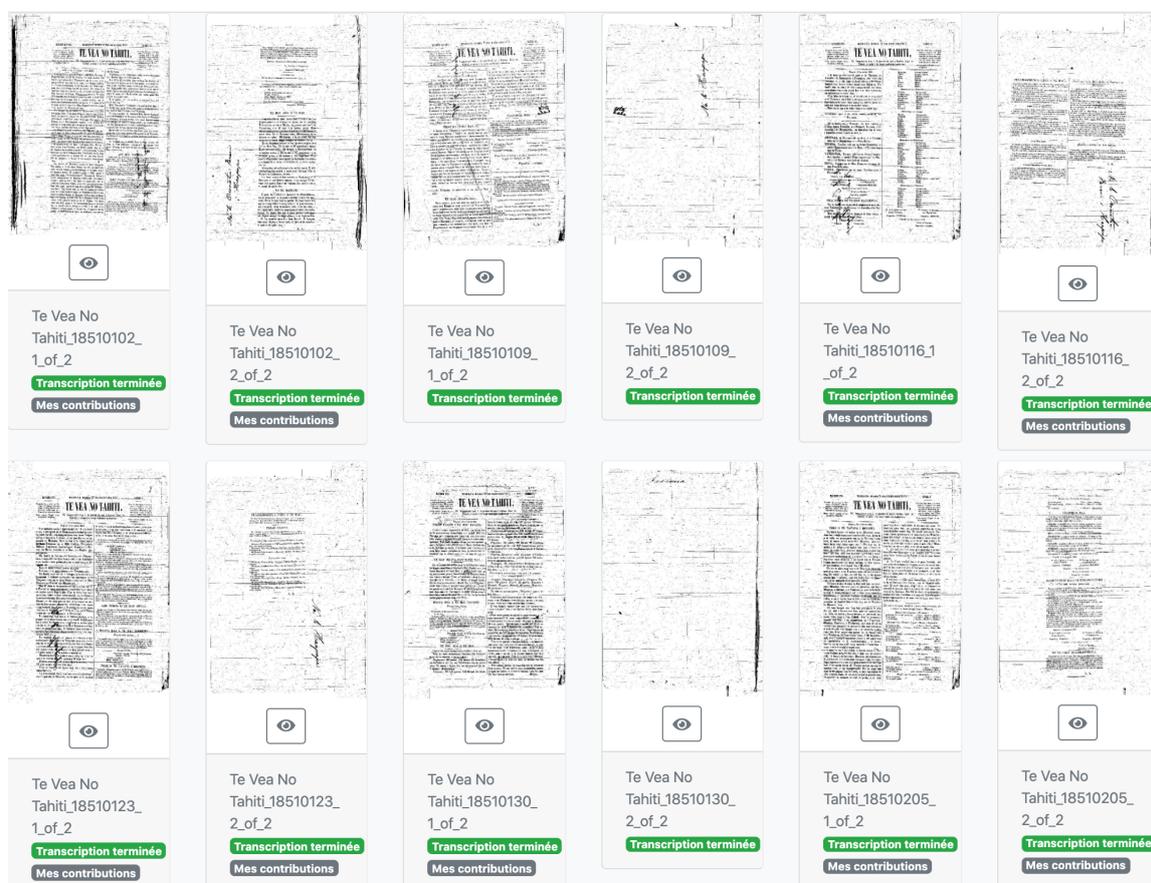


Figure 2 - Capture d'écran de la plateforme TACT. Extrait de l'index des numéros du TVNT.

Depuis septembre 2019, 8 étudiants⁴ (5 étudiantes et 1 étudiant de 3^e année de licence de Langues et civilisations polynésiennes et 2 étudiante de M1 du master Langues, cultures et

⁴ AUMERAND Marie-Ange, HUGUES Albert, PARENT Takurua, POHERUI Haiata, MAUATI Maeva, MOORIA Homai, RAI Taraina et TUAHIVA Pauline.

sociétés en Océanie), locuteurs du tahitien, ont été recrutés puis formés pour assurer la première étape de transcription anastatique sur la plateforme TACT (cf. Figure 3).

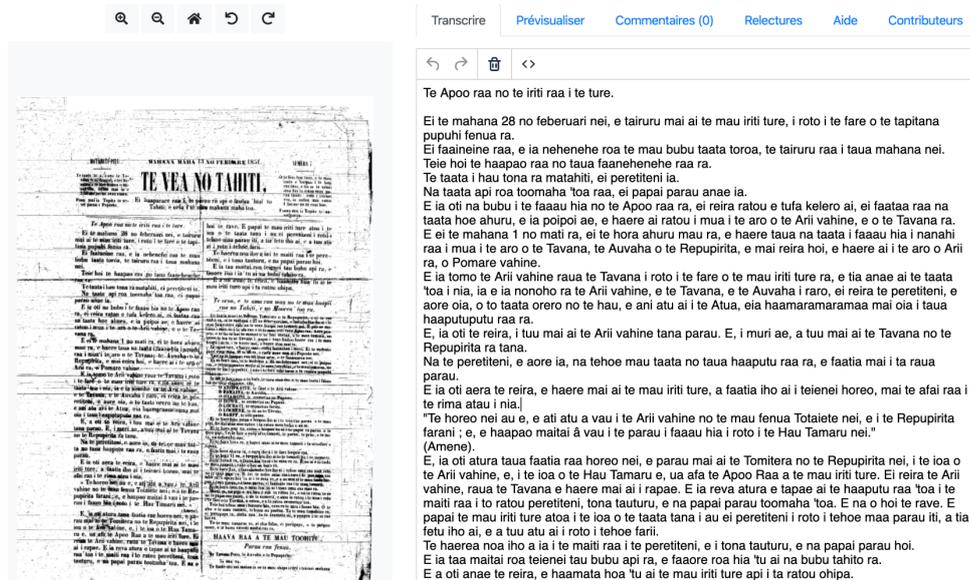


Figure 3 - Capture d'écran de la plateforme TACT. Dans la fenêtre de gauche, l'image d'une page du TVNT ; dans celle de droite, la transcription anastatique réalisée par un étudiant vacataire.

Le TVNT a ainsi été retranscrit en intégralité. Les versions textes sont disponibles sous la forme de fichiers XML exportables à partir de la plateforme TACT.

Production d'une seconde version dans l'orthographe contemporaine

Un outil informatique a été développé avec la société Tahiti ingénierie, grâce aux financements de la Maison des sciences de l'Homme du Pacifique et du laboratoire EASTCO de l'UPF. Implémenté dans la base de données *Anareo* qui héberge le dictionnaire en ligne de l'Académie tahitienne, cet outil permet de corréliser chaque mot d'un texte avec une entrée du dictionnaire, au numéro d'homonyme près, de manière semi-automatique, et de produire une version de ce texte convertie dans l'orthographe tahitienne officielle contemporaine. Plus précisément, l'outil permet de réaliser les tâches suivantes :

- Le texte cible est découpé en mots, chaque mot étant assigné à une ligne de travail.
- à partir d'une chaîne de caractères formant un mot dans la version anastatique du texte, l'algorithme recherche dans la base du dictionnaire les lemmes qui lui correspondent, à deux variables près : la présence ou l'absence d'une occlusive glottale représentée par une apostrophe ; la présence ou l'absence d'une longueur vocalique, représentée graphiquement par un macron. S'il n'y a qu'un seul candidat possible dans la base, le mot source est corrélé automatiquement à l'entrée du dictionnaire (cf. Figure 4, exemple de « fenua »). S'il y a plusieurs graphies possibles, un locuteur expert doit choisir la graphie contemporaine qui convient (cf. Figure 4, exemple de « papaa ») puis, le cas échéant, le numéro d'homonyme qui convient sémantiquement, s'il y a plusieurs homonymes possibles.

Détail Texte				
N°	Orthographe origine*	Cible correspondance	Orthographe corrigée	Lemme
642	te			
643	mau			
644	fenua			fenua
645	papaa			
646	O			pāpa'a
647	te			pāpā'a
648	aratai			papa'ā
649	hoi			ho'i
650	teienei			teienei
651	ture			ture

Figure 4 – Capture d'écran de l'outil de corrélation. Le texte original linéarisé se trouve dans la 2^e colonne (Orthographe origine). Les lignes en vert correspondent aux items reconnus automatiquement. À la ligne 644, le mot « fenua » a été corrélé automatiquement à l'entrée du dictionnaire car il n'existe pas d'allographes. À la ligne 645, la séquence « papaa » doit être corrélée manuellement à l'entrée pertinente du dictionnaire car il existe trois correspondants orthographiques contemporains possibles : *pāpa'a*, *pāpā'a* ou *papa'ā*.

Pour aider la sélection de l'entrée qui convient, l'utilisateur dispose de deux outils optionnels : 1. les variantes orthographiques du dictionnaire associées à la séquence de caractères, et leur définition, peuvent être visualisées ; 2. le contexte élargi d'occurrence du mot peut être affiché dans une fenêtre (cf. Figure 5).

The screenshot shows the 'Détail Texte' window with a list of words. A context window is open for the word 'papa'ā' (line 645). The context text reads: 'hoi i imi hia i te tapati i mairi aenei ; ua haapaa atu hia hoi ia , ei haamaramarama, e ei faaitiaifaro i te imi raa i te mau ture atos; no tei reira hoi mau ravea e maramarama i te rave raa ohipa a te feia iriti ture, e au atu ai hoi ratou i to te mau fenua papaa. O te aratai hoi teienei ture, o te aveia ia; e nehenehe ai te imi raa i te mau ture a muri ae. Ahiri hoi i imi noa hia te ture, mai te faanehenehe ore hia te aratai e maitai ai taua imi raa ra, oi haere noa ia na te poiiri, e oi ore ra oi nehenehe te ture,'.

Figure 5 – Capture d'écran de l'outil de corrélation. Les articles du dictionnaire des variantes orthographiques contemporaines associées à la forme orthographique de départ « papaa » apparaissent à droite, avec leur définition. Une fenêtre au centre permet de visualiser le contexte d'occurrence du mot cible.

- Une fois le travail de corrélation à l'homonyme près achevé, l'outil génère la seconde version du texte, dans la graphie officielle contemporaine (cf. Figure 6).

The screenshot shows the 'Visualisation' window with two columns: 'Graphie d'origine' and 'Graphie moderne'. The text in the 'Graphie d'origine' column is in a historical script, while the 'Graphie moderne' column shows the same text converted to modern orthography. The text is a historical document from 1851, starting with 'Papeete 13 nō māti 1851 AUA A TŌ PAUMOTU.'

Figure 6 - Capture d'écran de l'outil de corrélation. Dans la fenêtre de gauche, l'image du texte de départ ; dans celle de droite, la version convertie dans l'orthographe contemporaine, après traitement.

L'opération de corrélation-conversion nécessite une grande rigueur et un haut niveau de compétence en tahitien, à la fois en compréhension écrite et en analyse grammaticale. La langue tahitienne est de type isolant et n'a pratiquement pas de morphologie grammaticale (ni déclinaison, ni conjugaison), ce qui est un facteur facilitant pour les procédures d'appariement avec les entrées du dictionnaire car il n'y a pas à gérer la relation d'une entrée à plusieurs formes fléchies. En revanche, cette langue comporte de très nombreux homonymes, lesquels nécessitent un arbitrage expert. Une fois traités par les étudiants, les corrélations sont systématiquement vérifiées par l'équipe des enseignants-chercheurs du projet. Le rendement moyen du travail est de 600 mots traités à l'heure, soit 10 mots par minutes. Sachant qu'il y a en moyenne 700 mots par page du TVNT, il faut compter 1 heure et 10 minutes en moyenne pour le traitement de corrélation-conversion d'une page.

A ce jour, environ 200 pages du corpus de textes anastatiques ont été corrélées et converties dans la graphie contemporaine. Le travail se poursuit et nous prévoyons la conversion complète, avec relecture de validation, de tous les TVNT avant décembre 2021.

Implémentation d'une fonction de concordance

Après accord formel de l'Académie tahitienne⁵, Tahiti Ingénierie a implémenté dans le dictionnaire en ligne tahitien-français la fonction de concordance qui permet de révéler les occurrences d'une entrée du dictionnaire dans un corpus de textes.

Une fois l'entrée du dictionnaire sélectionnée, un lien actif s'affiche en bas de l'article avec la mention « Voir les concordances ». Le nombre d'occurrences représentées dans le corpus est précisé entre parenthèses (cf. Figure 7).

The screenshot shows a search interface with a search bar containing 'purumu' and an 'OK' button. Below the search bar are three dropdown menus: 'Rechercher en tahitien et français', 'Le mot ou locution complète', and 'Sans tenir compte des accents'. Below the search bar, the results are displayed as follows:

1 résultat trouvé dans les entrées en tahitien.
Aucun résultat dans les entrées en français.

Résultat trouvé dans les entrées en tahitien :
PURŪMU

n.c.
1°) balai, brosse. *'Ua mo'e tōna purūmu niho.* 'Il a perdu sa brosse à dents.' ; *PURŪMU NĪ'AU* 'balai fait avec des nervures de feuilles de cocotier' ; *PURŪMU PĒNI* 'pinceau'
2°) route, chemin, rue, passage. *'Ua api te purūmu i te pere'o'o.* 'La route est encombrée de voitures.' ; *PURŪMU ARA RAU* 'chaussée à plusieurs voies' ; *PURŪMU ARA RUA* 'chaussée à deux voies' ; *PURŪMU HA'APA'UMA* 'route à forte déclivité' ; *PURŪMU HĀTUA* 'route de ceinture' ; *PURŪMU TĀRAVA* 'route transversale' ; *PURŪMU PAEĀTAHI* 'chaussée à sens unique' ; *PURŪMU PAEĀRUA* 'chaussée à double sens' ; *PURŪMU 'ĀURI* 'chemin de fer' - Cf. *ara* (1), *'ē'a*, *arati'a*, *'ē'ati'a*, *haere'a*, *ta'iara*.
MA'I PURŪMU 'maladie vénérienne'
vi. balayer, brosser. *'Ua purūmuhia te fare.* 'La maison a été balayée.'
 Variante : *porūmu*, *purōmu*
 Emprunt : anglais *broom*

[Voir les concordances \(10 occurrences\)](#)

6 285 048 recherches

Figure 7 - Capture d'écran du dictionnaire tahitien-français en ligne de l'Académie tahitienne. En bas figure le lien vers la fonction de concordance.

En cliquant sur ce lien, l'utilisateur arrive sur une page où sont répertoriées les occurrences en contexte limité, de moins d'une ligne. En cliquant sur une ligne, l'occurrence du mot dans le contexte élargi précédé des références complètes de la source apparaissent dans une fenêtre (cf. Figure 8).

⁵ Courrier 2020/08 FD-hl du 02 juillet 2020 de Mme Flora Aurima Devatine, directrice de l'Académie tahitienne.

Concordance du mot purūmu

Les exemples suivants sont issus d'un corpus de textes authentiques dont certains sont déposés sur la Bibliothèque scientifique numérique polynésienne *Ana'ite*. La fonction de concordance vise à accroître les exemples disponibles pour illustrer l'emploi des mots tahitiens dans des documents authentiques avec un écart diachronique d'environ un siècle et demi, du XIXe au XXe siècle. Cette fonction a une visée descriptive. Les occurrences n'ont pas été systématiquement relues et validées par l'Académie tahitienne et peuvent présenter un écart dans leur forme ou leur usage par rapport aux indications normatives du dictionnaire de l'Académie tahitienne.

<p>Auteur : Traducteur : Année : 1851 Titre : Te Vea No Tahiti_1851/02/05_page 1 Titre de la publication : Ville : Papeete Editeur : Page(s) : 1 Titre abrégé : VNT18510205-1 Extrait : Titauhia mai e te ture, 'o tō Teavaro ana'e rā ho'i tei 'ore i t'ā. Te vahie rā ho'i, 'ua 'ohihia ia, 'e 'ua 'āpapa-māte-hia i tahatai, mai tei fa'a'itehia i te fa'a'era'a nō te mahana 6 nō Eperāra 1850. Te huru i'itoto n'i ato'a rā ho'i tahi pae mata'eina'a i Moorea i te raverā i te purūmu, 'oia ho'i 'o Maatea, 'o Haumi, 'e 'o Teavaro, e mea nehenehe mata'i te ma'a purūmu i oti i roto i tei reira tau mata'eina'a, e purūmu i'itō'itō'ara i'a mata'i ho'i tō reira, 'e 'ua 'ō-māte-hia te hī, e au tei reira tau purūmu 'a fa'a'irohia 'ei hi'ora'a, nā tō Tahiti 'e tō Moorea ato'a. Te mau ha'apitira'a ato'a ho'i, panapanu n'i noa iho ai, 'aita i fa'aru'e-roa-hia, mai tei mana'ohia ra. 'Aita rā rāto'u i huru 'ite re'a, 'e 'aita roa atu i huru fāto noa a'e</p>	
<p>Voir le document original Voir sa transcription</p>	
<p>'o Maatea, 'o Haumi, 'e 'o Teavaro, e mea nehenehe mata'i i te ma'a purūmu i oti i roto i tei reira tau mata'eina'a, e purūmu tū ai'arao u'a</p>	
<p>'o te ture ia i tu'uhia mai e Hoaere, nō te mau purūmu. 'Ua fa'a'itehia mai i roto i taua ture ra ē ; 'ia</p>	
<p>n'i a i taua purūmu ra. 'E 'ia 'etu ho'i te pua'a i te purūmu ra, nā te fatu pua'a ia e fa'anehenehe fa'ahou ; 'o te 'ohipa rarahi</p>	
<p>fa'a'itehia mai i roto i taua ture ra ē ; 'ia oti te purūmu i roto i te mata'eina'a i te ravehia e te Hau, nā</p>	
<p>vāere haere i te 'aihere n'i e tupu mai i ni'a i taua purūmu ra. 'E 'ia 'etu ho'i te pua'a i te purūmu ra, nā te</p>	

Le développement de la fonction de concordance a reçu le soutien des partenaires suivants : la Maison des Sciences de l'Homme du Pacifique (MSH-P), l'Équipe d'Accueil Sociétés Traditionnelles et Contemporaines en Océanie (EASTCO), la Bibliothèque universitaire et la Direction des systèmes d'information de l'Université de la Polynésie française, le Service du Patrimoine Archivistique et Audiovisuelle (SPA), l'Académie tahitienne, Tahiti Ingénierie, Coll'Ex-Persée.

Figure 8 - Capture d'écran d'une page de concordance.

En bas de la fenêtre de contexte élargi se trouvent deux mentions avec un lien actif. La première, « Voir le document original », permet de visualiser le document image déposé sur *Ana'ite*, la seconde, « Voir sa transcription », redirige l'utilisateur vers une page où sont restituées les deux transcriptions, l'une dans l'orthographe originelle, l'autre dans celle, officielle, de l'Académie tahitienne (cf. Figure 6).

Premières valorisations et perspectives

Les premiers documents déjà convertis, rendus manipulables par un traitement automatisé et une recherche plein texte performante, offrent de précieuses ressources dans l'étude des évolutions orthographiques, lexicales et grammaticales de la langue tahitienne. Une table des formes sans concordance avec le dictionnaire permet d'extraire les mots dont il conviendra plus tard de déterminer s'ils doivent y être ajoutés. Par exemple *tapitana* < fr. capitaine, auquel s'est substitué *tāpena* < ang. *captain* dans la langue moderne ; *ianuari* < ang. *january*, devenu *tēnuare* ; *asena* < fr. arsenal ou *fa'anēneva* 'prendre qqn pour un sot' ne figurent pas dans le dictionnaire.

Dans une perspective historique, la recherche d'occurrences de mots clés comme *purūmu* 'route', *ha'api'i* 'enseigner', *fa'a'apu* 'agriculture', *fa'ehau* 'soldat', révèle instantanément les sections du *Vea no Tahiti*, habile instrument d'information et de propagande, qui rendent compte des velléités des premiers administrateurs coloniaux du Protectorat de tracer un réseau routier, de développer la scolarisation, de promouvoir une agriculture plus intensive dédiée à l'export ou de constituer un corps militaire autochtone pour mater les insurrections locales. Ce corpus, encore insuffisamment exploité par les historiens, offre de nouvelles ressources pour une meilleure connaissance des premières heures de la colonisation française en Polynésie.

Alors que l'exemplification des entrées du dictionnaire en ligne de l'Académie tahitienne était jusqu'à présent le plus souvent puisée dans des transcriptions de la tradition orale antérieure

à la colonisation ou dans des textes religieux postérieurs à la conversion chrétienne, la première cible proposée pour le projet Concordance, le *Veà no Tahiti*, est remarquable par son caractère laïque et par son ancrage dans un quotidien marqué par les profondes mutations politiques, économiques et sociales des îles de la Société au moment de l'implantation française au XIX^e siècle.

D'un point de vue didactique, la conversion des textes dans l'orthographe de l'Académie tahitienne facilite leur exploitation pour l'enseignement de la langue tahitienne dans le premier et le second degrés.

Grâce aux outils développés et l'expertise acquise dans le traitement des documents, le programme de conversation et d'intégration de nouvelles ressources textuelles dans le corpus de Concordance se poursuit, sur financements propres de la Maison des sciences de l'Homme du Pacifique, du laboratoire EASTCO et de la Bibliothèque universitaire de l'UPF. À cette fin, de nouvelles numérisations ont déjà été réalisées — on citera par exemple les 7 000 pages en tahitien du périodique *Te Veà Porotetani* de l'Église protestante, publié depuis 1927, ou le périodique politique bilingue *Te Veà a te nunaa* (1977-1990) auquel ont contribué les intellectuels Henri Hiro et Turo Raapoto — ou sont en cours de numérisation, comme le tapuscrit original de la pièce *Maro Pūtoto* (« La ceinture royale ensanglantée »), adaptation de *Macbeth* de Shakespeare en tahitien par le dramaturge John Mairai, ou la revue littéraire *Littéramā'ohi*.

Pour dépasser certaines contraintes constatées lors de l'étape de transcription sur la plateforme TACT, notre équipe développe actuellement son propre outil collaboratif de transcription, nommé ReoScript, directement intégré à la base *Anareo* et dont nous pouvons faire évoluer les fonctionnalités selon les besoins spécifiques de l'analyse lexicographique, grammaticale et littéraire des langues polynésiennes. ReoScript dispose de fonctions corrélées aux ressources déjà contenues dans *Anareo*. Il permet, entre autres, la manipulation de différents types de fichiers image, l'assistance des dictionnaires, des lexiques de langues polynésiennes et des index taxinomiques disponibles dans *Anareo*, la gestion de la chaîne relationnelle du mot dans un dictionnaire jusqu'à ses occurrences dans l'image originelle du document et le travail de balisage et d'annotation des transcriptions. Le projet ReoScript a obtenu un financement interne de l'UPF de 10 000 € en 2021 (AAP « Actions incitatives » de l'UPF et participation financière de la Maison des sciences de l'Homme du Pacifique et du laboratoire EASTCO). Ces nouveaux développements ont été rendus possibles grâce à la première étape modélisante du projet Concordance-CollEx Persée et à ses résultats que nous pouvons désormais présenter à nos partenaires techniques et financiers.

La procédure complète de conversion et ses outils numériques peuvent à présent être déclinés dans d'autres langues polynésiennes. La prochaine langue cible sera le marquisien ou le pa'umotu.

On trouve une première valorisation de l'outil Concordance dans un entretien réalisé par Claire Camberlein et publié sur le Carreau de la BULAC (Bibliothèque universitaire des langues et civilisations) : « Le projet Concordance : “Étayer la connaissance de la langue tahitienne écrite sur un corpus de textes laïcs du XIXe et XXe siècles” » (bulac.hypotheses.org/31626).

La présentation officielle de Concordance aux médias et au grand public était programmée à l'occasion de l'intronisation de nouveaux membres de l'Académie tahitienne, lors d'une cérémonie au Conseil économique, social et environnemental de la Polynésie française, le vendredi 24 septembre 2021. La dégradation de la situation sanitaire locale en raison du variant Delta du Covid-19 depuis le mois d'août 2021 nous a contraint à reporter cette action de promotion à une date ultérieure.

Une présentation à la communauté des spécialistes des langues océaniques devait être réalisée à l'occasion du 12^e Colloque de linguistique océanique (bit.ly/cool12tahiti), organisé par le laboratoire EASTCO à l'UPF et prévu initialement en novembre 2021. Pour les mêmes raisons sanitaires, le colloque est repoussé à septembre 2022.

La fonction de concordance n'en est pas moins déjà opérationnelle et accessible à tous via le dictionnaire en ligne de l'Académie tahitienne, lequel reçoit environ 600 consultations quotidiennes⁶.

Rapport établi par Jacques Vernaudo et Claire Mouraby, le 31 août 2021

⁶ Source : Google analytics, consulté le 24 août 2021